# Systems Cancer Biology

# Systems Cancer Biology

By

Bor-Sen Chen

Cambridge
Scholars
Publishing

Systems Cancer Biology

By Bor-Sen Chen

This book first published 2021

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library



ISBN (10): 1-5275-6974-8
ISBN (13): 978-1-5275-6974-4

# CONTENTS

# PREFACE

Cancer is an extremely complex and heterogeneous disease that exhibits a high level of robustness against a range of therapeutic efforts. Cancer is considered as the emperor of all maladies by Siddhartha Mukherjee in his book of "A Biography of Cancer" and is driven by the somatic evolution of all lineages that have escaped control on replication and by the population level evolution of genes that influence cancer risk. Cancer risk appears to follow more or less inevitably from the combination of multicellularity, cell replacement and genetic and epigenetic changes that occurs a long period of time. The development of most cancers requires a series of nested mutations in caretaker, gatekeeper, landscaper and other genes. Hallmarks of cancer include (i) self-sufficiency of cells in signals controlling growth, (ii) loss of sensitivity to antigrowth signals, (iii) evasion of apoptosis via mutation or loss of gatekeeper genes, (iv) development of limitless replicative potential, (v) sustained angiogenesis, and (vi) tissue invasion and metastasis. The acquisition of these hall marks is an evolutionary and systematically developmental process involving selection among variant cells, the stabilization of gene expression patterns and heterochronic changes. Molecular biologists have long recognized carcinogenesis as an evolution process involving natural selection among renegade cells. Recently, the complex genetic and epigenetic network and the evolutionary forces in cancer have come under the focused scrutiny of systems molecular biologists, evolutionary biologists and ecologists, and this disciplinary crossover has begun to yield significant insights into the carcinogenesis process.

In this book, due to complex, heterogeneous and evolving nature of cancer, we will investigate the cellular mechanism of different cancer cells in the carcinogenetic process by genome-wide highthroughput data and big data mining from the systems biology perspective. Cell cycle is an important clue to unravel the mechanism of cancer cells. Therefore, it is more appealing to first construct a dynamic model for gene regulatory network (GRN) of cancer cell cycle to gain more insight into the infrastructure of gene regulatory mechanism of cancer by genome-wide microarray data.

Then we compare the protein-protein interaction (PPI) network in cancerous and normal cells to shed light on the molecular mechanisms of carcinogenesis. After constructing GRN and PPI network by microarray data, we could estimate their network robustness and sensitivity and then compare network robustness and sensitivity between cancer and normal cells from systematic point of view. It is found that the cancer network is more robust and less sensitive than normal cells and this systematic property is useful for robustness-based cancer drug design. Further, comparing the PPI networks between cancer and normal cells, we could find significant network marker for molecular mechanism investigation and as drug targets for diagnosis of cancer. We could also find core and specific network markers of carcinogenesis from multiple cancer samples. From different stages of cancer samples, we could find the evolution of network markers. The findings of core and specific network markers are new clues for targeted cancer therapy. Based on high throughput next generation sequence (NGS) data, the integrated genetic and epigenetic network is also constructed to investigate the dysfunctions of cellular system in cancer cells to shed light on the cellular systems mechanism of carcinogenesis from the genetic and epigenetic perspective. Therefore, we also provide a new method of how to select multiple drug targets from these genetic and epigenetic network markers of carcinogenic mechanism in cancer cell. According to these drug targets and drug target interaction (DTI) data, multi-molecule drug could be designed for cancer by combining several molecular drugs with adequate regulation ability and less side-effect on the drug targets. Recent analyses of the evolution of cancer at the population level show how rapid changes in human environments have augmented cancer risk. Finally, we discuss the systematic evolution of network robustness and response ability of GRN by time profile microarray data in the carcinogenesis process from the systems evolutionary biology perspective. Further, based on the nonlinear Poisson genetic variation process, the stochastic evolutionary Nash game between cancerous and healthy cells in the carcinogenesis process is also introduced to investigate the natural selection force on cell network of organ in the carcinogenesis process.

This monograph describes the current systems and evolutionary developments in cancer for systems biologists, molecular biologists, evolutionary biologists and cancer biologists. Cancer biology research is currently undergoing revolutionary changes due to integrating powerful technologies for this complex and heterogeneous disease. Systems biology sits at the heart of a new integrative 21$^{st}$ century paradigm. It integrates bioinformatics, physics, chemistry, mathematics, computer science, and

engineering for analyzing biological data, modeling dynamic equations to understand carcinogenesis mechanism of cancer biological systems, investigating network (multiple) markers for drug targets and designing multiple drugs. Obviously, systems cancer biology will provide a platform to integrate different fields of biologists, scientists, mathematicians and engineers for cancer research. Finally, I would like to thank Dr. Yung-Hao Wong and Ms. Chih-Ying Wang for their careful typing of this book.

*Bor-Sen Chen*
National Tsing Hua University

# CHAPTER 1

# INTRODUCTION TO SYSTEMS CANCER BIOLOGY

## 1.1 Introduction

Cancer is due to failures of the mechanisms that usually control the growth and proliferation of cells. Worldwide, between 100 and 350 of each 100,000 people die of cancer each year [1]. In industrial countries where the average life expectancy is high, cancer is one of the major causes of death. During normal development and through adult life, intricate genetic control systems regulate the balance between cell birth and death in response to growth signals, growth-inhibiting signal and death signals. In some adult tissues, cell proliferation occurs continuously as a constant tissue-renewal strategy. The cells in many adult tissues, however, normally do not proliferate except during healing processes [1]. The losses of cellular regulation give rise to most cases of cancer. When some errors occur in the control systems causing cells to proliferate continuously, tumor just comes into being.

However, cancer is a complex and heterogeneous (i.e. systematic) disease that exhibits high levels of robustness against various therapeutic interventions. It is a constellation of diverse and evolving disorders that are manifested by controlled proliferation of cells that may eventually lead to fatal dysfunction of the host system. Although some cancer subtypes can be cured by early diagnosis and specific treatment, no effective treatment is yet established for a significant portion of cancer subtypes [2]. Due to the complex, heterogeneous, and evolving nature of cancer, it is essential for a system-oriented view to be adopted for an in-depth understanding of cancer which shall eventually lead to better care and treatment for patients.

