Contract number .....
Grant number .....

Detection of protein associated with lung cancer by applying network clustering technique

By

Nilubon Kurubanjerdjit

Natthakan Iam-On

This research was supported by Mae Fah Luang University, 2014

ACKNOWLEDGEMENTS

First of all, our great gratitude goes to Mea Fah Luang University for giving us a support funding for this research. Special thank go to Prof.Ka-Lok Ng, Asia University, Taiwan for being such a good supervisor which providing us valuable advices throughout this research time.
We would also like to make a special reference the members of BioGrid Lab, Asia University, Taiwan in supporting us the valuable research information, consultation, analysis the research result.

Nilubon Kurubanjerdjit

EXCUTIVE SUMMARY

Lung cancer is one of the most common cancers in the world which will continue rising to reach a high rate of death in the year 2030. The two main types are small cell lung cancer and non-small cell lung cancer. The information of genetics mechanism on how genes or protein cause cancer is widely studied nowadays. The development of lung cancer is a multi-gene and extremely complex process that involves several biological processes such as oncogene activation, tumor suppressor gene mutation and tumor cell apoptosis suppression. The cancer diagnostics development relies on the understanding of cancer mechanisms; therefore, to identify novel cancer associated protein is an essential first step in cancer research development.

In this study, we identified the novel lung cancer associated proteins based on two different concepts of network clustering approach for discovering protein interaction dense regions (network motif). Firstly, K-Means clustering approach is adopted to cluster a group of protein-protein interaction into sub-clusters, and then clique percolation clustering method (CPM) is adopted to discover "significant network motif" of significant protein cluster resulted by K-Means. Secondly, the Molecular Complex Detection approach (MCODE) is also adopted in this work to be a candidate of the first algorithm in term of clustering efficiency. Then analyzing biological processes and KEGG pathways of proteins involved in same cluster was investigated. Besides, cancer protein types; tumor suppressor protein (TSP) and onco-protein (OCP) are also observed. Finally, the comparison of discovering accurate "protein complexes" among two different approaches is investigated by referring to known protein complexes from MIPS.

Our results indicated that associated proteins findings involved in crucial processes in cancer formation i.e. programmed cell death, apoptosis. Basically, there are two limitations of our methodology i) the cancer-associated protein prediction is limited by the quality of gene ontology and pathway information, and ii) limited by the number of known lung cancer proteins. This work can be the essential first step on discovering lung cancer associated proteins based on clustering analysis.

Further study will make more experiments in using different clustering algorithm to overcome trapping the result in increasing accuracy and precision of the prediction of lung cancer associated protein.

ABSTRACT

Discovering cancer-associated proteins is a major challenge in cancer research. Recently various techniques have been developed to identify novel cancer-associated proteins. Protein-protein interaction network and also protein clustering approaches are good predictors for cancer proteins. In this study, we implemented two different network clustering approaches on lung cancer protein-protein interaction network in order to identify novel lung cancer-associated proteins. Firstly, we adopted K-Means clustering technique to identify novel lung cancer associated proteins, and secondly, the Molecular Complex Detection approach (MCODE) was applied in this research work to detect significant proteins which related to lung cancer formation.

Enriched biological functions and KEGG pathways are determined, and results strongly suggest that most of predicted proteins involve in lung cancer formation. Also, based on the assumption that cancer proteins tend to interact with cancer proteins, we have identified several putative lung cancer proteins. It is expected that the approach developed in this work should be of value for identifying cancer-associated and cancer proteins.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 Lung Cancer

Lung cancer is one of the most common cancers in the world. The World Health Organization's Global Burden of Disease analyses 1,676,000 deaths from lung cancer worldwide in 2015. It predicts that this toll will continue to rise to reach a staggering 2,279,000 deaths in the year 2030 (World Health Organization Web site).

Lung cancer forms in tissues of the lung, usually in the cells lining air passages which are leading cause of cancer death in human. Cigarette smoking causes most lung cancers. Common symptoms of lung cancer include a cough that doesn't go away and gets worse over time, constant chest pain, coughing up blood, shortness of breath, wheezing, or hoarseness, repeated problems with pneumonia or bronchitis, loss of appetite or weight loss and fatigue. Lung cancer is the leading cause of cancer deaths because 84% of cases are diagnosed at an advanced stage, with a five-year survival rate of less than 15% (Okada M. 2005; Jemal A. 2008; Kassis ES. 2009). The two main types are small cell lung cancer and non-small cell lung cancer. Treatment depends on the types, stage, and how advanced it is. Treatments include surgery, chemotherapy, radiation therapy, and targeted therapy (NIH: National Cancer Institute).

1.2 Genetic Mechanism related Cancer

The information of genetic mechanisms on how genes cause cancer is widely studied nowadays. Genes come in pairs and work together to make a protein product. Proteins are very important molecules in living cells. They are involved in virtually all cell functions. Each protein has a specific role such as some proteins are involved in body movement, defense against germs, while others are involved in structural support. Protein are constructed from a set of twenty of amino acids, each amino acid has different three-dimensional shapes. There are many types of proteins and their functions; antibodies defend that body from germs, enzymes speed up chemical reactions and contractile proteins are responsible for movement.

When genes have error in their DNA code which is said to be "altered", they may not work properly in making protein that work in specific function in human body. An accumulation of many mutations in gene can lead to the development of cancer. The occurrence and development of lung cancer 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(Plebani M. 1995; Vielhaber S. 2006; Beane J. 2007).

1.3 Protein-Protein Interaction

Protein combinations are likely same as instrumental in the pathogenesis of human disease, for instance the defect in fusion of Bcr and Abl can leads to chronic myelogenous leukemia (Ren R. 2005) or the abnormal interactions acquired by the huntington protein in Huntington's Disease (Li SH. 2004).

The protein function can be expressed in terms of its interactions with other molecules. The cancer diagnostics development relies on the understanding of cancer

mechanisms; therefore, to identify novel cancer associated protein is an essential first step in cancer research development. There are two types of cancer protein which are Onco-Protein (OCP) and Tumor Suppressor Protein (TSP). OCP is the good protein that normally controls cell growth and it divides. If OCP mutates, it becomes a bad protein that can makes cell grows out of control, which can lead to cancer. TSP is normal protein that slow down cell division and control apoptosis or programmed cell death, cells can grow out of control which can also lead to cancer if TSP works not properly.

Currently, there are various methods have been developed to accelerate cancer protein discovery i.e. gene annotation and sequence based (Perez-Iratzeta C. 2002; Turner F.S. 2003), microarray expression data (de-Lichtenberg U. 2005), structural information (Zhang QC. 2012), domain composition (Xia K. 2008; Peng W. 2014), and network analysis based (George R.A. 2006; Lage K. 2006) which is generally connect gene networks with phenotype networks to infer gene-cancer relationships.

Many biological functions involve the formation of protein-protein complexes. Protein interactions appear to form a molecular network which usually contains small circuit patterns called "network motifs" which are known to have interesting dynamical properties. Motifs reveal the cores of functional modules in molecular networks. The dynamic modules or sub-networks of proteins may have leading roles in the cancer development and metastasis process. The static modules of protein may belong to the inherent components in a protein-protein interaction 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 protein-protein interaction network, and cells having these proteins are robust (Tang X. 2011).

1.4 Contribution of This Research Work

Discovering the relationship of proteins in protein-protein interaction network has been one of the major challenges in today era. In this study, we further explore proteins relationship, focusing on lung cancer protein in particular. The protein-protein interaction network was investigated in order to imply involvement of proteins in lung cancer. We aim to predict a novel set of lung cancer associated proteins based on various clustering techniques.

In this research, the novel lung cancer associated proteins are predicted based on various networking approaches to discover network motif or cluster which reveals the cores of functional modules in molecular networks. Proteins which appear in the same cluster are likely to have similar molecular functions. Therefore, we hypothesized that the proteins found in same cluster as lung cancer proteins might have a high probability in forming lung cancer as well.

Initially, we apply K-Means clustering approach to cluster a group of protein-protein interaction into sub-clusters, and then analyze the biological functions of proteins involved in same cluster and also their KEEG pathways (Kyoto Encyclopedia of Genes and Genomes: http://www.genome.jp/kegg/). KEGG pathways service is a database resource for understanding high-level functions and utilities of the biological system are observed. Besides, clique percolation clustering method (CPM) is adopted to discover "significant network motif" of clusters resulted by K-Means. The clique

motifs will help in revealing the significant proteins which involve in the core functional modules related to lung cancer. Furthermore, cancer protein types; tumor suppressor protein (TSP) and onco-protein (OCP) are also observed in this study.

Secondly, the Molecular Complex Detection approach (MCODE) is also adopted in this work to be a candidate of the first algorithm in term of clustering efficiency. The same input data set as K-Mean algorithm is submitted into MCODE algorithm to cluster protein-protein interaction network into sub-clusters. Then analyzing biological processes and KEGG pathways of proteins involved in same cluster was investigated. Besides, cancer protein types; tumor suppressor protein (TSP) and onco-protein (OCP) are also observed.

The comparison of discovering accurate "protein complexes" among K-Mean and MCODE algorithm is investigated by referring to known protein complexes from The MIPS Mammalian Protein-Protein Interaction Database (MIPS) (Pagel P. 2005). The web pages displaying the significant protein modules found from these two approaches are created. The web service is freely accessible at http://sit.mfu.ac.th/lungcancerproj/

CHAPTER 2
LITERATURE REVIEWS

Protein-Protein Interaction (PPI) plays a crucial role in determining the outcome of cellular processes. Protein networks have been used to further the study of molecular evolution, robustness of cells to perturbation and for discovery of new protein functions. The accuracy of interacting protein identification and their networks is important for obvious understanding the molecular mechanism within the cell. Many complex systems in nature can be described in terms of networks which makes the tangle connections among the units to be understandable. A key question is how to interpret the networks or sub-network (community) associated with more highly interconnected parts.

Interaction maps of entire genomes are useful for improving the understanding of cellular function. There are various attributes to be used in mapping network i.e. microarray expression data (de-Lichtenberg U. 2005), gene ontology (Mukhopadhyay A. 2012), structural information (Zhang QC. 2012) and domain composition (Xia K. 2008; Peng W. 2014).

Furthermore, several computational methods have been developed to evaluate and predict PPI, such as mRNA-co expression based on the assumption that proteins that are co-expressed are more likely to interact in comparison to proteins that are not co-expressed (Browne F. 2010). The Gene Ontology (GO) annotation (Wu X. 2006) implies that proteins found within the same biological process are more likely to interact than proteins from a different biological process. The Interolog approach involves PPI transferring from one organism to another using comparative genomics (Jansen R. 2003). With the protein domain interaction approach (Ng S.K. 2003), PPI could be inferred by recognizing protein domains and the interaction transfers by known domain-domain interactions (DDI). Also, it was proposed that PPI can be inferred from protein structural information (Ogmen U. 2005). Among those computational techniques, the interolog approach has been broadly used for PPI prediction (Von-Mering C. 2007). Also, the interolog approach has been justified to be reliable on exploring interaction sub-networks in cancer (Rhodes D.R. 2005).

Even though there are nowadays many techniques or methodologies help in capturing the role of molecular functions but the main drawback is that result datasets are often incomplete which cause high rate of false positive and false negative events (Satuluri V. 2010). The computational methods which are based on graph-based approaches are introduced such as Cfinder (Adamcsek B. 2006), Clique (Spinrin V. 2005), jClust (Pavlopoulos GA. 2009), MCODE (bader GD. 2003), SCAN (Mete M. 2008), PCP (Chua HN. 2008), LCMA (Li XL. 2005), DPClust (Alfaf-UI-Amin M. 2006), CMC (Liu G. 2009) and GIBA (Moschopoulos CN. 2009). These algorithms used graph theory to identify highly connected sub-networks. Otherwise, DMSP (maraziotis IA. 2007), GFA (Feng J. 2008) and MTISSE (Ulitsky I. 2007) are the methods to predict protein complex based on gene expression data, whereas others like STM (Cho YR. 2007), SWEMODE (Lubovac Z. 2006) and DECAFF (Li XL. 2007)adopt graph annotation information to make a prediction.

Adamcsek et al (Adamcsek B. 2006)developed an efficient tool names "Cfinder" for finding and visualizing the overlap, dense of node groups in undirected graphs. This

program can be used in discovery novel modules of protein associated network based on Clique Percolation Method (Palla G. 2005).
jClust (Pavlopoulos GA. 2009) an application that provides access to a widely used set of clustering algorithms and allows the interactive visualization of data. This toolbox supports a various supervised and unsupervised clustering analysis methods i.e. k-Means (MacQueen J.B. 1967), Spectral clustering (Paccanaro A. 2006), Affinity propagation (Frey BJ. 2007), Restricted neighborhood search cluster algorithms-RNSC (King AD. 2004), Markov clustering-MCL (Enright AJ. 2002) and MULIC (Andreopoulos B. 2007).
GIBA (Moschopoulos CN. 2009) is a clustering tool which implements various methods i.e. MCL, RNSC , Cluster Density, haircut operation, best neighbor, and cutting edge method.
Spinrin et al (Spinrin V. 2005) studied protein complexes in molecular networks. They presented molecular networks on the meso-scale level which focused on multibody interactions and discovered sets of proteins that have many interacting proteins among themselves. They analyzed a yeast PPI network, then analyzed functional annotation of these sub-networks and found that most of identified sub-networks correspond to either of the two types of cellular modules which are protein complexes or functional modules. Their work discovered highly connected clusters of proteins in a network of protein interactions and also the findings strongly support the suggested modular architecture of biological networks.
Bader and Hogue (bader GD. 2003) developed an automated method for discovering molecular complexes in large protein interaction networks names "Molecular Complex Detection (MCODE)". This method is based on vertex weighting by local neighborhood density and outward traversal from a locally dense seed protein to isolate the dense regions.
Structural Clustering Algorithm for Network (SCAN) is a new method for finding clusters or functional modules in complex network which was developed by Mete M. et al (Mete M. 2008). Their work adopted the budding yeast PPI for evaluating the effective of their algorithm. This method is based on common neighbors. Two vertices are assigned to a cluster according to how they share neighbors.
Chua et al (Chua HN. 2008) studied the indirect protein interactions between level-2 interaction. They proposed a method in both direct and indirect interactions and first weighted using topological weight to estimate the strength of functional association. Furthermore, they also proposed an algorithm for searching cliques in the modified network, and merge cliques to form clusters using a "partial clique merging" method. The findings from this work are i) indirect interactions and topological weight to augment protein-protein interactions improve the precision of clusters predicted by various clustering algorithms; and ii) this algorithm performs very well on interaction networks modified in this way.

The work of Li et al (Li XL. 2005) purposed algorithm to identify interaction graph using local clique merging. This algorithm aims to locate local cliques for each graph member (protein) and then merge the detected local cliques according to their affinity to form maximal dense regions.
Liu et al (Liu G. 2009) developed an algorithm called "Clustering-based on maximal cliques (CMC)" to find complexes from weighted PPI network. This algorithm

generates all the maximal cliques from the PPI networks, and then removes or merges highly overlapped clusters based on their interconnectivity. Their findings are (i) the iterative scoring method improve CMC performance (ii) the iterative scoring method reduce the impact of random noise on algorithm performance (iii) the iterative scoring method improve the performance of other protein complex prediction methods and reduce the impact of random noise on their performance; and (iv) this algorithm is an effective approach to protein complex prediction from protein interaction network.

The work of Maraziotis et al (maraziotis IA. 2007) presents algorithm that discovers biologically functional modules of PPI by integrating of two pieces of information which are protein interaction and microarray data. This approach firstly assigns gene expression information as weights onto the PPI network. The enriched PPI graph is observed to see their topology. This algorithm aims to reveal the functional module of the weighted graph by expanding a kernel protein set which originates from a given seed protein.

Feng et al (Feng J. 2008) purposed Graph Fragmentation Algorithm (GFA) for identifying protein complex. They combined PPI data and microarray gene expression profiles and then adapted a classical max-flow algorithm for discovering the densest sub-graphs (weight). This approach searches for large dense sub-graphs in a network of PPI, after that breaks each sub-graph into fragments iteratively by weighting its nodes in term of their corresponding log-fold changes in the microarray data until the fragment sub-graphs are sufficiently small.

