A Seed-based Approach to Identify Risk Disease Sub-networks in Human Lung Cancer

Yi-Bin Wang*, Yong-Mei Cheng, Shao-Wu Zhang, Wei Chen
School of Automation
Northwestern Polytechnical University
Xi'an, 710072, China
yibeen.wong@gmail.com

Abstract—Lung cancer is the leading cause of cancer deaths worldwide. The identification of lung cancer risk disease sub-networks not only helps to understand lung cancer mechanism better, but also provide the potential benefits for the early diagnosis and lead to important applications such as drug targeting. Although some researches are devoted to investigating the carcinogenic process of lung cancer, these approaches have still some limitation. In this paper, the differentially expressed genes are scored and ranked in according to the method of augmented fuzzy measure similarity for obtaining the seed genes. Then, the model of random walk with restarts is used to identify risk disease sub-networks in the PPI network. At last 37 risk disease sub-networks are exploited from the PPI network, which play an important potential role in the carcinogenic process of the lung cancer disease. In terms of the proof and comments in the existing literatures, the identified results show that the proposed method works well in identifying the significant lung cancer risk disease sub-networks, and it is also suitable to recognize other complex risk disease sub-networks.

Keywords-Augmenting Fuzzy Measure Similarity, Random Walk with Restarts, Seed Gene, Risk Disease Sub-network, Lung Cancer

I. INTRODUCTION

Cancer, the complex disease of uncontrolled cell growth, is one of the leading causes of human death worldwide and the deaths from cancer are projected to continue rising [1-2]. Identification and verification of genes/proteins which have a functional role in the patho-physiology of cancer remains an important goal as that can help uncovering the underlying molecular basis of cancer and cancer prevention, diagnosis and treatment. For the past few years, there is a common view that genes with a role in cancer tend to cluster together on well-connected sub-networks of protein-protein interactions [3]. This view has been widely accepted and applied to a variety of disease research, such as oral cancer [4], breast cancer [5], etc.. This suggests a hypothesis that the synergistic expression of multiple cancer-related genes at the level of mRNA can co-regulate the expression of proteins in their immediate "network neighborhood" [6]. So, these genes/proteins and their network neighborhood should provide an ideal starting place to search for sub-networks with a possible role in the disease.

The effectiveness of network-based approaches to the identification of multiple disease genes/proteins has been demonstrated in the context of various diseases, including congenital heart disease [7], Crohn's disease [8], and coronary artery disease [9]. In addition, these approaches are also widely used in other fields, such as metabolic [10], communication [11], etc.

Among all types of cancer, the mostly common causes of cancer deaths are lung cancer. More than 1.6 million people around the world are diagnosed as lung cancer each year [12]. This number continues to grow and it is estimated that 2.2 million people will be newly diagnosed as lung cancer in the year 2020. Lung cancer is the most prevalent cancer type and is associated with the highest mortality [13]. Previous researches have shown that the major types of lung cancer are associated with cigarette smoking [14-15]. In light of this, some research have been devoted to investigating the molecular alterations which ensued from cigarette smoking, and the mechanism that links cigarette smoking to lung cancer. Spira et al. have used DNA microarray to compare the gene expressions of large-airway epithelial cells from nonsmokers and smokers, and determine how cigarette smoking alters the transcriptome [16]. Recently, Wan et al. have selected differentially expressed genes from the survival genes and constructed gene co-expression networks for identifying the smoking-associated gene signature [17]. Takahashi et al. have showed that induction of IKKbeta-dependent and JNK1-dependent inflammation is likely to be an important contributor to the tumor promoting activity of tobacco smoke [18]. Fang X et al. have coupled a network-based approach with gene set enrichment analysis and identified some new candidate genes in lung cancer and smoking [19]. However, most of these methods are generally limited to mRNA expression data in terms of quantification of molecular expression. And only a few of independent new disease markers are identified in the most cases, let alone are considered to construct the new interaction sub-network modules for obtaining more information. For an identified set of seed genes/proteins, their corresponding risk disease sub-networks with PPI information and different function may have the potential to expand the information of lung cancer mechanism and cure this disease. Therefore, these potential sub-networks need to be given more attention, and

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

the identification of the sub-networks also need to try and in-depth study.

