Vishwesh V. Kulkarni · Guy-Bart Stan
Karthik Raman *Editors*

A Systems Theoretic Approach to Systems and Synthetic Biology II: Analysis and Design of Cellular Systems

Springer

A Systems Theoretic Approach to Systems and Synthetic Biology II: Analysis and Design of Cellular Systems

Vishwesh V. Kulkarni
Guy-Bart Stan · Karthik Raman
Editors

A Systems Theoretic Approach to Systems and Synthetic Biology II: Analysis and Design of Cellular Systems

Springer

Editors
Vishwesh V. Kulkarni
Electrical and Computer Engineering
University of Minnesota
Minneapolis, MN
USA

Guy-Bart Stan
Department of Bioengineering
Imperial College
London
UK

Karthik Raman
Department of Biotechnology
Bhupat and Jyoti Mehta School
of Biosciences
Indian Institute of Technology Madras
Chennai
India

ISBN 978-94-017-9046-8 ISBN 978-94-017-9047-5 (eBook)
DOI 10.1007/978-94-017-9047-5
Springer Dordrecht Heidelberg New York London

Library of Congress Control Number: 2014942544

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

श्रेयो हि ज्ञानमभ्यासाज्ज्ञानाद्ध्यानं विशिष्यते ।
ध्यानात् कर्मफलत्यागस्त्यागाच्छान्तिरनन्तरम् ॥१२-१२॥
– श्रीमद्भगवद्गीता

Understanding is superior to mere practice
Union with the subject matter supersedes that
Dispassion towards all results is better still
And manifests peace immediately

Bhagavat Gita (12:12)

To my father Vasant, brother Vinay, sister Ketki, and Prof. Peter Falb
To my mother in memory

Vishwesh V. Kulkarni

To my parents, Florentina and Stephan
To my wife Cristina and my daughter Eva-Victoria

Guy-Bart Stan

To my parents and teachers
In memory of Sunder Mama and Prof. E. V. Krishnamurthy

Karthik Raman

Foreword

There is no design template more versatile than DNA. Nor are any designs more consequential than those whose blueprints DNA encodes. This exquisite substance has been shaped over billions of years by the creative combination of mutation and selection. Yet in the very long history of this template, it is only during our times that complex living organisms are beginning to understand and manipulate the very template whose sequences define them. But how should we go about this understanding? And how can we use this understanding to more effectively and responsibly alter the DNA template?

The complexity and diversity of living organisms are daunting. *Systems biology* aims at reverse engineering biological complexity for the purpose of understanding their design principles. By measuring and characterizing interactions of key biological molecules in response to stimuli and perturbations, systems biology aims to construct models that capture the complexity of endogenous biological networks. Through the systematic understanding of such models, it is hoped that one will achieve a holistic understanding of biological networks and the way they achieve biological function.

At the same time, the versatility of DNA and the dramatic decrease in the cost of DNA synthesis is making it possible to economically design and test new complex genetic circuits. This has given impetus to a new field: *Synthetic biology*. In our quest to understand biological complexity, we have examined endogenous biological subsystems and ascribed functions and design principles to their components. But a true understanding of these biological design principles is demonstrated only when one can build such systems *de novo* and demonstrate their function. When these circuits do not exhibit behavior consistent with our models, further investigations will lead to a deeper understanding of the underlying biology. Synthetic biology, therefore, serves as an important testbed for our understanding of biological principles. But the promise of synthetic biology extends beyond scientific understanding. Whether it be the detection and interference with the course of disease through the introduction of designer circuits, the cost-effective synthesis of new bio-substances, or the development of improved food products, synthetic biology provides a tremendous opportunity to alleviate suffering and improve the quality of our lives.

In both systems and synthetic biology, challenges abound. Quantitative modeling, analysis, and design of biological networks must contend with difficulties arising from the inescapable fact that at its most basic level, biology involves complex dynamic interactions among nonlinear stochastic components, taking place at multiple temporal and spatial timescales. The complexity of network interconnections of such components and the crosstalk between them adds another level of difficulty.

System theory has emerged as a field to deal with the challenges and complexities emerging from the interconnection of engineered systems, many of which are shared with biological systems. Notions from system theory such as nonlinearity, stochasticity, feedback, loading, modularity, robustness, identifiability, etc., are needed for a deeper understanding of biological complexity and for a more reliable design of biological circuits. These concepts are now being utilized to help us expand our understanding of endogenous biological circuits and to design novel ones. The articles in this book make significant strides in this direction.

While system theory will undoubtedly aid our understanding and design of biological systems, there is no doubt that the study of biological designs that have evolved over billions of years will also shape the future of system theory. For example, evolution and development are two central themes in biology that have little analogy with engineered man-made systems. Through the study of these and other biological themes, new systems notions and insights will undoubtedly emerge, enriching system theory in the process. One need only look at the history of *feedback*, a predominant concept in system theory, to imagine what is possible. While its human discovery can be traced back a little over one millennium, it is likely that feedback was invented by nature more than three billion years earlier. Since then, it has been wildly successful as a biological design principle, as evidenced by its prevalence at every level of biological organization. One wonders if an early systematic understanding of this concept in its biological context could have sped up the course of our own technological development.

As the physical sciences helped us understand the physical world around us over the last few centuries, so will quantitative biological science help us understand who we are, how we function, and how we can effectively and responsibly synthesize this most consequential of substances, the DNA. I believe that system theory will be central to this understanding.

Zürich, September 2013 Mustafa Khammash

Preface

Underlying every living cell are billions of molecules interacting in a beautifully concerted network of pathways such as metabolic, signalling, and regulatory pathways. The complexity of such biological systems has intrigued scientists from many disciplines and has given birth to the highly influential field of *systems biology* wherein a wide array of mathematical techniques, such as flux balance analysis, and technology platforms, such as next generation sequencing, is used to understand, elucidate, and predict the functions of complex biological systems. This field traces its roots to the general systems theory of Ludwig von Bertalanffy and effectively started in 1952 with a mathematical model of the neuronal action potential for which Alan Hodgkin and Andrew Huxley received the Nobel Prize in 1963. More recently, the field of *synthetic biology*, i.e., *de novo* engineering of biological systems, has emerged. Here, the phrase 'biological system' can assume a vast spectrum of meanings: DNA, protein, genome, cell, cell population, tissue, organ, ecosystem, and so on. Scientists from various fields are focusing on how to render this *de novo* engineering process more predictable, reliable, scalable, affordable, and easy. Systems biology and synthetic biology are essentially two facets of the same entity. As was the case with electronics research in the 1950s, a large part of synthetic biology research, such as the *BioFab* project, has focused on reusable macromolecular "parts" and their standardization so that composability can be guaranteed. Recent breakthroughs in DNA synthesis and sequencing combined with newly acquired means to synthesize plasmids and genomes have enabled major advances in science and engineering and marked the true beginning of the era of synthetic biology. Significant industrial investments are already underway. For example, in 2009, Exxon Mobil set up a collaboration worth $600 million with Synthetic Genomics to develop next generation biofuels.