Ulitsky et al (Ulitsky I. 2007) purposed algorithm for identifying functional modules, firstly they computed pair-wise similarity of gene expression patterns from microarray data, then created a network of proteins and assigned similarity values between proteins in network, finally, search for sub-networks that reach high similarity.

Cho et al (Cho YR. 2007) developed semantic similarity and semantic interactivity metrics based on Gene Ontology annotation to measure the reliability of the interaction of proteins. Weighted graph is created by assigning the reliability values to each interaction as a weight.

SWEMODE (Lubovac Z. 2006) This work identify the core modules in protein interaction network by combining functional information with topological information of the network. The weight is used to represent the strengths of interactions between proteins, their semantic similarity is calculated which based on the Gene Ontology term of proteins. This algorithm can identifies dense sub-graphs containing functionally similarity proteins based on range of nodes; the highest ranked nodes are considered as seeds for candidate modules.

DECAFF (Li XL. 2007) propose a method name "Dense-neighborhood Extraction using Connectivity and confidence Features (DECAFF)" algorithm to discover dense sub-graphs of protein interaction networks. Their experiment result with yeast protein interaction data indicates that pair-wise protein interaction networks can be effectively discovered new protein complexes.

2.1 Protein-Protein Interaction Network in Diseases Research

Wachi (Wachi S. 2005) studied differentially expressed genes in lung cancer tissues by observing the degree of distribution and centrality of the set of differentially expressed genes in human PPI network based on interolog approach (Matthews LR. 2001). Their result supports the notion that topological analysis cancer genes using

protein interaction data may provide the rationales for therapeutic targets in cancer treatments.

The work of Jonsson and Bates (Jonsson PF. 2006) states that the network topology of human proteins translated from known cancer genes is different from the network topology of undocumented proteins as being mutated in cancer. Their work also indicates that cancer proteins tend to target in central hub proteins rather than peripheral proteins, furthermore they tended to reside in larger clusters and tended to participate in more clusters than undefined-cancer proteins. These evidences also support the work of Goh et al (Goh KI. 2007) that disease genes are likely to encode hub proteins; play a central role in the human interactome and are expressed widely in most tissues.

Besides, Chuang et al (Chuang HY. 2007) used protein network based approach to identify breast cancer metastasis. A human PPI network was created from metastatic and nonmetastatic patients' information. They found sub-network markers were more than single marker genes.

Efroni et al (Efroni S. 2007) performed a related study, in which they predicted pathways associated with cancer gene expression data sets. The expression data were adopted in being score of the interaction of known pathways and the scores were used as features for make prediction. Their work different with Chuang et al (Chuang HY. 2007) in which they adopted known pathways in prediction rather than sub-networks dynamically picked up from a protein network.

Furthermore, the work of Li et al (Li BQ. 2012) studied in identifying colorectal cancer related gene. Their work combined two computational methods to identify colorectal cancer-related genes which based on i) the gene expression profiles and ii) the shortest path analysis of functional protein association networks. They found that the genes identified from both methods have more cancer genes than the genes identified from the gene expression profiles alone, and this group of genes had greater functional similarity with the reported colorectal cancer genes than another group of genes.

In addition, Feizi and Bordel (Feizi A. 2013) also studied sub-networks of metabolic and protein interaction which controlling the growth rate of cancer cells. They analyzed gene expression profiles of 60 different cell lines using several genome-scale biological networks and new algorithms. Their findings are over 100 growth-correlated metabolic sub-networks have been identified which are a key role of simultaneous lipid synthesis and degradation in the energy supply of cancer cells growth.

The previous research works mentioned above clearly proved that proteins close to one another in a network cause similar diseases. This idea is becoming an interestingly and increasingly important factor in discovery of disease genes. Various approaches to be implemented in order to identify essential proteins, different approaches adopt different kind of data, but all of them involve known disease genes (proteins) and also candidate genes (proteins). The new approaches that do not depend on prior knowledge of disease genes (proteins) are needed to discovery novel disease related genes (proteins).

2.2 K-Means Clustering

K-means (Tapas K. 2002) is one of unsupervised learning algorithms that solve clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more. Finally, this algorithm aims at minimizing an objective function knows as squared error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\left\| x_i - v_j \right\| \right)^2$$

where, ||xi - vj|| is the Euclidean distance between xi and vj, ci is the number of data points in its cluster, c is the number of cluster centers.

Steps for K-Means clustering

Let X = {x1,x2,x3,……..,xn} be the set of data points and V = {v1,v2,…….,vc} be the set of centers.

1) Randomly select 'c' cluster centers.
2) Calculate the distance between each data point and cluster centers.
3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.
4) Recalculate the new cluster center using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

where, ci represents the number of data points in i cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.
6) If no data point was reassigned then stop, otherwise repeat from step 3).

2.3 Clique Percolation Method (CPM)

The clique Percolation Method (CPM) is one of the earliest overlapping communities finding methods widely used in several different networks based on the concept of mapping the connections among unit into a graph. The idea of representing a complex system with a network is frequently used in various fields including investigations on mobile phone networks (Onnela J.P. 2007; Lambiotte R. 2008; Seshadri M. 2008), e-mail networks(Ebel H. 2002) online social networks (Aiello L.M. 2010) and also including biology, economy, etc.

Clique percolation clustering is a well known approach for analyzing the overlapping community structure of networks. This method builds up the communities from k-cliques which is fully connected sub graphs of k nodes. Any two *k*-cliques are

adjacent if they share *k-1* common nodes. A *k*-clique community is constructed by merging all possible adjacent *k*-cliques. The main advantage of this approach is that it allows overlaps between the communities, as a given node can be a member of several clusters at the same time. This characteristic can be applied to discover significant proteins that involve in more than one community.

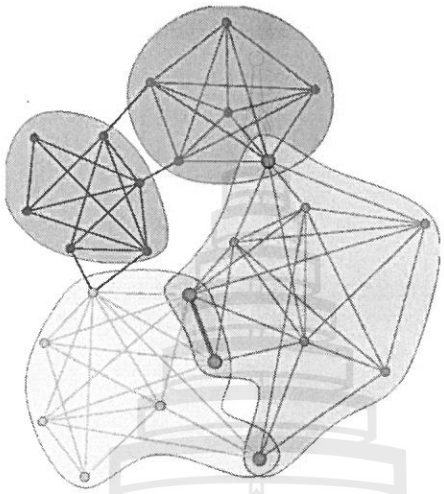

Figure 1 Illustration of the k-clique communities at k = 4.

In protein interaction networks, completely connected graphs, so-called cliques, have been found to have a high functional significance (Spirin V. 2003; Yeger-Lotem E. 2004). Motifs and cliques reveal the cores of functional modules in molecular networks. In this study, the lung cancer associated genes are predicted based on a clique percolation clustering approach to discover network motif or cluster which reveals the cores of functional modules in molecular networks. Proteins which appear in the same cluster are likely to have similar molecular functions. Therefore, we hypothesized that the proteins located in the same cluster as lung cancer proteins have a high probability in forming lung cancer as well.

Directed Clique Percolation Method (CPMd): The k nodes can be ordered such that between an arbitrary pair of them there exists a directed link pointing from the node with the higher rank towards the node with the lower rank. The directed Clique Percolation Method defines directed network communities as the percolation clusters of directed k-cliques.

2.4 MCODE (Molecular Complex Detection)

MCODE algorithm (bader GD. 2003) is a well-known automated method to find highly interconnected sub-graphs as molecular complexes or clusters in large protein-protein interaction networks. This algorithm detects densely connected regions in protein-protein interaction networks as protein complexes. Firstly, it weights every vertex based on their local neighborhood densities, and then selects seed vertices that high weights and then outward traversal (Depth-First-Search) from a dense seed protein with a high weighting value to include neighboring vertices whose weight satisfied some given threshold. The MCODE algorithm operates in three stages,

vertex weighting, complex prediction and optionally post-processing to filter or add proteins in the resulting complexes by certain connectivity criteria.
The first stage of MCODE, vertex weighting, weights all vertices based on their local network density using the highest *k*-core of the vertex neighborhood. A *k*-core is a graph of minimal degree *k* (graph G, for all *v* in G, deg(*v*) >= *k*). The highest *k*-core of a graph is the central most densely connected subgraph. We define here the term core-clustering coefficient of a vertex, *v*, to be the density of the highest *k*-core of the immediate neighborhood of *v* (vertices connected directly to *v*) including *v* (note that Ci does not include*v*). The core-clustering coefficient is used here instead of the clustering coefficient because it amplifies the weighting of heavily interconnected graph regions while removing the many less connected vertices that are usually part of a biomolecular interaction network, known to be scale-free. A scale-free network has a vertex connectivity distribution that follows a power law, with relatively few highly connected vertices (high degree) and many vertices having a low degree. A given highly connected vertex,*v*, in a dense region of a graph may be connected to many vertices of degree one (singly linked vertex). These low degree vertices do not interconnect within the neighborhood of *v* and thus would reduce the clustering coefficient, but not the core-clustering coefficient. The final weight given to a vertex is the product of the vertex core-clustering coefficient and the highest *k*-core level, *k*max, of the immediate neighborhood of the vertex. This weighting scheme further boosts the weight of densely connected vertices. This specific weighting function is based on local network density. Many other functions are possible and some may have better performance for this algorithm but these are not evaluated here.

The second stage, molecular complex prediction, takes as input the vertex weighted graph, seeds a complex with the highest weighted vertex and recursively moves outward from the seed vertex, including vertices in the complex whose weight is above a given threshold, which is a given percentage away from the weight of the seed vertex. This is the vertex weight percentage (VWP) parameter. If a vertex is included, its neighbours are recursively checked in the same manner to see if they are part of the complex. A vertex is not checked more than once, since complexes cannot overlap in this stage of the algorithm. This process stops once no more vertices can be added to the complex based on the given threshold and is repeated for the next highest unseen weighted vertex in the network. In this way, the densest regions of the network are identified. The vertex weight threshold parameter defines the density of the resulting complex. A threshold that is closer to the weight of the seed vertex identifies a smaller, denser network region around the seed vertex.

The third stage is post-processing. Complexes are filtered if they do not contain at least a 2-core (graph of minimum degree 2). The algorithm may be run with the 'fluff' option, which increases the size of the complex according to a given 'fluff' parameter between 0.0 and 1.0. For every vertex in the complex, *v*, its neighbors are added to the complex if they have not yet been seen and if the neighborhood density (including *v*) is higher than the given fluff parameter. Vertices that are added by the fluff parameter

are not marked as seen, so there can be overlap among predicted complexes with the fluff parameter set. If the algorithm is run using the 'haircut' option, the resulting complexes are 2-cored, thereby removing the vertices that are singly connected to the core complex. If both options are specified, fluff is run first, then haircut.

Resulting complexes from the algorithm are scored and ranked. The complex score is defined as the product of the complex subgraph, C = (V,E), density and the number of vertices in the complex subgraph (DC × |V|). This ranks larger more dense complexes higher in the results.

There are two important parameters of MCODE which are node score cutoff and fluff. Node score cutoff is used to control how new nodes are added to a module. The default value is set to 0.2, which means the new node score must be at least eighty percent that of the modules seed node score. A setting of 0.1 makes it harder for new nodes to join a module, therefore, creating smaller modules. Once a module is found, fluff parameter is set for adding nodes that have a node score of fifty percent of the original seed node score, and can be used to grow the module. The node score cutoff is the most important parameter for deciding one the module shape and size. The higher value of the node score cutoff, coupled with adding fluff parameter would be good to identify the pathway interacting modules. Since many proteins need only interact with just one member of a complex to phosphorylate its target.

Figure 2 shows the effect of adding Fluff to a module, the left module was defined from the human interactome using a node score threshold of 0.2. The right module is the same seed module, after applying fluff setting to 0.5, where the white nodes have been added. In general, the fluff nodes are connected to the seed module via a single edge. Figure 3 shows the effect of lowering the node score threshold from 0.2 to 0.1. The central module was defined from the human interactome using a node score threshold of 0.2. From this network, four tighter, more coherent sub-networks were identified when running MCODE with node score threshold is 0.2.

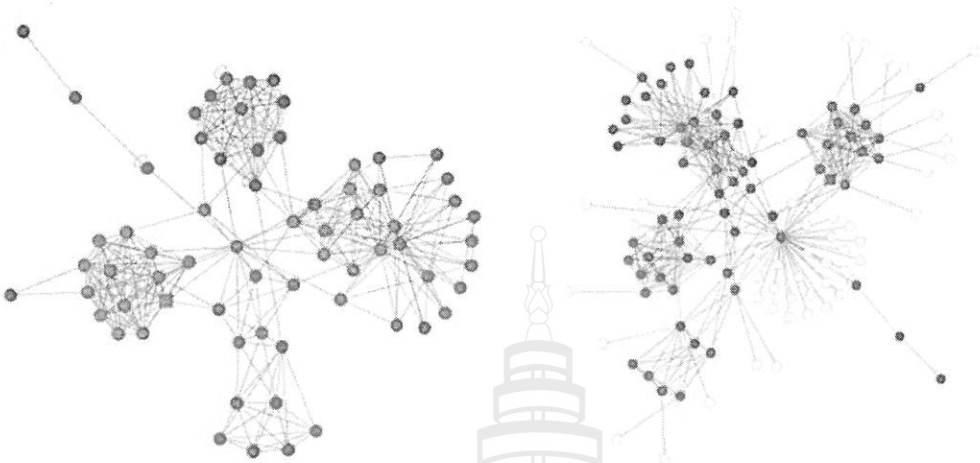

Figure 2 The effect of adding Fluff to a module

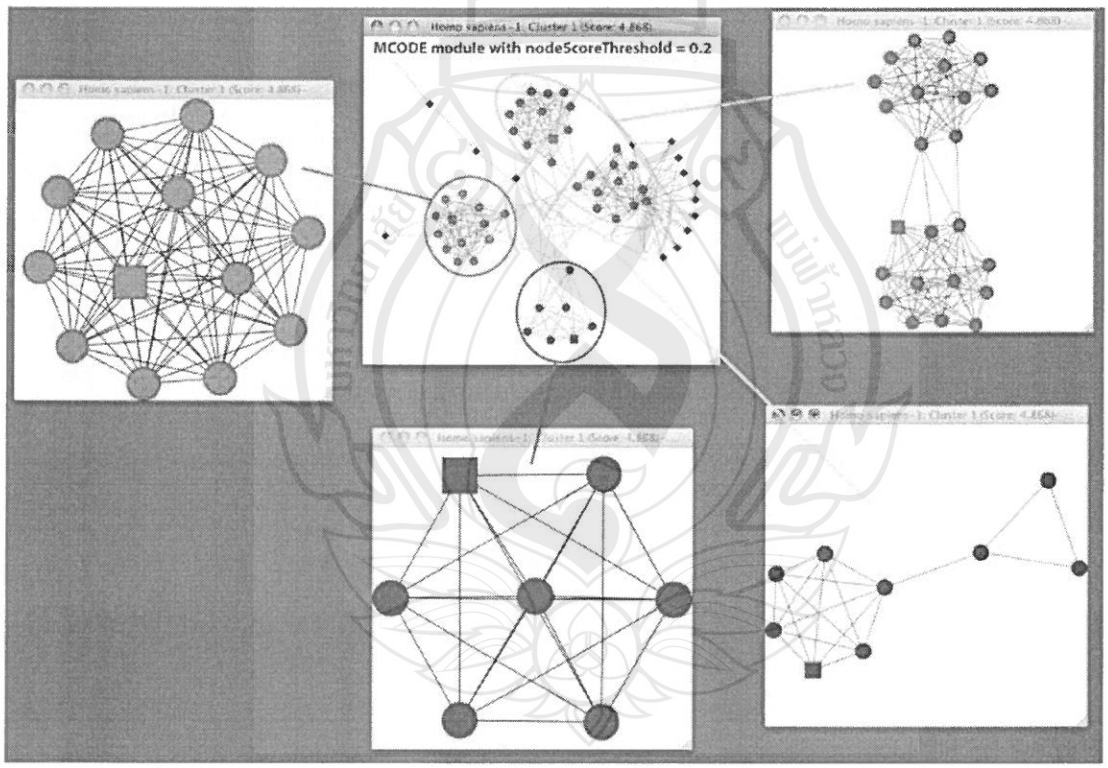

Figure 3 the effect of lowering node score ratio from 0.2 to 0.1

CHAPTER 3
METHODOLOGY

3.1 Data Source

A collection of experimentally confirmed lung cancer proteins was obtained from two resources i.e. Online Mendelian Inheritance in Man (OMIM: http://www.ncbi.nlm.nih.gov/omim) and Lung Cancer Database (Wang L. 2010) (HlungDB: (http://www.megabionet.org/bio/hlung/index.jsp). A total of experimentally confirmed human PPIs was obtained from BioGrid (Database of protein and genetic interactions: http://www.thebiogrid.org) (Stark C. 2005). The Onco-Protein (OCP) and Tumor Suppressor Protein (TSP) data are derived from the following three databases: (1) Tumor Associated Gene database of Taiwan national Cheng Kung University (http://www.binfo.ncku.edu.tw/TAG/), (2) Memorial Sloan-Kettering Cancer Center and (3) National Yang Ming University. This research collected 656 OCP and 1,024 TSP.