In this paper, we firstly incorporate Gene Ontology into microarray gene expression profiles information to score and rank the differentially expressed genes for selecting the seed genes with a method of augmenting fuzzy measure similarity (AFMS). And protein-protein interaction network information and the model of random walk with restarts are used to identify the risk disease sub-networks of lung cancer. The new comprehensive method of bioinformatics for sub-network identification can probe into deep genetic relationships between lung cancer and pathogenesis mechanism. Our study indicates that disease-specific sub-networks could provide more specific information for the development of precision prediction and therapies for lung cancer. The overall work flow for the present study is described in Fig. 1.

II. Material and Methods

A. Datasets

In this study, three kinds of data, namely microarray gene expression profile data, protein-protein interaction information and Gene Ontology (GO) terms, are used. The GSE4115 dataset of smoker microarray gene expression profiles is downloaded from NCBI Entrez Gene GEO site (http://www.ncbi.nlm.nih.gov/geo/). Data are collected from a total pool of 152 subjects and divided into two parts: 79 and 73 smokers with and without lung cancer in the data set. Protein-Protein interaction (PPI) data are extracted from the Human Protein Reference Database (HPRD) http://www.hprd.org/ (Version 9.0). The Gene Ontology project is a major bioinformatics initiative with the aim of standardizing the representation of gene and gene product attributes across species and databases. For certain genes that we are interested, their terms could be searched out in Gene Ontology project that provides a controlled vocabulary of terms for describing gene product characteristics (http://www.geneontology.org/). Currently, there are 36613 terms and 100% defined, including 22628 biological process, 2987 cellular component and 9368 molecular function.

B. Data preprocessing and selection of the differentially expressed genes

Before using the raw data, the each gene expression value is normalized to z-transformed score, so that for each gene the normalized expression value has mean equal to 0 and standard deviation equal to 1 over each sample. The scores for each gene are used to select differentially expressed genes using one-way analysis of variance (ANOVA). The genes with Bonferroni adjusted p-values less than 0.05 are selected as the results. At the same time, to reduce error, the proteins which interact with self in HPRD are removed, and a PPI network which consists of 9,502 proteins and 37,520 interact pairs are constructed.

C. Selection of the seed genes based on augmenting fuzzy measure similarity

Currently in analyzing the similarity between gene products, the mostly used features are the DNA sequence and the expression values. However, as for many gene products, additional information is available. One form of information is symbolic, taking the form of associated Gene Ontology (GO) terms [20]. In this paper, the method of Mihail Popescu et al. [21] are adopted, who proposed the AFMS to calculate the similarity between any two genes based on the GO terms.

Assuming $G = \{T_1, \cdots\cdots$ is a finite set of terms that describing a gene. The fuzzy measure, g, is a real valued function, satisfying the following properties:

$$g(\Phi) = 0 \; and \; g(G) = \quad (1)$$

$$g(A) \le g(B) \;\; if \;\; A, B \subseteq G \;\; with \;\; A \subseteq B \quad (2)$$

$$For \; all \; A, B \subseteq G \; with \; A \cap \; g(A \cup \quad A) + g(B) + \lambda g(A) g(B), \; for \; some \; \lambda > -1 \quad (3)$$

The $g^i = g(\{T_i\})$ is defined as the fuzzy density value, which is interpreted as the information source of the single term T_i in determining the similarity of two genes. For each terms in the GO, their corresponding density value could be calculated in a simple fashion, adapting the approach in [22].

The number of occurrences in the corpus of the term T_i and its children is counted, and is converted to a probability, i.e.,

$$p(T_i) = \frac{count(T_i + children \; of \; T_i \; in \; corpus)}{count(all \; GO \; terms \; in \; corpus)} \quad (4)$$

Where $1 \le k \le |GO|$.

Then, the density value g^i is calculated as:

$$g^i = -\ln(p(T_i)) / \max_{T_j \in GO} \{-\ln(p(T_j)\} \quad (5)$$

The denominator is used to scale the value into the interval [0, 1].

Once all the density values of terms for a gene are known, the value of λ can be uniquely determined for a finite set G using the property (3) and the facts $G = \bigcup$ and $g(G) = 1$, which leads to the following equation for:

$$1 + \lambda = \prod_{i=1}^{n} (1 + \lambda g^i \quad (6)$$

There is a unique solution $\lambda > -1$ for this equation [23]. For any one gene G_i which has been described by terms in the GO, its corresponding parameter λ_i could be obtained in according with the above description.