Recent advances in systems and synthetic biology clearly demonstrate the benefits of a rigorous and systematic approach rooted in the principles of systems and control theory—not only does it lead to exciting insights and discoveries but it also reduces the inordinately lengthy trial-and-error process of wet-lab experimentation, thereby facilitating significant savings in human and financial resources. So far, state-of-the-art systems and control-theory-inspired results in systems and synthetic biology have been scattered across various books and journals from various disciplines. Hence, we felt the need for an edited book that provides a

panoramic view and illustrates the potential of such systematic and rigorous mathematical methods in systems and synthetic biology.

Systems and control theory is a branch of engineering and applied sciences that rigorously deals with the complexities and uncertainties of interconnected systems with the objective of characterising fundamental systemic properties such as stability, robustness, communication capacity, and other performance metrics. Systems and control theory also strives to offer concepts and methods that facilitate the design of systems with rigorous guarantees on these fundamental properties. For more than 100 years, the insights and techniques provided by systems and control theory have enabled outstanding technological contributions in diverse fields such as aerospace, telecommunication, storage, automotive, power systems, and others. Notable examples include Lyapunov's theorems, Bellman's theory of dynamic programming, Kalman's filter, H^∞ control theory, Nyquist-Shannon sampling theorem, Pontryagin's minimum principle, and Bode's sensitivity integral. Can systems and control theory have, or evolve to have, a similar impact in biology? The chapters in this book demonstrate that, indeed, systems and control theoretic concepts and techniques can be useful in our quest to understand how biological systems function and/or how they can be (re-)designed from the bottom up to yield new biological systems that have rigorously characterized robustness and performance properties.

Several barriers must be overcome to contribute significantly in this exciting journey. One of these is the language barrier, e.g., what a systems theorist means by the word *sensitivity* is different from what a biologist means by it. Another one is the knowledge barrier as, traditionally, systems and control theorists and biologists are not well versed with each other's knowledge base (although that scenario is now fast changing for the better with the introduction of bioengineering courses in systems and control theory at the undergraduate and graduate levels). A third barrier is due to the sheer volume of *big data*: the European Bioinformatics Institute in Hixton, UK, which is one of the world's largest biological data repositories, currently stores 20 petabytes of data and backups about genes, proteins and small molecules, and this number is more than doubling every year. Finally, a fourth barrier comes from the effort required to produce timely contributions based on currently available models. As an example of this last barrier, the systems and control theory community could have played a greater role than it did in two of the most significant technological advances of the last 50 years: VLSI and Internet. In retrospect, besides the fact that the systems and control theorists caught on the Internet too late, by which time infrastructures based on TCP/IP were already in place, the main difficulty posed by the Internet for the systems and control theory community was a lack of *good* models of the underlying networked system. This lack-of-good-models barrier is even more daunting in biology since some of the currently available *big data* are not guaranteed to be reproducible. As Prof. M. Vidyasagar illustrates and observes in the September 2012 issue of IEEE Lifesciences, one of the major challenges to the application of systems and control

theory concepts in biology comes from "the fact that many biological experiments are not fully repeatable, and thus the resulting datasets are not readily amenable to the application of methods that people like us [i.e., systems and control theorists] take for granted."

The chapters in this book serve to propose ways to overcome such barriers and to illustrate that biologists as well as systems and control theorists can make deep and timely contributions in life sciences by collaborating with each other to solve important questions such as how to devise experiments to obtain models of biological systems, how to obtain predictive models using information extracted from experimental data, how to choose components for (re-)engineering biological networks, how to adequately interconnect biological systems, and so on. Furthermore, and as Prof. Mustafa Khammash observes in his foreword, this research will fundamentally enrich systems and control theory as well by forcing it to investigate currently open questions that are specific to living biological systems, e.g., Why do biological systems naturally evolve the way they do? Can the evolvability of biological systems be consciously exploited for (re-)design and optimization purposes?

This book is intended for (1) systems and control theorists interested in molecular and cellular biology, and (2) biologists interested in rigorous modelling, analysis, and control of biological systems. We believe that research at the intersection of these disciplines will foster exciting discoveries and will stimulate mutually beneficial developments in systems and control theory and systems and synthetic biology.

The book consists of 12 chapters contributed by leading researchers from the fields of systems and control theory, systems biology, synthetic biology, and computer science. Chapters 1–6 highlight some state-of-the-art methods used to address currently open questions in systems biology. Chapters 7–12 discuss frameworks and methods required to enable a bottom-up design of synthetic biology systems of increasing complexity. These chapters are organized into two main parts as follows.

- **Part I—Systems Biology:** Chapters 1–6 focus on specific problems in modelling biological systems. Examples of such problems include: characterization and synthesis of memory, understanding how homoeostasis is maintained in the face of shocks and relatively gradual perturbations, understanding the functioning and robustness of biological clocks such as those at the core of circadian rhythms, and understanding how the cell cycles can be regulated, among others. A brief summary of each chapter is as follows.
  - Chapter 1: Today, several approaches used to identify biomarkers for a specific disease rely on genome-wide gene expression profiles without an explicit regard for how the genes are correlated. Wang and Chen present a network biomarker construction scheme that integrates microarray gene expression profiles and protein–protein interaction information so as to enable molecular investigation and diagnosis of lung cancer.

- Chapter 2: In this overview chapter, Lal and Seshashayee discuss next-generation sequencing techniques and illustrate how these have been used, at the scale of the whole bacterial genome, to investigate a variety of problems, from the analysis of gene expression and protein–DNA interactions to that of bacterial community function and evolution.
- Chapter 3: Given a biological network of oscillators, such as circadian rhythm networks for instance, how do biological parameter variations affect the oscillation characteristics? Sacré and Sepulchre present a novel and scalable approach to characterize the parameter sensitivity of models of oscillators, and illustrate its use on a circadian rhythm network model.
- Chapter 4: Osmosis facilitates the basic mechanism by which water is transported into and out of cells. Montefusco et al. demonstrate how a control theoretic analysis of the osmosis regulation system of *Saccharomyces cerevisiae* can be used to explain how cells maintain homoeostasis in the face of osmotic perturbations.
- Chapter 5: State synchronization is a recurring theme in neuronal networks and coupled networks of genetic clocks, among others. Hamadeh et al. explain how incremental dissipativity theory can be used to systematically analyse and/or synthesize feedback interconnections that ensure state synchronization in networks of identical oscillators and illustrate its use in the context of realising synchronization in a genetic repressilator network.
- Chapter 6: Multistability is a key property of biological systems that characterizes salient phenotypes such as memory. Salerno et al. present a systematic approach to characterize bistability and explain its utility in characterizing the memory of the galactose regulatory system of *Saccharomyces cerevisiae*.

