Translational Bioinformatics 4
Series Editor: Xiangdong Wang, MD, PhD, Prof

Bairong Shen *Editor*

Bioinformatics for Diagnosis, Prognosis and Treatment of Complex Diseases

Springer

Translational Bioinformatics

Series editor

Xiangdong Wang, MD, Ph.D.
Professor of Clinical Bioinformatics, Lund University, Sweden
Professor of Medicine, Fudan University, China

Aims and Scope

The Book Series in Translational Bioinformatics is a powerful and integrative resource for understanding and translating discoveries and advances of genomic, transcriptomic, proteomic and bioinformatic technologies into the study of human diseases. The Series represents leading global opinions on the translation of bioinformatics sciences into both the clinical setting and descriptions to medical informatics. It presents the critical evidence to further understand the molecular mechanisms underlying organ or cell dysfunctions in human diseases, the results of genomic, transcriptomic, proteomic and bioinformatic studies from human tissues dedicated to the discovery and validation of diagnostic and prognostic disease biomarkers, essential information on the identification and validation of novel drug targets and the application of tissue genomics, transcriptomics, proteomics and bioinformatics in drug efficacy and toxicity in clinical research.

The Book Series in Translational Bioinformatics focuses on outstanding articles/chapters presenting significant recent works in genomic, transcriptomic, proteomic and bioinformatic profiles related to human organ or cell dysfunctions and clinical findings. The Series includes bioinformatics-driven molecular and cellular disease mechanisms, the understanding of human diseases and the improvement of patient prognoses. Additionally, it provides practical and useful study insights into and protocols of design and methodology.

Series Description

Translational bioinformatics is defined as the development of storage-related, analytic, and interpretive methods to optimize the transformation of increasingly voluminous biomedical data, and genomic data in particular, into proactive, predictive, preventive, and participatory health. Translational bioinformatics includes research on the development of novel techniques for the integration of biological and clinical data and the evolution of clinical informatics methodology to encompass biological observations. The end product of translational bioinformatics is the newly found knowledge from these integrative efforts that can be disseminated to a variety of stakeholders including biomedical scientists, clinicians, and patients. Issues related to database management, administration, or policy will be coordinated through the clinical research informatics domain. Analytic, storage-related, and interpretive methods should be used to improve predictions, early diagnostics, severity monitoring, therapeutic effects, and the prognosis of human diseases.

Translational Bioinformatics

Series editor: Xiangdong Wang, MD, Ph.D., Professor of Clinical Bioinformatics,
Lund University, Sweden
Professor of Medicine, Fudan University, China

Recently Published and Forthcoming Volumes

Applied Computational Genomics
Editor: Yin Yao Shugart
Volume 1

Pediatric Biomedical Informatics
Editor: John Hutton
Volume 2

Bioinformatics of Human Proteomics
Editor: Xiangdong Wang
Volume 3

For further volumes:
http://www.springer.com/series/11057

Translational Bioinformatics
Volume 4

Bioinformatics for Diagnosis, Prognosis and Treatment of Complex Diseases

Editor: Bairong Shen

Springer

Editor
Bairong Shen
Center for Systems Biology
Soochow University
Suzhou
People's Republic of China

ISSN 2213-2775 ISSN 2213-2783 (electronic)
ISBN 978-94-007-7974-7 ISBN 978-94-007-7975-4 (eBook)
DOI 10.1007/978-94-007-7975-4
Springer Dordrecht Heidelberg New York London

Library of Congress Control Number: 2013953602

© Springer Science+Business Media Dordrecht 2013
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

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

Contents

Part I Bioinformatics for Complex Diseases: The Basics

Part II Network Based Diagnosis of Complex Diseases

Part III Applications in Detection and Treatment of Complex Diseases

Part I
Bioinformatics for Complex Diseases: The Basics

Chapter 1
Molecular Diagnostic Techniques

Feng Guo

Abstract Prior to the rapid development of molecular techniques for clinical diagnostics, cancer was evaluated mainly according to the traditional morphology under a microscope. The mapping of the human genome refreshes the way disease is diagnosed and treated, which is indeed a huge accomplishment in the field of medicine. With the emergence of novel molecular diagnosis technologies, it helps physicians to answer multiple clinical questions that are inadequately answered now. Understanding the behaviors of specific genes will not only facilitate us detect cancer earlier, but it will provide important clues as to how to manage the disease more efficiently. A selection of techniques that are currently available to describe a detailed molecular characterization of various cancers is summarized in this volume.

Keywords Molecular diagnostic · Targeted therapy · PCR · FISH · Mutation analysis

1.1 Introduction

Analysis of nucleic acid (DNA and RNA) in samples forms the foundation of molecular diagnostics or nucleic acid-based diagnostics, which is the most rapidly growing area of laboratory medicine. These methods are extensively used in the diagnosis and monitoring of diverse inherited genetic, infectious, and neoplastic diseases.

Understanding the behaviors of specific genes not only helps us detect cancer earlier, but also provides important clues to better manage certain disease.

F. Guo (✉)
Central Lab, The First Affiliated Hospital of Soochow University, Shizi Road 188,
Suzhou 215006, China
e-mail: guofeng27@suda.edu.cn

B. Shen (ed.), *Bioinformatics for Diagnosis, Prognosis and Treatment of Complex Diseases*, Translational Bioinformatics 4, DOI: 10.1007/978-94-007-7975-4_1,
© Springer Science+Business Media Dordrecht 2013

Molecular diagnosis facilitates choose appropriate therapies (molecularly targeted therapy) that target specific gene products, which are characteristic of a given tumor have been developed. Therapeutic agents, which are currently used in clinical practice and can be dramatically effective in cancer patients, are Gefitinib (IRESSA; AstraZeneca) and Erlotinib (Tarceva; Roche), inhibitors of receptor tyrosine kinase (TKIs) that are specifically effective in advanced non-small cell lung cancer (NSCLC) patients harboring epidermal growth factor receptor genes (*EGFR*) mutations (Lynch et al. 2004; Paez et al. 2004); Trastuzumab (Herceptin; Roche/Genentech), a monoclonal antibody that targets the *HER-2/neu* antigen overexpressed in about 18–20 % of breast cancers (Piccart-Gebhart et al. 2005); and STI571 (Gleevec; Novartis), a TKI that targets the fusion gene product BCR-ABL common to chronic myelogenous leukemia (CML) (Druker et al. 2001; Kelloff and Sigman 2012). In August 2011, Crizotinib (Xalkori; Pfizer) was approved for the treatment of patients with late stage NSCLC patients whose tumors have an *ALK* gene rearrangement (Kwak et al. 2010).

Here, we aim to review of some principles and applications of molecular diagnostic techniques, for instance polymerase chain reaction (PCR), real-time PCR, fluorescent in situ hybridization (FISH), and DNA sequencing.

1.2 Polymerase Chain Reaction and Real-Time PCR

1.2.1 PCR

PCR was initially developed by Dr. Kary Mullis in 1983 (Saiki et al. 1985). As the most frequently used technique in the field of molecular diagnostics, PCR permits specific and exponential synthesis and analysis of targeting DNA regions in samples.

A typical PCR reaction mix includes targeting DNA, forward and reverse primers (short DNA fragments) flanking a location of interest, deoxyribonucleotide triphosphates (dNTPs)/magnesium ions/buffer component, and heat-stable DNA polymerase. In general, PCR is carried out in a volume of 10–200 μl in small reaction tubes (0.2–0.5 ml vol.), in a thermal cycler. Each PCR cycle contains three basic steps including denaturing, annealing, and polymerization. During denaturing, DNA is melted by incubating at 95–98 °C. Primers are bind to the complementary sequences on the single-stranded DNA by decreasing the temperature to the calculated annealing temperature of the primer pair used (usually between 45 and 65 °C) in the reaction (annealing or hybridization). Subsequently, PCR extension occurs by increasing the temperature to the optimal temperature of the DNA polymerase (usually around 70 °C) (extension or polymerization). Incubation times for each step vary between 30 s and 2 min. Consequently, two new helixes consist of one original strand and the newly synthesized complementary strand. The whole process is usually repeated 30–40 times and the amount of the targeted genetic material is doubled after each PCR cycle.

PCR can also be used to amplify an RNA target. RNA is not an appropriate target for the heat-stable DNA polymerases used in PCR assays; therefore, the RNA must first be reverse transcribed (RT) to a double-stranded nucleic acid sequence (cDNA) prior to PCR reaction using special reverse transcriptase enzymes. The cDNA sequence can then be amplified using the same PCR cycles described before. The whole procedure is termed as RT-PCR.

Detection of the *BCR-ABL* fusion transcript by RT-PCR is used clinically to confirm CML or to detect and monitor the presence of minimal residual disease (MRD) in leukemic patients following treatment (Campana and Pui 1995). The sensitivity of a RT-PCR assay enables the detection of one positive cell within a background of 10^5–10^7 normal cells. The detection of the echinoderm microtubule-associated protein-like 4 (*EML4*)-*ALK* fusion transcript by RT-PCR helps select appropriate NSCLC patients for Crizotinib treatment (Kwak et al. 2010).

The visualization of PCR amplification products is facilitated by agarose gel electrophoresis. The agarose gel is stained with ethidium bromide (EB) or other DNA intercalating dyes, which can be detected by fluorescence during exposure to ultraviolet (UV) light. The size of PCR products is determined by comparison with a molecular weight marker, which contains DNA fragments of known size. If primers are labeled with a fluorescent dye, the PCR product can be detected by a capillary electrophoresis system, which tracks the fluorescence of the identical PCR sequences as they migrate (Netto et al. 2003). This detection system results in unsurpassed sensitivity, single base resolution, and differential product detection.

The detection of clonality in a suspected lymphoproliferation using frozen and paraffin-embedded tissues is valuable in the diagnosis of malignant lymphoma. The stepwise rearrangement processes during early lymphocytes maturation in the immunoglobulin heavy chain gene (*IGH*) or the T cell receptor (*TCR*) gene join *V*-, *D*-, and *J*-gene segments. In the procession of gene rearrangement, nucleotides are deleted and randomly inserted at the joining sites, resulting in an enormous diversity of antigen receptors. PCR-based assays amplify the DNA between primers that that target the conserved framework (FR) and the *J* gene regions. Reactive lymphoproliferations therefore, have polyclonally rearranged *IGH* or *TCR* genes, whereas malignant lymphomas have clonal rearrangements (van Krieken et al. 2007).