Any two genes, e.g. the gene G_1 and the gene G_2, are considered as being represented by collections of terms $G_1 = \{T_{11}, T_{12}, \cdots$ and $G_2 = \{T_{21}, T_{22}, \cdots$, where $T_{1i}, T_{2j} \in GO$. The purpose of the proposed method is to augment each set as:

$$G_1^+ = G_1 \cup \{ T_{1 \; j} \} \; and \; G_2 = G_2 \cup G_2 \{ \; i \; . \quad (7)$$

Where $\{T_{1i,2j}\}$ is the set of nearest common ancestors (NCA) of every pair (T_{1i}, T_{2j}). And the parameters λ_1^+ or λ_2^+ can be obtained by the Eq. (4), (5) and (6) based on the augmented

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

set G_1^+ or G_2^+. At the same time, the resulting augmented intersection is:

$$[G_1 \cap G_2]^+ = G_1^+ \cap G_2^+ = [G_1 \cap G_2] \cup \{T_{1i,2j}\} \quad (8)$$

Using the augmented intersection, the augmenting fuzzy measure similarity denoted by $S_{AFMS}(G_1, G_2)$, is defined as:

$$S_{AFMS}(G_1, G_2) = \frac{g_1^+([G_1 \cap G_2]^+) + g_2^+([G_1 \cap G_2]^+)}{2} \quad (9)$$

Where g_k^+ is the fuzzy measure compute on the result of Eq. (8), which using Eq. (3) and parameter λ_k^+, $k = \{1,2\}$.

For example, the calculation of the AFMS for two gene products: COL21A1 gene (GenBank ID AAL02227) and COL27A1 gene (GenBank ID BAB13947). They are described as respectively:

COL21A gene:

$$G_1 = \{T_1 = 5198(\textit{structural molecular activity}), T_2 = 7155(\textit{cell adhesion}), T_3 = 5201(\textit{extracellular matrix structural constituent}),\}$$

COL27A1 gene:

$$G_2 = \{T_3 = 5201(\textit{extracellular matrix structural constituent}), T_4 = 5581(\textit{collagen})\}$$

Since NCA $(T_1, T_3) = T_1$, the rest of NCA=0, $G_1^+ = \{T_1, T_2, T_3\}$ and $G_2^+ = \{T_1, T_3, T_4\}$, the augmented intersection is $[G_1 \cap G_2]^+ = \{T_1, T_3\}$. The set of related densities are calculated for G_1^+ using Eq. (4) and (5): $\{0.42, 0.44, 0.58\}$, then the Eq.(6) is solved as:

$$1 + \lambda = (1 + 0.42\lambda) \times (1 + 0.44\lambda) \times (1 + 0.58\lambda)$$

$$\Rightarrow \lambda_1^+ = -0.72$$

And using Eq.(3), the fuzzy measure compute is:

$$g_1^+([G_1 \cap G_2]^+) = g_1^+(\{T_1, T_3\}) = 0.42 + 0.58 + (-0.72) \times 0.42 \times 0.58 = 0.82$$

The same process used in G_2^+ and the densities and parameter are obtained: $\{0.42, 0.58, 0.65\}$, $\lambda_2^+ = -0.75$. Then

$$g_2^+([G_1 \cap G_2]^+) = g_2^+(\{T_1, T_3\}) = 0.42 + 0.58 + (-0.75) \times 0.42 \times 0.58 = 0.81$$

Hence, the AFMS is:

$$S_{AFMS}(G_1, G_2) = \frac{g_1^+([G_1 \cap G_2]^+) + g_2^+([G_1 \cap G_2]^+)}{2} = \frac{0.82 + 0.81}{2} = 0.82$$

As for these pervious selected differential expressed genes, a further screening work is carried out for ensuring that the selected seed genes are the most important and reliable ones.

According to the method of AFMS, the similarity scores of any two differential expressed genes are calculated. And then corresponding similarity scores matrix $SS(s_{ij})_{n \times n}$ is also constructed. Where $s_{ij} = S_{AFMS}(G_i, G_j)$ or 1 if $i = j$, n denotes the number of the differential expression genes.