- **Part II—Synthetic Biology:** Chapters 7–12 focus on how biomacromolecules, platforms, and scalable architectures should be chosen and synthesized in order to build programmable *de novo* biological systems. For example, a standardization of the components used is a necessary step in the modular design of large scale systems and presents an opportunity to develop *in silico* design tools that optimize these systems with respect to a set of formal specifications. What are the types of constrained optimization problems encountered in this process and how can these be solved efficiently? Should DNA be used as the basic macromolecule in synthesising artificial biological networks or should it be used with other macromolecules to enable certain applications? This set of chapters aims at answering such questions. A brief summary of each chapter is as follows.

  - Chapter 7: Modern nucleic acid biochemistry extensively uses protein enzymes to manipulate nucleic acids. However, predictive modification of the behavior of protein enzymes remains a very difficult problem. Chandran et al. show how meta-biochemical systems offer the possible advantage of being far easier in terms of re-engineering and programming. They show how a biochemical system can be synthesized based entirely on strands of DNA as the

only component molecule. These *meta-DNAs* have the same pairing mechanism as DNA but have a much larger alphabet of bases, thereby providing an increased power of base addressing.
- Chapter 8: An open challenge today is to specify synthetic biological systems using high level languages. Chen and Cai choose a rule-based modelling framework that was originally developed for systems biology and extend it to synthetic biology. They introduce a new model-specification language that facilitates the swift generation of mathematical models that encode the phenotypic behaviors of biological systems.
- Chapter 9: Krishnan and Liu address how bistable and monostable switches give rise to irreversible transitions and decision making in cell cycles. They propose a modular framework to address such questions for binary signalling mechanisms, outlining some of the design principles of signalling networks, which can be exploited in synthetic biology.
- Chapter 10: In standardising the components for a scalable design, an important constrained optimization problem concerns the selection of kinetic parameters and protein abundances. Koeppl et al. explain how this inverse problem can be solved more elegantly, by linearising the forward operator that maps parameter sets to specifications, and then inverting it locally, rather than relying on a brute force random sampling approach.
- Chapter 11: Marchisio and Stelling demonstrate how concepts and algorithms from electrical engineering can be exploited to set up a framework for the computation-based automated design of genetic Boolean gates and devices. They also explain how the Karnaugh algorithm used in the design of electrical circuits can be modified when it comes to the design of genetic circuits.
- Chapter 12: Kim and Franco focus on how to synthesize and couple transcriptional circuits by exploiting the modular architecture of nucleic acid templates as well as the catalytic power of natural enzymes. They illustrate the programmability of dynamic behaviors for elementary circuits such as adapters, bistable switches, and oscillators. They also present insulating and amplifying devices as a solution for the scaling-up of biomolecular networks.

The burgeoning fields of systems biology and synthetic biology have thrown up a very large number of interesting research problems. As the pre-eminent computer scientist Donald Knuth put it, "biology easily has 500 years of exciting problems to work on." The chapters in this book address but a small fraction of these interesting challenges. Nevertheless, we believe this book can serve as a good introduction on some of the currently open problems and on some of the state-of-the-art concepts and techniques available to propose solutions to such problems.

We are very grateful to all authors for their invaluable time and contributions and to Prof. Mustafa Khammash (ETH Zürich) for his stimulating foreword. We are also grateful to our institutions: University of Minnesota (Minneapolis, USA), Imperial College (London, UK), and Indian Institute of Technology Madras (Chennai, India) for their support and for providing a stimulating work

environment. Finally, we thank and acknowledge the financial support of our respective funding agencies: the National Science Foundation, the UK Engineering and Physical Sciences Research Council, and the Ministry of Human Resource and Development of the Government of India.

Minneapolis, MN, USA, September 2013 — Vishwesh V. Kulkarni
London, UK — Guy-Bart Stan
Chennai, India — Karthik Raman

Contents

Part I
Systems Biology

Chapter 1
Network Biomarker Construction for Molecular Investigation and Diagnosis of Lung Cancer via Microarray Data

Yu-Chao Wang and Bor-Sen Chen

Abstract Lung cancer is the leading cause of cancer deaths worldwide. Many studies have investigated the carcinogenic process and identified the biomarkers for signature classification. However, those biomarkers are mainly identified based only on analysis of genome-wide expression profiles, that is, the identification method cannot elucidate how the different genes in the biomarker gene set are related to each other. Therefore, from the systems perspective, we developed a network biomarker construction scheme, which integrated microarray gene expression profiles and protein-protein interaction information, for molecular investigation and diagnosis of lung cancer. The network biomarker consisted of two protein association networks constructed for cancer samples and non-cancer samples. Based on the network biomarker, a total of 40 significant proteins were identified with carcinogenesis relevance values (CRVs) to gain insights into the lung carcinogenesis mechanism. In addition, the network biomarker was also acted as the diagnostic tool, demonstrated to be effective to diagnose the smokers with lung cancer. Taken together, the network biomarker not only successfully sheds light on the mechanisms in lung carcinogenic process but also provides potential therapeutic targets to combat against cancer.

Keywords Network biomarker · Lung cancer · Protein association network · Microarray data · Analysis of variance (ANOVA) · Protein–protein interaction (PPI) · Akaike's information criterion

B.-S. Chen (✉)
Laboratory of Control and Systems Biology, Department of Electrical Engineering,
National Tsing Hua University, Hsinchu 30013, Taiwan
e-mail: bschen@ee.nthu.edu.tw

Y.-C. Wang
Institute of Biomedical Informatics,
National Yang-Ming University, Taipei 11221, Taiwan
e-mail: yuchao@ym.edu.tw

V. V. Kulkarni et al. (eds.), *A Systems Theoretic Approach to Systems and Synthetic Biology II: Analysis and Design of Cellular Systems*, DOI: 10.1007/978-94-017-9047-5_1,

1.1 Introduction

Cancer, the complex disease of uncontrolled cell growth, is the leading cause of human death worldwide and the deaths from cancer are projected to continue rising [19, 49]. Among all types of cancer, the mostly diagnosed and the most common cause of cancer deaths are lung cancer and the mortality rate within 5 years is as high as 80–85 % [49, 59]. Lung cancer can be divided into two main types, small cell lung carcinoma (SCLC) and non-small cell lung carcinoma (NSCLC). NSCLCs are further divided into three main subtypes: squamous cell carcinoma, adenocarcinoma, and large cell carcinoma [6]. Previous study has shown that all these major histological types of lung cancer are associated with cigarette smoking [33]. Therefore, many researchers devoted themselves to investigate the molecular alterations resulted from cigarette smoking and the mechanism that links cigarette smoking and lung cancer. Spira et al. used DNA microarray to compare the gene expressions of large-airway epithelial cells from nonsmokers and smokers and to define how cigarette smoking alters the transcriptome [58]. Hecht indicated that many tobacco smoke carcinogens such as polycyclic aromatic hydrocarbons and nicotine-derived nitrosamine ketone are likely to play major roles in lung cancer induction [23]. Recently, Takahashi et al. showed that induction of IKKβ- and JNK1-dependent inflammation is likely to be an important contributor to the tumor-promoting activity of tobacco smoke [61].