1.2.2 Real-Time PCR

Real-time PCR, also called quantitative real-time PCR (qPCR), refers to PCR amplification that is detected and measured continuously during each cycle of PCR process. The basic goal of this technique is to precisely distinguish and simultaneously quantify nucleic acid sequences in a sample even in a very small quantity. This is a new approach compared to the conventional PCR, where the products are detected at the end (plateau). qPCR results can either be qualitative (the presence or absence of a sequence) or quantitative (copy number). The main advantage of qPCR is to determine the amount of starting DNA in the sample before the PCR amplification with accuracy and high sensitivity over a wide dynamic range.

In the case of qRT-PCR, RNA is reverse transcribed into cDNA, which is followed by a subsequent qPCR amplification using cDNA as template. To fulfill monitoring the progress of DNA amplification in real time, specific chemistries and instrumentation are required.

All available qPCR instruments measure the progress of amplification by monitoring fluorescence changes within the PCR tube. A threshold for detection fluorescence is set slightly above background. A signal that is detected above the threshold is considered as a real signal, which can be used to define the threshold cycle (Ct) or quantification cycle (Cq). The Ct is a basic principle of qPCR and is an essential component in producing accurate and reproducible data. The amplicon doubles every cycle and the amount of fluorescence increases exponentially beyond the threshold. Therefore, the amount of fluorescence is directly proportional to the number of amplicons produced in the samples. A universal method of DNA quantification by qPCR is to plot fluorescence against the number of cycles on a logarithmic scale (Fig. 1.1). During amplification, how quickly the fluorescent signal reaches a threshold level correlates with the amount of original target sequence, thereby enabling quantification (Valasek and Repa 2005). If a large amount of target DNA template is present at the start of the reaction, few cycles (low Ct) are required to accumulate enough products to give a fluorescence signal above background. In contrast, if a small amount of template is present more amplification cycles (high Ct) are needed for the fluorescence signal to rise above background.

In general, two approaches are used to obtain a fluorescent signal from the synthesis of product in qPCR. One depends on the property of non-specific fluorescent dyes such as SYBR green I, which intercalates with any double-stranded DNA and undergoes a conformational change resulting in an increase in their

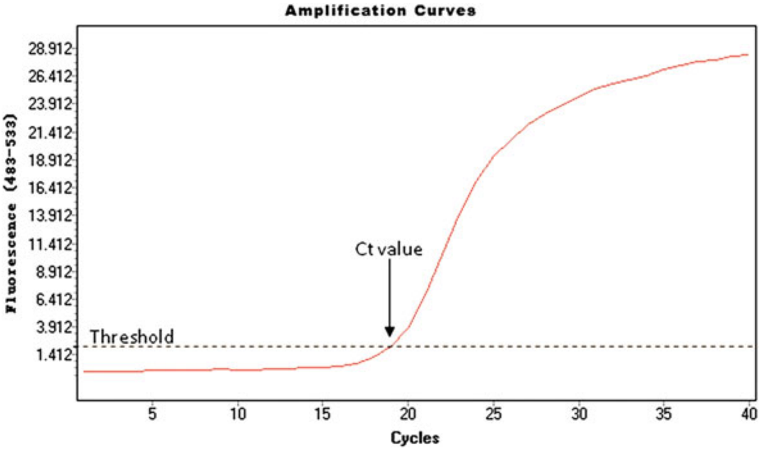


Fig. 1.1 Amplication curve. The number of PCR cycles is shown on the x-axis, and the fluorescence from the amplification reaction is shown on the y-axis

fluorescence (Wittwer et al. 1997). SYBR Green I has an excitation and emission maxima of 494 and 521 nm, respectively. During the extension phase, more and more SYBR Green I can bind to the PCR product, resulting in an increased fluorescence. Consequently, during each subsequent PCR cycle more fluorescence signal will be detected (van der Velden et al. 2003).

Another approach is to use fluorescent resonance energy transfer (FRET), which was first described over 50 years ago and is being used more and more in biomedical research and drug discovery today. FRET is a quantum phenomenon occurring between two dye molecules. FRET relies on the distance-dependent transfer of energy from a donor molecule to an acceptor molecule. Excitation is transferred from a donor molecule to an acceptor fluorophore through dipole–dipole interaction without the emission of a photon (Didenko 2001). As a result, the donor molecule fluorescence is quenched, and the acceptor molecule is excited. Therefore, the intensity of the donor fluorescence decreases while the fluorescence intensity of the acceptor increases. The sequence-specific DNA probes are labeled with a fluorescent reporter which permits detection only after hybridization of the probe with its complementary sequences. DNA-based FRET probes are applied in monitoring various types of DNA or RNA reactions including PCR, hybridization, ligation, recombination, and synthesis. The FRET probes used for PCR amplification are either cleaved in the reaction (as TaqMan® probes), incorporated into amplified DNA (as Scorpion primers), or undergo a conformation change in the presence of a complementary DNA target (as molecular beacons).

The qPCR method has been used to monitor MRD using leukemia-specific marker such as the *BCR-ABL* fusion gene and the *WT1* gene (Dolken 2001; Inoue et al. 1996). The mRNA expression of the excision cross complementation group 1 (*ERCC1*), ribonucleotide reductase subunit M1 (*RRM1*), thymidylate synthetase (*TYMS*), and class III *β*-tubulin (*TUBB3*) genes in tumor tissue, detected by qRT-PCR, has been linked to the chemo-sensitivity in multiple cancers. Median survival time in NSCLC patients with low *ERCC1* expression is significantly longer as well as in patients with low *RRM1* expression. Among cisplatin-treated patients, low *ERCC1* levels are highly predictive of a longer survival (Ceppi et al. 2006; Simon et al. 2007). High *ERCC1* mRNA levels are predictive of poor response to platinum-based chemotherapy in ovarian cancer, lung cancer, and chronic lymphocytic leukemia (Rosell et al. 2003). Low *TYMS* mRNA levels are significantly associated with a longer survival and/or a strong response to 5-fluorouracil (5-FU)-based chemotherapy in different cancers (Park and Lenz 2006). The *TUBB3* overexpression confers resistance to paclitaxel and vinorelbine, whereas downregulation of *TUBB3* renders cells more sensitive to the two drugs (Stengel et al. 2010).

1.2.3 FISH

In situ hybridization (ISH) is a powerful technique for detecting specific targets on the genome in tissues and cells, and gaining temporal and spatial information about genetic loci and gene expression. Two basic ways to visualize the RNA or DNA targets are fluorescence (FISH) and chromogenic (CISH) detection, which both use a labeled, target-specific oligonucleotide probes that are hybridized with the complementary DNA target in the sample.

The main advantage of FISH is that it permits the use of both dividing and non-dividing cells as targets, and enables a large number of cells to be evaluated. The high sensitivity and specificity, and the speed with which FISH can be performed make it a useful cytogenetic technique in the diagnosis of blood disorders and cancers.

FISH involves the specific hybridization of a fluorescence-labeled (e.g., Texas red, FITC spectra, Rhodamine) nucleic acid probe to complementary gene sequences, and subsequent visualization the colored signals at the hybridization site in bone marrow or peripheral blood smears, or fixed and sectioned tissue by fluorescence microscopy. Both the labeled nucleic acid probe and the DNA target are denatured to a single-stranded state and permitted to hybridize to each other.

Three types of probes are widely used, such as painting, centromeric, and allele-specific probes. Painting probes are many separate region-specific probes that bind along a single chromosome. Chromosome painting allows the highly sensitive and specific visualization of individual chromosomes in metaphase or interphase cells, and the identification of both numerical and structural chromosomal aberrations. Centromeric probes identify the centromeric region of a specific chromosome, which are generally used as control for chromosome enumeration. Allele-specific oligonucleotides (ASO) are synthetic DNA oligonucleotides complementary to the targeting sequence.

FISH plays an important role in the molecular analysis of many hematopoietic disorders and cancer, and detects numerical and structural chromosomal abnormalities. Currently, the evaluation of the *HER2/neu* gene amplification by FISH in breast and gastric cancer is widely used in clinical practice. Breast cancer can be classified as being HER2 positive or HER2 negative. In normal cells, there are two copies of the *HER2* gene, one on each of two copies of chromosome 17. About 18–20 % breast cancer have *HER2* gene amplification, which are considered more aggressive (Slamon et al. 1987). Trastuzumab (Herceptin) is effective in the treatment of HER2-positive early stage and metastatic breast cancer (Piccart-Gebhart et al. 2005).

HER2/neu FISH testing measures the number of copies of the *HER2* gene present in each tumor cells and is reported as either positive or negative (Fig. 1.2). In a typical *HER2/neu* FISH testing, a centromeric chromosome 17 probe (green signal) and an allele-specific probe for the *HER2/neu* oncogene (red signal) are included. The ratio of the *HER2/neu* gene to chromosomes 17 in 60 tumor cells is determined. The *HER2* gene/chromosome 17 ratio in a normal, non-dividing cell should be 1. The *HER2* gene/chromosome 17 ratio can increase up to 2 in cells during certain stages of normal cell division. A FISH ratio (*HER2* gene signals to chromosome 17 signals) more than 2.2 is reported as HER2 positive (Wolff et al. 2007). Concordance between immunohistochemistry (IHC) and FISH results has been extensively studied. In one study conducted in 2,963 samples using FISH as the

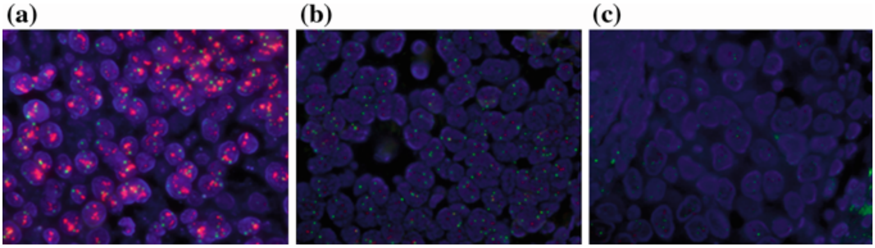


Fig. 1.2 Fluorescent in situ hybridization analysis for the *HER2/neu* gene in breast cancer. **a** High *HER2* amplification (*large clusters*). **b** HER2 negative. **c** Polysomy

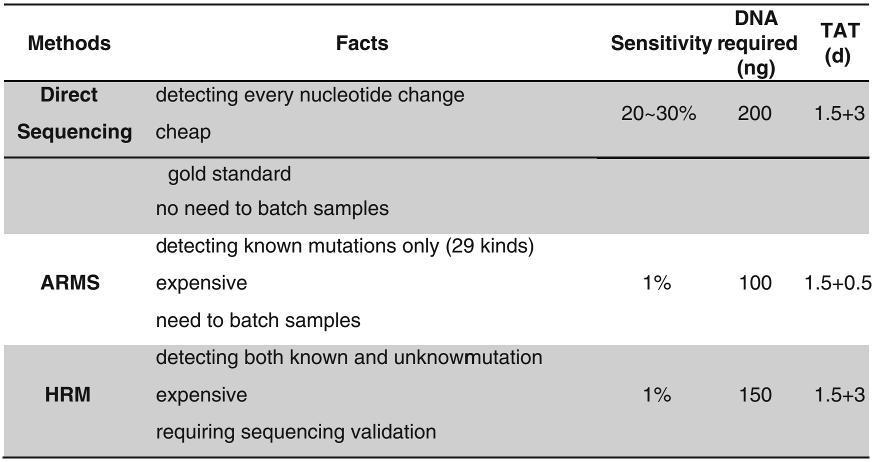
standard method, the positive predictive value of an IHC 3+ result was 91.6 %, and the negative predictive value of an IHC 0 or 1+ result was 97.2 % (Yaziji et al. 2004). About 25 % breast cancer patients with an IHC 2+ result have a positive FISH result.