## 1.2 Why systems cancer biology

The Human Genome project and high-throughput experimental methodologies such as microarray Chromatin-immunoprecipitation DNA

Chips (CHIP-chips) have led to the development of biology as an increasingly information-rich science encompassing transcriptomes, proteomes, metabolomes, interactomes, and so forth [3-4, 16-17]. Although large-scale experiments are now being deployed, there are practical limitations of how much they do to convey the reality of cancer pathology and progression with the patient's body. Therefore how to achieve an in-depth yet systematic understanding of cancer dynamics is an important topic in systems cancer biology. The systems biology approach with system-oriented thinking and understanding may complement the limitations of an experimental approach in the era of large-scale experimental data. Systematic studies not only provide us with new insight from the system modeling and large scale experimental data, but also enable us to perceive what are the systematic characteristics of cancer under some system modeling. It is an engine of thoughts and proving grounds of various system models and hypotheses on how cancer may behave as well as how molecular mechanisms work within anomalous conditions. Systems biology is not only just a computing strategy that helps us understand and then fight against cancer, but also a systematic approach to be combined with a proper theoretical framework that enables us to perceive "cancer" as complex dynamical and evolvable systems that entail a robust yet fragile nature [2, 5].

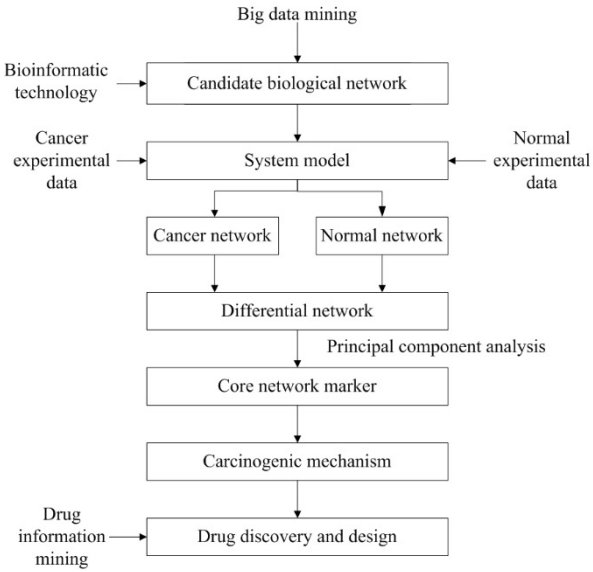


**Figure 1.1 The flowchart of systems cancer biology**

## 1.3 Outline of systems cancer biology

In this book, we hope to enlighten the reader on the systems biology approaches with application to systems cancer research. These approaches offer new promising systems insights into dynamic mechanism of carcinogenesis to defeat cancer. The first chapter introduces the general context of the book and the spirit of the book is described to provide reading guidelines. Before introducing the systems biology approaches to cancer biological network, some preliminary mathematics for system modeling and system identification methods for biological networks based on experimental data are introduced in Chapter 2. In this book, based on big data mining, we could construct candidate gene regulatory network (GRN) or candidate protein-protein interaction network (PPIN) for different kinds of cancers. Then we want to prune false positive regulations in candidate GRN or false positive interactions in candidate PPIN from the corresponding cancer microarray or NGS data to obtain the real GRN or PPIN of the cancer. In this situation, we need to develop the gene regulatory model for candidate GRN or protein interaction model for candidate PPIN. Then by the parameter estimation algorithm (i.e. the so-called reverse engineering method), we could identify the regulatory parameters in GRN and interaction parameters in PPN. Then, based on the system order detection method to determine the true number of gene regulations in GRN or the true number of protein interactions in PPIN, we could prune the false positive gene regulations out of true number of gene regulations to obtain real GRN in cancer or the false positive protein interactions out of true number of protein interactions in candidate PPIN to obtain real PPIN of cancer. For the systematic analyses of GRN and PPIN in the carcinogenesis process, some system characteristics such as stability, robustness and transductivity of biological systems of cancer cells are also introduced in Chapter 2 and investigated in Chapter 5 for their changes in carcinogenesis.

In Chapter 3, since cell cycle is an important clue to unravel the mechanism of cancer cells, it is more appealing to construct a dynamic model for gene regulatory network of cancer cell cycle to gain more insight into the infrastructure of gene regulatory mechanism of cancer cell via microarray data and database mining. Since cancer is caused by genetic abnormalities, such as mutations of oncogenes or tumor suppressor genes, to alter downstream signal transduction pathways and protein-protein interactions (PPIs), comparison of the protein interactions of PPINs in cancerous and normal cells can shed light on the mechanisms of carcinogenesis [6]. In Chapter 4, a cancer-perturbed PPIN is constructed by comparing PPIN of cancerous cells with PPIN of normal cells, which are

constructed from the corresponding microarray data. The rules in the cancer-perturbed PPIN of apoptosis can provide insight into the mechanism of apoptosis in the carcinogenesis process and allow the identification of potential drug targets [7].

Since robustness is defined as the ability to uphold system performance in face of perturbations and uncertainties, and sensitivity is a measure performance in deviation generated by perturbation to the system, robustness and sensitivity are two powerful systems biology approaches for systems analysis in cancer cells through mathematical model and experimental data. While cancer appears as a robust but fragile system, few computational and quantitative evidences demonstrate robustness tradeoffs in cancer. In Chapter 5, the comparisons of network robustness and sensitivity between cancer and normal cells are given by their corresponding microarray data. We provide several efficient computational methods based on system and control theory to compare network robustness and sensitivity between cancer and normal cells via microarray data. These methods provide network models for the robustness-based cancer drug design [8].