3.2 Research Overview System Flowchart

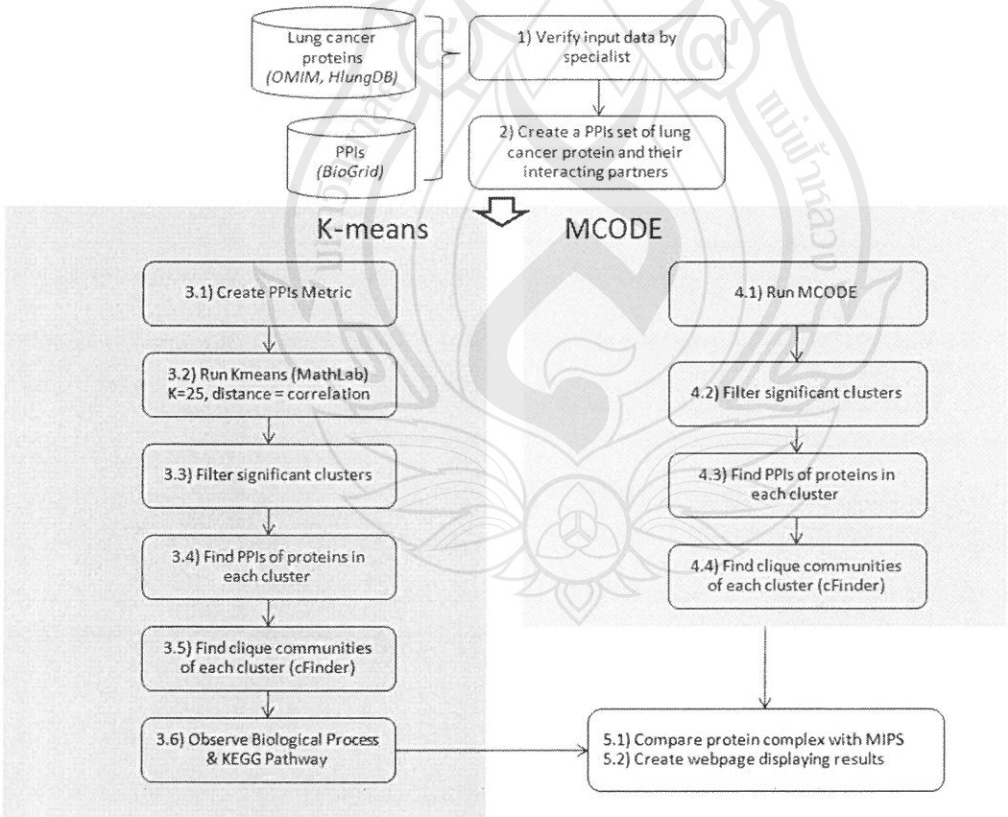

Figure 4 System Flowchart

As a guideline of describing the methodology section, a brief content summarizing each of the following steps is given. A flowchart depicting a methodology structure is presented in figure 4.

1. Verification and Pre-Processing of Input Data
2. Construct a set of lung cancer protein-protein interaction
3. K-Means Clustering Process
4. MCODE Clustering Process
5. Identification of Protein Complex
6. Gene Set Enrichment Analysis (GSEA)
7. Identification of cancer-related proteins

3.3 Verification and Pre-Processing of Input Data

A set of lung cancer proteins (2,683 proteins) was extracted from two different sources which are OMIM and HLungDB, besides, a set of 159,840 homo-sapiens protein-protein interactomes was gathered from another BioGrid. Before clustering processes, we verified a set of lung cancer proteins whether each identified lung cancer protein was proved by at least two literature references. We focused in this process to make sure that our initial input data set is reliable by literature search and credible by specialists. This process is able to protect garbage in garbage out problem.

3.4 Construct a set of lung cancer protein-protein interaction

Lung cancer proteins were merged with their protein interacting partners from bioGrid, and then a set of 76,360 lung cancer protein-protein interactions was obtained. Figure5 shows how to merge lung cancer proteins with their interacting partners.

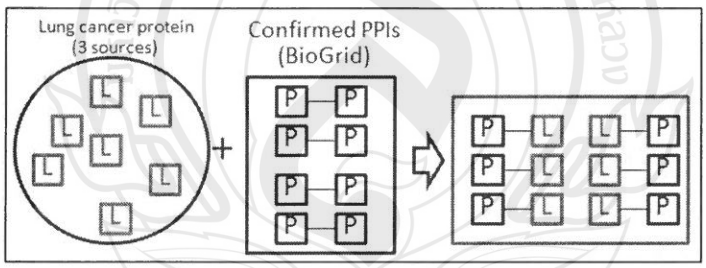

Figure 5 Lung Cancer protein are merged with their interacting partners

3.5 K-means Clustering Process

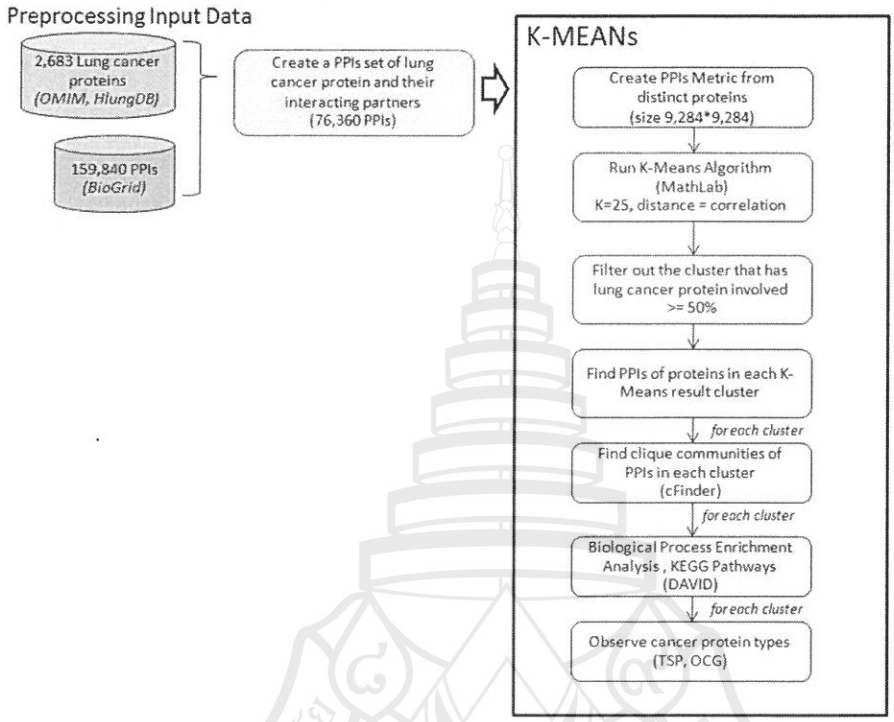

Figure 6 K-MEANs Flowchart

1) Construct Lung PPI Metric; to call K-means algorithm of MATLAB needs to re-format PPI input data in a metric format. Java programming language was adopted to create the PPIs metric. The metric of 9,284 in size was obtained from the JAVA programming script. Figure 7 shows the original file format of Lung cancer PPI data composing of protein A which interacts to protein B. Figure 8 indicates the metric format of PPI data. Zero (0) represents non-binding between two proteins in such row and column, one (1) represents such pair of protein interact to each other.

Protein A	Protein B
A2M	ADAMTS1
A2M	APOE
A2M	IL10
A2M	IL4
A2M	LCAT
A2M	LEP
A2M	NGF
A2M	PAEP
ABCB1	DHX9
ABCB1	PIM1
ABCB1	UBC
ABCB7	FECH
ABCE1	RNASEL
ABCE1	UBASH3A
ABL1	ABI1
ABL1	ABI2
ABL1	ABL1
ABL1	ABL2

Figure 7 the original file format of protein-protein interaction data

Figure 8 the metric file format of protein-protein interaction data

2) Run K-Means algorithm; K-Means algorithm was adopted for clustering Lung cancer PPI into sub-clusters, measurement for clustering is set to correlation distance. We set distance measurement to be "correlation" which consist to the work of Goele Hollanders (Hollanders 2005). This work did an experiment on comparison of clustering performance of the K-means algorithm run with two different distance measurements which are squared Euclidean distance and correlation distance on microarray data of gene interaction. Their result indicates that the centroids obtained from the correlation distance give good indications of the different type of influences in a genetic regulatory system. This evidence support our goal in which we aimed to distinguish a set of proteins associated to lung cancer from other proteins.
The value of K or the number of output sub-cluster was set to 25 which is the maximum number of cluster based on our input data. We also further did an experiment to prove whether the correlation distance gives the best clustering performance on the assumption that balancing the number of member in each cluster which is indicating efficiency in clustering more than imbalance of such number. By doing this, the number of output cluster was fix to 10, then various types of distance method; city-block, Euclidian, correlation, hamming, and cosine were set for running on input data. To avoid bias, we run each method for ten times, and then take an average of the number of each cluster. From our evidence as table 1 indicates that correlation distance give the best variance value.

Table 1 Variation in the number of member in various methods

method	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5	cluster 6	cluster 7	cluster 8	cluster 9	cluster 10	sum	variance $*10^{-5}$
Cityblock	7256	1	89	1642	37	204	35	5	13	2	9284	46.78
EU	6607	3	165	49	1091	1026	304	1	1	37	9284	37.40
correlation	839	836	1520	836	622	803	814	862	734	1418	9284	0.77
Hamming	7152	42	1	137	19	34	289	14	16	1580	9284	45.13
Cosine	1934	1174	1311	412	173	716	1245	547	991	781	9284	2.37

3) Filter out the output clusters which have the involvement of Lung cancer protein more than 50% of the proteins in a cluster.
4) Find PPIs of proteins in each cluster; protein members in each cluster were merged with their protein interacting partners that present in same cluster. Ideally, proteins which are grouped in same cluster might not have the linkages to proteins in a cluster, some have the linkages, and some don't have. Thus, to find significant set of protein-protein interaction in a cluster, we filtered out only the proteins that have their interacting partners present in same cluster. Figure 9 shows protein-protein interaction in same cluster.

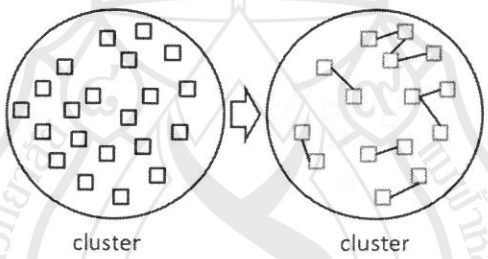

Figure 9 protein-protien interaction in same cluster

(blue squres indicate proteins with their interacting partners, yellow squres indicate non interacting proteins)

5) Find Clique community in result cluster; we adopted cFinder software to discover the community of protein in cluster. cFinder software applies the concept of Clique Percolation Clustering Method (CPM) to find dense region in a protein-protein interaction network. A set of PPI of each cluster was submitted to cFinder software to search for clique communities.
6) Observe enriched biological process by DAVID; a list of protein of each cluster was submitted to DAVID for identifying their enriched biological processes and also KEGG Pathways.
7) Observe cancer protein types (tumor suppressor protein or onco-protein) involved in a cluster.

3.6 MCODE Clustering Process

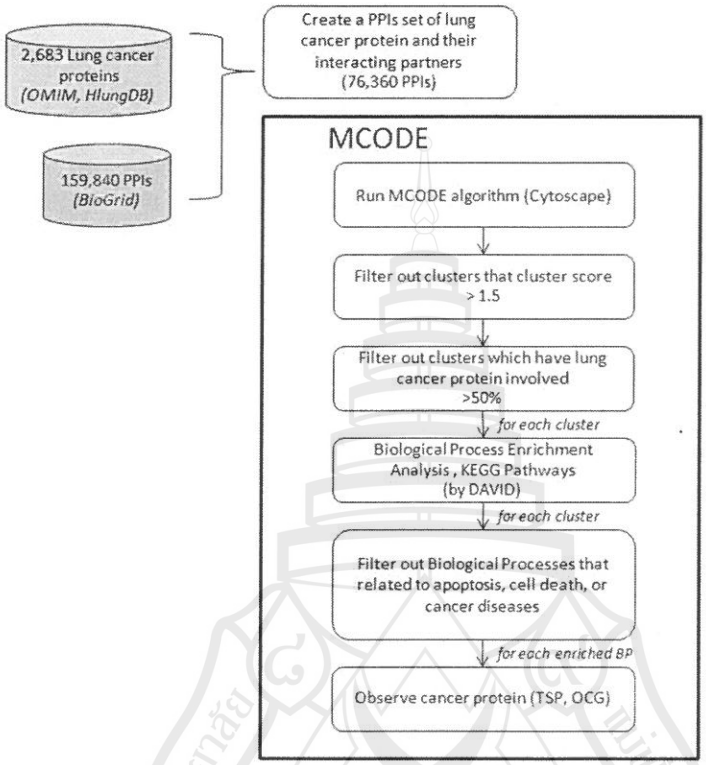

Figure 10 MCODE Flowchart

1) Initially, a list of 76, 360 PPIs lung cancer protein-protein interaction was adopted in this calculation, this data set was submitted to allegro-MCODE plugin of Cytoscape software (bader GD. 2003) to find highly interconnected regions or cluster in a network.
2) Resulting complexes from MCODE are scored and ranked. The complex score of each cluster is the product of the complex subgraph, C= (V, E), density and the number of nodes in the complex sub-graph. Figure 11 shows list of clusters resulted by MCODE algorithm.

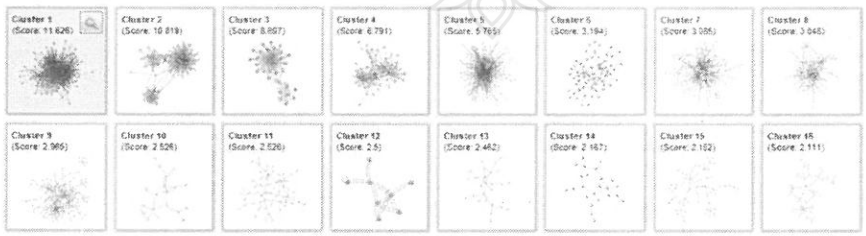

Figure 11 cluster list resulted from MCODE algorithm

3) Clusters with their score greater than 1.5 was filter out to be determined in our experiment.
4) Clusters with associated by greater than 50% of involving lung cancer proteins was filter out to be determined in our experiment.
5) Enrichment biological process and KEGG pathway analysis by DAVID software was evaluated for protein-protein interaction in each cluster. By doing this, the experimentally confirmed protein partners of each protein in a cluster was mapped to their partners, then submitted all protein-protein interaction pairs of a cluster into DAVID to identify their enriched biological process and KEGG pathways with p-value set to 0.005.
6) After we got a list of biological process or KEGG pathway of each cluster that satisfy 0.005 of p-value, only the biological processes or KEGG pathways that related to apoptosis, cell death, or any processes reported related to cancer were extracted to be determined in our experiment.
7) Undefined lung cancer proteins in cluster were observed their linkage to cancer protein.
8) For each biological process (or KEGG pathway), we observed protein type of involving proteins (tumor suppressor protein or onco-protein).