1.3 Mutation Analysis

Molecular markers of cancer can be products of altered genes. DNA mutations include gene rearrangements such as translocations, inversions, point mutations, and insertions/deletions. Detection of mutations in cancer is important for understanding the disease process. Mutation analysis is also a precursor to targeted therapy, which is the standard of care for certain tumor types. For example, Gefitinib and Erlotinib for advanced NSCLC patients with *EGFR* mutation, and Cetuximab and Panitumumab for metastatic colon cancer patients with the wild-type *KRAS* gene.

Mutations in exons 18–21 of the *EGFR* gene, which encode tyrosine kinase domain, enhance the activity of the intracellular signaling pathway and confer the oncogenic properties of EGFR (Sharma et al. 2007). In-frame deletions in exon 19 and a specific point mutation in exon 21 (p.L858R) are the most prevalent *EGFR* mutations. Mutations associated with TKI-resistance include a point mutation (p.T790 M) and insertions (e.g. p.D770_N771insNPG) in exon 20, and a point mutation (p.D761Y) in exon 19.

A variety of techniques are existing for mutation analysis of the *EGFR* gene and classified into screening methods that indentify all mutations, and targeted methods that distinctively detect known and pre-determined mutations (Ellison et al. 2013). Among diverse screening methods, direct DNA sequencing using Sanger method has been successfully used to detect mutations for many years and considered as the ‘gold standard’. With direct sequencing, there is no requirement for batching of samples and it provides better contamination control since the exact, specific mutation will be presented. Direct sequencing, detects all existing



Methods	Facts	Sensitivity	DNA required (ng)	TAT (d)
Direct Sequencing	detecting every nucleotide change cheap gold standard no need to batch samples	20~30%	200	1.5+3
ARMS	detecting known mutations only (29 kinds) expensive need to batch samples	1%	100	1.5+0.5
HRM	detecting both known and unknowmutation expensive requiring sequencing validation	1%	150	1.5+3

Fig. 1.3 Comparison among direct sequencing, ARMS, and HRM assay

mutations, but it is time-consuming and successful when viable tumor cells constitute at least 25 % of the tissues (Fig. 1.3).

Capillary electrophoresis-based Sanger sequencing (the chain-termination method) was developed by Dr. Sanger and colleagues in 1975 (Sanger and Coulson 1975; Sanger et al. 1977). Sanger sequencing, which is also referred to as dideoxy sequencing or chain termination, is based on the use of dideoxynucleotides (ddNTP) in addition to the normal nucleotides (NTP). ddNTP are essentially the same as nucleotides except they contain a hydrogen group on the 3′ carbon instead of a hydroxyl group (OH), which prevent the addition of further nucleotides when integrated into a sequence. This occurs because a phosphodiester bond cannot form between the dideoxynucleotide and the next incoming nucleotide, and thus the DNA chain is terminated.

Alternative screening methods to direct sequencing include high resolution melting (HRM), pyrosequencing, and denaturing high pressure liquid chromatography (dHPLC) analysis. As an alternative to direct sequencing, HRM is an in-tube, fast, and sensitive screening method that detects sequence variation by monitoring the melting curve of PCR amplicons. HRM is able to detect mutant genes at levels of 1–10 % (Wittwer 2009). Nevertheless, any DNA alteration due to single nucleotide polymorphism (SNP) interference or formalin fixation may produce an abnormal melting point curve, which must be confirmed by sequencing. In addition, the amplification product is usually designed to be short in length and does not cover the whole exon. The requirement for sequencing validation increases turn-around-time (TAT) and reduces the value of high sensitivity, suggesting HRM assay as a screening method prior to DNA sequencing.

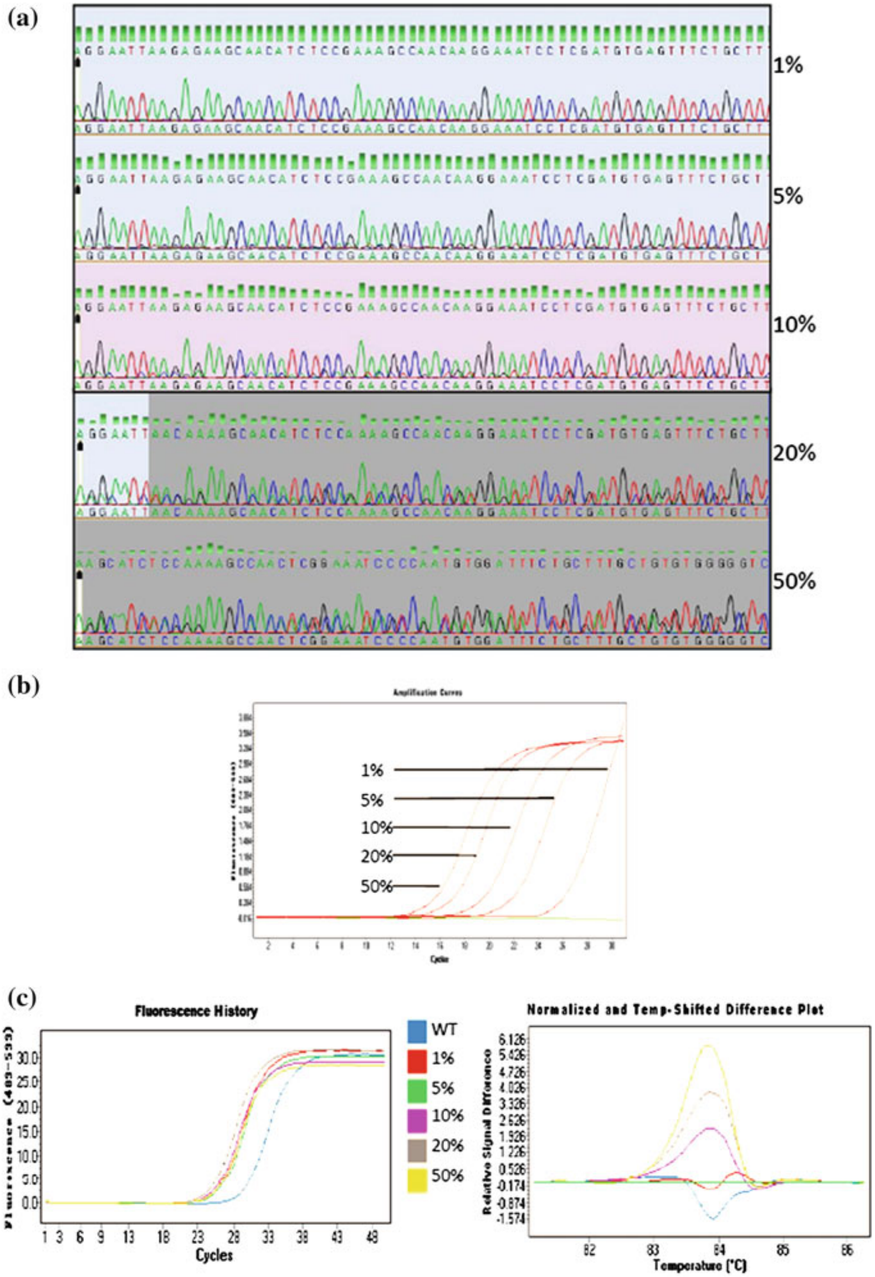
Scorpion amplification refractory mutation system (ARMS) falls into targeting method category and has been used successfully to analyze *EGFR* mutation status in the phase III Iressa Pan-Asia Study (IPASS) clinical trial (Newton et al. 1989).

ARMS is more sensitive than direct sequencing, however, detects known mutations only. Additionally, ARMS requires to batch samples and the reagents are expensive. ARMS, discriminating between mutant and wild-type DNA by selectively amplifying mutation-specific target sequences, detects mutations in samples with mutation frequency as low as 0.1–1 % (Zhuang et al. 2013). Other targeted method, such as peptide nucleic acid (PNA)-locked PCR clamping, mutant-enriched PCR, and SNAPshot PCR, are also used in clinical practice (Ikeda et al. 2012). Using cell lines heterozygous for *EGFR* mutations, we found that the sensitivities of direct sequencing, ARMS, and HRM in our experiment setting were 10, 1, and 1 % (Fig. 1.4, unpublished data).

In fact, none of the available methods can provide the entire information of *EGFR* mutation with the best sensitivity and specificity. The concordance rate was 73.68 % between direct sequencing and ARMS assay. The concordance rate between HRM and sequencing was 78.67 % (unpublished data). To decrease the frequency of false negatives and not to lose any opportunity for a potential EGFR-TKIs treatment, a workflow for the *EGFR* mutation examination according to the specimen quality and quantity (tumor load and DNA yield) is now be proposed. If specimens containing tumor tissue are sufficient (e.g. surgical biopsy), direct DNA sequencing is ready to perform at first. Macro-dissection to enrich DNA from tumor tissue is suggested when tumor load is lower than 10 %. The hematoxylin and eosin staining for FF-PETs slide is performed to guide tissue macro-dissection. The subsequent ARMS assay is necessary to perform in order to rule out false negatives due to limited sensitivity of sequencing. In fact, several reports showed that around 20–30 % samples that were negative for sequencing were detected somatic *EGFR* mutations by ARMS assay (Ellison et al. 2010; Liu et al. 2011). If specimens are insufficient, such as biopsy of transbronchial needle aspiration (TBNA) or surrogate tissue including bronchial alveolar lavage, plasma or pleural fluid, ARMS assay is recommended to be the first and the best choice. Sequencing is suggested to further exclude any possibilities for an uncommon mutation, if *EGFR* mutation is negative from ARMS assay and if there are enough DNA left (Fig. 1.5). In the future, we might be benefit from the incorporation of next-generation sequencing into daily practice.