In addition to the investigation of carcinogenesis, many studies identified the cancer biomarkers through analysis of genome-wide expression profiles [2, 21]. The biomarkers are used as diagnostic evaluation to determine the patient with or without the cancer or used as prognostic indicator to evaluate the patient's prognosis. In lung cancer, Spira et al. used gene expression profile from patient samples to identify an 80-gene biomarker that distinguishes smokers with and without lung cancer [59]. The 80-gene biomarker could be beneficial for the decrease of the high mortality rate since the poor prognosis of lung cancer is closely related to the fact that there is no effective screening tool to diagnose the disease at an early stage [26, 59]. However, the biomarker identification method based only on gene expression profiles cannot elucidate how the different genes in the biomarker gene set are related to each other, i.e., the biomarkers are not identified from the systems perspective. Further, the gene lists obtained for the same clinical types of patients by different groups differ widely and have only very few genes in common [17].

Due to these kinds of limitation and the widely accepted opinion that cancer is a disease of pathways [22, 68], protein-protein interaction (PPI) and pathway information are integrated for biomarker identification. Chuang et al. developed a protein-network-based approach that identifies biomarkers not as individual genes but as subnetworks extracted from protein interaction databases. They showed that the subnetwork classification could achieve higher accuracy in the signature discrimination and are informative of the network structure [11]. Many other network-based approaches were also developed for prioritizing disease genes and protein interaction subnetworks that are discriminative of disease signature [9, 45, 46, 65]. In addition, the dynamic structure of the human protein interaction network was examined to

predict breast cancer prognosis, suggesting that the network modularity might be a defining feature of tumor phenotype [62].

Network analysis has shown that under different cellular states or in response to diverse stimuli, transcription factors alter their interactions to regulate different genes, thereby rewiring the network [41]. The same situation happens for protein interaction networks [62, 76]. Motivated by the dynamic structure of human protein interaction network and the observation that interacting proteins tend to result in similar disease phenotypes when dysregulated [47], we develop a computational framework to construct the network biomarker for molecular investigation and diagnosis of lung cancer via microarray data. The network biomarker consists of two protein association networks for cancer and non-cancer smokers. Based on the concept of network comparison [72], 40 significant proteins that may play important roles in lung carcinogenesis are identified. With the help of the network biomarker, the smokers suspect with cancer can be classified into smokers with cancer or without cancer, making the network biomarker a useful tool for molecular diagnosis. Hopefully, the proposed method can help understand the lung carcinogenesis and provide potential drug targets for humans to combat against lung cancer.

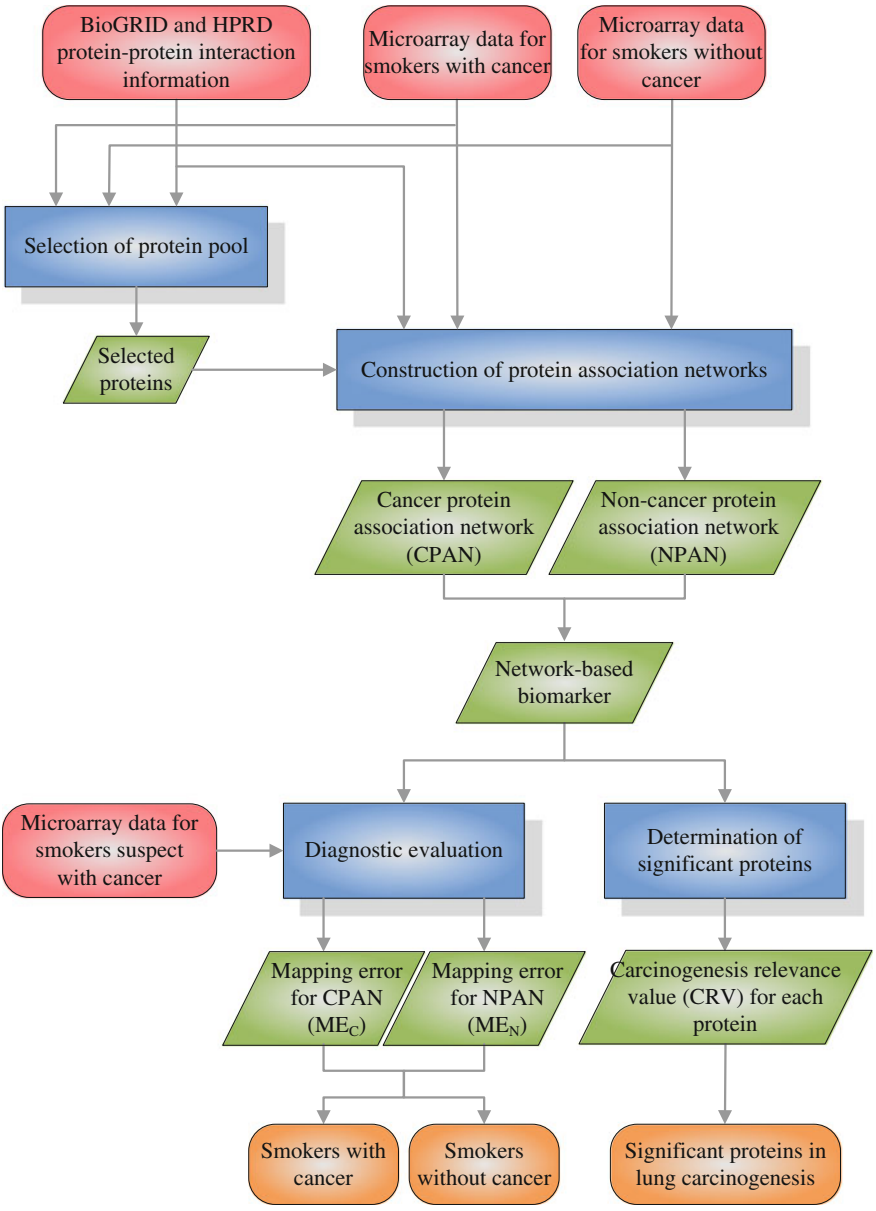
1.2 Methods

1.2.1 Overview of the Network Biomarker Approach for Lung Cancer Investigation

The overall flowchart of the proposed network biomarker approach is shown in Fig. 1.1. Our goal is to investigate the lung cancer by the construction of network biomarker which is composed of protein association networks for smokers with and without cancer. Microarray gene expression profiles of patient samples and protein-protein interaction information were integrated for protein selection and network construction. Two protein association networks with quantitative protein association abilities for cancer and non-cancer smokers were constructed respectively. Based on the comparison of two protein association networks within the network biomarker, a score named carcinogenesis relevance value (CRV) was computed to correlate proteins with significance of lung carcinogenesis. A higher score suggests that the particular protein plays a more critical role in lung carcinogenesis. According to the CRV for each protein and the statistical assessment, a set of significant proteins was selected. Furthermore, given the microarray data for the smokers suspect with cancer, mapping errors can be computed for diagnostic evaluation that the smokers with or without cancer.