Lung cancer is the leading cause of cancer deaths worldwide. Many studies have investigated the carcinogenic process and identified the biomarker for signature classification. However, based on the research dedicated to this field, there is no highly sensitive network-based method for carcinogenesis characterization and diagnosis from the systems perspective. In chapter 6, a systems biology approach integrating microarray gene expression profiles and protein-protein interaction information from database mining was proposed to develop a network-based biomarker through network sensitivity during carcinogenesis for molecular investigation into the network mechanism of lung carcinogenesis and diagnosis of lung cancer. Based on the network-based biomarker, a total of 40 significant proteins in lung carcinogenesis were identified with carcinogenesis relevance values [9]. After identifying the network marker of lung cancer in chapter 6, since non-small-cell lung cancer (NSCLC) is the predominant type of lung cancer in the world and lung adenocarcinoma (LADC) and lung squamons cell carcinoma (LSCC) are subtypes of NSCLC, further investigation of genetic and epigenetic progression mechanism of two subtypes of NSCLC is necessary. Since the differences of genetic and epigenetic progression mechanism between LADC and LSCC are complicated to analyze, systems biology method is used in chapter 7 to construct their genetic and epigenetic networks by NGS data to investigate their genetic and epigenetic mechanisms of LADC and LSCC for the identification of their significant genetic and epigenetic biomarkers for drug design. In chapter 8, a systems biology approach is further used to construct the PPIN

of four cancers, i.e. bladder cancer, colorectal cancer, liver cancer, lung cancer and non-cancer by their corresponding microarray data, PPI modeling and big database-mining. By comparing PPI networks between cancers and non-cancer samples to find significant proteins with large PPI changes during carcinogenesis, core and specific protein network markers are identified by the intersection and difference between significant proteins of core networks in cancer and non-cancer cells. From these core and specific protein network markers, we could not only gain an insight into crucial common and specific pathways in carcinogenesis, but also obtain a high promising PPI drug target for cancer therapy [10]. After investigating core and specific biomarkers among different cancers and between different stages of a cancer in chapter 8, multiple-molecule drug targeting on common core network marker is investigated in chapter 9 to give us a direction for cancer therapy. Then, we will focus on the evolution of network biomarkers from the early to late stage bladder cancer in chapter 10. By comparing the networks of these two stages of the carcinogenesis process, we find that both cancer networks showed very significantly different evolutionary mechanisms of cancers [11]. To investigate the evolution of network biomarkers in the carcinogenesis of bladder cancer, primary pathway analysis is employed to find significant pathways related to cancer mechanisms at the early and late stage bladder cancer [12].

Epigenetic and microRNA (miRNA) regulation are found to be associated with carcinogenesis and development of cancer. In chapter 11, we will identify genome-wide genetic and epigenetic network (GWGEN) of bladder cancer by next-generation sequential (NGS) data and then investigate network biomarkers of different stages of carcinogenesis of bladder cancer. Then a systematic design of multiple-molecule drug for each carcinogenesis stage of bladder cancer with minimum side-effects is also introduced in this chapter.

Thyroid cancer is the most common endocrine cancer, particularly, papillary thyroid cancer (PTC). Up to now, there are few researches about the pathogenesis and progression mechanisms of PTC from viewpoint of systems biology. In chapter 12, systems biology approach is introduced to investigate genetic and epigenetic molecular progression mechanisms for identifying multiple biomarkers as drug targets in PTC. Finally based on the identified multiple biomarkers, we suggest potential candidate multiple-molecule drug to prevent the progression of PTC with querying Connectivity Map (CMap). Since colorectal cancer (CRC) is the third most diagnosed cancer all over the world, the mechanisms leading to the development and progression of CRC are complicated and implicated in both genetic and epigenetic dysregulation. In chapter 13, systems biology

method and genome-wide data are employed to construct GRNs and PPINs to be combined as GWGENs of each stage of CRC. By principle network projection (PNP) method, core GWGENs are extracted from GWGENs and then projected to KEGG pathways to obtain the core signaling pathways in each progression stage of CRC. By comparing these core pathways of neighboring stages, we could find carcinogenic biomarkers as drug targets in each stage of CRC. By drug data mining, several multiple-molecule drugs, are selected at each stage to prevent the progression of CRC.

Hepato cellular carcinoma (HCC) is the fifth most common type of cancer and the third leading cause of cancer-related deaths worldwide. The mechanisms leading to the development and progression of HCC are complicated and regulated genetically and epigenetically. In chapter 14, the approach to HCC carcinogenesis involves analyzing GRNs, PPINs and epigenetic networks at different stages of HCC, which are identified by the corresponding NGS data and big data mining method. Core genetic and epigenetic networks are extracted by PNP and projected to KEGG pathways, we could find the carcinogenic biomarkers as drug targets of HCC at different stages for the therapeutic treatment of each stage HCC.

## 1.4 Conclusion

Cancer is a complex, heterogeneous and evolving disease. In order to focus on these natural properties of cancer for an in-depth understanding of cancer, it is essential from a system-oriented view. Therefore, we investigate cellular mechanism of cancer cells of different kinds of cancers in the carcinogenesis process through their genome-wide high throughput data by systems biology method. After investigating cellular carcinogenetic mechanism of cancer cells, we select significant multiple biomarkers as drug targets. Then based on drug targets and big drug data mining, candidate multiple-molecule drugs are selected for better therapeutic treatment of cancer patients.

# CHAPTER 2

# SYSTEM MODELING AND PARAMETER IDENTIFICATION


## Abstract

Since cancers are mainly due to the dysfunctions in signal transduction pathways (STPs) and their downstream gene-regulatory networks (GRNs), the system models of STPs and GRNs and their parameter estimation methods are introduced in this chapter for the convenience of further system identification analysis in the following chapters. In this chapter, we first introduce linear dynamic biological networks. Then we give the fundamental parameter estimation methods of GRNs and STPs in biological networks by microarray data. Finally, system order detection method is introduced to prune false positives in candidate network constructed by database mining to obtain true network by microarray data.