3.7 Identification of Protein Complexes

In this study, we compared the clustering results with known protein complexes obtained from The MIPS Mammalian Protein-Protein Interaction Database (MIPS) (Pagel P. 2005) which is a database of high-quality published experimental evidence of protein interaction data in mammals in order to identify realistic cancer-related protein modules. Subunits from k-community are compared with the MIPS protein complexes. The Jaccard Index (JI) is a quantity which is used to quantify the similarity between two sets, hence, given two modules A and B the JI is given by:

$$JI(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

where $|A \cap B|$ and $|A \cup B|$ denote the cardinality of $|A \cap B|$ and $|A \cup B|$ respectively. It is noted that JI lies between 0% and 100%.

3.8 Gene Set Enrichment Analysis (GSEA)

The functional annotation of the lung cancer PPIs is given by implementing The Database for Annotation, Visualization and Integrated Discovery, i.e. DAVID (huang W.da 2009). DAVID provides functional annotation tools which mainly provide typical batch annotation and gene ontology (GO) term enrichment analysis to highlight the most relevant GO terms associated with a given gene list.

In order to investigate the enriched biological processes of proteins in clusters, the proteins of each cluster was submitted into DAVID, and then the tool clustered redundant annotation terms of the protein list. The proteins with an enriched biological processes e-value less than or equal to 0.05 of each were examined in this work.

3.9 Identification of cancer-related proteins

The oncoprotein (OCP) and tumor suppressor protein (TSP) data are derived from the following three databases: (1) Tumor Associated Gene database of Taiwan national Cheng Kung University (http://www.binfo.ncku.edu.tw/TAG/), (2) Memorial Sloan-Kettering Cancer Center and (3) National Yang Ming University.

CHAPTER 4
RESULTS

4.1 K-Means

There are 25 output clusters from K-Means algorithm, but there is only one cluster that has lung cancer protein involved more than fifty percent. This cluster involves 1,056 proteins in total, 983 proteins are lung cancer proteins (93% involvement of lung cancer protein). Therefore, we focused in this cluster for further study.

4.1.1 Clique Percolation Clustering Network Analysis

There are 18 clique communities from a K-Mean clustering process. A list of non-Lung cancer protein involved those communities are UBC, COPS6, CUL2, NRP1, and SH3GL3. See detail in table2.

UBC was recorded in uniProt that it involves in DNA damage response, by inducing the cell cycle regulator phosphoprotein p53 in response to the detection of DNA damage and resulting in the stopping or reduction of cell cycle rate.

The work of Park et al (Park SW. 2009) indicates that somatic mutation of CUL2 occurs in a fraction of colorectal cancers and this protein may play a central role in HIF1alph activation in gastric, colorectal, breast, lung and hepatocellular carcinomas, and acute leukemias.

SH3GL3 was reported in the work of Fang et al (Fang WJ. 2012) that it exhibited hypermethylation in its promoter region; this evidence support previous studies that SH3GL3 is significantly associated with colorectal cancer.

Table 2 Clique Community List

community ID	involved protein	lung cancer protein
1	UBC	No
	CD74	Yes
	MIF	Yes
2	UBC	No
	CSTB	Yes
	CTSH	Yes
3	UBC	No
	ADAMTS1	Yes
	VEGFA	Yes
	CTGF	Yes
4	UBC	No
	FGF2	Yes
	GPC3	Yes
	RPS19	Yes
7	UBC	No
	MVP	Yes
	PARP4	Yes
8	VEGFA	Yes
	PGF	Yes
	NRP1	No
9	UBC	No
	SPINT1	Yes
	ST14	Yes
10	UBC	No
	BCL2	Yes
	ITM2B	Yes
11	UBC	No
	DNAJB4	Yes
14	UBC	No
	HLA-DMA	Yes
	HLA-DRB1	Yes
15	UBC	No
	CIT	Yes
	RHOC	Yes
16	UBC	No
	FABP5	Yes
	S100A7	Yes
17	UBC	No
	LAMP1	Yes
	LAPTM5	Yes
18	UBC	No
	AKAP12	Yes

	SDC4	Yes
5	UBC	No
	GSTM1	Yes
	GSTM2	Yes
6	UBC	No
	COPS6	No
	IFI27	Yes
	SKP2	Yes
	CUL2	No
	TCEB1	Yes

	PABPN1	Yes
12	SH3GL1	No
	DPYSL4	Yes
	PTPRO	Yes
	SH3GL3	No
13	UBC	No
	FAS	Yes
	PDCD6	Yes
	HEBP2	Yes

	FHL1	Yes

Biological Process Enrichment Analysis and KEGG Pathway Analysis

There are 1,193 PPIs in observing cluster (cluster number 20); the list of distinct proteins was submitted to DAVID software for investigating the enriched biological processes of this protein list. From our evidence as table 3 indicates that most of biological process that related to cancers involved almost 100% of lung cancer proteins.

Table 3 List of enriched biological Process of clique protein community

Category	Term	Count	%	PValue	Involved proteins
GOTERM_ BP_FAT	GO:0010941~regulation of cell death	70	13.11	6.68E-12	XRCC4, HRAS, TP63, PAWR, TGFB1, TGFB2, ACVR1B, PCGF2, CASP3, NOD2, DYNLL1, CASP9, CD44, PCBP4, APOE, HMOX1, RHOA, PIK3CA, NOS3, FAS, FGF2, API5, TERT, PRKCA, IRAK1, PRAME, CYCS, PRKCI, SKP2, FADD, PIM2, ECT2, ARHGEF11, TNFRSF10A, TNFRSF10B, BTG2, RASGRF1, UNC13B, NMNAT1, MCL1, CLU, CALR, CD74, MIF, PEA15, ERCC5, PPP2CB, THBS1, NEFL, BMP4, TXNIP, PTPRC, SMAD6, KLF10, LGALS1, BIRC6, BIRC5, MALT1, SOD1, TAX1BP1, SOD2, ATF5, NRAS, CASP10, DUSP1, BAX, PLCG2, ID3, BMP7, PDCD6
GOTERM_ BP_FAT	GO:0043067~regulation of programmed cell death	69	12.92	1.57E-11	XRCC4, HRAS, TP63, PAWR, TGFB1, TGFB2, ACVR1B, PCGF2, CASP3, NOD2, DYNLL1, CASP9, CD44, PCBP4, APOE, HMOX1, RHOA, PIK3CA, NOS3, FAS, FGF2, API5, TERT, PRKCA, IRAK1, PRAME, CYCS, PRKCI, SKP2, FADD, PIM2, ECT2, ARHGEF11, TNFRSF10A, TNFRSF10B, BTG2, RASGRF1, UNC13B, NMNAT1, MCL1, CLU, CALR, CD74, MIF, PEA15, ERCC5, PPP2CB, THBS1, NEFL, TXNIP, PTPRC, SMAD6, KLF10, LGALS1, BIRC6, BIRC5, MALT1, SOD1, TAX1BP1, SOD2,

					ATF5, NRAS, CASP10, DUSP1, BAX, PLCG2, ID3, BMP7, PDCD6
GOTERM_ BP_FAT	GO:0042981~regulation of apoptosis	67	12.54	7.84E-11	XRCC4, HRAS, TP63, PAWR, TGFB1, TGFB2, ACVR1B, PCGF2, CASP3, NOD2, DYNLL1, CASP9, CD44, PCBP4, APOE, HMOX1, RHOA, PIK3CA, NOS3, FAS, API5, TERT, PRKCA, IRAK1, PRAME, CYCS, PRKCI, SKP2, FADD, PIM2, ECT2, ARHGEF11, TNFRSF10A, TNFRSF10B, BTG2, RASGRF1, UNC13B, NMNAT1, MCL1, CLU, CALR, CD74, MIF, PEA15, ERCC5, PPP2CB, THBS1, NEFL, TXNIP, PTPRC, SMAD6, KLF10, LGALS1, BIRC6, BIRC5, MALT1, SOD1, TAX1BP1, SOD2, ATF5, NRAS, CASP10, DUSP1, BAX, ID3, BMP7, PDCD6
GOTERM_ BP_FAT	GO:0043069~negative regulation of programmed cell death	39	7.30	1.27E-09	XRCC4, HRAS, MCL1, CLU, TP63, CD74, MIF, PEA15, PCGF2, ERCC5, CASP3, APOE, HMOX1, PPP2CB, RHOA, PIK3CA, NOS3, FAS, THBS1, NEFL, API5, TERT, IRAK1, PRAME, SMAD6, PRKCI, SKP2, BIRC6, MALT1, BIRC5, PIM2, SOD1, TAX1BP1, SOD2, ATF5, NRAS, BTG2, BAX, PLCG2
GOTERM_ BP_FAT	GO:0060548~negative regulation of cell death	39	7.30	1.37E-09	XRCC4, HRAS, MCL1, CLU, TP63, CD74, MIF, PEA15, PCGF2, ERCC5, CASP3, APOE, HMOX1, PPP2CB, RHOA, PIK3CA, NOS3, FAS, THBS1, NEFL, API5, TERT, IRAK1, PRAME, SMAD6, PRKCI, SKP2, BIRC6, MALT1, BIRC5, PIM2, SOD1, TAX1BP1, SOD2, ATF5, NRAS, BTG2, BAX, PLCG2
GOTERM_ BP_FAT	GO:0043066~negative regulation of apoptosis	38	7.11	3.00E-09	XRCC4, HRAS, MCL1, CLU, TP63, CD74, MIF, PEA15, PCGF2, CASP3, ERCC5, APOE, HMOX1, PPP2CB, RHOA, PIK3CA, NOS3, FAS, THBS1, NEFL, API5, TERT, IRAK1, PRAME, SMAD6, PRKCI, SKP2, BIRC6, MALT1, BIRC5, PIM2, SOD1, TAX1BP1, SOD2, ATF5, NRAS, BTG2, BAX
GOTERM_ BP_FAT	GO:0006916~anti-apoptosis	24	4.49	9.88E-07	IRAK1, MCL1, CLU, PRKCI, SKP2, BIRC6, TP63, MALT1, BIRC5, PIM2, SOD1, TAX1BP1, SOD2, ATF5, PEA15, APOE, HMOX1, BAX, PIK3CA, NOS3, FAS, THBS1, API5, TERT
GOTERM_ BP_FAT	GO:0010942~positive regulation of cell death	33	6.17	6.76E-05	TP63, PAWR, TGFB1, TGFB2, ACVR1B, CASP3, CD44, CASP9, DYNLL1, APOE, PCBP4, HMOX1, FAS, TXNIP, BMP4, PRKCA, PTPRC, KLF10, SKP2, FADD, SOD1, ECT2, ARHGEF11, TNFRSF10A, CASP10, TNFRSF10B, DUSP1, RASGRF1, BAX, ID3, BMP7, UNC13B, PDCD6
GOTERM_	GO:0043065~positive regulation	32	5.99	1.27E-04	TP63, PAWR, TGFB1, TGFB2, ACVR1B, CASP3,

BP_FAT	of apoptosis				CD44, CASP9, DYNLL1, APOE, PCBP4, HMOX1, FAS, TXNIP, PRKCA, PTPRC, KLF10, SKP2, FADD, SOD1, ECT2, ARHGEF11, TNFRSF10A, CASP10, TNFRSF10B, DUSP1, RASGRF1, BAX, ID3, BMP7, UNC13B, PDCD6
GOTERM_ BP_FAT	GO:0043068~positive regulation of programmed cell death	32	5.99	1.43E-04	TP63, PAWR, TGFB1, TGFB2, ACVR1B, CASP3, CD44, CASP9, DYNLL1, APOE, PCBP4, HMOX1, FAS, TXNIP, PRKCA, PTPRC, KLF10, SKP2, FADD, SOD1, ECT2, ARHGEF11, TNFRSF10A, CASP10, TNFRSF10B, DUSP1, RASGRF1, BAX, ID3, BMP7, UNC13B, PDCD6
GOTERM_ BP_FAT	GO:0008219~cell death	54	10.11	2.43E-07	HRAS, TP63, GJA1, PAWR, PDCD4, TGFB1, TGFB2, CASP3, CASP9, DYNLL1, CXCR4, GSN, HMOX1, RHOB, FAS, FGF2, API5, YARS, CYCS, FADD, IL24, PIM2, ECT2, BCAP31, ARHGEF11, TNFRSF10A, TNFRSF10B, RASGRF1, BUB1B, CTSD, SIAH2, GADD45B, MCL1, ALDOC, CLU, ITGB2, ARF6, PEA15, PEG10, THBS1, LGALS1, GARS, BIRC6, BIRC5, SOD1, ITPR1, TAX1BP1, SOD2, ATXN1, NRAS, CASP10, BAX, PARP4, PDCD6
GOTERM_ BP_FAT	GO:0016265~death	54	10.11	2.99E-07	HRAS, TP63, GJA1, PAWR, PDCD4, TGFB1, TGFB2, CASP3, CASP9, DYNLL1, CXCR4, GSN, HMOX1, RHOB, FAS, FGF2, API5, YARS, CYCS, FADD, IL24, PIM2, ECT2, BCAP31, ARHGEF11, TNFRSF10A, TNFRSF10B, RASGRF1, BUB1B, CTSD, SIAH2, GADD45B, MCL1, ALDOC, CLU, ITGB2, ARF6, PEA15, PEG10, THBS1, LGALS1, GARS, BIRC6, BIRC5, SOD1, ITPR1, TAX1BP1, SOD2, ATXN1, NRAS, CASP10, BAX, PARP4, PDCD6
GOTERM_ BP_FAT	GO:0006915~apoptosis	46	8.61	1.44E-06	HRAS, MCL1, ALDOC, CLU, GJA1, TP63, ARF6, ITGB2, PAWR, PDCD4, PEA15, CASP3, PEG10, CASP9, DYNLL1, CXCR4, GSN, RHOB, FAS, THBS1, FGF2, API5, YARS, LGALS1, CYCS, BIRC6, BIRC5, FADD, IL24, PIM2, SOD1, ECT2, TAX1BP1, BCAP31, ARHGEF11, SOD2, TNFRSF10A, CASP10, NRAS, TNFRSF10B, RASGRF1, BAX, BUB1B, SIAH2, GADD45B, PDCD6
GOTERM_ BP_FAT	GO:0012501~programmed cell death	46	8.61	2.16E-06	HRAS, MCL1, ALDOC, CLU, GJA1, TP63, ARF6, ITGB2, PAWR, PDCD4, PEA15, CASP3, PEG10, CASP9, DYNLL1, CXCR4, GSN, RHOB, FAS,

					THBS1, FGF2, API5, YARS, LGALS1, CYCS, BIRC6, BIRC5, FADD, IL24, PIM2, SOD1, ECT2, TAX1BP1, BCAP31, ARHGEF11, SOD2, TNFRSF10A, CASP10, NRAS, TNFRSF10B, RASGRF1, BAX, BUB1B, SIAH2, GADD45B, PDCD6
GOTERM_ BP_FAT	GO:0060284~regulation of cell development	20	3.74	1.15E-04	BMP4, XRCC4, CDH2, CALR, TTC3, TGFB1, TGFB2, THY1, ACTR3, ATF5, CCND2, APOE, BAX, NTRK2, RHOA, AGRN, BMP7, FGF2, NEFL, DBN1

4.1.2 Identification of proteins interacting to OCP and TSP

The cause of cancer is closely related to the gain of OCP function or the lost of TSP function. The cause of disease is associated with many proteins and there are great chances that these proteins are regulated in biological processes or functions. Previous researches have suggested that if two proteins involving in the same PPI have highly similarity in their biological function, therefore, if a protein is associated to Lung cancer forming, then its partners in PPI are also likely connected to the lung cancer.