The each gene score based the $SS(s_{ij})$ is defined as:

$$s_i = \frac{2(\sum_{j=1}^{n} s_{ij} - 1)}{\sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij} - n} \quad (10)$$

Finally, these scores are ranked in accordance with the principle of high to low, $s_n > s_{n-1} > \cdots$, the preceding m genes whose sum scores accounted for 90% of all sum scores are viewed as the seed genes.

D. Identification of the lung cancer risk disease sub-networks

These seed genes are mapped to the PPI network and are found out their corresponding proteins, they are called proteomic seeds. Each of them, there is a candidate sub-network consisting of it and its direct interactors, termed as the interactor sub-network is selected.

Based on the protein's proximity and connectivity to the seeds, each protein in HPRD are assigned a score. If the score is significant (p<0.001), and it is an interactor for a seed, the protein is called a crosstalker. The seed and their crosstalkers together constitute a crosstalker sub-network, it is also identified as a risk disease sub-network. In other words, we use an information flow based algorithm based on random walk with restarts [24], identified a new smaller-scale sub-network from each interactor sub-network.

Let $G = (V, E)$ be a PPI network. For convenience, $N(v)$ is defined as the set of interacting partners of protein $v \in V$, i.e., $N(v) = \{u \in V : uv \in E\}$, and defined $S \subseteq V$ as the set of proteomic seeds. Our objective is to compute a score $\alpha(v)$ for each protein $v \in V$, to quantify the network crosstalk between v and the seeds in S. Here, network crosstalk is used an indicator of functional association between proteins.

The random walk starts at a randomly chosen protein in S. At each step, when the random walk is at a protein v, it either moves to an interacting partner of v with the probability $1-r$, or it restarts at a protein in S with probability r. The parameter $0 \le r \le 1$ is called the restart probability.

For each move, the interacting partner to be moved to is selected uniformly at random from $N(v)$. If $u \in N(v)$, the probability of a move from v to u defined as:

$$P(u, v) = \varphi(u, v) / \sum_{u' \in N(v)} \varphi(u', v) \quad (11)$$

otherwise 0. $\varphi(u, v)$ denotes the reliability of the interaction between u and v. Similarly, for each restart the protein to be restarted is selected uniformly at random from S, if $u \in S$, the probability of restart at $u \in N(v)$ defined as:

$$\rho(u) = z_p(u) / \sum_{u' \in S} z_p(u') \quad (12)$$

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

Otherwise 0. $z_p(u)$ denotes the z-score of the fold change of u with respect to the phenotype of interest, based on proteomic seeds screening.

Based on the above described random walk model, the probability between the proteins in S and each protein $v \in V$ is defined. Let α_t denote a $|V|$-dimensional vector, such that $\alpha_t(v)$ is equal to the probability that the random walk will be at protein v at step t, where $\|\alpha_t\| = 1$. Let P denote the stochastic matrix derived from network $G = (V, E)$, where $P(u,v) = 1/|N(v)|$, if $uv \in E$, 0 otherwise. Then, we have

$$\alpha_{t+1} = (1 - r)P\alpha_t + r\rho \quad (13)$$

Where ρ denotes the restart vector with $\rho(u) = 1/|S|$ for $u \in S$, and 0 otherwise. In the initial step, let $\alpha_0 = \rho$, the vector containing the crosstalk scores for each node in the network is given by $\alpha = \lim_{t \to \infty} \alpha_t$.

Since the moving process of the seed protein is random and the probability is uniform in general case, the parameter is set as $r = 0.5$. At the same time, the parameter t must be sufficiently large. In our experiment, we has selected $t = 1000$, $t = 2000$ and $t = 3000$ for testing the related performance respectively. And the obtained results adopting the different parameters show that the score vectors are almost consistent. Moreover, more experimental results show that the values of the score vector are almost stable with the increase of the parameter t. Here, we just choose $t = 1000$ to carry out the experiments. Then $\mu = \sum_{i=1}^{n} \alpha^{(i)}(v)/n$ and the $\sigma^2 = \sum_{i=1}^{n} (\alpha^{(i)}(v) - \mu)^2/(n-1)$ of the scores distribution which using in this sample are computed, under the assumption of normally distributed the scores. Subsequently, the scores are adjusted to the standard normal distribution for each protein:

$$z^{(i)}(v) = (\alpha^{(i)}(v) - \mu)/\sigma \quad (14)$$

Once all proteins in the PPI network are scored and adjusted according to these rules, the crosstalker sub-networks are selected as follow: for each proteomic seed u, the sub-network induced by the proteins in $N(u)$ that have significant adjusted scores with respect to S is considered a crosstalker sub-network, the crosstalker sub-network is defined as:

$$C(S) = \{N^*(u) : u \in S\} \quad (15)$$

Where $N^*(u) = \{v \in N(u) : z^{(i)}(v) > z^*\}$. Here, z^* denotes the cut-off for adjusted score to be considered significant and in our experiments, we used a p-value cut-off of 0.001.