## 2.1 Introduction to linear dynamic biological networks

Since the cellular dysfunctions of signal transduction pathways play important roles in the carcinogenic process, the signal flow of transduction pathways will be discussed at first. Consider the coupling signal transduction pathways in Fig 2.1 throughout cellular stress responses. The receptors in the cell membrane sense extracellular signals (ligands) $u_i(t)$. $y_i(t)$ denotes the expression level of the $i$th protein in the coupling signal transduction pathways. They are commuted to intracellular signals and sequences of reactions. Different external changes or events outside the cell may stimulate signalings. Typical extracellular signals are hormones, pheromones, heat, cold, light, osmotic pressure, pathogen, and appearance or concentration change of substance such as glucose, potassiumion, calciumion or cAMP. The extracellular signals $u_i(t)$ for $i$=1,…,$l$ are perceived by a transmembrane receptor, as depicted in Fig 2.1. The receptor changes its own state from susceptible to active and then triggers subsequent processes within the cell. The active receptor stimulates an internal

signaling cascade. This cascade frequently includes a series of changes in protein phosphorylation states. The sequence of state changes crosses the nuclear membrane. Eventually the transcription factor (TF) is activated or deactivated to change its binding activity to a set of genes. In the signal flow chart of simple coupling signal transduction pathways in Fig 2.1, $y_{13}(t)$, $y_{14}(t)$, $y_{15}(t)$ and $y_{16}(t)$ represent the expression levels of terminal TFs in the simple coupling signal transduction pathways.

For the purpose of parameter estimation of the coupling signal transduction pathways in Fig 2.1, a linear regression model for the expression level of the $i$th protein at time $t$+1 can be given as

$$\begin{aligned} y_i(t+1) &= C_{i,1}y_1(t) + \cdots + C_{i,i-1}y_{i-1}(t) + C_{i,i}y_{i+1}(t) + \cdots \\ &\quad + C_{i,M}y_M(t) + h_i + \sum_{i=1}^{l} b_{i,j}u_j + w_i(t), \quad \text{for } i = 1,...,M \end{aligned} \tag{2.1}$$

where $y_i(t)$ indicates the expression level of the $i$th protein at time $t$; $C_{i,j}$ denotes the interaction ability between protein $i$ and protein $j$; $h_i$ denotes the based level of the $i$th protein; and $b_{i,j}$ denotes the binding ability of extracellular signal $u_i(t)$ to the $i$th protein, $w_i(t)$ denotes the model residue or measurement noise.

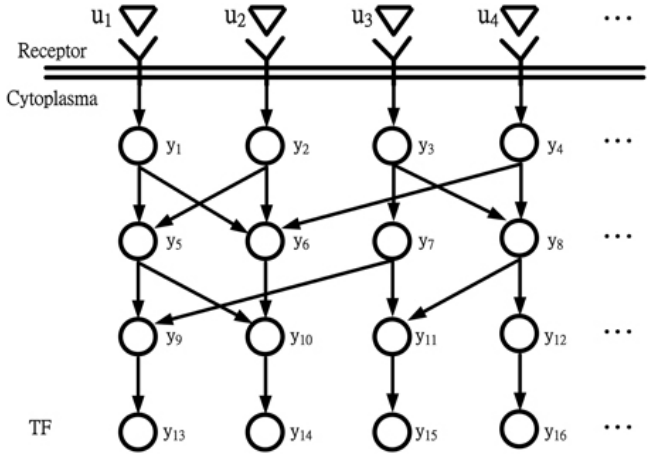


**Figure 2.1** The coupling signal transduction pathways: $u_1(t)$, $u_2(t)$, … denote the extracellular signals; $y_{13}(t)$, $y_{14}(t)$, … are the expression levels of terminal transcription factors (TFs). The protein-protein interaction network (PPIN) can be modeled by linear dynamic system in (2.2) or (2.4) or static system in (2.25)-(2.27)

In general, extracellular signals always bind to the receptor proteins on

the membrance. Let us denote the state vector and system matrix of coupling signal transduction pathways of Fig. 2.1 in (2.1) as follows:

$$y(t)=\begin{bmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_M(t) \end{bmatrix}, C=\begin{bmatrix} C_{1,1} & \cdots & C_{1,M} \\ \vdots & & \\ C_{M,1} & \cdots & C_{M,M} \end{bmatrix}, H=\begin{bmatrix} h_1 \\ \vdots \\ h_M \end{bmatrix}, B=\begin{bmatrix} b_{1,1} & \cdots & b_{1,l} \\ \vdots & & \\ b_{M,1} & \cdots & b_{M,l} \end{bmatrix},$$

$$u(t)=\begin{bmatrix} u_1(t) \\ \vdots \\ u_l(t) \end{bmatrix}, w(t)=\begin{bmatrix} w_1(t) \\ \vdots \\ w_M(t) \end{bmatrix}$$

Then the protein-protein interaction network (PPIN) of coupling signal transduction pathways in Fig. 2.1 is represented by

$$y(t+1)=Cy(t)+H+Bu(t)+w(t) \tag{2.2}$$

If we want to know the information flow from $u(t)$ to any protein $i$, then the coupling signal transduction dynamical system in (2.2) is represented by

$$\begin{aligned} y(t+1)&=Cy(t)+H+Bu(t)+w(t) \\ y_i(t)&=D_i y(t)=[0\cdots 010\cdots 0]y(t) \end{aligned} \tag{2.3}$$

and the information flow from $u(t)$ to the protein $y_i(t)$ is given by

$$y_i(t)=\sum_{k=0}^{t} D_i C^{t-k} Bu(k)$$

where $y_i(t)$ denotes the expression of the $i$th protein and $D_i$ is a row vector with all zeros except 1 at the $i$th element.

**Remark 2.1**

In the continuous time case, the linear dynamic system in (2.2) is modified as

$$\dot{y}(t)=Cy(t)+H+Bu(t)+w(t) \tag{2.4}$$

And (2.3) is modified as

$$\begin{aligned} \dot{y}(t)&=Cy(t)+H+Bu(t)+w(t) \\ y_i(t)&=D_i y(t) \end{aligned} \tag{2.5}$$

Then the information flow $u(t)$ to the $i$th protein is given by [19]

$$y_i(t)=D_i\int_0^t e^{C(t-\tau)}Bu(\tau)d\tau$$

## 2.2 Parameter estimation methods of PPIN

In the following, we want to estimate the system matrices *C, H, B* of PPIN from the time-profile protein expression data or protein real-time PCR data. In general, we do not identify *C, H, B* from PPIN in (2.3) directly owing to its complex computation with much more round off error in the parameter identification process. We want to estimate these parameters protein-by-protein from (2.1). From (2.1), these protein interaction equations are represented by the following regression form:

$$\begin{aligned} y_i(t+1) &= [y_1(t)\cdots y_M(t) \quad u_1(t)\cdots u_l(t) \quad 1]\begin{bmatrix} c_{i,1} \\ \vdots \\ c_{i,M} \\ b_{i,1} \\ \vdots \\ b_{i,l} \\ b_i \end{bmatrix} + w_i(t) \\ &= \phi(t)\theta_i + w_i(t), \quad \text{for } i = 1,...,M \end{aligned} \tag{2.6}$$

with regression vector $\phi(t)$ and parameter vector $\theta_i$.