The Sanger sequencing method is considered as a first-generation technology, and newer methods are referred to as next-generation sequencing (NGS). NGS, or second-generation sequencing, is an innovative and extremely sensitive platform, which performs massively parallel sequencing and offers new diagnostic opportunities. In the past decade, several NGS platforms have been developed and provided low-cost, high-throughput sequencing. The major NGS platforms that enter the market include Ion Torrent Personal Genome Machine (PGM) and SOLiD from Life Technologies, HiSeq sequencing system from Illumina/Solexa, and 454 GS Junior System from Roche Applied Sciences.

NGS encompasses several different methodologies that allow the investigation of genomics, transcriptomics, and epigenomics (Braggio et al. 2013). Although each NGS platform is different in how sequencing is accomplished, the whole procedure of NGS is generally includes template preparation, sequencing and


(a)
AGGAATTAAGAGAAGCAACATCTCCGAAAGCCAACAAGGAAATCCTCGATGTGAGTTTCTGCTT
1%
5%
10%
20%
AAGCATCTCCAAAAGCCAACTCGGAAATCCCCAATGTGGATTTCTGCTTTGCTGTGTGGGGTC
50%
(b)
Amplification Curves
1%
5%
10%
20%
50%
Cycle
(c)
Fluorescence History
Fluorescence (483-533)
30.0
25.0
20.0
15.0
10.0
5.0
0.0
1 3 6 9 13 18 23 28 33 38 43 48
Cycles
WT
1%
5%
10%
20%
50%
Normalized and Temp-Shifted Difference Plot
Relative Signal Difference
6.126
5.426
4.726
4.026
3.326
2.626
1.926
1.226
0.526
-0.174
-0.874
-1.574
82
83
84
85
86
Temperature (°C)

◀ **Fig. 1.4** The sensitivity testing for *EGFR* mutations using serial dilutions of PC-9/A549 DNA. PC-9 cells harbor in-frame deletions in exon 19 of the *EGFR* gene (heterozygous for c.2235_2249del15). A549 cells are wild-type for the *EGFR* gene. The gDNA of PC-9 cells are serially diluted into A549 gDNA at ratios of 100, 40, 20, 10, and 2 % to give mutant allele frequencies of 50, 20, 10, 5, and 1 %. **a** Direct sequencing. At least 10 % mutant DNA is necessary to detect *EGFR* mutations. **b** ARMS. 1 % mutant DNA is ready to be indentified from wild-type DNA. **c** HRM. 1 % mutant DNA is ready to be plotted differently from wild-type DNA

imaging, genome alignment and assembly methods, and data analysis (Grada and Weinbrecht 2013; Metzker 2010). The template (DNA or cDNA) is first fragmented into a library of small segments that can be uniformly and accurately sequenced in millions of parallel reactions. The newly identified strings of bases, called reads, are then reassembled by aligning to a reference genome. The full set of aligned reads reveals the entire sequence of each chromosome in the sample.

Currently, NGS is widely used in many fields related to biological sciences and is particularly successful in the application of whole-exome sequencing and targeted sequencing. The whole-exome sequencing, sequencing of the protein-encoding parts of all the genes, is proposed as a method for detecting disease-causing sequence variations in complex human disease. High-throughput sequencing of the human genome facilitates the discovery of genes and regulatory elements associated with disease. Several successful cases have been reported recently. Homozygosity mapping, followed by the whole exome sequencing, has

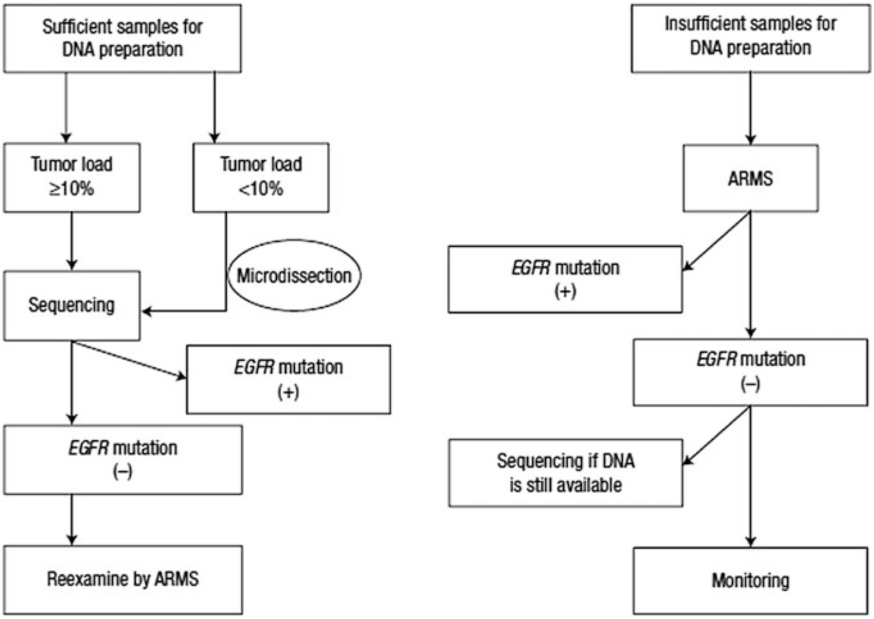


Fig. 1.5 Proposed algorithm of a sequential method for EGFR mutation detection (Zhuang et al. 2013)

identified that the *SLC45A2* and *G6PC3* genes are associated with neutropenia (Cullinane et al. 2011). Other genes such as *FLNA*, *RPL21*, *STAT1*, *WDR35*, and *c16orf57*, detected by NGS, have been linked to diverse inherited skin disorders (Lai-Cheong and McGrath 2011). NGS also improves our knowledge of the genetic basis of multiple hematological malignancies and solid tumors (Braggio et al. 2013). As matter of fact, with NGS, clinicians are provided a fast, affordable, a thorough way to determine the genetic cause of a disease.

Targeted sequencing allows the identification of disease-causing mutations for diagnosis of pathological evaluations. With targeted sequencing, only a subset of genes or defined regions in a genome is sequenced, allowing researchers to focus time, expenses, and data storage on the genomic regions of interest. The ability to batch samples and obtain high sequence coverage during a single reaction allows NGS to identify rare, novel mutations that are missed, or too expensive to identify, using first-generation sequencing methods (Grada and Weinbrecht 2013; Metzker 2010).

More and more innovative technologies in molecular biology are gradually applied into the clinical laboratory as validated diagnostic tests. Molecular diagnostics eventually helps establish a definitive diagnosis and classification of cancers based on the recognition of unique molecular alterations that occur in specific cancer types.

References

Braggio E, Egan JB, Fonseca R, Stewart AK. Lessons from next-generation sequencing analysis in hematological malignancies. Blood Cancer J. 2013;3:e127.

Campana D, Pui CH. Detection of minimal residual disease in acute leukemia: methodologic advances and clinical significance. Blood. 1995;85:1416–34.

Ceppi P, Volante M, Novello S, Rapa I, Danenberg KD, Danenberg PV, Cambieri A, Selvaggi G, Saviozzi S, Calogero R, et al. ERCC1 and RRM1 gene expressions but not EGFR are predictive of shorter survival in advanced non-small-cell lung cancer treated with cisplatin and gemcitabine. Ann Oncol. 2006;17:1818–25.

Cullinane AR, Vilboux T, O'Brien K, Curry JA, Maynard DM, Carlson-Donohoe H, Ciccone C, Markello TC, Gunay-Aygun M, Huizing M, et al. Homozygosity mapping and whole-exome sequencing to detect SLC45A2 and G6PC3 mutations in a single patient with oculocutaneous albinism and neutropenia. J Invest Dermatol. 2011;131:2017–25.

Didenko VV. DNA probes using fluorescence resonance energy transfer (FRET): designs and applications. Biotechniques. 2001:31:1106–16, 1118, 1120–1.

Dolken G. Detection of minimal residual disease. Adv Cancer Res. 2001;82:133–85.

Druker BJ, Talpaz M, Resta DJ, Peng B, Buchdunger E, Ford JM, Lydon NB, Kantarjian H, Capdeville R, Ohno-Jones S, et al. Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. N Engl J Med. 2001;344:1031–7.

Ellison G, Donald E, McWalter G, Knight L, Fletcher L, Sherwood J, Cantarini M, Orr M, Speake G. A comparison of ARMS and DNA sequencing for mutation analysis in clinical biopsy samples. J Exp Clin Cancer Res. 2010;29:132.

Ellison G, Zhu G, Moulis A, Dearden S, Speake G, McCormack R. EGFR mutation testing in lung cancer: a review of available methods and their use for analysis of tumour tissue and cytology samples. J Clin Pathol. 2013;66:79–89.

Grada A, Weinbrecht K. Next-generation sequencing: methodology and application. J Invest Dermatol. 2013;133:e11.

Ikeda T, Nakamura Y, Yamaguchi H, Tomonaga N, Doi S, Nakatomi K, Iida T, Motoshima K, Mizoguchi K, Nagayasu T, et al. Direct comparison of 3 PCR methods in detecting EGFR mutations in patients with advanced non-small-cell lung cancer. Clin Lung Cancer. 2012

Inoue K, Ogawa H, Yamagami T, Soma T, Tani Y, Tatekawa T, Oji Y, Tamaki H, Kyo T, Dohy H, et al. Long-term follow-up of minimal residual disease in leukemia patients by monitoring WT1 (Wilms tumor gene) expression levels. Blood. 1996;88:2267–78.

Kelloff GJ, Sigman CC. Cancer biomarkers: selecting the right drug for the right patient. Nat Rev Drug Discov. 2012;11:201–14.

Kwak EL, Bang YJ, Camidge DR, Shaw AT, Solomon B, Maki RG, Ou SH, Dezube BJ, Janne PA, Costa DB, et al. Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer. N Engl J Med. 2010;363:1693–703.

Lai-Cheong JE, McGrath JA. Next-generation diagnostics for inherited skin disorders. J Invest Dermatol. 2011;131:1971–3.

Liu Y, Liu B, Li XY, Li JJ, Qin HF, Tang CH, Guo WF, Hu HX, Li S, Chen CJ, et al. A comparison of ARMS and direct sequencing for EGFR mutation analysis and tyrosine kinase inhibitors treatment prediction in body fluid samples of non-small-cell lung cancer patients. J Exp Clin Cancer Res. 2011;30:111.