1.2.2 Data Selection and Preprocessing

Here, two kinds of data, microarray gene expression profile and protein-protein interaction information, were integrated. The microarray data was


BioGRID and HPRD protein-protein interaction information
Microarray data for smokers with cancer
Microarray data for smokers without cancer
Selection of protein pool
Selected proteins
Construction of protein association networks
Cancer protein association network (CPAN)
Non-cancer protein association network (NPAN)
Network-based biomarker
Microarray data for smokers suspect with cancer
Diagnostic evaluation
Determination of significant proteins
Mapping error for CPAN (ME_C)
Mapping error for NPAN (ME_N)
Carcinogenesis relevance value (CRV) for each protein
Smokers with cancer
Smokers without cancer
Significant proteins in lung carcinogenesis

◀ **Fig 1.1** The flowchart of constructing the network biomarker for lung cancer investigation and diagnosis The figure indicates the flowchart of the proposed method. The *red rounded rectangles* represent the data needed. The *blue rectangles* denote the processing steps of the approach. The *green parallelograms* are the processed results for each step and the *orange rounded rectangles* are the overall results for the whole method. In summary, two kinds of data, microarray data and PPI information, are needed for the proposed method. These data are used for protein pool selection. Then the selected proteins and the input data are used for protein association network construction, resulting in cancer protein association network (CPAN) and non-cancer protein association network (NPAN). The two constructed protein association networks constitute the network biomarker, which can be used for either determination of significant proteins or diagnostic evaluation. With the help of the network biomarker, carcinogenesis relevance value (CRV) is computed for each protein and significant proteins in lung carcinogenesis are determined based on the CRVs. These significant proteins provide targets for further characterization. On the other hand, given the microarray data for smokers suspect with cancer, mapping errors for CPAN and NPAN can be computed, respectively, which help diagnose the smokers with cancer or without cancer

downloaded from GEO database http://www.ncbi.nlm.nih.gov/geo/ (accession number GSE4115). Spira et al. performed gene expression profiling in histologically normal large-airway epithelial cells obtained at bronchoscopy from current and former smokers. Each individual was followed after bronchoscopy until a final diagnosis of lung cancer or not lung cancer was made [59]. Data was collected from a total of 187 subjects and was divided into primary and prospective data sets (79 smokers with lung cancer and 73 smokers without lung cancer in the primary data set; 18 smokers with lung cancer and 17 smokers without lung cancer in the prospective data set). The primary data set was used for network biomarker construction and the prospective data set was used for diagnostic evaluation. Protein-protein interaction (PPI) data was extracted from BioGRID http://thebiogrid.org/ and HPRD http://www.hprd.org/databases. The Biological General Repository for Interaction Datasets (BioGRID) database was developed to house and distribute collections of protein and genetic interactions from major model organism species. BioGRID currently contains over 340,000 interactions as derived from both high-throughput studies and conventional focused studies [60]. The Human Protein Reference Database (HPRD) is a database that integrates a wealth of information relevant to human proteome, including protein-protein interactions, post-translational modifications, disease associations, and tissue expression [50]. Prior to further processing, the gene expression value g_{ij} is normalized to z-transformed scores z_{ij} so that for each gene i the normalized expression value has mean $\mu_i = 0$ and standard deviation $\sigma_i = 1$ over sample j.

1.2.3 Selection of Protein Pool and Construction of Network Biomarker

To integrate the gene expression and PPI information data and construct the network biomarker consisting of protein association networks, the expression value of each gene was first overlaid on its corresponding protein. The gene expression for

each protein was then used to select differentially expressed protein using one-way analysis of variance (ANOVA) where the null hypothesis is that the average expression levels for the protein are the same for smokers with and without cancer [48]. The proteins with Bonferroni adjusted p-values less than 0.05 were selected in the protein pool. Since we aimed at investigating the lung cancer using the network biomarker, the differentially expressed proteins without interaction information were excluded from the protein pool. In addition to the proteins that differentially expressed, the proteins which are highly connected with the proteins in the protein pool based on the PPI information were also included into the pool. In other words, the protein pool consists of both differentially expressed proteins and the proteins that are highly connected with them. On the basis of the protein pool and the PPI information, the rough PPI network can be easily constructed by linking the proteins that have interactions among them. One thing should be noted is that since the data for cancer and non-cancer samples are limited, the number of proteins selected for rough PPI network construction is also restricted. That is, in order to avoid overfitting in network construction, the maximum degree of the proteins in the rough PPI network should be less than the cancer/non-cancer sample number, thereby restricting the size of the rough PPI network.

From the process above, we have selected a protein pool and constructed a rough PPI network among them. The rough PPI network comprises all possible protein interactions under all kinds of experimental conditions. Consequently, the network should be further pruned using microarray data to indicate the effective protein associations for samples with and without lung cancer. Here, a simple linear regression model was applied to prune the rough PPI network to obtain the protein association networks independently for samples with and without cancer, according to their respective data sets. For a target protein i in the rough PPI network, the protein was described by the following protein association model [73].

$$y_i[n] = \sum_{k=1}^{N_i} \alpha_{ik} y_{ik}[n] + \varepsilon_i[n] \tag{1.1}$$

where $y_i[n]$ represents the gene expression level of the target protein i for the sample n, α_{ik} denotes the association ability between the target protein i and its kth interactive protein, which quantifies the expression relation between the interactive proteins and can be identified using the data we have, $y_{ik}[n]$ indicates the gene expression level of the kth protein that interacts with the target protein i for the sample n, N_i is the number of proteins interacting with the target protein i and can be obtained from the rough PPI network, $\varepsilon_i[n]$ denotes the stochastic noises due to other factors or model uncertainty. The biological meaning of Eq. (1.1) is that the expression level of the target protein i is associated with the expression levels of the proteins interacted with it. For each protein in the protein pool, a protein association model was constructed.