Assume that there are $n$ parameters in $\theta_i$ to fit the linear system model of the $i$th protein in (2.1) to $N$ time profile data, for example, by real-time PCR data. In order to avoid the overfitting in the parameter estimation procedure, $N$ should be at least 2 times than the number $n$ of parameters to be estimated. The parameter estimation problem is to find an estimate $\hat{\theta}_i$ of the parameter vector $\theta_i$ from the following $N$ time profile data $\{y_i(t)\}$ and $\{\phi(t)\}$, i.e.

$$\begin{aligned} y_i(1) &= \phi(0)\theta_i + w_i(0) \\ y_i(2) &= \phi(1)\theta_i + w_i(1) \\ &\vdots \\ y_i(N) &= \phi(N-1)\theta_i + w_i(N-1) \end{aligned} \tag{2.7}$$

In matrix notation, we obtain the vectors of $N$ time profile data $Y_i(N)$, the regression matrix $\Phi(N)$, and noise $W_i(N)$, respectively, as follows

$$Y_i(N)=\begin{bmatrix} y_i(1) \\ y_i(2) \\ \vdots \\ y_i(N) \end{bmatrix}, \Phi(N)=\begin{bmatrix} \phi(0) \\ \phi(1) \\ \vdots \\ \phi(N-1) \end{bmatrix}, W_i(N)=\begin{bmatrix} w_i(0) \\ w_i(1) \\ \vdots \\ w_i(N-1) \end{bmatrix} \tag{2.8}$$

Finally, the resulting parameter estimation model for the linear regression system in (2.7) is

$$Y_i(N)=\Phi(N)\theta_i+W_i(N) \tag{2.9}$$

### 2.2.1 Least square parameter estimation method

By the least square parameter estimation method [19], we select parameter $\hat{\theta}_i$ to minimize the following square estimation error

$$V(\hat{\theta}_i)=\frac{1}{2}(Y_i(N)-\Phi(N)\hat{\theta}_i)^T(Y_i(N)-\Phi(N)\hat{\theta}_i) \tag{2.10}$$

The optimal estimate $\hat{\theta}_i$ to minimize the square estimation error in (2.10) is obtained as [19]

$$\begin{aligned} \hat{\theta}_i &= (\Phi^T(N)\Phi(N))^{-1}\Phi^T(N)Y_i(N) \\ &= \Phi^{\dagger}(N)Y_i(N) \end{aligned} \tag{2.11}$$

where $\Phi^{\dagger}(N)$ denotes the pseudo inverse of $\Phi(N)$.

The covariance of parameter estimation error $\tilde{\theta}_i=\theta_i-\hat{\theta}_i$ is [19]

$$Cov(\tilde{\theta}_i)=\sigma_{W_i}^2(\Phi^T(N)\Phi(N))^{-1} \tag{2.12}$$

where $\sigma_{W_i}^2$ is the covariance of noise vector $W_i(N)$, i.e. $Cov(W_i(N))=\sigma_{W_i}^2 I$.

The estimate of noise covariance $\sigma_{W_i}^2$ is [19]

$$\sigma_{W_i}^2=\frac{2}{N-n}V(\hat{\theta}_i) \tag{2.13}$$

**Remark 2.2**

(i) For the parameter estimate of *C, H, B* in linear dynamic system of PPIN in (2.2), it is more appealing to estimate *C, H, B* one row by one row as the parameter estimation method in (2.11), i.e. $\hat{\theta}_i$, $i$=1,…,$M$ for each protein and finally $[C, H, B] = [\hat{\theta}_1 \hat{\theta}_2 ... \hat{\theta}_M]^T$ for whole biological network, i.e. one by one, we could identify the whole PPIN iteractively. In general, since the scale of PPIN is very large, it is not suitable to estimate [*C*, *H*, *B*] of PPIN directly because of its complexity with much more round off error in the parameter estimation process. Actually, the two results are equivalent.

(ii) At present, it is still lacking in genome-wide protein expression data for parameter estimation in (2.11). In this situation, genome-wide gene expression data in microarray, which is more available could be employed to replace the corresponding genome-wide protein expression data in (2.6) for parameter estimation in (2.11). If there exists some ratio between protein $y(t)$ and gene expression data $x(t)$, for example $y(t)=kx(t)$ for some $k$, then the constant $k$ could be canceled in both sides of (2.6). Therefore we could use gene expressions in microarray to replace protein expression in the parameter identification of protein-protein interaction network.

(iii) The block least square parameter estimation in (2.11) is equivalent to the following recursive least square parameter estimation algorithm [19]

$$\hat{\theta}_i(t+1) = \theta_i(t) + P_i(t)\phi_i(t)\left(y_i(t) - \phi_i^T(t)\hat{\theta}_i(t-1)\right)$$

$$P_i(t) = P_i(t-1) - \frac{P_i(t-1)\phi_i(t)\phi_i^T(t)P_i(t-1)}{1+\phi_i^T(t)P_i(t-1)\phi_i(t)}, \text{ for } i = 1,...,M \qquad (2.14)$$

$\hat{\theta}_i(o)$ and $P_i(o)$ are given.

This recursive least square parameter estimation can use time-profile protein data to update the system parameters protein-by-protein. Therefore, it can be used for real-time parameter estimation. If the number of time-profile data $y_i(t)$ is small, we can repeat several rounds of the recursive parameter estimation algorithm in (2.14) with the previous result as initial parameter estimate $\hat{\theta}_{ii}(o)$ and initial $P_i(o)$ to achieve the optimal parameter estimate.