Lynch TJ, Bell DW, Sordella R, Gurubhagavatula S, Okimoto RA, Brannigan BW, Harris PL, Haserlat SM, Supko JG, Haluska FG, et al. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. N Engl J Med. 2004;350:2129–39.

Metzker ML. Sequencing technologies—the next generation. Nat Rev Genet. 2010;11:31–46.

Netto GJ, Saad RD, Dysert PA 2nd. Diagnostic molecular pathology: current techniques and clinical applications, part I. Proc (Bayl Univ Med Cent). 2003;16:379–83.

Newton CR, Graham A, Heptinstall LE, Powell SJ, Summers C, Kalsheker N, Smith JC, Markham AF. Analysis of any point mutation in DNA. The amplification refractory mutation system (ARMS). Nucleic Acids Res. 1989;17:2503–16.

Paez JG, Janne PA, Lee JC, Tracy S, Greulich H, Gabriel S, Herman P, Kaye FJ, Lindeman N, Boggon TJ, et al. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. Science. 2004;304:1497–500.

Park DJ, Lenz HJ. Determinants of chemosensitivity in gastric cancer. Curr Opin Pharmacol. 2006;6:337–44.

Piccart-Gebhart MJ, Procter M, Leyland-Jones B, Goldhirsch A, Untch M, Smith I, Gianni L, Baselga J, Bell R, Jackisch C, et al. Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer. N Engl J Med. 2005;353:1659–72.

Rosell R, Taron M, Alberola V, Massuti B, Felip E. Genetic testing for chemotherapy in non-small cell lung cancer. Lung Cancer. 2003;41(Suppl 1):S97–102.

Saiki RK, Scharf S, Faloona F, Mullis KB, Horn GT, Erlich HA, Arnheim N. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science. 1985;230:1350–4.

Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol. 1975;94:441–8.

Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA. 1977;74:5463–7.

Sharma SV, Bell DW, Settleman J, Haber DA. Epidermal growth factor receptor mutations in lung cancer. Nat Rev Cancer. 2007;7:169–81.

Simon G, Sharma A, Li X, Hazelton T, Walsh F, Williams C, Chiappori A, Haura E, Tanvetyanon T, Antonia S, et al. Feasibility and efficacy of molecular analysis-directed individualized therapy in advanced non-small-cell lung cancer. J Clin Oncol. 2007;25:2741–6.

Slamon DJ, Clark GM, Wong SG, Levin WJ, Ullrich A, McGuire WL. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. Science. 1987;235:177–82.

Stengel C, Newman SP, Leese MP, Potter BV, Reed MJ, Purohit A. Class III beta-tubulin expression and in vitro resistance to microtubule targeting agents. Br J Cancer. 2010;102:316–24.

Valasek MA, Repa JJ. The power of real-time PCR. Adv Physiol Educ. 2005;29:151–9.

van der Velden VH, Hochhaus A, Cazzaniga G, Szczepanski T, Gabert J, van Dongen JJ. Detection of minimal residual disease in hematologic malignancies by real-time quantitative PCR: principles, approaches, and laboratory aspects. Leukemia. 2003;17:1013–34.

van Krieken JH, Langerak AW, Macintyre EA, Kneba M, Hodges E, Sanz RG, Morgan GJ, Parreira A, Molina TJ, Cabecadas J, et al. Improved reliability of lymphoma diagnostics via PCR-based clonality testing: report of the BIOMED-2 Concerted Action BHM4-CT98-3936. Leukemia. 2007;21:201–6.

Wittwer CT. High-resolution DNA melting analysis: advancements and limitations. Hum Mutat. 2009:30:857–9.

Wittwer CT, Herrmann MG, Moss AA, Rasmussen RP. Continuous fluorescence monitoring of rapid cycle DNA amplification. Biotechniques. 1997:22:130–1, 134–8.

Wolff AC, Hammond ME, Schwartz JN, Hagerty KL, Allred DC, Cote RJ, Dowsett M, Fitzgibbons PL, Hanna WM, Langer A, et al. American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer. Arch Pathol Lab Med. 2007;131:18–43.

Yaziji H, Goldstein LC, Barry TS, Werling R, Hwang H, Ellis GK, Gralow JR, Livingston RB, Gown AM. HER-2 testing in breast cancer using parallel tissue-based methods. JAMA. 2004;291:1972–7.

Zhuang Y, Xu J, Ma H, Zhu W, Guo L, Kang S, Guo F. A sequential method of epidermal growth factor receptor mutation detection reduces false negatives: a new case with doublet mutations of L833V and H835L in China. Clin Lung Cancer. 2013;14:295–300.

Chapter 2
Identifying Biomarkers with Differential Analysis

Xing-Ming Zhao and Guimin Qin

Abstract The initiation and development of diseases is a complex process, involving genetic mutations and environmental influences. Disease biomarkers (biological markers) are biological characteristics of pathogenic processes, which can help make diagnostic or prognostic decisions so that necessary interventions can be adopted to prevent the development of diseases. In the post-genomic era, with the accumulation of various kinds of omics data, it is possible to identify molecular biomarkers that can help diagnosis and develop efficient therapies. In this chapter, we summarize the recent progress on identifying biomarkers with differential analysis based on different types of omics data. Differential analysis is a very powerful and widely used approach in biology, which identifies biomarkers by comparing molecular datasets generated under different conditions. In particular, we focus on the approaches that identify biomarkers based on molecular networks that take into account the differences between different physiological conditions together with the network topology structure.

Keywords Gene biomarker · Gene set biomarker · Omics data · Pathway biomarker · Network biomarker

X.-M. Zhao (✉) · G. Qin
School of Electronics and Information Engineering, Tongji University,
4800 Caoan Highway, Shanghai 201804, China
e-mail: xm_zhao@tongji.edu.cn

G. Qin
e-mail: gmqin@mail.xidian.edu.cn

B. Shen (ed.), *Bioinformatics for Diagnosis, Prognosis and Treatment of Complex Diseases*, Translational Bioinformatics 4, DOI: 10.1007/978-94-007-7975-4_2,
© Springer Science+Business Media Dordrecht 2013

2.1 Introduction

Diseases are generally caused by genetic mutations or/and environmental influences, involving various biological processes. Early diagnosis of disease risks can help prevent the development of diseases, and precise prognosis of disease states can avoid unnecessary treatments for good outcomes while adopt timely intervention for poor outcomes. Disease biomarkers (biological markers) are biological characteristics of pathogenic processes, which can help make diagnostic or prognostic decisions. Biomarkers are useful for predicting disease risks of certain populations so that timely intervention can be adopted to prevent the disease. Furthermore, biomarkers can help identify subtypes of heterogeneous diseases, e.g. breast cancer, so that appropriate therapeutic strategies can be adopted. In the past decades, with the development of molecular biology and biotechnology, a huge amount of molecular data are publically available, which enables the identification of specific molecules that can serve as biomarkers. For example, the hormone receptors ER and PR can be used as the biomarkers to predict the response of patients to endocrine therapy, while the HER2 oncogene can serve as a biomarker of invasive breast cancer and predicts survival of patients (Ross 2009).

Despite the success of molecular biomarkers, it is not an easy task to identify reliable and useful biomarkers considering more than 20,000 genes encoding about 30,000 proteins within the human genome, where complex interactions can be found among proteins. Recently, with the rapid progress in biotechnologies, especially in high-throughput techniques, genome-wide screening is making it possible to identify molecular biomarkers in an efficient way. In particular, the accumulation of various kinds of '-omics' (e.g. genomics, transcriptomics and proteomics) data enables one to identify potential gene biomarkers that can predict disease risks (Joyce and Palsson 2006). For example, the genome-wide association study (GWAS) is able to provide genetic variants associated with diseases based on the comparison of disease population against normal/control population. In the landmark Wellcome Trust Case Control Consortium (WTCCC) (2007) study, many DNA variants and genes were identified to be associated with seven common diseases. The transcriptome profiles enable the monitoring of expression of tens of thousands of genes, where those genes that are differentially expressed between different physiological conditions are generally regarded as potential biomarkers for diagnosis and prognosis. In their pivotal work, Golub et al. (1999) identified gene biomarkers that can successfully discriminate acute myeloid leukemia (AML) from acute lymphoblastic leukemia (ALL) based on gene expression profiles.

Although the gene biomarkers identified based on the omics data achieve some success, most of the gene biomarkers are not reliable and have low reproducibility, where the biomarkers identified from one dataset sometimes fail to work in another dataset for the same disease. This phenomenon arises since many diseases, especially complex diseases, are well recognized as the results of dysregulation of biological systems instead of the mutations of individual genes, whereas the gene biomarkers are generally assumed to be functionally independent of each other. Therefore, it is necessary to identify biomarkers from a systematic perspective. The molecular networks, including protein–protein interaction network, gene regulation network and metabolic network, can describe the biological systems in an accurate way (Barabasi and Oltvai 2004), thereby providing an alternative way to predict biomarkers at systematic levels. Biomarkers identified from the molecular networks can provide insights into the molecular underpinnings of diseases, and help develop efficient therapeutic strategies (Barabasi et al. 2011). For example, with the network biomarkers identified for cancer, Chen et al. (2011) successfully predicted the breast cancer metastasis.

In this chapter, we survey the recent progress on biomarker identification with differential analysis based on different types of omics data, where biomarkers are identified by comparing molecular datasets generated under different conditions. Here, biomarkers range from genes to gene sets, pathways, and networks. In particular, we focus on the approaches that identify biomarkers from molecular networks that take into account the differences between different physiological conditions together with the network topology structure.

2.2 Differential Analysis in Biology

Differential analysis is a widely used approach to identify biomarkers in biology, where the differences of biological characteristics, e.g. genes or blood pressure, across different species or conditions are generally investigated and those significantly changed biological markers will be treated as biomarkers. In this chapter, the biomarkers are referred to as molecular biomarkers, ranging from genes to gene sets/pathways and networks.