From our evidence, we found that some non-lung cancer proteins interact to OCP and TSP, therefore, those non lung cancer proteins might have significant role in associating in causing disease as well. Table 4 lists some of non-lung cancer proteins that are interacting partners of OCP and TSP.

Interestingly, our result shows that UBC, COPS6 and SH3GL3 interact to set of OCP and TSP, this evidence supports these proteins might have importance role in lung cancer formation. Besides, it was found that CUL2 is defined as TSP and NRP1 is defined as TSP and OCP, this evidence also supports that these two proteins have highly possibility to associate in lung cancer as well.

Table 4 Protein Type of Interacting Proteins

Non-lung cancer protein	Interacting protein	Interacting protein type OCP (Onco-Protein) TSP (Tumor Suppressor Protein)
UBC	VEGFA	OCP
	CTGF	TSP/OCP
	FGF2	OCP
	GPC3	TSP
	GSTM1	TSP
	TCEB1	TSP
	MVP	TSP
	ST14	TSP
	BCL2	TSP/OCP
	DNAJB4	TSP
	FAS	TSP
	RHOC	OCP
	AKAP12	TSP
ADAMTS1	VEGFA	OCP
FHL1	AKAP12	TSP
ITM2B	BCL2	TSP/OCP
CIT	RHOC	OCP
COPS6	TCEB1	TSP
PABPN1	DNAJB4	TSP
DPYSL4	PTPRO	TSP
PDCD6	FAS	TSP
RPS19	FGF2	OCP

SDC4	FGF2	OCP
GSTM2	GSTM1	TSP
PARP4	MVP	TSP
PGF	NRP1	TSP/OCP
SH3GL3	PTPRO	TSP
SH3GL1	PTPRO	TSP
SKP2	CUL2	TSP
SKP2	TCEB1	TSP
SPINT1	ST14	TSP
PGF	VEGFA	OCP
VEGFA	NRP1	TSP/OCP

4.2 MCODE

There are 383 output clusters from MCODE algorithm, out of 127 clusters satisfy 1.5 of cluster score, out of 32 clusters satisfy 50% of the involvement of lung cancer proteins. Among these 32 clusters, there are only 12 clusters which satisfy 0.005 of p-value in enriched biological processes predicted by DAVID.

There are 7 significant clusters out of 12 clusters which have related cancer biological processes. Therefore, we focused on those clusters to insight observe proteins involved in related cancer biological processes.

4.2.1 MCODE Clustering Network

In a protein-protein interaction network, proteins are represented as nodes, some nodes interact with many more partners than average; these proteins are called hubs (Albert R. 2005). The work of Sun and Zhao (Sun J. 2010) states that cancer-related protein tended to have higher degree of connecting to other proteins, and also higher in betweeness, shortest-path distance. Their result imply that hub protein or protein which has highly interconnection help in identification of cancer candidate protein prioritization and verification, biomarker discovery and to reveal insight system biological system of cancer protein.

MCODE algorithm identified the seed of each cluster which the node that densely connecting to other nodes. Table 4 lists all seed of result clusters.

PTPN11 protein (tyrosine phosphatase) encoding SHP2 was reported by that SHP2 is a drugable target for the treatments of PTPN11-associated diseases (Xu D. 2013). Besides, the work of Tartaqlia et al (Tartaqlia M. 2001) also reports that the mutation of PTPN11 cause Noonan syndrome which is an autosomal dominant disorder characterized by dysomorphic facial features, proportionate short stature and heart disease.

The work of Giri et al (Giri K. 2014) reported that silencing of PPA1 by the siRNA approach significantly inhibited proliferation of ovarian cancer cells.

RICTOR was studied by Dao et al (Gao D. 2010), this work found that the Rictor/Cullin-1 E3 ligase activity is regulated by a signal that relay cascade and the error-regulation of this mechanism may contribute the overexpression of SGK1 in various human cancers.

NOTCH2 was studied by Baumgart et al (Baumgart A. 2014) , their result highlights the role of this protein in lung cancer.
USP8 was studied by (Byun S. 2013), their result show that the inhibitor of USP8 activity or reduction in USP8 expression can kill NSCLC (non small cell lung cancer) cells and their suggest USP8 as a potential therapeutic target for gefitinib-resistant and sensitive NSCLC cells.

Table 5 Seed protein in clusters

Cluster No.	Seed protein	Cancer Protein	Node Density	Node Score Ratio	Node Score
1	PTPN11	No	0.219723183	0.523152022	13.5625
4	HIF1A	Yes	0.123371056	0.656240165	17.0127551
8	PPA1	No	0.59	0.298944013	7.75
10	KRT6B	No	0.530612245	0.188045427	4.875
12	B2M	Yes	0.226666667	0.157379558	4.08
17	RICTOR	No	0.354166667	0.195010073	5.055555556
21	BAZ1A	Yes	0.3075	0.285729044	7.407407407
22	FLT4	Yes	0.345679012	0.216975493	5.625
23	USP50	No	0.32	0.1758948	4.56
25	BAMBI	Yes	0.298611111	0.154293684	4
26	ZNF579	No	0.24691358	0.120006199	3.111111111
27	UNC13B	Yes	0.177514793	0.086404463	2.24
32	METTL18	No	0.194444444	0.246869894	6.4
43	NOTCH2	No	0.28125	0.120006199	3.111111111
46	USP8	No	0.231111111	0.226811715	5.88
49	IPO8	Yes	0.208888889	0.160722588	4.166666667
60	CAV2	Yes	0.378698225	0.231440526	6
63	F13A1	Yes	0.26446281	0.168758717	4.375

4.2.2 Biological Process Enrichment Analysis and KEGG Pathway Analysis

Clustering protein-protein interaction networks can be useful for discovering groups of interacting proteins that participate in the same biological processes or perform together in specific biological functions. The functional annotation of our protein-protein interaction was given by the DAVID (huang W.da 2009) which accepts batch annotation and conducts GO term enrichment analysis. Sets of proteins involved in the network were submitted to DAVID for clustering of the annotation terms. With such the enriched biological processes related to protein list were obtained. Table 5 lists the enriched biological process of proteins involved in our significant clusters.

Table 6 Cluster 1: Protein-Protein Interaction Clustering Networks

Cluster 1					
Category	Term	Count	%	PValue	Involved proteins
GOTERM _BP_FAT	GO:0042981~regulation of apoptosis	34	18.99	1.99E-09	*Lung Cancer Protein:* PRKDC, HSPA1A, HSPE1, HSPA5, MYC, RASA1, CFLAR, CEBPB, MSH2, CREB1, ACTN1, YWHAE, TP73, TNFRSF10A, TNFSF10, HDAC1, JUN, HSPD1 *Predicted Lung Cancer Protein:* DPF2, ERBB2, NR3C1, DAXX, SART1, RPS3, VDR, PPP2CA, SOS1, BAG3, RXRA, VAV1, CUL4A, HIPK2, UBC, PSME3
GOTERM _BP_FAT	GO:0043067~regulation of programmed cell death	34	18.99	2.55E-09	*Lung Cancer Protein:* PRKDC, HSPA1A, HSPE1, HSPA5, MYC, RASA1, CFLAR, CEBPB, MSH2, CREB1, ACTN1, YWHAE, TP73, TNFRSF10A, TNFSF10, HDAC1, JUN, HSPD1 *Predicted Lung Cancer Protein:* DPF2, ERBB2, , NR3C1, DAXX, SART1, RPS3, VDR, PPP2CA, SOS1, BAG3, RXRA, VAV1, , CUL4A, , HIPK2, UBC, PSME3
GOTERM _BP_FAT	GO:0010941~regulation of cell death	34	18.99	2.80E-09	*Lung Cancer Protein:* PRKDC, HSPA1A HSPE1, HSPA5, MYC, RASA1, CFLAR, CEBPB, MSH2, CREB1 ACTN1 YWHAE, TP73, TNFRSF10A, TNFSF10, HDAC1 JUN *Predicted Lung Cancer Protein:* DPF2, ERBB2, , NR3C1, DAXX, SART1, RPS3, VDR, PPP2CA, SOS1, BAG3, RXRA, , VAV1, , CUL4A, , HIPK2, UBC, PSME3, HSPD1
GOTERM _BP_FAT	GO:0010628~positive regulation of gene expression	33	18.43	1.96E-12	*Lung Cancer Protein:* ING2, THRB, PPARG, PRKDC MYC CEBPA, CEBPB, CREB1 HMGA1, TP73, STAT3 RB1 YWHAH, HDAC2 HDAC1, JUN DNMT1 *Predicted Lung Cancer Protein:* SMARCAD1, TBP, SMARCD1, , RUNX2, , RXRA, MTA2, SMAD2, TOPORS, ARID1B, DDX5, HDAC4, EP300, SP1, HIPK2, UBC, PIAS2
GOTERM _BP_FAT	GO:0008219~cell death	28	15.64	4.30E-07	*Lung Cancer Protein:* FUS PRKDC TOP1, TSC22D3 MYC, CFLAR, MSH2 HSPE1 YWHAE TP73, TNFRSF10A, TNFSF10 PKM2, JUN HSPD1, GADD45A *Predicted Lung Cancer Protein:* DPF2, TBP, DAXX, RPS3, SOS1, BAG3, TOPORS, VAV1, EP300, HIPK2, UBC, PSME3
GOTERM _BP_FAT	GO:0016265~death	28	15.64	4.93E-07	*Lung Cancer Protein:* FUS PRKDC TOP1 TSC22D3 HSPE1 MYC CFLAR, MSH2 YWHAE TP73,

					TNFRSF10A, TNFSF10 PKM2, JUN HSPD1, GADD45A *Predicted Lung Cancer Protein:* DPF2, TBP, DAXX, RPS3, SOS1, BAG3, TOPORS, VAV1, EP300, HIPK2, UBC, PSME3
GOTERM _BP_FAT	GO:0012501~programmed cell death	26	14.52	2.48E-07	*Lung Cancer Protein:* PRKDC TOP1, TSC22D3 HSPE1, MYC, CFLAR, MSH2 YWHAE, TP73, TNFRSF10A PKM2, JUN HSPD1, GADD45A *Predicted Lung Cancer Protein:* DPF2, DAXX, RPS3, SOS1, BAG3, TOPORS, VAV1, TNFSF10, EP300, HIPK2, UBC, PSME3
GOTERM _BP_FAT	GO:0043065~positive regulation of apoptosis	23	12.84	2.99E-08	*Lung Cancer Protein:* CFLAR, CEBPB PRKDC YWHAE, TP73, TNFRSF10A, TNFSF10 JUN, HSPD1, MYC *Predicted Lung Cancer Protein:* DPF2, RXRA, NR3C1, VAV1, DAXX, SART1, RPS3, VDR, CUL4A, PPP2CA, SOS1, HIPK2, UBC
GOTERM _BP_FAT	GO:0043068~positive regulation of programmed cell death	23	12.84	3.39E-08	*Lung Cancer Protein:* CFLAR, CEBPB PRKDC YWHAE, TP73 TNFRSF10A TNFSF10 JUN HSPD1 MYC *Predicted Lung Cancer Protein:* DPF2 , RXRA, NR3C1, VAV1, DAXX, SART1, RPS3, ,VDR, CUL4A, PPP2CA, SOS1, HIPK2, UBC
GOTERM _BP_FAT	GO:0010942~positive regulation of cell death	23	12.84	3.68E-08	*Lung Cancer Protein:* CFLAR, CEBPB PRKDC YWHAE, TP73, TNFRSF10A TNFSF10 JUN, HSPD1, MYC *Predicted Lung Cancer Protein:* DPF2, RXRA, NR3C1, VAV1, DAXX, SART1, RPS3, , VDR, , CUL4A, PPP2CA, SOS1, HIPK2, UBC
GOTERM _BP_FAT	GO:0006915~apoptosis	23	12.84	8.68E-06	*Lung Cancer Protein:* CFLAR, MSH2 YWHAE, TP73 TNFRSF10A, TNFSF10, TSC22D3 JUN, HSPE1 HSPD1, MYC, GADD45A *Predicted Lung Cancer Protein:* DPF2, TOPORS, VAV1, DAXX, RPS3, EP300, SOS1, BAG3, HIPK2, UBC, PSME3
GOTERM _BP_FAT	GO:0006917~induction of apoptosis	18	10.05	7.35E-07	*Lung Cancer Protein:* CFLAR, CEBPB YWHAE TP73, TNFRSF10A TNFSF10 MYC *Predicted Lung Cancer Protein:* DPF2, VAV1, DAXX, SART1, RPS3, VDR, CUL4A, PPP2CA, SOS1, HIPK2, UBC
GOTERM _BP_FAT	GO:0012502~induction of programmed cell death	18	10.05	7.68E-07	*Lung Cancer Protein:* CFLAR, CEBPB YWHAE TP73 TNFRSF10A TNFSF10, MYC *Predicted Lung Cancer Protein:* DPF2, VAV1, DAXX,

					SART1, RPS3, VDR, CUL4A, PPP2CA, SOS1, HIPK2, UBC
GOTERM _BP_FAT	GO:0043066~negative regulation of apoptosis	14	7.82	6.30E-04	*Lung Cancer Protein:* CFLAR, CEBPB, MSH2, ERBB2, HSPA1A, TP73, HDAC1, HSPD1, HSPA5, MYC, RASA1 *Predicted Lung Cancer Protein:* BAG3, HIPK2, UBC
GOTERM _BP_FAT	GO:0043069~negative regulation of programmed cell death	14	7.82	7.17E-04	*Lung Cancer Protein:* CFLAR, CEBPB, MSH2, ERBB2, HSPA1A, TP73, HDAC1, HSPD1, HSPA5, MYC, RASA1 *Predicted Lung Cancer Protein:* BAG3, HIPK2, UBC
GOTERM _BP_FAT	GO:0060548~negative regulation of cell death	14	7.82	7.36E-04	*Lung Cancer Protein:* CFLAR, CEBPB, MSH2, ERBB2, HSPA1A, TP73, HDAC1, HSPD1, HSPA5, MYC, RASA1 *Predicted Lung Cancer Protein:* BAG3, HIPK2, UBC
KEGG_PA THWAY	hsa05200:Pathways in cancer	25	13.96	9.06E-08	*Lung Cancer Protein:* HSP90AB1, GRB2, ERBB2, PPARG PIK3R3, MYC, PIK3R2, CEBPA, MSH2 CDK6, RB1, STAT3 HDAC2, , HDAC1, JUN, *Predicted Lung Cancer Protein:* SOS1, RARA, NOS2, RXRA, CBL, SMAD2, EP300, PIAS4, PLCG1 PIAS2
KEGG_PA THWAY	hsa05223:Non-small cell lung cancer	9	5.027	1.81E-05	*Lung Cancer Protein:* GRB2, ERBB2, CDK6, RB1, PIK3R3, PIK3R2 *Predicted Lung Cancer Protein:* PLCG1, SOS1, RXRA
KEGG_PA THWAY	hsa05222:Small cell lung cancer	9	5.02	4.40E-04	*Lung Cancer Protein:* CDK6 RB1, PIK3R3, MYC, PIK3R2 *Predicted Lung Cancer Protein:* PIAS4, RXRA, , PIAS2, NOS2
KEGG_PA THWAY	hsa05215:Prostate cancer	9	5.02	6.51E-04	*Lung Cancer Protein:* HSP90AB1, GRB2, ERBB2, CREB1, RB1, PIK3R3, PIK3R2 *Predicted Lung Cancer Protein:* EP300, SOS1
KEGG_PA THWAY	hsa05210:Colorectal cancer	8	4.46	0.002218	*Lung Cancer Protein:* GRB2, MSH2, JUN, PIK3R3, MYC, PIK3R2 *Predicted Lung Cancer Protein:* SOS1, SMAD2
KEGG_PA THWAY	hsa05214:Glioma	7	3.91	0.002335	*Lung Cancer Protein:* GRB2 CDK6, RB1, PIK3R3, PIK3R2 *Predicted Lung Cancer Protein:* PLCG1, SOS1
KEGG_PA THWAY	hsa05211:Renal cell carcinoma	7	3.91	0.003983	*Lung Cancer Protein:* GRB2, JUN PIK3R3, PIK3R2 *Predicted Lung Cancer Protein:* EP300, SOS1, PTPN11
Cluster 4					
Category	Term	Count	%	PValue	Involved proteins
GOTERM	GO:0010629~negative regulation	13	19.40	4.18E-06	*Lung Cancer Protein:* E2F1, SOX2, TP53,