III. Result and Discussion

Using the ANOVA, 59 differentially expressed genes are identified from GSE4115 dataset, and according to the method of AFMS, each differentially expressed gene are scored and ranked, 37 genes are finally chosen as the seed genes. Subsequently, their corresponding proteins are found out, called them proteomic seeds. For each one of them, a crosstalker sub-network consisting of itself and its significant interactors is obtained, resulting in a total of 37 crosstalker sub-networks.

These sub-networks could be considered as different disease function modules and play an important potential role in the carcinogenic process of lung cancer. For example, the MAPK1 crosstalker sub-network, which is illustrated with Fig.2 and consists of 81 genes, some of these genes have been confirmed to involve in the pathogenesis of lung cancer, such as VAV1, DUSP6, TIP30, HMGA1, EGFR [25-29] etc. The MAPK cascade is a highly conserved module that is identifiable in many cancers. Some MAPK-related proteins annotated with cell growth function, they are identified as significant proteins in lung carcinogenesis. Elevated expression of activated MAPK1 has been observed in NSCLC (Non Small Cell Lung Cancer) [30] and may play a role in lung metastasis [31].

TP53 is a well-studied tumor suppressor protein and plays important roles in anti-cancer mechanisms and cell cycle function, the TP53 crosstalker sub-network is illustrated with Fig.3. Its activation is induced by a number of stress signals such as DNA damage, oxidative stress and activated oncogenes. Activated TP53 induces cell cycle arrest, apoptosis and inhibition of angiogenesis and metastasis function. Once damaged, tumor suppression is severely reduced, resulting in uncontrolled proliferation of the cell. Due to the importance in carcinogenesis, it is no surprise that TP53 was found to be significantly mutated in lung adenocarcinoma as well as in squamous cell carcinoma and SCLC (small cell lung cancer) [32-34].

Large numbers of methods on discovering discriminative lung cancer candidate genes are proposed, including gene set enrichment analysis [19], protein parameter estimation model [35] and enriched sera protein profiling method[36]. The biological database used in those method relatively simple, and the algorithm used in those method relatively more complicated. The most critical thing is that there is no information on the sub-networks of biomarkers, since those studies mainly included some novel protein biomarkers. Instead of existed methods, we take a more direct approach to identify new seed genes according to the similarity between two differential expression genes which are screened out previously. Notice that the seed genes or proteins are viewed as the center, by means of the random walk with restarts model, the new sub-networks can be discovered finally. In order to display the identified results of proposed method, Fang's method [19] which is under the condition of the disease stage or smoking is chosen for the purpose of comparison. The 32 common genes involved in smoking and lung cancer related networks are identified by Fang's method. While, 37 novel seed genes with corresponding crosstalker sub-networks are discovered by the presented method, which can provide us more potential and useful information. All of the seed genes are listed in Table 1.

If some proteins with unknown functions are discovered

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

in these sub-networks, that will be a great challenge to confirm the expression of those identified genes in the formation of lung cancer, the production of correspondent proteins in the tissue, or interaction between identified genes and targeted proteins within the network. At the same time, we also note that the number of the crosstalker sub-networks is not very much, one potential explanation for this observation is that current human PPI network capture only a very small fraction of all protein relationships in the human interactome [37], and therefore cannot be expected to reveal a significant sub-network for every gene.