As shown in Fig. 2.1, molecular biomarkers can be identified based on different kinds of data, where the resultant biomarkers range from individual genes to gene sets and networks. Right now, a huge amount of omics data on distinct major diseases are publically available. For example, the gene expression data for patients can be retrieved from Gene Expression Omnibus (Barrett et al. 2009) and ArrayExpress (Parkinson et al. 2009), protein–protein interaction data can be freely available at BioGrid (Stark et al. 2006) and STRING (von Mering et al. 2005) databases, and pathway knowledge can be found at KEGG (Kanehisa and Goto 2000) and Gene Ontology (Ashburner et al. 2000). Inspired by the wealth of the publically available data, a lot of computational approaches have been proposed to identify biomarkers by conducting differential analysis. In this chapter, we focus on the differential analysis of transcriptome data and protein–protein interactions. Those readers that are interested in identifying biomarkers from genomic and metabolic data are referred to the review papers on identifying biomarkers based on GWAS (Manolio 2013) and metabolic profiling (Spratlin et al. 2009). For different types of data, the biomarkers identified are different. For example, gene biomarkers can be obtained with differential expression analysis,

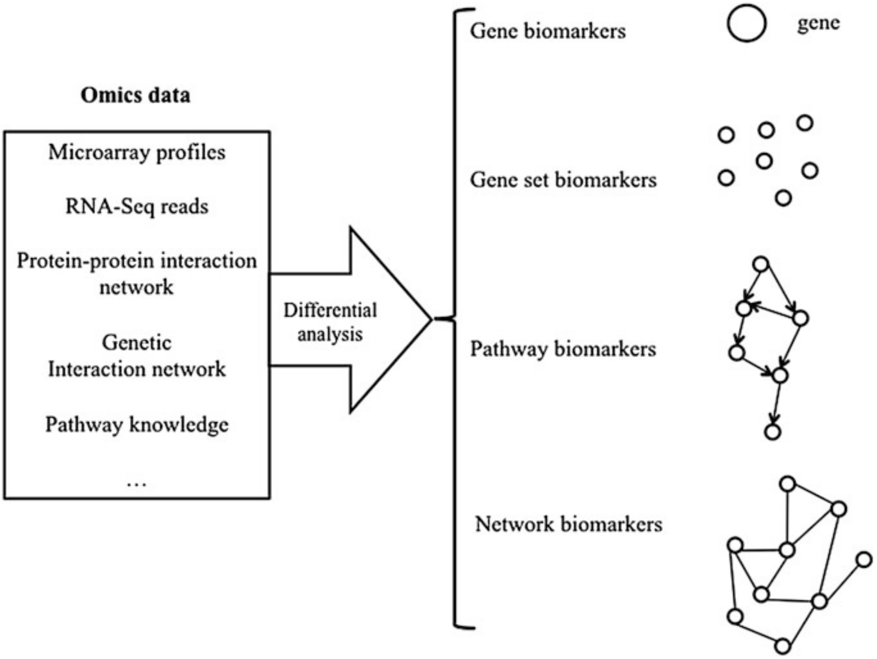


Fig. 2.1 Biomarkers identified based on different data

gene set biomarkers are identified by considering a set of genes as an entity, while pathway and network biomarkers are generally detected by taking into account the functional interactions among genes.

In the following sections, different computational approaches for differential analysis on distinct types of data will be introduced. Especially, these computational approaches are introduced based on the type of biomarkers they identify.

2.3 Gene Biomarkers

With the accumulation of huge amount of gene expression data deposited in public databases, e.g. GEO, it is becoming easy to identify genome-wide genes that are significantly differentially expressed between case and control samples (de la Fuente 2010) or between different disease stages (Weigelt et al. 2005). These differentially expressed genes are generally regarded as potential biomarkers. On the other hand, those genes that are able to discriminate samples of different conditions are also regarded as important genes and used as biomarkers.

Early approaches for identifying gene biomarkers generally detect differentially expressed genes by setting a threshold, where those genes whose expression changes above the threshold are used as gene biomarkers. For example, DeRisi et al. (1997) detected differentially expressed genes by setting a two-fold change threshold. Unfortunately, the noise inherited in the gene expression data makes it a challenging task to detect reliable differentially expressed genes with such an arbitrarily set threshold. Therefore, a lot of statistical approaches have been proposed to detect more reliable differential genes, e.g. the nonparametric approach (Pan 2003) and the empirical Bayesian method (Efron et al. 2004), where most of the approaches are based on statistical tests. The Significance Analysis of Microarrays (SAM) statistical approach proposed by Tusher et al. (2001) is one of the most widely used tools for determining the significance of the changes in expression and has shown good performance. SAM assigns a score to each gene based on its expression change relative to the standard deviation of repeated measurements for that gene, where genes with scores above a threshold are regarded as statistical significant. Later, an improved SAM statistics was proposed by Wu (2005), which utilizes the penalized linear regression model to prevent overfitting considering the large number of genes and relatively small number of samples. Both SAM and its improved version can be seen as a shrinkage of ordinary *t*-statistics, which are generally used for comparing two conditions with replication of samples. With more than two conditions, the analysis of variance (ANOVA) will be more appropriate and powerful by taking into account multiple factors and/or several sources of variation (Pavlidis 2003). More details about statistical tests for detection of differentially expressed genes are referred to a review paper by Cui and Churchill (2003).

Beyond statistical tests, the identification of gene biomarkers can be regarded as a feature/variable selection problem that is well studied in machine learning field, which is also known as gene selection in bioinformatics. In gene selection, the aim is to select a small set of genes that lead to good discrimination between diseases and normal or between different conditions. For example, Golub et al. (1999) identified a set of genes that are most correlated with the class distinctness between acute myeloid leukemia and acute lymphoblastic leukemia, and obtained a high accuracy when used together with self-organizing maps (SOMs). Guyon et al. (2002) proposed a new method for gene selection by utilizing Support Vector Machine (SVM) based on Recursive Feature Elimination (RFE), which is able to eliminate gene redundancy while get a more compact and reasonable gene set. When applied to real cancer data sets, SVM-RFE yields better classification performance and the genes identified are found to be more biologically relevant to cancer. Later, Zhang et al. (2006) developed a recursive support vector machine (R-SVM) algorithm for gene selection, which shows better performance compared with SVM-RFE. Li et al. (2001) presented a hybrid intelligent approach that combines Genetic Algorithm (GA) and *k*-Nearest Neighbor (KNN) method to identify genes capable of discriminating different classes of samples. Random forest is a recently developed algorithm for classification that utilizes an ensemble of classification trees with each tree built with a bootstrap sample of the data (Breiman 2001). Random forest has shown excellent performance even with noisy variables and is able to return measures of variable importance. When applied to gene selection, random forest shows comparable performance to other popular classification methods while identifying a small set of genes (Diaz-Uriarte and Alvarez de Andres 2006). More details about gene selection techniques are referred to the recent review papers (Duval and Hao 2010; Saeys et al. 2007).

Recently, with the descending cost of next-generation sequencing, more and more RNA-Seq data are being available. RNA-Seq is able to discover unanticipated transcripts, and detect fewer false positive transcripts compared with microarrays (McIntyre et al. 2011). Unfortunately, the well-established methods for detecting differentially expressed genes in microarray are not immediately transferable to the analysis of RNA-Seq data due to the difference between the microarray data and the RNA-Seq data. Encouragingly, a lot of tools are being introduced for this purpose, e.g. DESeq (Anders and Huber 2010), Cuffdiff 2 (Trapnell et al. 2013) and edgeR (Robinson et al. 2010). Interested readers are referred to a recent comprehensive comparison of different tools (Soneson and Delorenzi 2013).

2.4 Gene Set Biomarkers

The gene biomarkers identified above generally correlate very well with the phenotype of interest and are easy to interpret. However, the noise inherited in the data and the parameters involved in the model for identifying differential genes may lead to false positives and false negatives. For example, there is no standard criterion to set a threshold when detecting the differentially expressed genes. Pan et al. (2005) showed that different choices of the threshold values may lead to completely different biological conclusions. Although those genes with significant expression change are more likely to be related to the phenotype of interest, there are also many important genes without large enough expression changes are discarded but these genes are indeed related to the phenotype (Ben-Shaul et al. 2005; Breslin et al. 2004).

Under the circumstances, gene set analysis that investigates groups of genes instead of individual genes is becoming a trend in interpreting gene expression data, where the genes in the same group are more likely to be associated with the same biological processes. The pioneering knowledge-based approach Gene Set Enrichment Analysis (GSEA) is among such gene set analysis approaches, which scores the enrichment of predefined gene sets that share common biological functions based on the Kolmogorov–Smirnov statistic (Subramanian et al. 2005). The significance of the score is evaluated with an empirical permutation test that corrects for multiple hypothesis testing. Compared with single gene biomarkers, the gene sets identified by GSEA are pathways or processes that are more reasonable for the interpretation of the data. Furthermore, instead of focusing on significant differential genes, GSEA can detect those important genes with modest expression changes. Thereinafter, a lot of variants of GSEA have been proposed, including non-parametric enrichment statistics (Barry et al. 2005; Hänzelmann et al. 2013; Tian et al. 2005), battery testing (Dorum et al. 2009; Efron and Tibshirani 2007; Irizarry et al. 2009), and focused gene set testing (Jiang and Gentleman 2007; Wu et al. 2010a). Among these variant versions of GSEA, the

Simpler Enrichment Analysis (SEA) approach proposed by Irizarry et al. (2009) estimates enrichment based on a one-sample *t* test by assuming gene independence, which has shown better performance than GSEA. However, the gene independency assumption has its limitations as shown in (Kim and Volsky 2005; Nam et al. 2006; Tamayo et al. 2012; Wang et al. 2008). More statistical methods for the analysis of gene set enrichment can be found in the recent review papers (Chen et al. 2007; Dopazo 2009; Goeman and Buhlmann 2007; Liu et al. 2007; Nam and Kim 2008; Song and Black 2008).

Recently, it is noticed that the inter-gene correlation affects the tests and leads to Type I error. To overcome this problem, two new approaches, namely Correlation Adjusted MEan RAnk gene set test (CAMERA) (Wu and Smyth 2012) and Quantitative Set Analysis of Gene Expression (QuSAGE) (Yaari et al. 2013), have been proposed to account for inter-gene correlations and shown better performance. In the future, more reliable methodologies are believed to appear.

2.5 Pathway Biomarkers

Although the gene set biomarkers consider groups of genes that are related to the same functions or processes and are able to detect important genes with modest changes, they generally treat a gene set as a union of individual genes and assume they are functionally independent. A molecular pathway represents the interactions among a set of functionally related genes, and are most interested to biologists rather than the gene sets. It is well recognized that, instead of the mutations of individual genes, the dysfunction of molecular pathways leads to the initiation and development of diseases, especially complex diseases. Therefore, it is more reasonable to identify those dysfunctional pathways underlying diseases, i.e. pathway biomarkers, which can improve the robustness and accuracy of diagnosis compared with gene biomarkers and gene set biomarkers. Furthermore, the pathway biomarkers are more easier to interpret for the development of diseases. With more pathway knowledge being comprehensive in public databases, such as Reactome (Joshi-Tope et al. 2005) and KEGG (Kanehisa and Goto 2000), and Pathway Interaction Database (PID) (Schaefer et al. 2009), as well as the wealth of the transcriptome data that describes the activities of genes, it is possible to detect those aberrantly functioned pathways in patients.