After the protein association model of the rough PPI network was constructed, the association parameters in Eq. (1.1) were identified using maximum likelihood estimation method [8, 32] by microarray data (see Appendix 1 for details). Since

there are two data sets of microarray data (smokers with and without cancer), the association parameters were separately identified for cancer data set and non-cancer data set, resulting in $\alpha_{ik,\mathrm{C}}$ and $\alpha_{ik,\mathrm{N}}$. In this case, for each protein in each phenotype, i.e., with cancer and without cancer, a mathematical description was constructed to characterize the expression association, respectively. Once the association parameters for all proteins in the rough PPI network were identified, the significant protein associations were determined based on the estimated association abilities $\alpha_{ik}{}'$s. Akaike Information Criterion (AIC) [1, 32] and student's t-test [48] were employed for both model order selection and significance determination of protein associations (see Appendix 2 for details). In this way, the rough PPI network was pruned and the protein association networks for smokers with and without cancer were constructed, respectively.

On the basis of the identified protein association abilities, two matrices were established to represent the cancer protein association network (CPAN) and the non-cancer protein association network (NPAN).

$$C = \begin{bmatrix} \alpha_{11,\mathrm{C}} & \alpha_{12,\mathrm{C}} & \cdots & \alpha_{1K,\mathrm{C}} \\ \alpha_{21,\mathrm{C}} & \alpha_{22,\mathrm{C}} & \cdots & \alpha_{2K,\mathrm{C}} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{K1,\mathrm{C}} & \alpha_{K2,\mathrm{C}} & \cdots & \alpha_{KK,\mathrm{C}} \end{bmatrix}$$
$$N = \begin{bmatrix} \alpha_{11,\mathrm{N}} & \alpha_{12,\mathrm{N}} & \cdots & \alpha_{1K,\mathrm{N}} \\ \alpha_{21,\mathrm{N}} & \alpha_{22,\mathrm{N}} & \cdots & \alpha_{2K,\mathrm{N}} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{K1,\mathrm{N}} & \alpha_{K2,\mathrm{N}} & \cdots & \alpha_{KK,\mathrm{N}} \end{bmatrix} \tag{1.2}$$

where $\alpha_{ij,\mathrm{C}}$ and $\alpha_{ij,\mathrm{N}}$ indicate the quantitative protein association ability between protein i and protein j for CPAN and NPAN, respectively, and K is the number of proteins in the protein association network. For any protein i and protein j in the protein association network, the association ability α_{ij} quantifies the expression relation between the interactive proteins. If the estimated protein association ability α_{ij} equals to zero, it means that there is no association between protein i and protein j. In addition, we said that protein i is associated with protein j means that the expression level changes of protein i account for the expression level changes of protein j and vice versa. As a consequence, when the estimated protein association ability α_{ij} does not equal to α_{ji}, the one which has larger absolute value would be selected as the association ability between protein i and protein j, i.e., $\alpha_{ij} = \alpha_{ji}$. The resulting cancer and non-cancer protein association networks (CPAN and NPAN) constituted the network biomarker, which was used for determining the significant proteins playing important roles in lung carcinogenesis and for diagnostic evaluation.

1.2.4 Determination of Significant Proteins in Lung Carcinogenesis via the Network Biomarker

According to equations (1.1) and (1.2), the protein association models for CPAN and NPAN can be represented as the following equations.

$$\begin{aligned} Y_{\mathrm{C}} &= CY_{\mathrm{C}} + E_{\mathrm{C}} \\ Y_{\mathrm{N}} &= NY_{\mathrm{N}} + E_{\mathrm{N}} \end{aligned} \tag{1.3}$$

where $Y_{\mathrm{C}} = \left[y_{1,\mathrm{C}}[n]\ y_{2,\mathrm{C}}[n] \ldots y_{K,\mathrm{C}}[n] \right]^T$, $Y_{\mathrm{N}} = \left[y_{1,\mathrm{N}}[n]\ y_{2,\mathrm{N}}[n] \ldots y_{K,\mathrm{N}}[n] \right]^T$ denotes the vectors of expression levels; E_{C} and E_{N} indicate the noise vectors in cancer case and non-cancer case, respectively. A matrix indicating the difference between two protein association networks is defined as $C - N$ [73].

$$\begin{aligned} D &= \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1K} \\ d_{21} & d_{22} & \cdots & d_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ d_{K1} & d_{K2} & \cdots & d_{KK} \end{bmatrix} \\ &= \begin{bmatrix} \alpha_{11,\mathrm{C}} - \alpha_{11,\mathrm{N}} & \alpha_{12,\mathrm{C}} - \alpha_{12,\mathrm{N}} & \cdots & \alpha_{1K,\mathrm{C}} - \alpha_{1K,\mathrm{N}} \\ \alpha_{21,\mathrm{C}} - \alpha_{21,\mathrm{N}} & \alpha_{22,\mathrm{C}} - \alpha_{22,\mathrm{N}} & \cdots & \alpha_{2K,\mathrm{C}} - \alpha_{2K,\mathrm{N}} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{K1,\mathrm{C}} - \alpha_{K1,\mathrm{N}} & \alpha_{K2,\mathrm{C}} - \alpha_{K2,\mathrm{N}} & \cdots & \alpha_{KK,\mathrm{C}} - \alpha_{KK,\mathrm{N}} \end{bmatrix} \end{aligned} \tag{1.4}$$

where d_{ij} denotes the protein association ability difference between CPAN and NPAN among protein i and protein j. Using the matrix D to show the difference of network structure between CPAN and NPAN, a score named carcinogenesis relevance value (CRV) was then presented to quantify the correlation of each protein with significance of lung carcinogenesis. To identify the significant proteins for lung carcinogenesis, two important issues were taken into consideration. First, the magnitude of the association abilities α_{ij}'s denotes the significance of one protein to the other one. A higher absolute value of α_{ij} implies that the two proteins are more tightly associated. Second, if a protein plays more crucial roles in lung carcinogenesis, the difference of association numbers linked to the protein for CPAN and NPAN would be larger. For instance, if one protein associates with a lot of proteins in CPAN but associates with no protein in NPAN, it would be more likely involved in lung carcinogenesis. As a result, the CRV was determined based on the difference of protein association abilities as the following equation.

$$\mathrm{CRV}_i = \sum_{j=1}^{K} \left| d_{ij} \right| \tag{1.5}$$

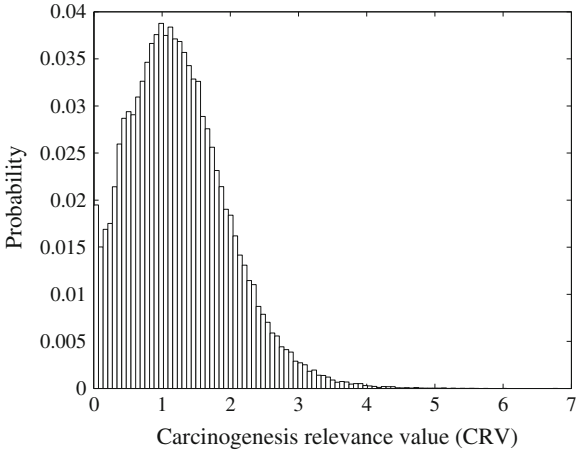


Fig. 1.2 Distribution of carcinogenesis relevance values (CRVs) of random networks. The null distribution of CRVs is generated by 100,000 randomly permuted network structures

For the ith protein in the network biomarker, the implication of Eq. (1.5) is that the CRV quantifies the extent of protein associations that differentiate CPAN from NPAN in the network biomarker.