_BP_FAT	of gene expression				UBE2I, ILF3, RBBP7, RPS14, MDM2, SMARCA2, NCOR2, SMARCA4 *Predicted Lung Cancer Protein*: KDM1A, SIN3A
GOTERM _BP_FAT	GO:0010628~positive regulation of gene expression	13	19.40	1.76E-05	*Lung Cancer Protein:* E2F1,CREBBP, SOX2, TP53, ILF3, HIF1A, ILF2, SMARCA2, ING1, SMARCA4 *Predicted Lung Cancer Protein:* RELA, SMARCB1, SMARCC1
GOTERM _BP_FAT	GO:0042127~regulation of cell proliferation	14	20.89	7.77E-05	*Lung Cancer Protein*: EGFR, ERBB3, SOX2, STAT1 HIF1A, MDM2,SMARCA2, ING1 *Predicted Lung Cancer Protein*: ERBB4, , RELA, TP53, RPS9, CBLB, EIF2AK2
GOTERM _BP_FAT	GO:0008284~positive regulation of cell proliferation	8	11.94	0.0036	*Lung Cancer Protein*: EGFR, HIF1A, SOX2, MDM2, STAT1 *Predicted Lung Cancer Protein*: ERBB4, RELA, , RPS9
GOTERM _BP_FAT	GO:0008284~positive regulation of cell proliferation	8	11.94	0.0036	*Lung Cancer Protein*: EGFR, HIF1A, SOX2, MDM2, STAT1 *Predicted Lung Cancer Protein*: ERBB4, RELA, RPS9
KEGG_PA THWAY	hsa05200:Pathways in cancer	12	17.91	5.08E-06	*Lung Cancer Protein* : E2F1, EGFR, HIF1A, HSP90AA1, CREBBP, TP53, MDM2, STAT1, CRK *Predicted Lung Cancer Protein*: CBLB, CRKL, RELA
KEGG_PA THWAY	hsa05215:Prostate cancer	7	10.44	2.20E-05	*Lung Cancer Protein:* E2F1, EGFR, HSP90AA1, CREBBP, TP53, MDM2 *Predicted Lung Cancer Protein:* RELA
KEGG_PA THWAY	hsa04012:ErbB signaling pathway	6	8.95	2.46E-04	*Lung Cancer Protein:* EGFR, ERBB3, CRK *Predicted Lung Cancer Protein:* CBLB, CRKL, ERBB4
KEGG_PA THWAY	hsa05212:Pancreatic cancer	5	7.46	0.0012	*Lung Cancer Protein:* E2F1, EGFR, TP53, STAT1 *Predicted Lung Cancer Protein:* RELA
Cluster 8					
Category	Term	Count	%	PValue	Involved proteins
GOTERM _BP_FAT	GO:0042981~regulation of apoptosis	29	17.68	4.66E-08	*Lung Cancer Protein:* HMGB1, TP63, EIF5A, SFN, HSPA1B, PTEN, AKT1, MAGED1, MAP3K1, TPT1, TERT, WWOX, SKP2, BIRC5, ESR2, BIRC2, EEF1E1, TNFAIP3, ABL1 *Predicted Lung Cancer Protein:* DEDD,

					BCAR1, BAG1, ARHGEF2, AARS, ADNP, MAPK9, FAF1, , BARD1 RAC1
GOTERM _BP_FAT	GO:0043067~regulation of programmed cell death	29	17.68	5.74E-08	*Lung Cancer Protein:* HMGB1, TP63, EIF5A, SFN, HSPA1B, PTEN, AKT1, MAGED1, MAP3K1 TPT1, TERT, WWOX, SKP2, BIRC5, ESR2, BIRC2, EEF1E1, TNFAIP3, ABL1 *Predicted Lung Cancer Protein:* BARD1 DEDD, BCAR1, BAG1, , RAC1, ARHGEF2, AARS, ADNP, MAPK9, FAF1
GOTERM _BP_FAT	GO:0010941~regulation of cell death	29	17.68	6.21E-08	*Lung Cancer Protein:* HMGB1, TP63, EIF5A, SFN, HSPA1B, PTEN, AKT1, MAGED1, MAP3K1, TPT1, TERT, WWOX, SKP2, BIRC5, ESR2, BIRC2, EEF1E1, TNFAIP3, ABL1 *Predicted Lung Cancer Protein:* DEDD, BCAR1, BAG1, ARHGEF2, AARS, ADNP, MAPK9, FAF1, BARD1, RAC1
GOTERM _BP_FAT	GO:0043065~positive regulation of apoptosis	17	10.36	2.09E-05	*Lung Cancer Protein:* SKP2, TP63, EIF5A, SFN, PTEN, AKT1, MAGED1, EEF1E1, MAP3K1, ABL1, WWOX *Predicted Lung Cancer Protein:* ARHGEF2, DEDD, RAC1, MAPK9, FAF1, BARD1
GOTERM _BP_FAT	GO:0043068~positive regulation of programmed cell death	17	10.36	2.28E-05	*Lung Cancer Protein:* SKP2, TP63, EIF5A, SFN, PTEN, AKT1, MAGED1, EEF1E1, MAP3K1, ABL1, WWOX *Predicted Lung Cancer Protein:* ARHGEF2, DEDD, RAC1, MAPK9, FAF1, BARD1
GOTERM _BP_FAT	GO:0010942~positive regulation of cell death	17	10.36	2.41E-05	*Lung Cancer Protein:* SKP2, TP63, EIF5A, SFN, PTEN, AKT1, MAGED1, EEF1E1, MAP3K1, ABL1, WWOX *Predicted Lung Cancer Protein:* ARHGEF2, DEDD, RAC1, MAPK9, FAF1, BARD1
GOTERM _BP_FAT	GO:0043066~negative regulation of apoptosis	15	9.14	3.64E-05	*Lung Cancer Protein:* HMGB1, SKP2, TP63, BIRC5, HSPA1B, ESR2, PTEN, AKT1, TPT1, TNFAIP3, TERT *Predicted Lung Cancer Protein:* BARD1, AARS, ADNP, BAG1
GOTERM _BP_FAT	GO:0043069~negative regulation of programmed cell death	15	9.14	4.24E-05	*Lung Cancer Protein:* HMGB1, SKP2, TP63, BIRC5, HSPA1B, ESR2, PTEN, AKT1, TPT1, TNFAIP3, TERT *Predicted Lung Cancer Protein:* AARS, ADNPBAG1, BARD1
GOTERM _BP_FAT	GO:0060548~negative regulation of cell death	15	9.14	4.37E-05	*Lung Cancer Protein:* HMGB1, SKP2, TP63, BIRC5, HSPA1B, ESR2, PTEN, AKT1, TPT1,

					TNFAIP3, TERT *Predicted Lung Cancer Protein:* BARD1, AARS, ADNP, BAG1
GOTERM _BP_FAT	GO:0008219~cell death	19	11.58	9.35E-04	*Lung Cancer Protein:* TP63, BIRC5, ITGB2, SFN, PTEN, BIRC2, MAGED1, AKT1, MAP3K1, CYFIP2 , TNFAIP3 *Predicted Lung Cancer Protein:* ARHGEF2, DEDD, SMN1, BAG1, PAK2, ATXN10, RAC1, FAF1
GOTERM _BP_FAT	GO:0006915~apoptosis	17	10.36	9.76E-04	*Lung Cancer Protein:* TP63, BIRC5, ITGB2, SFN, PTEN, BIRC2, AKT1, MAGED1, MAP3K1, CYFIP2, TNFAIP3 *Predicted Lung Cancer Protein:* ARHGEF2, DEDD, , BAG1, PAK2, RAC1, FAF1
GOTERM _BP_FAT	GO:0016265~death	19	11.58	0.0010	*Lung Cancer Protein:* TP63, BIRC5, ITGB2, SFN, PTEN, BIRC2, , MAGED1, AKT1, MAP3K1, CYFIP2, TNFAIP3 *Predicted Lung Cancer Protein:* ARHGEF2, DEDD, SMN1, BAG1, PAK2, ATXN10, RAC1, FAF1
GOTERM _BP_FAT	GO:0012501~programmed cell death	17	10.36	0.0011	*Lung Cancer Protein:* TP63, BIRC5, ITGB2, SFN, PTEN, BIRC2, AKT1, MAGED1, MAP3K1, TNFAIP3, CYFIP2 *Predicted Lung Cancer Protein:* ARHGEF2, DEDD, BAG1, PAK2RAC1, FAF1
GOTERM _BP_FAT	GO:0051301~cell division	11	6.70	0.0015	*Lung Cancer Protein:* PLK1, BIRC5, CDK4, WEE1 *Predicted Lung Cancer Protein:* ARHGEF2, BCAR1, PPP1CC, CDK2, STAG2, SMC4, CDK3
KEGG_PA THWAY	hsa05212:Pancreatic cancer	8	4.87	1.87E-04	*Lung Cancer Protein:* AKT1, SMAD4, MAPK9, CDK4 *Predicted Lung Cancer Protein:* MAP2K1, RAC1, MAPK10, RAD51
KEGG_PA THWAY	hsa04012:ErbB signaling pathway	7	4.26	0.0033	*Lung Cancer Protein:* AKT1, PAK2, MAP2K1, ABL1 *Predicted Lung Cancer Protein:* PAK6, MAPK9, MAPK10
Cluster 12					
Category	Term	Count	%	PValue	Involved proteins
GOTERM _BP_FAT	GO:0006955~immune response	6	85.71	1.96E-06	*Lung Cancer Protein:* CD8A, TAP2, B2M *Predicted Lung Cancer Protein:* TAP1, HLA-A, TAPBP

Cluster 17					
Category	Term	Count	%	PValue	Involved proteins
GOTERM _BP_FAT	GO:0010627~regulation of protein kinase cascade	6	15.00	3.16E-04	*Lung Cancer Protein:* BST2, GJA1 *Predicted Lung Cancer Protein:* MAP3K5, GRIN2B, RICTOR, TAB2
Cluster 22					
Category	Term	Count	%	PValue	Involved proteins
GOTERM _BP_FAT	GO:0006952~defense response	5	45.45	7.13E-04	*Lung Cancer Protein:* CD44, CD46, THBS1 *Predicted Lung Cancer Protein:* F2, ITGB1
GOTERM _BP_FAT	GO:0042981~regulation of apoptosis	5	45.45	0.0019	*Lung Cancer Protein:* CD44, TIAM1, THBS1 *Predicted Lung Cancer Protein:* F2, PLG
GOTERM _BP_FAT	GO:0043067~regulation of programmed cell death	5	45.45	0.0020	*Lung Cancer Protein:* CD44, TIAM1, THBS1 *Predicted Lung Cancer Protein:* F2, PLG
GOTERM _BP_FAT	GO:0010941~regulation of cell death	5	45.45	0.0020	*Lung Cancer Protein:* CD44, TIAM1, THBS1 *Predicted Lung Cancer Protein:* F2, PLG
Cluster 23					
Category	Term	Count	%	PValue	Involved proteins
GOTERM _BP_FAT	GO:0009967~positive regulation of signal transduction	5	20	0.001188	*Lung Cancer Protein:* ENG, PEBP1 *Predicted Lung Cancer Protein:* ACVR2B, MYD88, TRAF6

4.2.3 Protein-Protein Interaction Network in cancer related biological processes and pathways

Figure 12 and 13 indicates that there is a group of proteins that involved in the process of regulation of apoptosis and regulation of programmed cell death, we found non lung cancer proteins i.e. UBC, NR3CR1, DAXX, CUL4A, and BAG3 have high degree of the link to cancer proteins in both sub-network. This evidence indicates that those proteins may be significant proteins induced lung cancer. DAXX was reported in UniProt that it involves in programmed cell death, and also our evidence indicates that this type of protein interacts to five lung cancer proteins; TP73, CREB1, CFLAR, CEBPB and HDAC1. It is highly possible that DAXX involves in lung cancer forming.

Interestingly, UBC (Ubiquitin) has highest dense of connection to lung cancer proteins, UBC was recorded in uniProt that it involves in DNA damage response, by inducing the cell cycle regulator phosphoprotein p53 in response to the detection of DNA damage and resulting in the stopping or reduction of cell cycle rate. NR3C1 (Nuclear receptor subfamily 3 group C member 1) is another protein with high number of lung cancer protein found in our evidence as figure 10 and 11. This protein encodes a receptor for glucocorticoids that can act as both a transcriptionfactor and as a regulator of other transcription factors. It can also found in heteromeric cytoplasmic complexes along with heat shock factors and immunophilins.

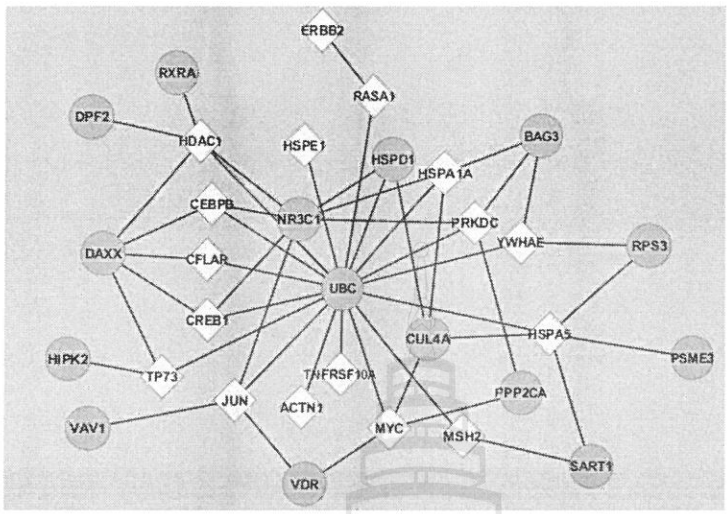

Figure 12 a group of protein-protein interactions involved in biological process of regulation of apoptosis

(diamond element indicate lung cancer protein, circle element indicate non lung cancer protein)

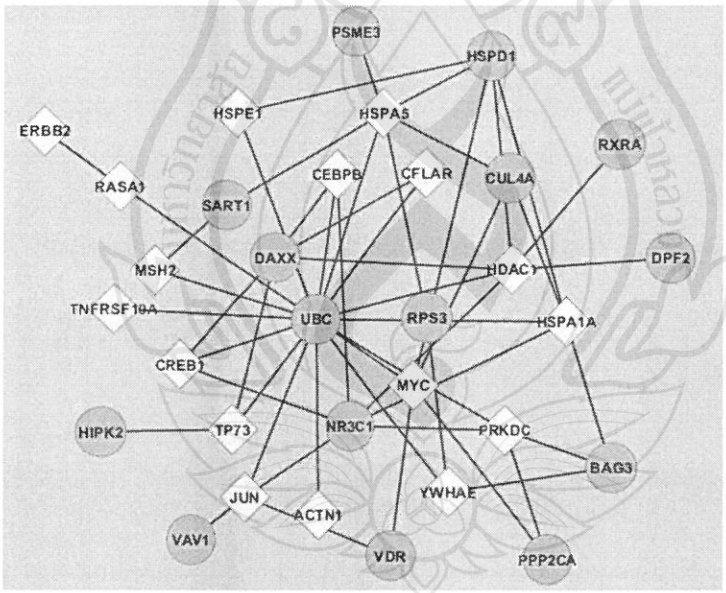

Figure 13 a group of protein-protein interaction involved in biological process of programmed cell death

(diamond element indicate lung cancer protein, circle element indicate non lung cancer protein)

EP300, SP1 and SMAD2 proteins are found highly connected link to lung cancer proteins as figure 14. There are many previous evidences supported our result that SMAD family are components of the transforming growth TBF-β signaling pathway that is deregulated in a variety of cancer types (Xu J. 2000; Singh P. 2011; Fleming

NI. 2013). The work of Hsu et al (Hsu TI. 2012) studied the role of SP1 expression in lung cancer, their work shows that the SP1 protein was highly increased and required for lung tumor growth in transgenic mice bearing Kras-induced lung tumors under to control of doxycycline and also this protein was highly up-regulated in lung ademoncarcinoma cells with low invasiveness and in patients with stage I lung cancer. Furthermore, the work of Szalad et al (Szalad A. 2009) investigated function of SP1 in tumor invasiveness under normoxic and hypoxic conditions. They found that SP1 binds to the ADAM17 promoter and it regulates ADAM17 protein expression under hypoxia, regulates glioma invasiveness.
Besides, the work of Roelfsema et al (Roelfsema JH. 2005) studied the effect of EP300 mutation, they found that the mutation of EP300 cause congenital disorder. Furthermore, result from the work of Gayther et al (Gayther SA. 2000) shows that EP300 is mutated in epithelial cancers and provide the first evidence that it behaves as a classical tumor-suppressor gene.