IV. CONCLUSION

Lung cancer is a complex disease and carcinogenesis in humans is a multistep process that transforms normal cells into malignant derivatives. In the recent decade, a lot of researchers have investigated the underlying mechanisms that prompt the uncontrolled cell proliferation and metastasis. They have successfully identified some key components of the various steps in the carcinogenesis and some therapeutic interventions have been developed to at least slow down the carcinogenic process. However, because of the complexity and lack of comprehensive information, the detection and therapy that targets some specific molecules is only partially effective. Therefore, investigation of the disease mechanism from the systems perspective is inevitable. The identification of potential risk disease sub-networks can provide more additional new biological information about lung cancer and help understanding lung cancer mechanism, further diagnosing and curing this disease. In this paper, we propose a seed-based method to identify sub-networks of lung cancer. The algorithm of augmented fuzzy measure similarity in conjunction with microarray gene expression profile and GO information are used to score and rank differentially expressed genes for selecting the seed genes. Then each of proteins is scored in the PPI network via the random walk with restarts model and the topological structure information of the protein-protein interaction network. According to the predetermined significant *p*-value (0.001), the crosstalker sub-networks can be identified eventually. Experimental results indicate that our method is superior to Fang's method in term of the number of new candidate genes, and potential and useful information. At the same time, base on the proof and comments which has been record in the existing literature, our method is good at identifying the risk disease sub-networks of lung cancer, and it is also suitable for recognizing other complex disease sub-networks.

However, as an uncovering method in disease areas, the sophisticated relationships of genes/proteins and the possibility of various parameters make this approach go a long way before it becomes mature. For example, it is still weak in mathematics and the difference of selected database may lead to diverse results. As for the new risk sub-networks identified in proposed work, the relationships between the known pathogenic genes/proteins of lung cancer and risk sub-networks are not considered in depth. Moreover, a web server need to be given for implementing our algorithm, and sufficient evidence are lacked to prove the accuracy and reliability of the results. It is necessary for us to construct a web server and pay more attention to the relationship between the entire genome or proteome as well as find other new evidence in the future work.

ACKNOWLEDGMENT

This research is supported by the Doctorate Foundation of Northwestern Polytechnical University (No. cx201017).

REFERENCES

[1] Parkin DM, Bray F, Ferlay J, Pisani P, "Global cancer statistics," CA Cancer J Clin, vol. 55, pp. 74-108, 2005

[2] Kanavos P, 'The rising burden of cancer in the developing world," Ann Oncol, vol. 17(Suppl 8), pp. viii15-viii23, 2006

[3] Ideker T, Sharan R. "Protein networks in disease," Genome Res, vol.18, pp. 644–52, 2008

[4] Jie Zhang, Knobloch, T, et al. "Identifying smoking associated gene co-expression networks related to oral cancer initiation," 2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW). pp. 1039-1041, 2011

[5] Van den Akker, Erik B. , Verbruggen, Bas, et al. "Integrating Protein-Protein Interaction Networks with Gene-Gene Co-Expression Networks improves Gene Signatures for Classifying Breast Cancer Metastasis," Journal of Integrative Bioinformatics. vol. 8(2), DOI: 10.2390/biecoll-jib-2011-188, 2011

[6] Rod K. Nibbe, M Koyutürk, et al. "An Integrative -omics Approach to Identify Functional Sub-Networks in Human Colorectal Cancer," PLoS Computational Biology. vol.6(1), e1000639, 2010

[7] Zhi-Ping Liu, Luonan Chen. "Identification of dysfunctional modules and disease genes in congenital heart disease by a network-based approach," BMC Genomics, vol.12, 592 doi:10.1186/1471-2164-12-592, 2011

[8] Insuk Lee, U. Martin Blom, et al. "Prioritizing candidate disease genes by network-based boosting ofgenome-wide association data," Genome Res. Vol. 21, pp. 1109-1121, 2011

[9] Liangcai Zhang, Xu Li, et al. "Predicting Candidate Genes Based on Combined Network Topological Features: A Case Study in Coronary Artery Disease," PLoS ONE,. vol. 7(6), e36542, 2012

[10] S.C. Jangaa, J. Javier Díaz-Mejía, et al. "Network-based function prediction and interactomics: The case for metabolic enzymes," Metabolic Engineering, vol. 13(1), pp. 1-10, 2011

[11] Rajagopal, R.. "Network-based Consensus Averaging With General Noisy Channels," IEEE Transactions on Signal Processing., vol. 59(1), pp.373-385, 2011

[12] C.L. Granger, C.F. McDonald, et al. "Exercise intervention to improve exercise capacity and health related quality of life for patients with Non-small cell lung cancer: A systematic review," Lung Cancer, vol.72(2), pp. 139-153, 2011

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

[13] J. Ferlay, H.R. Shin, et al. GLOBOCAN 2008, "Cancer Incidence and Mortality Worldwide: IARC Cancerbase No. 10," International Agency for Research on Cancer. 2010