For each protein, in addition to the CRV assigned, an empirical p-value was also computed to determine the significance of the CRV. To determine the p-value for an observed CRV, a null distribution of CRVs (Fig. 1.2) was generated by repeatedly permuting the network structure of the rough PPI network and computing the CRV for each random network structure. The permutation of the network structure was performed by keeping the network size, i.e., the proteins with which a particular protein interacted were permuted without changing the total number of protein interactions. The process was repeated 100,000 times and the p-value of the corresponding CRV was estimated as the fraction of random network structures whose CRV is at least as large as the CRV of the real network structure. The CRVs with p-value ≤ 0.05 were determined as significant CRVs and the corresponding proteins were identified as significant proteins in lung carcinogenesis.

1.2.5 Diagnostic Evaluation by the Network Biomarker

An important feature of the proposed network biomarker approach is that it can not only be used for investigation of significant proteins for lung cancer, but also for diagnosis of smokers suspect with lung cancer. Given the new microarray expression data for the smoker, we can classify the sample into smoker with or without cancer based on CPAN and NPAN within the network biomarker. The idea comes from the similarity comparison of new sample data between CPAN and NPAN.

Specifically, if a sample data is more similar to the network structure of CPAN than of NPAN, it would be regarded as the smoker with lung cancer, and vice versa. Since only one sample cannot be used for network construction, the new sample data was mapped to the CPAN and NPAN identified above and the mapping error would be employed as the criteria of classification. Suppose that we had a new sample data $Z = \left[z_1 \; z_2 \; \ldots \; z_K \right]^T$ from a smoker, based on Eqs. (1.1) and (1.3), the mapping errors for CPAN and NPAN are respectively defined as

$$\begin{aligned} \mathrm{ME_C} &= \| Z - C \cdot Z \|_2 \\ \mathrm{ME_N} &= \| Z - N \cdot Z \|_2 \end{aligned} \tag{1.6}$$

where $\| P \|_2 = \left(\sum_{i=1}^{K} p_i^2 \right)^{1/2}$ when $P = \left[p_1 \; p_2 \; \ldots \; p_K \right]^T$. The mapping errors can be considered as the similarity measurement of the new sample Z to the systems CPAN and NPAN. The smaller the mapping error is, the more matching the sample data is to the protein association network. Consequently, if $\mathrm{ME_C} < \mathrm{ME_N}$, the new sample Z is more similar to the cancer system and is classified into the smokers with cancer category, and vice versa. The criteria of mapping errors have simultaneously taken account of the protein association network structures with quantitative association abilities and the expression levels of the proteins. Further, since the modeling error is regarded as the criterion of classification, it is a classification more dependent on network structure than data only and therefore could be also suitable for classification with independent data. We believe that the kind of classification approach can provide new perspective for diagnostic evaluation.

1.3 Results

1.3.1 Construction of Network Biomarker and Determination of Significant Proteins in Lung Carcinogenesis

We applied the proposed network biomarker approach for molecular investigation and diagnosis of lung cancer. The primary data set (79 smokers with lung cancer and 73 smokers without lung cancer) of GSE4115 downloaded from GEO database http://www.ncbi.nlm.nih.gov/geo/ was used for construction of network biomarker. Based on the classical statistical method ANOVA, 199 proteins which have PPI information were identified as the differentially expressed proteins and were selected in the protein pool. In addition, the proteins that linked to three differentially expressed proteins in the protein pool according to PPI information were also included in the pool. In this case, the protein pool consisted of 339 proteins. Then, the proteins that have PPI information among them were linked together, resulting in the rough PPI network. The expression profiles for smokers with and without cancer and the

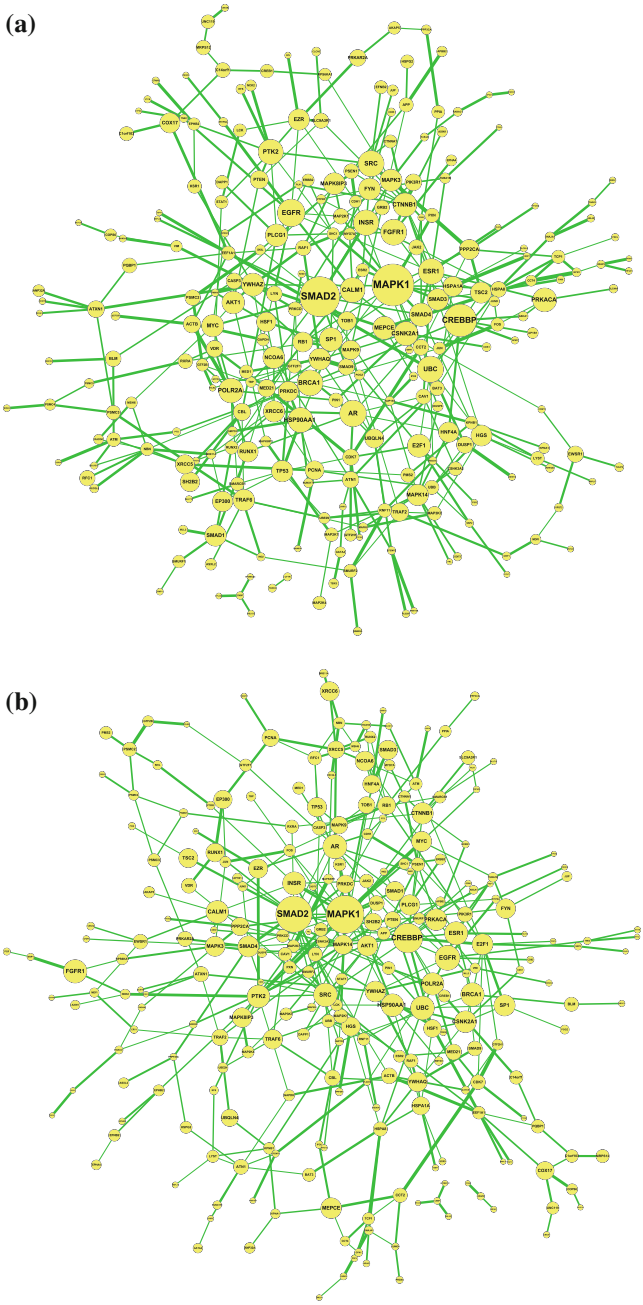


Fig. 1.3 The constructed network biomarker. **a** Cancer protein association network (CPAN). **b** Non-cancer protein association network (NPAN). The node size is proportional to the CRV for each protein and the edge width represents the magnitude of the association ability among two proteins. The figures are created using cytoscape [56]