### 2.2.2 Maximum likelihood parameter estimation method

Another famous parameter estimation method for $\theta_i$ in (2.9) is maximum likelihood parameter estimation algorithm, i.e. to estimate parameter $\theta_i$ from the system output measurement $Y_i(N)$. In general, the value $\theta_i$ should be consistent with the largest probability under the measurement $Y_i(N)$, i.e. the estimate of $\theta_i$ should satisfy the maximum probability [19]

$$\max_{\theta_i} P(\theta_i / Y_i(N)) \tag{2.15}$$

By the Bayes' condition probability

$$P(\theta_i / Y_i(N)) = \frac{P(\theta_i)P(Y_i(N) / \theta_i)}{P(Y_i(N))} \tag{2.16}$$

In general, it is difficult to calculate $P(\theta_i / Y_i(N))$ because we need three probability density functions $P(\theta_i)$, $Y_i(N)$ and $P(Y_i(N) / \theta_i)$. Since $P(Y_i(N))$ is not explicit with $\theta_i$, it does not influence the optimal estimation in (2.15). $P_i(\theta_i)$ denotes the probability density of $\theta_i$, i.e. the priori information of $\theta_i$.

In general, we have no information of $P(\theta_i)$ and always assume $P(\theta_i)$=constant, i.e., all value $\theta_i$ may occur with the same probability. In this situation the optimal parameter estimation method (i.e. Maximum a prior (MAP)) in (2.15) is equivalent to the following maximum likelihood estimation (MLE) method

$$\max_{\theta_i} P(Y_i(N) / \theta_i) \tag{2.17}$$

i.e., how to select $\theta_i$ so that the probability of $Y_i(N)$ is maximum to exploit the appearance of $Y_i(N)$ being with the largest probability.

Since $P(Y_i(N) / \theta_i)$ denotes the likelihood function, it means that the selection of $\theta_i$ should make the output data $Y_i(N)$ must be consistent (likelihood) with the largest probability to exploit the occurrence of output data. From (2.9), it is seen that the probability density function $P(Y_i(N))$ of $Y_i(N)$ is mainly due to the white noise of $W_i(N)$. Therefore, $Y_i(N)$ and $W_i(N)$ have the same probability density function. If we assume $W_i(N)$ is zero-mean Gaussian noise with covariance $\sigma_{W_i}^2 I$, then

$$P(W_i(N)) = \frac{1}{(2\pi\sigma_{W_i}^2 I)^{N-\frac{1}{2}}} \exp\left(-\frac{1}{2\sigma_{W_i}^2} W_i^T(N) W_i(N)\right)$$
$$= \frac{1}{(2\pi\sigma_{W_i}^2)^{N-\frac{1}{2}}} \exp\left(-\frac{1}{2\sigma_{W_i}^2} (Y_i(N) - \Phi(N)\theta_i)^T (Y_i(N) - \Phi(N)\theta_i)\right) \tag{2.18}$$

In general, the notation $\theta_i$ in (2.17) of the likelihood function is always neglected. Therefore the maximum likelihood method in (2.17) is equivalent to how to select $\theta_i$ to maximize log$P(W_i(N))$. The maximum likelihood estimation method in (2.17) is equivalent to the following likelihood optimization method

$$\max_{\theta_i, \sigma_{W_i}^2} \log L(\theta_i, \sigma_{W_i}^2) \tag{2.19}$$

where the log-likelihood function is denoted as

$$\log L(\theta_i, \sigma_{W_i}^2) = \log P(W_i(N)) \tag{2.20}$$

i.e., the log-likelihood method in (2.19) could simplify the parameter estimation procedure for (2.17).

Here we expect the log-likelihood function to have the maximum value at $\theta_i = \hat{\theta}_i$ and $\sigma_{W_i}^2 = \hat{\sigma}_{W_i}^2$. The necessary conditions of the maximum likelihood estimate $\hat{\theta}_i$ and $\hat{\sigma}_{W_i}^2$ must consist of

$$\frac{\partial \log L(\theta_i, \sigma_{W_i}^2)}{\partial \theta_i} = 0 \tag{2.21}$$

$$\frac{\partial \log L(\theta_i, \sigma_{W_i}^2)}{\partial \sigma_{W_i}^2} = 0 \tag{2.22}$$

After some computational arrangements from (2.21) and (2.22), the estimated maximum likelihood parameters $\hat{\theta}_i$ and $\hat{\sigma}_{W_i}^2$ are

$$\hat{\theta}_i = (\Phi_i^T(N)\Phi(N))^{-1} \Phi^T(N) Y_i(N) \tag{2.23}$$

$$\hat{\sigma}_{W_i}^2 = \frac{1}{N-1}(Y_i(N) - \Phi_i(N)\hat{\theta}_i)^T (Y(N) - \Phi_i(N)\hat{\theta}_i), \quad \text{for } i = 1, 2, ..., M \tag{2.24}$$

Under the Gaussian distribution of noise $W_i(N)$ in (2.9), the maximum likelihood parameter estimation $\hat{\theta}_i$ in (2.23) is equivalent to the least square parameter estimate $\hat{\theta}_i$ in (2.11).

## 2.3 Linear signal transduction models and their parameter identification

If the protein data for measuring the coupling STPs in Figure 2.1 is of one time-point data from different samples, for example, from $K$ patients, the regression model for protein expression level of the $i$th protein can not be represented by the discrete dynamic model in (2.1) but can be represented by the following linear static regression from

$$y_i(k) = \sum_{m=1}^{M_i} c_{i,m} y_m(k) + h_i + \sum_{l=1}^{l_i} b_{i,j} u_j(k) + w_i(k), \quad \text{for } i = 1, ..., M \tag{2.25}$$

where $k$=1,…,$K$ denote the samples of protein data from $K$ patients, $M_i$ denotes the number of proteins which have interactions with the $i$th protein; $M$ denotes the total number of proteins in STPs.