[14] Matthew A. Steliga, Carolyn M. Dresler. "Epidemiology of Lung Cancer: Smoking, Secondhand Smoke, and Genetics," Surgical Oncology Clinics of North America, vol.20(4), pp. 605-618, 2011

[15] Suresh H. Moolgavkar, Theodore R. Holford. et al. "Impact of Reduced Tobacco Smoking on Lung Cancer Mortality in the United States During 1975–2000," Journal of the National Cancer Institute. Vol.104(7), pp. 541-548, 2012

[16] Spira A, Beane J, Shah V, Liu G, Schembri F, Yang X, Palma J, Brody JS. "Effects of cigarette smoke on the human airway epithelial cell transcriptome," Proc Natl Acad Sci USA, vol. 101, pp. 10143-10148, 2004

[17] Ying-Wooi Wan, Changchang Xiao et al. "Network-based identification of smoking-associated gene signature for lung cancer," Bioinformatics and Biomedicine (BIBM), pp. 479-484, 2010

[18] Takahashi H, Ogata H, Nishigaki R, Broide DH, Karin M, "Tobacco smoke promotes lung tumorigenesis by triggering IKKbeta-and JNK1-dependent inflammation," Cancer Cell, vol.17, pp.89-97, 2010

[19] Fang X, Netzer M, et al. "Genetic network and gene set enrichment analysis to identify biomarkers related to cigarette smoking and lung cancer. Cancer Treatment Reviews," Available online. http://dx.doi.org/10.1016/j.ctrv.2012.06.001, 2012

[20] P.W. Lord, R.D. Stevens, A. Brass et al. "Semantic Similarity Measure as a Tool for Exploring the Gene Ontology," Proc. Pacific Symp. Biocomputing. pp. 601-612, 2003

[21] Mihail Popescu, James M. Keller et al. "Fuzzy Measures on the Gene Ontology for Gene Product Similarity," IEEE/ACM Transaction On Computational Biology And Bioinformatica,vol. 3, pp. 263-274,2006

[22] P. Resnik. "Semantic Similarity in a Taxonomy: An information-Base Measure and Its Application to Problems of Ambiguity in nature Language," J. Artificial Intelligence Research (JAIR), vol.11, pp. 95-130,1999

[23] M.Grabisch et al. "Fuzzy Measure and Integrals: Theory and Applications," Springer-Verlag, 2000

[24] Tong H, Faloutsos C, Pan J.-Y. "Random walk with restarts: fast solutions and applications," Knowledge and Information Systems,vol.14(3), pp. 327-346,2008

[25] Lena Ilan, Shulamit Katzav. "Human Vav1 Expression in Hematopoietic and Cancer Cell Lines Is Regulated by c-Myb and by CpG Methylation," PLoS One, vol.7(1), e29939,2012

[26] Zhenfeng Zhang,Susumu Kobayashi et al. "Dual specificity phosphatase 6 (DUSP6) is an ETS-regulated negative feedback mediator of oncogenic ERK signaling in lung cancer cells," Carcinogenesis, vol.31(4), pp. 577–586, 2010

[27] Xin Tong, Kai Li. et al. "Decreased TIP30 Expression Promotes Tumor Metastasis in Lung Cancer," Am J Pathol,vol. 174(5), pp. 1931–1939, 2009

[28] Joelle Hillion, Lisa J. Wood et al. "Up-regulation of MMP-2 by HMGA1 Promotes Transformation in Undifferentiated Large Cell Lung Cancer," Mol Cancer Res,vol.7(11), pp. 1803-1812, 2009

[29] Vladimir Ratushny, Igor Astsaturov et al, "Targeting EGFR resistance networks in Head and Neck Cancer," Cell Signal, vol.21(8), pp. 1255–1268, 2009

[30] Vicent S, Lopez-Picazo JM, et al. ERK1/2 is activated in non-small-cell lung cancer and associated with advanced tumours. Br J Cancer . 90:1047-1052, 2004

[31] Hu K, Gan YH, Li SL et al. "Relationship of activated extracellular signal-regulated kinase 1/2 with lung metastasis in salivary adenoid cystic carcinoma," Oncol Rep, vol. 21, pp. 137-143, 2009

[32] Ding L, Getz G, et al. "Somatic mutations affect key pathways in lung adenocarcinoma. Nature," vol. 455, pp. 1069-1075, 2008.