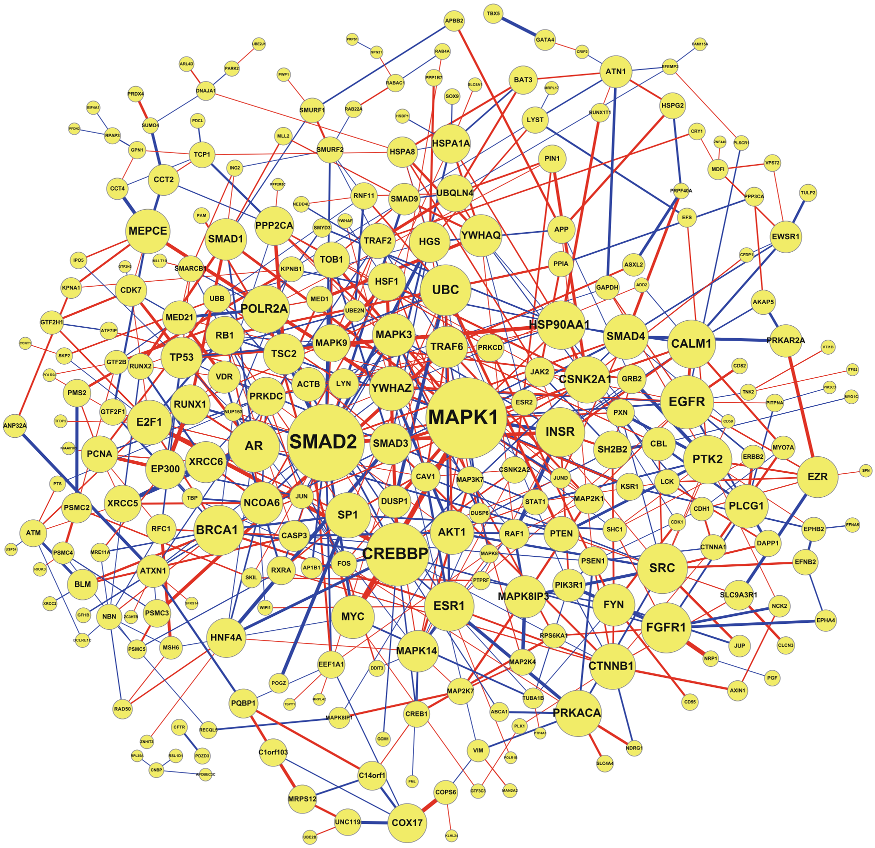


Fig. 1.4 The difference between CPAN and NPAN. The node size is proportional to the CRV for each protein and the edge width represents the magnitude of the association ability among two proteins. Red and blue edges indicate the positive and negative values of d_{ij}'s in Eq. (1.4), respectively. The figure is created using cytoscape [56]

protein association model Eq. (1.1) were further employed to prune the rough PPI network. The CPAN and NPAN, which consisted of 399 and 393 protein associations respectively, would constitute the network biomarker of lung cancer (Fig. 1.3). The difference between CPAN and NPAN was further shown in Fig. 1.4. According to the CPAN and NPAN with quantitative association abilities, the CRVs for each protein were computed and the significance of these CRVs was determined. Consequently, 40 proteins were identified to play significant roles in lung carcinogenesis and were shown in Table 1.1.

Table 1.1 The identified significant proteins in lung carcinogenesis

Protein symbol[a]	CRV	*p*-value	Functional annotation[b]			Literature evidence[c]
			Cell growth	Cell survival	Cell migration	
MAPK1	8.3418	<1e-5	+	+	+	[29, 67]
SMAD2	7.7901	<1e-5	+	+	+	[5]
CREBBP	5.7870	0.00002	+			[34]
EGFR	4.3635	0.00086	+	+	+	[16, 25, 38]
AR	4.0966	0.00159	+	+	+	[64]
UBC	4.0331	0.00180				
SRC	3.9446	0.00218	+	+	+	[7, 44]
FGFR1	3.9227	0.00237	+		+	[4]
BRCA1	3.9049	0.00243	+	+		[74]
ESR1	3.8409	0.00295	+	+	+	[24]
INSR	3.7946	0.00329	+		+	[13]
PTK2	3.6758	0.00432	+	+	+	[43, 44]
HSP90AA1	3.6732	0.00436	+	+	+	[20]
CALM1	3.6363	0.00482		+		
POLR2A	3.5701	0.00547				
CSNK2A1	3.4128	0.00761	+	+		[69]
PRKACA	3.3688	0.00856		+		
CTNNB1	3.2935	0.00994	+	+	+	[3]
SP1	3.2397	0.01133	+	+		[15]
SMAD4	3.1947	0.01266	+	+	+	[5]
E2F1	3.1382	0.01407	+	+		[30]
YWHAZ	3.1212	0.01467	+			[39]
MEPCE	3.0968	0.01545				
AKT1	3.0193	0.01857	+	+	+	[75]
PLCG1	2.9654	0.02069			+	[54]
MYC	2.8987	0.02385	+	+		[77]
MAPK3	2.8545	0.02654	+	+	+	[29, 67]
NCOA6	2.8132	0.02892	+	+		
FYN	2.7833	0.03089	+		+	[10]
MAPK8IP3	2.7746	0.03141			+	
YWHAQ	2.7582	0.03242	+			[70]
TRAF6	2.7150	0.03535		+		[31]
SMAD1	2.6940	0.03697	+	+	+	[37]
SMAD3	2.6815	0.03815	+	+	+	[5]
MAPK14	2.6727	0.03894	+	+	+	[66]
TP53	2.6522	0.04056	+	+	+	[16, 25, 28]
XRCC6	2.6270	0.04263		+		
EZR	2.6213	0.04314			+	[14]
TSC2	2.6116	0.04401	+	+	+	[40]
HGS	2.5730	0.04744	+			

[a]The full names of these proteins according to UniProt database http://www.uniprot.org/ are listed in Appendix 3

[b]The functional annotations are from the Gene Ontology database http://www.geneontology.org/ and literatures

[c]The literature evidences indicate that overexpression/dysregulation of the specific protein or mutation of the corresponding gene would result in carcinogenesis