Then the regression model for the protein expression level of coupling STPs in Figure 2.1 can be represented by the following linear static regression form

$$y(k) = Cy(k) + H + Bu(k) + w(k), \ k = 1, ..., K \tag{2.26}$$

where $y(k) = [y_1(k)...y_M(k)]^T$ ; $u(k) = [u_1(k)...u_l(k)]^T$ ; $w(k) = [w_1(k)...w_M(k)]^T$ ; $C$, $H$, and $B$ are defined as in (2.2)

In this situation,

$$y(k) = (I - C)^{-1} Bu(k) + (I - C)^{-1} H + (I - C)^{-1} w(k) \tag{2.27}$$

From (2.27), it is seen that the transfer function $T$ from extracellular signals $u(k)$ to transcription factors $y(k)$ is

$$T = (I - C)^{-1} B \tag{2.28}$$

Therefore, if we want to estimate the signal transfer function $T$ in (2.28), we need to estimate system matrices $C$ and $B$ in (2.26) from sample protein expression data. From the linear regression model in (2.25), we get

$$\begin{aligned} y_i(k) &= [y_1(k)...y_M(k), u_1(k)...u_l(k), 1] \begin{bmatrix} c_{i,j} \\ \vdots \\ c_{i,M} \\ b_{i,1} \\ \vdots \\ b_{i,l} \\ h_i \end{bmatrix} + w_i(k) \\ &= \phi_i(k)\theta_i + w_i(k), \text{ for } k = 1, 2, ..., K \end{aligned} \tag{2.29}$$

The estimate of $\theta_i$ could follow the similar procedure (2.7)-(2.11) with time profile $\phi(t)$ replaced by sample profile $\phi(k)$ of different patients.

The recursive least square parameter identification for $\theta_i$ of the $i$th protein in (2.29) with $K$ sample protein data from $K$ patients is given as (2.14) by [19]

$$\begin{aligned} &\hat{\theta}_i(k) = \hat{\theta}_i(k-1) + P_i(k)\phi_i(k)\varepsilon_i(k), \ \hat{\theta}_i(o) \text{ and } P_i(o) \text{ are given} \\ &\varepsilon_i(k) = (y(k) - \phi_i^T(k)\hat{\theta}_i(k-1)) \\ &P_i(k) = P_i(k-1) - \frac{P_i(k-1)\phi_i(k)\phi_i^T(k)P_i(k-1)}{1 + \phi_i^T(k)P_i(k-1)\phi_i(k)}, \\ &\text{for } i = 1, ..., M, \ k = 1, ..., K \end{aligned} \tag{2.30}$$

If the sample number of protein expression data is small, we can repeat several rounds of the recursive parameter estimation algorithm in (2.30) with previous results as initial conditions $\hat{\theta}_i(o)$ and $P_i(o)$ to achieve the optimal parameter estimate. After the parameters $\theta_i$ for $i$=1,…,$M$ are estimated by the recursive least square estimation algorithm in (2.30) through protein expression data with $K$ samples, we can estimate $C$, $H$ and $B$ in (2.26). Then the transfer function $T$ from extracellular signals to transcription factors in (2.28) can be calculated through sample microarray data.

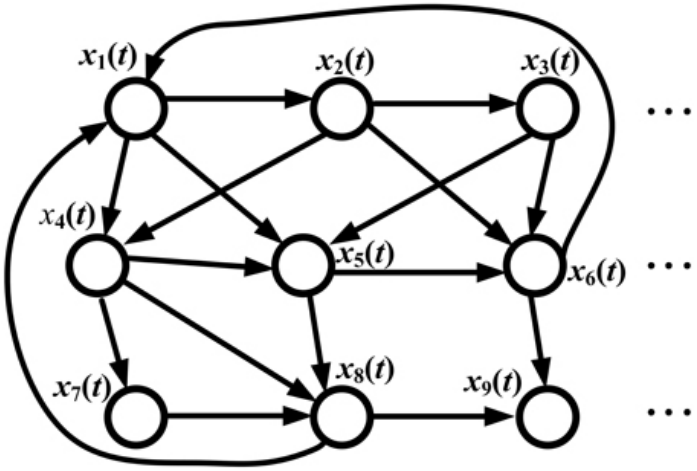


**Figure 2.2** A gene regulatory network (GRN): $x_i(t)$ denotes the gene expression of the $i$th gene. The GRN can be represented by the linear dynamic model in (2.33) or (2.37), or nonlinear dynamic model in (2.40). The GRN can be also modeled by the static system in (2.43)

## 2.4 System identification of GRNs by time profile microarray data

Consider the GRN Figure 2.2, the regulatory dynamic model of the $i$th gene can be represented by the following regression equation

$$x_i(t+1) = a_{i,1}x_1(t) + ... + a_{i,n}x_n(t) + v_i(t) \tag{2.31}$$

where $x_i(t)$ denotes the gene expression level of the $i$th gene; $v_i(t)$ denotes measurement noise and residue and $a_{i,j}$ denotes the regulatory ability of gene $j$ on gene $i$. Actually, the gene $j$ cannot regulate gene $i$ directly, but through its protein (TF) indirectly after its transcription and translation process.

**Remark 2.3**

(i) In general, a positive value of $a_{i,j}$ means the $j$th gene is an active regulator while a negative value of $a_{i,j}$ means the $j$th gene is an inhibitive regulator.

(ii) The regulator parameter $a_{i,j}$, for $j$=1,2,…,$n$ in (2.31) can be identified by the least square parameter estimation algorithm (2.11) or recursive least square parameter estimation algorithm in (2.14) through the corresponding time-profile microarray or NGS data $x_i(t)$, for $i$=1,…,$n$;

$t$=1,…,$N$.