[33] Herbst RS, Heymach JV, Lippman SM. Lung cancer. N Engl J Med, vol.359, pp. 1367-1380, 2008

[34] Hollstein M, Sidransky D, Vogelstein B, Harris CC. "p53 mutations in human cancers. Science,"vol. 253, pp.49-53, 1991

[35] Wang YC, Chen BS. "A network-based biomarker approach for molecular investigation and diagnosis of lung cancer," BMC Med Genomics. vol.4:2, 2011

[36] Emanuela Monari1, Christian Casali2, et al. "Enriched sera protein profiling for detection of non-small cell lung cancer biomarkers. Proteome Science," vol. 9, pp. 55, 2011

[37] Stumpf MP, Thorne T, de Silva E, Stewart R, An HJ, et al. "Estimating the size of the human interactome," Proc Natl Acad Sci U S A vol.105,pp. 6959–6964, 2008

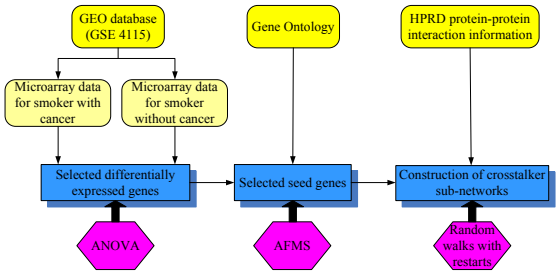

Fig. 1. The flowchart of constructing the crosstalker sub-networks in lung cancer and cigarette smoking.

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

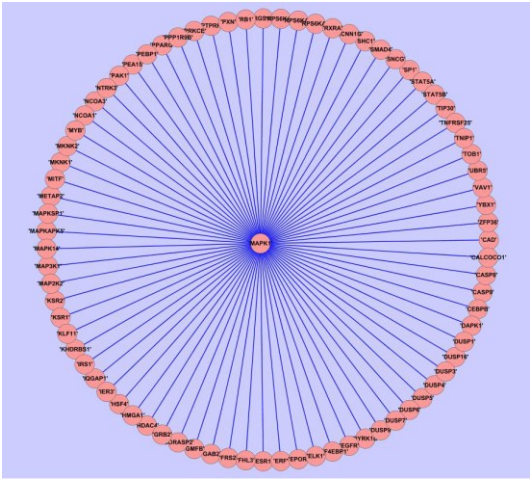

Fig. 2. The crosstalker sub-network of MAPK1 protein

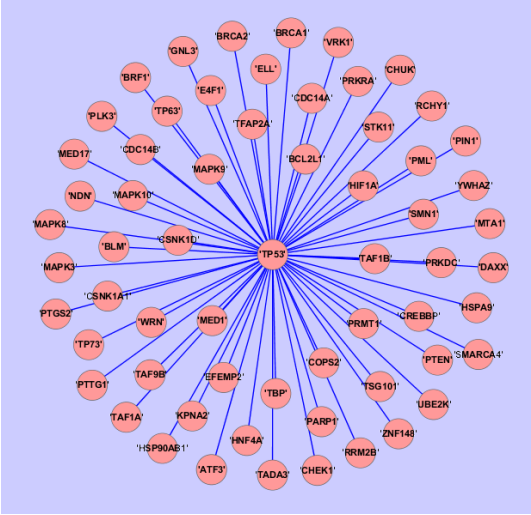

Fig. 3. The crosstalker sub-network of TP53 protein

TABLE I. THE SEED GENES OF IDENTIFIED CROSSTALKER SUB-NETWORKS WHICH RELATED LUNG CANCER

The seed genes of crosstalker sub-networks
LRRFIP1, NEDD9, AR, MSH6, FYN, SOX9, ECD, CD82, PLEKHA5, BCAS1, ESR1, PPM1D, TSC2, XRCC4, SGCB, ACVR2A, SP1, E2F1, AGPS, TP53, CORO2A, SLC5A1, MAPK1, STAM2, MED21, PRUNE, GOSR2, CTNNR1, SFRS14, MAN1A2, MED28, EIF2B3, SMAD2, ERFR, PTK2, HSP90AA1, AKT1

978-1-4673-4398-5/12/$31.00 ©2012 IEEE