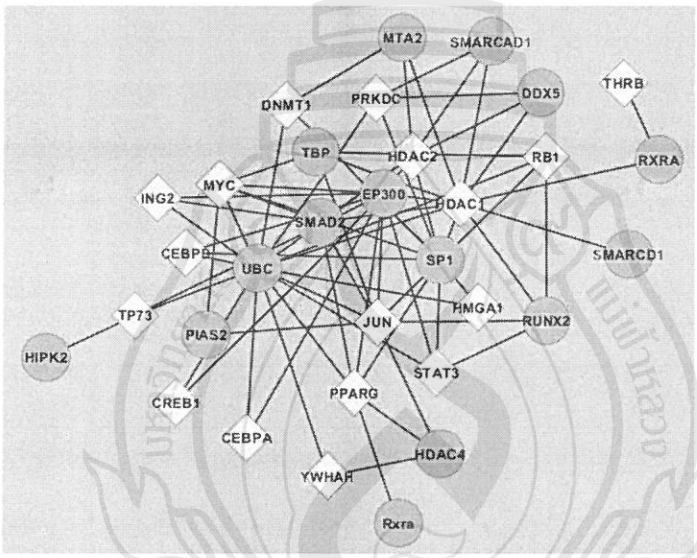

Figure 14 a group of protein-protein interaction involved in biological process of positive regulation of gene expression

(diamond element indicate lung cancer protein, circle element indicate non lung cancer protein)

TNFSF10 and BAG3 protein have two degree of cancer linkage as figure 15; TNFSF10 protein is reported by NCBI that induces apoptosis in transformed and tumor cells, but not appear to kill normal cells although it is expressed at a significant level in most normal tissues. Also it is reported from the work of Kuribayashi et al (Kuribayashi K. 2008) that it is a a53 target gene that mediates p53-dependent cell death. Beside, Rosati et al (Rosati A. 2011) studied various functions of BAG3 protein in major cell pathway, they reported that this protein involved in apoptosis and leukemias.

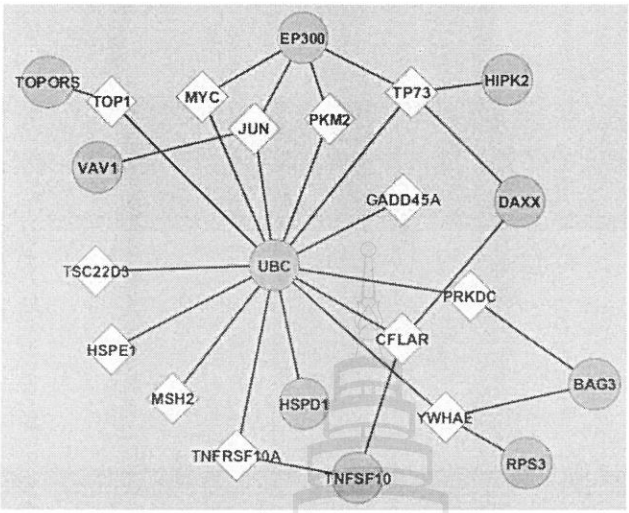

Figure 15 a group of protein-protein interaction involved programmed cell death

(diamond element indicate lung cancer protein, circle element indicate non lung cancer protein)

From our evidence as the figure 16, it is found that CUL4A and VDR involve in positive regulation of apoptotic process which is any process that activates or increases the frequency, rate or extent of cell death by apoptotic process. CUL4A was reported in the work of Puneet and Alo (Puneet S 2014) that it is attacked by several viral proteins and it overexpresses in a common feature of many human cancers. This research work presents that CUL4A is an attractive target for drug discovery efforts, especially, for further studies of a drug target for various types of cancer disease. Furthermore, the work of Ren et al (Ren S. 2012) which is about the relation of CUL4A and thalidomide treatment in prostate cancer also supports the work that mentioned above. They reported that sensitivity to thalidomide is positively correlated with the expression of CUL4A, the ectopic expression of CUL4A greatly increased sensitivity to thalidomide, while its down-regulation implies resistance to this drug.
Data suggest that Calcidiol or 25(OH)D interacts with VDR (vitamin D receptor) to decrease proliferation and increase apoptotic, the work of Hendrickson et al (Hendrickson WK. 2011) reported that high VDR protein expression in prostate tumors has significant relation with a reduced risk of lethal cancer, this evidence implies that vitamin D has crucial role in cancer progression.

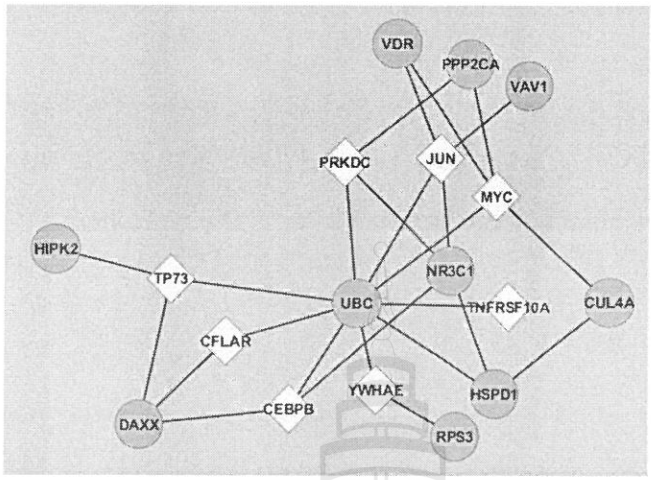

Figure 16 a group of protein-protein interaction involved in positive regulation of apoptotic process

(diamond element indicate lung cancer protein, circle element indicate non lung cancer protein)

Figure 17 shows the interaction of proteins that involved in KEGG cancer pathway. CBL (E3 ubiquitin-protein ligase CBL) is reported by Paolino et al (Paolino M. 2014) that CBL-b and TAM receptor regulates cancer metastasis via natural killer cells. PIAS2 is a protein inhibitor of activated STAT2, STAT2 increase appears to be an early detectable cellular event in cervical cancer progression (Liang Z. 2012).
Data suggests that RARA (Retinoic acid receptor alpha) is a marker of tanoxifen resistance in breast cancer, and it may be a target and predictive factor for oestrogen receptor alpha-positive breast cancer patients treated with adjuvant tamoxifen (Hentrik J. 2013).

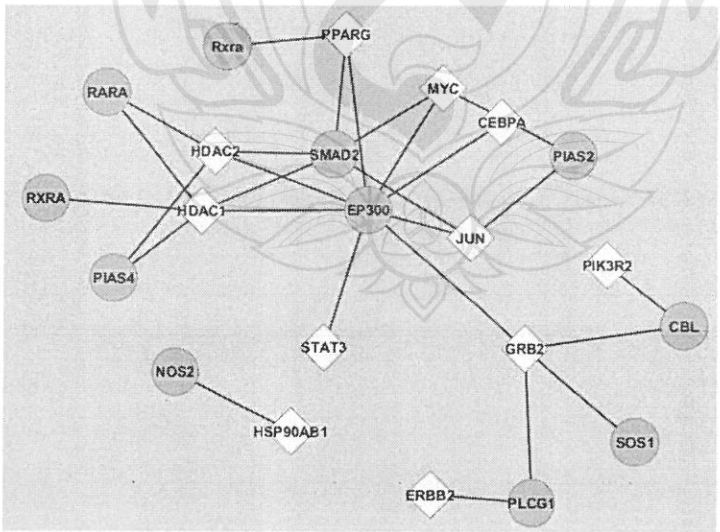

Figure 17 a group of protein-protein interaction involved in KEGG Pathways in Cancer

(diamond element indicates lung cancer protein, circle element indicates non lung cancer protein)

4.2.4 Identification of proteins interacting to OCP and TSP

The cause of cancer is closely related to the gain of OCP function or the lost of TSP function. The cause of disease is associated with many proteins and there are great chances that these proteins are regulated in biological processes or functions. Previous researches have suggested that if two proteins involving in the same PPI have highly similarity in their biological function, therefore, if a protein is associated to Lung cancer forming, then its partners in PPI are also likely connected to the lung cancer. From our evidence, we found that some proteins interact to OCP and TSP as table 7.

Table 7 A list of protein interacting to OCP, TSP

Protein	Interacting protein	Interacting protein type
AP2A1	ABL1	TSP/OCP
AP2B1	ABL1	TSP/OCP
BCAR1(TSP)	ABL1	TSP/OCP
CBL(OCP)	ABL1	TSP/OCP
CBLB(OCP)	ABL1	TSP/OCP
CREB1(TSP)	ABL1	TSP/OCP
CRK(TSP/OCP)	ABL1	TSP/OCP
CRKL(OCP)	ABL1	TSP/OCP
EGFR(TSP/OCP)	ABL1	TSP/OCP
ERBB2(TSP/OCP)	ABL1	TSP/OCP
ERBB3(TSP/OCP)	ABL1	TSP/OCP
ERBB4(OCP)	ABL1	TSP/OCP
GRB2	ABL1	TSP/OCP
HSP90AA1(OCP)	ABL1	TSP/OCP
HSPD1	ABL1	TSP/OCP
INPPL1	ABL1	TSP/OCP
MAPT	ABL1	TSP/OCP
MDM2(TSP/OCP)	ABL1	TSP/OCP
MUC1	ABL1	TSP/OCP
NCK1	ABL1	TSP/OCP
PIK3R2	ABL1	TSP/OCP
PLCG1	ABL1	TSP/OCP
PRKDC	ABL1	TSP/OCP
RAD51	ABL1	TSP/OCP
RB1(TSP/OCP)	ABL1	TSP/OCP
TP53(TSP/OCP)	ABL1	TSP/OCP
TP73(TSP/OCP)	ABL1	TSP/OCP
UBC	ABL1	TSP/OCP
VAV1	ABL1	TSP/OCP
ACTA1	DHX9	TSP
ACTL6A	EWSR1	OCP

ACTL6A	SMARCA2	TSP
ACTL6A	SMARCA4	TSP
ACTL6A	SMARCE1	TSP
ACTL6A	TP53	TSP/OCP
ACTL6A	TRRAP	TSP
ACVR2B	PEG10	OCP
ACVR2B	SMAD2	TSP
ADNP	SMARCA4	TSP
AIP(TSP)	HSP90AA1	OCP
STIP1	AIP	TSP
GAG	ANXA2	OCP

4.3 Clustering Performance Comparison among two Algorithms

4.3.1 Identification of protein complexes

A major problem in dealing with protein-protein interaction network is the high false positive rate in high throughput experiments, false positive in a network are error interaction, while false negatives are missing interactions. To evaluate the reliability of our protein-protein interaction network, we adopted known complexes from MIPS database (Mewes HW. 2006). The compassion of the overlaps of our protein-protein interaction with the MIPS protein complexes data set was evaluated. Table 8 shows results for the overlaps of our protein-protein interaction networks with the MIPS protein complexes data set. The eighteen clique communities of K-Means protein complexes and twelve of MCODE protein complexes were compared with MIPS' 1818 protein complexes records and their maximum JI values were computed. Among the protein communities clustered by K-Means, there are 14 out of 18 protein complexes (77%) have non-zero JI values but not fully covered. Four communities do not correspond to any real protein complexes. For the protein complexes identified by MCODE, there are 10 out of 12 protein complexes (83%) have non-zero JI values but not fully covered. Only two protein complexes do not correspond to any real protein complexes from MIPS. Protein complexes identified by K-Means ranges from 0.03 to 0.71, while protein complexes identified by MCODE ranges from 0.34 to 0.75. These results indicated that the clusters predicted by MCODE have high coverage ratio than K-Means.

Table 8 The results of JI value for protein complexes

Method	JI (%)
K-Means	3.12-71.4%
MCODE	34.50-75.5%

4.3.2 Identification of predicted novel lung cancer associated protein

From our experiment, K-Means is able to extract only one big cluster of proteins (1,056 proteins) which has involvement of lung cancer protein more than 50%. This cluster involves 93% of lung cancer proteins, while MCODE is able to extract seven differently significant clusters in same condition.

In term of identifying novel lung cancer associated protein, K-Means identified five predicted lung cancer associated protein in total (UBC, COPS6, CUL2, NRP1, and SH3GL3), however, there were nine proteins are identified as seed node (PTN11, PPA1, KRT6B, RICTOR, USP50, ZNF579, METTL18, NOTCH2, and USP8) and many proteins are predicted by MCODE as novel lung cancer associated protein that involving in cancer biological processes or cancer KEGG pathways.

CHAPTER 5
CONCLUSION

In this study, we identified the novel lung cancer associated proteins based on the concept of network clustering approach to discover protein interaction dense regions (network motif). The proteins which located in the same motif as lung cancer proteins have a high probability in forming lung cancer. We first adopted K-Means clustering approach to cluster a group of protein-protein interaction into sub-clusters, and then clique percolation clustering method (CPM) is adopted to discover "significant network motif" of significant protein cluster resulted by K-Means. Secondly, the Molecular Complex Detection approach (MCODE) is also adopted in this work to be a candidate of the first algorithm in term of clustering efficiency. The same input data set as the first algorithm is submitted into MCODE algorithm to cluster protein-protein interaction network into sub-clusters. Then analyzing biological processes and KEGG pathways of proteins involved in same cluster was investigated. Besides, cancer protein types; tumor suppressor protein (TSP) and onco-protein (OCP) are also observed. Finally, the comparison of discovering accurate "protein complexes" among two different approaches is investigated by referring to known protein complexes from MIPS.

Our results indicated that associated proteins findings involved in crucial processes in cancer formation i.e. programmed cell death, apoptosis. Basically, there are two limitations of our methodology i) the cancer-associated protein prediction is limited by the quality of gene ontology and pathway information, and ii) limited by the number of known lung cancer proteins. This work can be the essential first step on discovering lung cancer associated proteins based on clustering analysis.

Further study will make more experiments in using different clustering algorithm to overcome trapping the result in increasing accuracy and precision of the prediction of lung cancer associated protein.

REFERENCES

Adamcsek et al. (2006). "CFinder:locating cliques and overlapping modules in biological networks." Bioinfor 22(8): 1021-1023.

Aiello et al. (2010). "Link creation and profile alignment in the aNobii social network." In Proceedings of the Second IEEE International Conference on Social Computing ScialCom: 249-256.

Albert R. (2005). "Scale-free networks in cell biology." J Cell Sci 1(118): 4947-4957.

Alfaf-UI-Amin M. et al. (2006). "Development and implementation of an algorithm for detection of protein complexes in large interaction networks." BMC Bioinformatics 7(207).

Andreopoulos B. et al. (2007). "Clustering by common friends finds locally significant proteins mediating modules." Bioinfo 23(9): 1124-1131.

bader GD., H. C. (2003). "An automated method for finding molecular complexes in large protein interaction network." BMC Bioinformatics 4(2).

Baumgart A. et al. (2014). "Opposing role of Notch1 and Notch2 in a KrasG12D-driven murine non-small cell lung cancer model." Oncogene: doi:10.1038/onc.2013.1592.