Inspired by this, some computational approaches have been developed to identify dysfunctional pathways associated with diseases. For example, Tarca et al. (2009) proposed a signaling pathway impact analysis (SPIA) approach to measure the impact of perturbations on a given pathway under a given condition. When applied to cancer datasets, SPIA outperforms GSEA and successfully identifies pathways known to be involved in cancers. Later, Vaske et al. (2010) developed a probabilistic graphical-based model known as PARADIGM to identify patient-specific pathway activities in glioblastoma multiforme (GBM). PARADIGM is able to integrate different types of omics data and identify those pathways whose activities change significantly in patients, and detects fewer false-positives compared with SPIA. Most recently, Haynes et al. (2013) proposed a new approach entitled as Differential Expression Analysis for Pathways (DEAP) to identify disease associated pathways. Compared with other existing approaches, DEAP is able to detect the most differentially expressed portion of the pathway. DEAP successfully identified pathways related to chronic obstructive pulmonary disease and interferon treatment, some of which are generally ignored by existing approaches.

In biology, it has been observed that tumor associated alterations recurrently occur in patients but are mutually exclusive within the same molecular pathways (Ciriello et al. 2012). Based on this phenomenon, Vandin et al. (2012) proposed two novel algorithms, entitled as De novo Driver Exclusivity (Dendrix), to identify driver pathways underlying cancer from somatic mutation data. When applied to different cancer datasets, they successfully identified known tumor related pathways. Formulating the identification of driver pathways as a maximum weight submatrix problem, Zhao et al. (2012) developed two approaches for this purpose. The results on several cancer datasets demonstrate the efficiency of their approaches. Later, Leiserson et al. (2013) introduced the Multi-Dendrix algorithm for the simultaneous identification of multiple driver pathways de novo from the somatic mutation data. Benchmarking on cancer datasets, Multi-Dendrix is much faster than the iterative version of Dendrix, and gives more flexible optimal solutions for candidate pathways.

Generally, the above mentioned approaches treat pathways as independent functional units, whereas there are extensive cross-talks between distinct pathways. Similarly, the initiation and development of many diseases involve the cross-talks between pathways. Therefore, it is expected that more robust and reliable pathway biomarkers will be obtained if the cross-talks between pathways could be taken into account. Inspired by this, we proposed a novel approach to identify dysregulated pathways in cancer based on a pathway interaction network (Liu et al. 2012). Unlike traditional molecular networks, the pathway interaction network consists of pathways and their cross-talks, where each node represents a pathway and each edge represents the cross-talk between a pair of pathways. Based on the pathway interaction network, the dysregulated pathways in cancer are identified with feature selection techniques. Benchmarking on several distinct cancer datasets, the pathway biomarkers identified by our method are more reliable and accurate compared with other state of the art methods.

2.6 Network Biomarkers

Despite pathway biomarkers take into account the functional dependency among genes and are therefore more reliable, the scarceness of pathway knowledge limits the identification of pathway biomarkers. Furthermore, our current knowledge about pathways is only about their static topological structures defined based on different experiments, whereas the pathway activity is a dynamic process with different components involved under distinct conditions. On the other hand, the molecular networks can give a more global view about the biological systems while preserve the pathway structures within the network, thereby removing the limitations of prior pathway knowledge. Moreover, along with the high-throughput data, e.g. time-course gene expression, that can describe the activities of individual molecules, the molecular networks are able to characterize the dynamics of the biological systems. In addition, many diseases, especially complex diseases, are caused due to the dysfunction of multiple genes, where these genes have been found to tend to interact with each other compared with non-disease genes (Chen et al. 2013b; Goh et al. 2007). Therefore, a lot of computation approaches have been developed to identify subnetworks or modules from the molecular networks, and these subnetworks or modules have discriminative ability of separating different conditions and can therefore serve as biomarkers. Hereinafter, such predictive subnetworks or modules are called network biomarkers. Most approaches identify network biomarkers based on the analysis of differential networks that integrate the differences of single genes between distinct conditions with network topology. Based on the networks they used, these approaches can be categorized into gene association network based methods and protein–protein interaction network based approaches.

In the gene association networks, the nodes are genes and an edge is laid between a pair of genes if their coexpression correlation, typically Pearson correlation coefficient, is above a threshold. By constructing different association networks for distinct conditions based on gene expression data, the co-expression patterns associated with diseases can be extracted which are otherwise ignored by the detection of differentially expressed genes. For example, Chu et al. (2011) described an association network with Graphical Gaussian Models, and detected those edges that may rewire across two disease states by comparing the posterior probabilities of the connections in two disease conditions. Applied to breast cancer datasets, they successfully identified biomarkers consist of gene sets or pathways, which are able to separate different histological grades of breast cancer. Zhang et al. (2009) proposed a differential dependency network (DDN) analysis approach to detect statistically significant topological changes in the association networks corresponding to different conditions, and successfully detected those gene regulations that are inhibited by drug ICI. Gambardella et al. (2013) developed a new Differential Network Analysis (DINA) approach to identify condition-specific active pathways with the assumption that genes belonging to the same pathways tend to be co-regulated. DINA has been successfully utilized to detect tissue-specific pathways and identify dysregulated hepatocarcinoma-specific metabolic and transcriptional pathway. Skinner et al. (2011) developed a tool DAP finder to identify Differentially Associated Pairs (DAPs), and identified a network biomarker that is able to discriminate between oligodendroglioma (ODG) and glioblastoma multiforme (GBM) tumors.

Despite of the advantage of association networks over individual genes, it is not easy to select an appropriate threshold when constructing an association network.

Therefore, the experimentally determined protein–protein interactions (interactome) provide an alternative way to investigate the network biomarkers. Taylor et al. (2009) proposed a novel framework to detect network modules that rewire in different conditions by examining the dynamic structure of the human interactome based on gene expression data. Applied to a cohort of breast cancer patients, they found some genes that do not have significant changes in their expression but these genes have different interaction partners in surviving patients and those with poor outcomes. Furthermore, these genes can serve as a prognostic signature to predict outcomes and survival. Wu and Stein (2012) proposed a semi-supervised algorithm to discover network modules consist of interacting genes involved in the disease process. They identified novel network module signatures of 31 and 75 genes respectively for breast cancer and ovarian cancer, where the gene signatures are significantly related to cancer survival and outperform other well-known prognostic signatures. Recently, West et al. (2012) proposed to explore cancer with network entropy, and found cancer cells are characterised by the increase in network entropy. Through differential network analysis, the interaction patterns that are associated with certain diseases can be extracted from the networks. Recently, we developed a novel approach for identifying differential interactions for gastric cancer, where these interactions consist of potential disease genes were found to form network modules (Liu et al. 2012). By combining gene expression data generated under different stages of gastric cancer with human interactome, we successfully identified cancer associated network modules that serve as predictive biomarkers capable of discriminating tumors from normal samples. Benchmarking on real gastric cancer datasets, our identified module biomarkers have better performance in discriminating the tumors from normal samples compared with known biomarkers detected for gastric cancer. Investigating the dynamic structures of the module biomarkers, we noticed that the network modules have different topological structures in different gastric cancer stages as well as normal states, which provide insights into the molecular underpinnings of gastric cancer.

The above mentioned approaches generally explore the differential networks with some statistics, and the identified network modules have limited discriminative power. Therefore, some computational approaches have been proposed to identify network biomarkers by transforming the problem into a feature selection problem explicitly. For example, in their pivotal work, Chuang et al. (2007) proposed a novel approach to extract subnetworks from interactome, and the subnetworks are more reproducible biomarkers that achieve higher accuracy than individual gene biomarkers in the classification of metastatic versus non-metastatic tumors. Lee et al. (2008) proposed a novel Pathway Activity inference using Condition-responsive genes (PAC) approach to identify diagnostic biomarkers based on gene expression data, where the biomarkers are subsets of condition-responsive co-functional genes instead of individual genes or static literature-curated pathways. With defined pathway activity, their identified biomarkers outperform other pathway based approaches. Chen et al. (2013a) developed a new method based on bagging Markov random field (BMRF) to identify network biomarkers for breast cancers from human interactome. When applied to breast cancer progression and/or tamoxifen resistance, their identified biomarkers can lead to higher accuracy and are more biologically meaningful.

There are also some optimal approaches that have been proposed to identify the subnetworks that are especially active under certain conditions. For example, Kim et al. (2011) developed a novel computational method to simultaneously identify causal genes and their downstream dysregulated pathways based on a circuit flow algorithm that mimics the current flow in an electric circuit. Results on glioblastoma multiforme (GBM) demonstrate that this approach is able to identify both causal genes and causal pathways that underlie complex diseases. Lan et al. (2011) presented a tool ResponseNet to identify possible pathways that response to stimuli from molecular interaction networks based on a flow algorithm. We have developed an integer linear programming model to identify the subnetworks linking between membrane proteins and transcriptional factors based on interactome and gene expression data, which has been successfully applied to identify the yeast MAPK signaling pathways (Zhao et al. 2008). We also proposed an improved network flow model to detect the active pathways that response to stimuli (Zhao et al. 2009), and a variant of the model has been successfully used to detect network modules that response to drugs (Wu et al. 2010b).

2.7 Conclusions and Perspective

In this chapter, we introduced recent progress on computational approaches, especially differential analysis, that have been developed to detect biomarkers, ranging from gene biomarkers to gene set biomarkers, pathway biomarkers and network biomarkers. With the accumulation of various types of omics data, the intuitive differential analysis is becoming a powerful approach for detecting biomarkers, and is widely used in the community. The differential analysis based computational approaches developed for the identification of molecular biomarkers can help narrow down the search space of possible biomarkers and provide guidelines for future biological and medical experiments. Among different biomarkers, the gene biomarkers are easy to interpret and can help design targeted therapy, while the gene set/pathway/network biomarkers are more biological reasonable and have better performance since diseases are rarely caused due to the aberrant variation of single genes. Although gene set/pathway/network biomarkers generally perform better than gene biomarkers, it depends on the problem of interest to choose which type of biomarkers one should identify since pathway/network biomarkers may not perform better than gene biomarkers in some cases (Staiger et al. 2012). Considering more and more different types of omics data are being available, computational approaches that are able to integrate these multidimensional data in an efficient way are highly demanded. It is expected that more efficient computational approaches will arise to identify biomarkers that are more robust and accurate.