Therefore, the GRN in Figure 2.2 can be represented as follows:

$$\begin{bmatrix} x_1(t+1) \\ \vdots \\ x_i(t+1) \\ \vdots \\ x_n(t+1) \end{bmatrix} = \begin{bmatrix} a_{1,1} & \cdots & a_{1,j} & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{i,1} & \cdots & a_{i,j} & \cdots & a_{i,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{n,1} & \cdots & a_{n,j} & \cdots & a_{n,n} \end{bmatrix} \begin{bmatrix} x_1(t) \\ \vdots \\ x_j(t) \\ \vdots \\ x_n(t) \end{bmatrix} + \begin{bmatrix} v_1(t) \\ \vdots \\ v_i(t) \\ \vdots \\ v_n(t) \end{bmatrix} \tag{2.32}$$

or

$$X(t+1) = AX(t) + V(t) \tag{2.33}$$

Suppose we want to calculate the regulatory function (regulatory information flow) from gene $j$ to gene $i$ in a GRN. In general, it is difficult to solve the regulatory function (information flow) problem for GRN in Figure 2.2 from the graph theory perspective [20], especially for a diagraph (directed graph) like Figure 2.2. In this section, an input/output state space method is proposed to solve this difficult regulatory information flow problem of the digraph as follows. First, the dynamic model of GRN in (2.32) can be represented by the following input/output dynamic state space equation

$$\begin{bmatrix} x_1(t+1) \\ \vdots \\ x_i(t+1) \\ \vdots \\ x_n(t+1) \end{bmatrix} = \begin{bmatrix} a_{1,1} & \cdots & a_{1,i} & \cdots & a_{1,j} & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{i,1} & \cdots & a_{i,i} & \cdots & (a_{i,j}-1) & \cdots & a_{i,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{n,1} & \cdots & a_{n,i} & \cdots & a_{n,j} & \cdots & a_{n,n} \end{bmatrix} \begin{bmatrix} x_1(t) \\ \vdots \\ x_i(t) \\ \vdots \\ x_j(t) \\ \vdots \\ x_n(t) \end{bmatrix} + \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} x_j(t) + \begin{bmatrix} v_1(t) \\ \vdots \\ v_i(t) \\ \vdots \\ v_n(t) \end{bmatrix}$$

$$x_i(t) = \begin{bmatrix} 0 & \cdots & 1 & \cdots & 0 \end{bmatrix} \begin{bmatrix} x_1(t) \\ \vdots \\ x_i(t) \\ \vdots \\ x_n(t) \end{bmatrix} \tag{2.34}$$

where $a_{i,j}$ in (2.32) is changed to be $(a_{i,j}-1)$ because an extra $x_j(t)$ is added in (2.34) as input.

In the input/output dynamic state space system (2.34), $x_j(t)$ is considered as an input signal and $x_i(t)$ as an output signal. Therefore, (2.34) is simply

represented by

$$\begin{aligned} X(t+1) &= A_j X(t) + B_j x_j(t) + V(t) \\ x_i(t) &= D_i X(t) \end{aligned} \tag{2.35}$$

which is similar to the signal transduction dynamic system in (2.3) but with the gene expression level $x_j(t)$ replacing the extracellular signal $u(t)$.

By system theory [21], the regulatory function or regulatory information flow from gene $j$ to gene $i$ in the GRN in Figure 2.2 is given by

$$x_i(t) = \sum_{k=0}^{t} D_i A_j^{t-k} B_j x_j(k) \tag{2.36}$$

**Remark 2.4**

(i) With the similar procedure from (2.34) to (2.36), we could also obtain the protein interaction information flow from protein $j$ to protein $i$ of PPIN in (2.2) as follows

$$y_i(t) = \sum_{k=0}^{t} D_i C_j^{t-k} B_j y_j(k)$$

where $C_j$ is the modification of $C$ with the element $C_{i,j}$ being replaced by $(C_{i,j} - 1)$ as $A_j$ being the modification of A in (2.34)

(ii) In the continuous dynamic model, the linear regulatory in (2.33) is modified as

$$\dot{X}(t) = AX(t) + V(t) \tag{2.37}$$

and (2.35) is modified as

$$\begin{aligned} \dot{X}(t) &= A_j X(t) + B_j x_j(t) + V(t) \\ x_i(t) &= D_i X(t) \end{aligned} \tag{2.38}$$

and (2.36) should be modified as

$$x_i(t) = D_i \int_0^t e^{A_j(t-\tau)} B_j x_j(\tau) d\tau$$

(iii) If the gene regulatory functions in (2.31) are nonlinear as follows

$$x_i(t+1) = a_{i,1} f(x_1(t)) + a_{i,2} f(x_2(t)) + \ldots + a_{i,n} f(x_n(t)) + v_i(t) \tag{2.39}$$

where $f_i(x_i(t))$, $i$=1,…,$n$ are nonlinear regulatory functions from different genes, then the gene regulatory dynamic of GRN in (2.33) should be

modified as

$$X(t+1) = Af(X(t)) + V(t) \tag{2.40}$$

where $f(X(t))$ denotes the nonlinear regulatory vector of GRN in Figure 2.2. The gene regulatory functions in (2.39) can be represented by

$$\begin{aligned} x_i(t+1) &= [f_1(x_1(t))...f_i(x_i(t))...f_n(x_n(t))]\begin{bmatrix} a_{i,1} \\ \vdots \\ a_{i,i} \\ \vdots \\ a_{i,n} \end{bmatrix} + v_i(t) \\ &= \phi(x(t))\theta_i + v_i(t) \end{aligned} \tag{2.41}$$

In this situation, the least square parameter estimation algorithm in (2.11) or the recursive least square parameter estimation algorithm in (2.14) can be also employed to estimate parameters of nonlinear regulatory system in (2.39) or (2.41).

## 2.5 System identification of GRNs by samples data

If the microarray data or NGS data for regulatory information of GRN in Figure 2.2 are of one time-point data from different samples of different patients, then the regression model for gene regulation in (2.31) is modified to the following

$$x_i(k) = a_{i,1}x_1(k) + ... + a_{i,n}x_n(k) + v_i(k), \text{ for } k = 1, 2, ..., K \tag{2.42}$$

where $x_i(k),\ldots, x_n(k)$ denote the gene expression levels of $n$ genes of GRN in Figure 2.2 at the $k$th sample. Therefore, the whole GRN in Figure 2.2 can be represented by

$$X(k) = AX(k) + V(k), \ k = 1, ..., K \tag{2.43}$$

where