Beane J. et al. (2007). "Reversible and permanent effects of tobacco smoke exposure on airway epithelial gene expression." Genome Biol: 8:R201.

Browne F et al. (2010). "From experimental approaches to computational techniques: a review on the prediction of protein-protein interaction." Adv Art Int 2010: 924529.

Byun S. et al. (2013). "USP8 is a novel target for overcoming gefitinib resistance in lung cancer." Clin Cancer Res 19(14): 3894-3904.

Cho YR. et al. (2007). "Semantic integration to identify overlapping functional modules in protien interaction networks." BMC Bioinformatics 8(265).

Chua HN. et al. (2008). "Using indirect protein-protein interactions for protein complex prediction." J Bioinform Comput Biol 6(3): 435-4666.

Chuang HY et al. (2007). "Network-based classification of reast cancer metastasis." Mol.Syst.Biol 3(140): doi:10.1038/msb4100180.

de-Lichtenberg U et al. (2005). "Dynamic Complex Formation During the Yeast Cell Cycle." Science 307: 724-727.

Ebel H. et al. (2002). "Scale-free topology of e-mail networks." Phys.Rev. E66, 035103.

Efroni S. et al. (2007). "Identificaiton of key processes underlying cancer phenotypes using biologic pathway analysis." PLoS One 2(e425): doi:10.1371/journal.pone.0000425.

Enright AJ et al. (2002). "An efficient algorithm for large-scale detection of protein families." Nucleic Acids Res 30(7): 1575-1584.

Fang WJ. et al. (2012). "Genome-wide analysis of aberrant DNA methylation for identification of potential biomarkers in colorectal cancer patients." Asian Pac J Cancer Prev 13(5): 1917-1921.

Feizi A. et al. (2013). "Metabolic and protein interaction sub-networks controlling the proliferation rate of cancer cells and their impact on patient survival." Science Reports(3041): doi:10.1038/srep03041.

Feng J. et al. (2008). "A max-flow based approach to the identification of protein complexes using protein interaction and microarray data." Comput Syst Bioinformatics Conf 7: 51-62.

Fleming NI. (2013). "SMAD2, SMAD3 and SMAD4 mutations in colorectal cancer." Cancer Research 73(2): 725-735.

Frey BJ. et al. (2007). "Clustering by passing messages between data points." Science 315(972-976).

Gao D. et al. (2010). "Rictor forms a complex with Cullin-1 to promote SGK1 uniquitination and destruction." Mol Cell 39(5): 797-808.

Gayther SA. et al. (2000). "Mutations truncating the EP300 acetylase in human cancers." NAT GENET 24(3): 300-303.

George R.A. (2006). "Analysis of protein sequence and interaction data for candidate disease gene prediction." Nucleic Acids Res 34: e130.

Giri K. et al. (2014). "Understanding Protein-Nanoparticle Interactin: a new gateway to disease therapeutics." Bioconjug Chem PMID: 24831101.

Goh KI et al. (2007). "The human disease network." Proc. Natl. Acad.Sci USA 104: 8685-8690.

Hendrickson WK. (2011). "Vitamin D receptor protein expression in tumor tissue and prostate cancer progression." J Clin Oncol 29(doi:10.1200/JCO2010.30.9680).

Hentrik J. (2013). "Retinoic acid receptor alpha is associated with tamoxifen resistance in breast cancer." Nature Commnications 4(2175): doi:1038/ncomms3175.

Hollanders G. (2005). Reconstruction of Gene-Interaction Networks from microarray timeseries: a comparison of two methods on real data sets. Computer Science, section Operations Mathematics, transnational University Limburg. Master of Science.

Hsu TI. et al. (2012). "SP1 expression regulates lung tumor progression." Oncogene 31(35): 3973-3988.

huang W.da et al. (2009). "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources." Nat Protoc 4: 44–57.

Jansen R. et al. (2003). "A bayesian networks approach for predicting protein-protein interactions from genomic data." Science 302(5644): 449-453.

Jemal A. et al. (2008). "Cancer statistics." CA-Cancer J Clin 58: 71-96.

Jonsson PF. et al. (2006). "Global topological features of cancer proteins in the human inteactome." Bioinfo 22: 2291-2297.

Kassis ES. et al. (2009). "Application of the revised lung cancer staging system (IASLC Staging Project) to a cancer center population." J Thorac Cardiovasc Surg 138: 412-418.

King AD. et al. (2004). "Protein complex prediction via cost-based clustering." Bioinfo 20(17): 3013-3020.

Kuribayashi K. et al. (2008). "NTFSF10 (TRAIL), a p53 target gene that mediates p53-dependent cell death." Cancer Biol Ther 7(12): 2034-2038.

Lage K. (2006). "A human phenome-interactome network of protein complexes implicated in genetic disorders." Nat.Biotechnol 25: 309-316.

Lambiotte R. et al. (2008). "Geographical dispersal of mobile communication networks." Phys. A. Stat. Mech Appl 387: 5317-5325.

Li BQ. et al. (2012). "Identification of Colorectal cancer related genes with mRMR and Shortest path in protein-protein interaction network." PLoS One(DOI: 10.1371/journal.pone.0033393).

Li SH. et al. (2004). "Huntingtin-protein interactions and the pathogenesis of Huntington's disease." Trends Genet 20: 146-154.

Li XL. et al. (2005). "Interaction graph mining for protein complexes using local clique merging." Genome Inform 16(2): 260-269.

Li XL. et al. (2007). "Discovering protien complexes in dense reliable neighborhoods of protein interaction networks." Comput Syst Bioinformatics Conf 6: 157-168.

Liang Z. et al. (2012). "Detection of STAT2 in early stage of cervical premalignancy and in cervical cancer." Asian Pac J Troc Med 5(9): 738-742.

Liu G. et al. (2009). "Complex discovery from weighted PPI networks." Bioinfo 25(15): 1891-1897.

Lubovac Z. et al. (2006). "Combining functional and topological properties to identify core modules in protein interaction networks." Protiens 64(4): 948-958.

MacQueen J.B. (1967). "Some Methods for classification and Analysis of Multivariate Observations." Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability Berkeley, University of California 1: 181-297.

maraziotis IA. et al. (2007). "Growing functinal modules from a seed protein via integration of protein interaction and gene expression data." BMC Bioinformatics 8(408).

Matthews LR. et al. (2001). "Identificaiton of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs"." Genome Res. 11: 2120-2126.

Mete M. et al. (2008). "A structure approach for finding functional medules form large biological networks." BMC Bioinformatics 9 (Suppl 9) (S19).

Mewes HW. et al. (2006). "MIPS: analysis and annottion of proteins from whole genomes in 2005." Nucleic Acids Res 34: D169-D172.

Moschopoulos CN. et al. (2009). "a clustering tool for detecting protein complexes." BMC Bioinformatics 10 (suppl 6)(S11).

Mukhopadhyay A. et al. (2012). "Detectin protein complexes in a PPI network: a gene ontology based multi-objective evolutionary approach." Mol Biosyst 8(11): 3036-3048.

Ng S.K. et al. (2003). "Integrative approach for computationally inferring protein domain interactions." BMC Bioinformatics 19(8): 923-929.

Ogmen U. et al. (2005). "Prism: protein interactions by structural matching." Nucleic Acids Research 33: W331-W336.

Okada M. et al. (2005). "Effect of turmor size on prognosis in patients with non-small cell lung cancer: the role of segmentectomy as a type of lesser resection." J Thorac Cardiovasc Surg 129: 87-93.

Onnela J.P. et al. (2007). "Structure and tie strengths in mobile communication networks." Proc. Natl. Acad.Sci USA 104: 7332-7336.

Paccanaro A. et al. (2006). "Spectral clustering of protein sequences." Nucleic Acids Res 17(34): 1571-1580.

Pagel P. et al. (2005). "The MIPS Mammalian Protein-Protein Interaction Database." Bioinfo 21(6): 832-834.

Palla G. et al. (2005). "Uncovering the overlapping community structure of complex networks in nature and society." nature 435: 814-818.

Paolino M. (2014). "The E3 ligase Cbl-b and TAM receptors regulate cancer metastasis via natural killer cells." Nature 507(7433): 508-512.

Park SW. et al. (2009). "Mutational analysis of hyposia-related genes HIF1alpha and CUL2 in common human cancers." APMIS 117(12): 880-885.

Pavlopoulos GA. et al. (2009). "jClust: a clustering and visualization toolbox." Bioinfo 25(15): 1994-1996.

Peng W. et al. (2014). "Improving protein function prediction using domain and protein complexes in PPI networks." BMC Systems Biology 8(35): doi:10.1186/1752-0509-1188-1135.

Perez-Iratzeta C. et al. (2002). "Association of genes to genetically inherited diseases using data mining." Nat.Genet 31: 316-319.

Plebani M. et al. (1995). "Clinical evaluation of seven tumor markers in lung cancer diagnosis: can any combination improve the result?" Br J Cancer 72(170-173).

Puneet S. et al. (2014). "CUL4A ubiquitin ligase: a promising drug target for cancer and other human diseases." open Biology 12: doi:10.1098/rsob.130217.

Ren R. (2005). "Mechanisms of BCR-ABL in the pathogenesis of chronic myelogenous leukaemia." Nat Rev Cancer 5: 172-183.

Ren S. et al. (2012). "Oncogenic CUL4A determines the response to thalidominde treatment in prostate cancer." J Mol Med (Berl) 90(10): 1121-1132.

Rhodes D.R. et al. (2005). "Probabilistic model of the human protein-protein interaction network." Nat Biotechnol 23(8): 951-959

Roelfsema JH et al. (2005). "Genetic heterogeneity in Rubinstein-Taybi syndrome: mutations in both the CBP and EP300 genes casue disease." Am J Hum Genet 76(4): 572-580.

Rosati A. et al. (2011). "BAG3: a multifaced protein that reguates major cell pathways." Cell Death and Disease 2(4): e141.

Satuluri V. et al. (2010). Markov Clustering of Protein Interaction Networks with Improved Balance and Scalability. ACM-BCB 2010, Niagara Falls, NY, USA.

Seshadri M. et al. (2008). "Mobile call graphs: beyond power-law and lognormal distributions." In:Proceedins of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'08, New York: 596-604.

Singh P. et al. (2011). Indian Journal of Cancer 48(3): 315-360.

Spinrin V. et al. (2005). "Protein complexes and functional modules in molecular networks." Proc Natl Acad Sci USA 100(21): 12123-12128.

Spirin V. et al. (2003). "Protein complexes and functional modules in molecular networks." Proc Natl Acad Sci USA 100: 12123-12128.

Stark C. et al. (2005). "BioGRID: a general repository for interaction datasets." Nucleic Acids Research 34: D535-D539.

Sun J.et al. (2010). "A comparative study of cancer proteins in the human protein-protein interaction network." BMC Genomics 11(S5): doi:10.1186/1471-2164-1111-S1183-S1185.

Szalad A. et al. (2009). "Transcription factor SP1 induces ADAM17 and contributes to tumor cell invasiveness under hypoxia." Clinical Cancer Research 28(129): doi:10.1186/1756-9966-1128-1129.

Tang X. et al. (2011). "A comparison of the functional modules indentified from time course and static PPI network data." BMC Bioinformatics 12: 339.

Tapas K. (2002). "An efficient K-Means Clustering Algorithm: analysis and Implementation." IEEE Transactions on pattern analysis and machine intelligence 24(7): 881-892.

Tartaqlia M. et al. (2001). "Mutation in PTPN11, encoding the protein tyrosine phosphatase SHP-2, cause Noonan syndrome." NAT GENET 29(4): 465-468.

Turner F.S. et al. (2003). "POCUS: mining genomic sequence annotation to predict disease genes." Genome Biol 4: R75.

Ulitsky I. et al. (2007). "Identificaiton of functional modules using network topology and high-throughput data." BMC Syst Biol 1(8).

Vielhaber S. et al. (2006). "Brain 1H magnetic resonance spectroscopic differences in myotonic dystrophy type 2 and type 1." Muscle Nerve 34: 145-152.

Von-Mering C. et al. (2007). "String7-recent developments in the integration and prediction of protein interactions." Nucleic Acids Research 35: 358-362

Wachi S. et al. (2005). "Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues." Bioinfo 21: 4205-4208.

Wang L. (2010). "HLungDB: an integrated database of human lung cancer research." Nucleic Acids Research 38: D665-D669.

Wu X., S. L., Guo J., Zhang D.Y., Lin K., (2006). "Prediction of yeast protein-protein interaction network: insights from the gene ontology and annotations." Nucleic Acids Research 34(7): 2137-2150.

Xia K. et al. (2008). "Impacts of protein-protein interaction domains on organism and network complexity." Genome Res. 18(9): 1500–1508.

Xu D. et al. (2013). "Activating Mutations in Protein Tyrosine Phosphatase PTPN11 (Shp2) enhance reactive oxygen species production that contributes to myeloproliferative disorder." PLOS one: DOI: 10.1371/journal.pone.0063152.

Xu J. et al. (2000). "Mutation in the tumor suppressors Smad2 and Smad4 inactivate transforming growth factor B signling by targeting Sams to the ubitin-proteasome pathway." Proceedings of the National Academy of Sciences of USA 97(9): 4820-4825.

Yeger-Lotem E. (2004). "Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction." Proc Natl Acad Sci USA 101: 5934-5939.

Zhang QC. et al. (2012). "PrePPI: a structure-informed database of protein-protein inteactions." Nucleic Acids Res doi: 10.1093/nar/gks1231.

BIOGRAPHY

Dr.Nilubon Kurubanjerdjit

Dr.Nilubon Kurubanjerdjit received B.S. degree in Computer Science from Maejo University, Chiangmai, Thailand in 2001, the M.S. degree in Information Technology from King Mongkutt's University of Technology Thonburi (KMUTT), Bangkok, Thailand in 2006 and the Ph.D. degree in Biomedical Informatics from Asia University, Taiwan in 2014. From 2007 to 2009, she joined the faculty at the Department of Information Technology, Kasetsart University at Kamphaengsaen campus, Thailand. Currently, she is a lecturer at the School of Information Technology, Mea Fah Luang University, Chiangrai, Thailand. Her research interests includes PPI network, host-pathogen PPI studies, cancer-related proteins, and Bio-big data analysis with cloud computing.

Dr.Natthakan Iam-On

Dr.Natthakan Iam-On received B.S. degree and M.S. in Computer Science from Chiangmai University, Chiangmai, Thailand and Ph.D. degree in Computer Science from University of Wales, Aberystwyth, UK in 2011. Her Ph.D. dissertation was awarded by National Research Council of Thailand as the Good Thesis work. Dr.Natthakan has published articles in highly ranked journals. Currently, she is a lecturer in School of Information Technology, Mae Fah Luang University, Chiangrai, Thailand. Her research interests include Database Systems, Data Warehousing, Data mining, Data clustering, Ensemble methodology, Bioinformatics and Medical data analysis.

Dr.Ka-Lok Ng

Dr.Ka-Lok Ng received the Honours diploma in physics from Hong Kong Beptist College in 1983, and the ph.D. degree in theoretical physics from the Vanderbuilt University at USA in 1990. He is a professor at the Department of Biomedical Informatics, Asia University, Taiwan, sicne August 2008. Beginning from December 2009, he serves on the Editorial board of several international journals. He is the Editor-in-Chief, Associated Editor, Reviewer Editor and Guest Editor of the WSEAS Transactions of Biology and Biomedicine, IST Transactions of Biomedical Sciences and Engineering, Frontiers in Genomic Assay Technology and Current Bioinformatics respectively. Furthermore, he is also actively involved in reviewing manuscripts for international journals. His is the PI and co-PI of more than 15 national funded research grants in the last ten years in the area of bioinformatics research.

Dr.Ng has published articles in highly ranked journals, in the areas of PPI network, robustness study of biological networks, domain-domain interactions, non-coding RNA, protein function prediction and DNA data hiding method. His research interests include PPI network, mRNA-microRNA expression profile study, cancer-related microRNAs, physio-chemical properties of protein complexs, time series microarray data analysis and host-pathogen PPI studies.