References

Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11:R106.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet. 2000;25:25–9.

Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet. 2004;5:101–13.

Barabasi AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet. 2011;12:56–68.

Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Muertter RN, Edgar R. NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Res. 2009;37:D885–90.

Barry WT, Nobel AB, Wright FA. Significance analysis of functional categories in gene expression studies: a structured permutation approach. Bioinformatics. 2005;21:1943–9.

Ben-Shaul Y, Bergman H, Soreq H. Identifying subtle interrelated changes in functional gene categories using continuous measures of gene expression. Bioinformatics. 2005;21:1129–37.

Breiman L. Random forests. Mach Learn. 2001;45:5–32.

Breslin T, Eden P, Krogh M. Comparing functional annotation analyses with Catmap. BMC Bioinf. 2004;5:193.

Chen JJ, Lee T, Delongchamp RR, Chen T, Tsai CA. Significance analysis of groups of genes in expression profiling studies. Bioinformatics. 2007;23:2104–12.

Chen L, Xuan J, Riggins RB, Clarke R, Wang Y. Identifying cancer biomarkers by network-constrained support vector machines. BMC Syst Biol. 2011;5:161.

Chen L, Xuan J, Riggins RB, Wang Y, Clarke R. Identifying protein interaction subnetworks by a bagging Markov random field-based method. Nucleic Acids Res. 2013a;41:e42.

Chen WH, Zhao XM, Noort Vv, Bork P. Human monogenic disease genes have frequently functionally redundant paralogs. PLoS Comput Biol. 2013b;9:e1003073.

Chu JH, Lazarus R, Carey VJ, Raby BA. Quantifying differential gene connectivity between disease states for objective identification of disease-relevant genes. BMC Syst Biol. 2011;5:89.

Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. Mol Syst Biol. 2007;3:140.

Ciriello G, Cerami E, Sander C, Schultz N. Mutual exclusivity analysis identifies oncogenic network modules. Genome Res. 2012;22:398–406.

Cui X, Churchill GA. Statistical tests for differential expression in cDNA microarray experiments. Genome Biol. 2003;4:210.

de la Fuente A. From 'differential expression' to 'differential networking'—identification of dysfunctional regulatory networks in diseases. Trends Genet. 2010;26:326–33.

DeRisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science. 1997;278:7.

Diaz-Uriarte R, Alvarez de Andres S. Gene selection and classification of microarray data using random forest. BMC Bioinf. 2006;7:3.

Dopazo J. Formulating and testing hypotheses in functional genomics. Artif Intell Med. 2009;45:97–107.

Dorum G, Snipen L, Solheim M, Saebo S. Rotation testing in gene set enrichment analysis for small direct comparison experiments. Stat Appl Genet Mol Biol. 2009;8 Article34.

Duval B, Hao JK. Advances in metaheuristics for gene selection and classification of microarray data. Brief Bioinform. 2010;11:127–41.

Efron B, Tibshirani R. On testing the significance of sets of genes. Ann Stat. 2007;1:107–29.

Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. Ann Stat. 2004;32:407–99.

Gambardella G, Moretti M, de Cegli R, Cardone L, Peron A, di Bernardo D. Differential network analysis for the identification of condition-specific pathway activity and regulation. Bioinformatics. 2013;29:1776–85.

Goeman JJ, Buhlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics. 2007;23:980–7.

Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL. The human disease network. Proc Natl Acad Sci USA. 2007;104:8685–90.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999;286:531–7.

Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn. 2002;46:389–422.

Hänzelmann S, Castelo R, Guinney J. GSVA gene set variation analysis for microarray and RNA-Seq data. BMC Bioinf. 2013;14:7.

Haynes WA, Higdon R, Stanberry L, Collins D, Kolker E. Differential expression analysis for pathways. PLoS Comput Biol. 2013;9:e1002967.

Irizarry RA, Wang C, Zhou Y, Speed TP. Gene set enrichment analysis made simple. Stat Methods Med Res. 2009;18:565–75.

Jiang Z, Gentleman R. Extensions to gene set enrichment. Bioinformatics. 2007;23:306–13.

Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath GR, Wu GR, Matthews L, Lewis S, Birney E, Stein L. Reactome: a knowledgebase of biological pathways. Nucleic Acids Res. 2005;33:D428–32.

Joyce AR, Palsson BO. The model organism as a system: integrating 'omics' data sets. Nat Rev Mol Cell Biol. 2006;7:198–210.

Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28:27–30.

Kim SY, Volsky DJ. PAGE: parametric analysis of gene set enrichment. BMC Bioinf. 2005;6:144.

Kim YA, Wuchty S, Przytycka TM. Identifying causal genes and dysregulated pathways in complex diseases. PLoS Comput Biol. 2011;7:e1001095.

Lan A, Smoly IY, Rapaport G, Lindquist S, Fraenkel E, Yeger-Lotem E. ResponseNet: revealing signaling and regulatory networks linking genetic and transcriptomic screening data. Nucleic Acids Res. 2011;39:W424–9.

Lee E, Chuang HY, Kim JW, Ideker T, Lee D. Inferring pathway activity toward precise disease classification. PLoS Comput Biol. 2008;4:e1000217.

Leiserson MD, Blokh D, Sharan R, Raphael BJ. Simultaneous identification of multiple driver pathways in cancer. PLoS Comput Biol. 2013;9:e1003054.

Li L, Weinberg CR, Darden TA, Pedersen LG. Gene selection for sample classification based on gene expression data study of sensitivity to choice of parameters of the GAKNN method. Bioinformatics. 2001;17:1131–42.

Liu Q, Dinu I, Adewale AJ, Potter JD, Yasui Y. Comparative evaluation of gene-set analysis methods. BMC Bioinf. 2007;8:431.

Liu KQ, Liu ZP, Hao JK, Chen L, Zhao XM. Identifying dysregulated pathways in cancers from pathway interaction networks. BMC Bioinf. 2012;13:126.

Manolio TA. Bringing genome-wide association findings into clinical use. Nat Rev Genet. 2013;14:549–58.

McIntyre LM, Lopiano KK, Morse AM, Amin V, Oberg AL, Young LJ, Nuzhdin SV. RNA-seq: technical variability and sampling. BMC Genomics. 2011;12:293.

Nam D, Kim SY. Gene-set approach for expression pattern analysis. Brief Bioinform. 2008;9:189–97.

Nam D, Kim SB, Kim SK, Yang S, Kim SY, Chu IS. ADGO: analysis of differentially expressed gene sets using composite GO annotation. Bioinformatics. 2006;22:2249–53.

Pan W. On the use of permutation in and the performance of a class of nonparametric methods to detect differential gene expression. Bioinformatics. 2003;19:1333–40.

Pan KH, Lih CJ, Cohen SN. Effects of threshold choice on biological conclusions reached during analysis of gene expression by DNA microarrays. Proc Natl Acad Sci USA. 2005;102:8961–5.

Parkinson H, Kapushesky M, Kolesnikov N, Rustici G, Shojatalab M, Abeygunawardena N, Berube H, Dylag M, Emam I, Farne A, Holloway E, Lukk M, Malone J, Mani R, Pilicheva E, Rayner TF, Rezwan F, Sharma A, Williams E, Bradley XZ, Adamusiak T, Brandizi M, Burdett T, Coulson R, Krestyaninova M, Kurnosov P, Maguire E, Neogi SG, Rocca-Serra P, Sansone SA, Sklyar N, Zhao M, Sarkans U, Brazma A. ArrayExpress update–from an archive of functional genomics experiments to the atlas of gene expression. Nucleic Acids Res. 2009;37:D868–72.

Pavlidis P. Using ANOVA for gene selection from microarray studies of the nervous system. Methods. 2003;31:282–9.

Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40.

Ross JS. Breast cancer biomarkers and HER2 testing after 10 years of anti-HER2 therapy. Drug News Perspect. 2009;22:93–106.

Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007;23:2507–17.

Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH. PID: the pathway interaction database. Nucleic Acids Res. 2009;37:D674–9.

Skinner J, Kotliarov Y, Varma S, Mine KL, Yambartsev A, Simon R, Huyen Y, Morgun A. Construct and compare gene coexpression networks with DAPfinder and DAPview. BMC Bioinf. 2011;12:286.

Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinf. 2013;14:91.

Song S, Black MA. Microarray-based gene set analysis: a comparison of current methods. BMC Bioinf. 2008;9:502.

Spratlin JL, Serkova NJ, Eckhardt SG. Clinical applications of metabolomics in oncology: a review. Clin Cancer Res. 2009;15:431–40.

Staiger C, Cadot S, Kooter R, Dittrich M, Müller T, Klau GW, Wessels LFA. A critical evaluation of network and pathway-based classifiers for outcome prediction in breast cancer. PLoS ONE. 2012;7:e34796.

Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 2006;34:D535–9.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA. 2005;102:15545–50.

Tamayo P, Steinhardt G, Liberzon A, Mesirov JP. The limitations of simple gene set enrichment analysis assuming gene independence. Stat Methods Med Res. 2012;0962280212460441.

Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim JS, Kim CJ, Kusanovic JP, Romero R. A novel signaling pathway impact analysis. Bioinformatics. 2009;25:75–82.

Taylor IW, Linding R, Warde-Farley D, Liu Y, Pesquita C, Faria D, Bull S, Pawson T, Morris Q, Wrana JL. Dynamic modularity in protein interaction networks predicts breast cancer outcome. Nat Biotechnol. 2009;27:199–204.

Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ. Discovering statistically significant pathways in expression profiling studies. Proc Natl Acad Sci USA. 2005;102:13544–9.

Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol. 2013;31:46–53.

Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA. 2001;98:5116–21.

Vandin F, Upfal E, Raphael BJ. De novo discovery of mutated driver pathways in cancer. Genome Res. 2012;22:375–85.

Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, Haussler D, Stuart JM. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics. 2010;26:i237–45.

von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen MA, Bork P. STRING: known and predicted protein–protein associations, integrated and transferred across organisms. Nucleic Acids Res. 2005;33:D433–7.

Wang L, Zhang B, Wolfinger RD, Chen X. An integrated approach for the analysis of biological pathways using mixed models. PLoS Genet. 2008;4:e1000115.

Weigelt B, Hu Z, He X, Livasy C, Carey LA, Ewend MG, Glas AM, Perou CM, Van't Veer LJ. Molecular portraits and 70-gene prognosis signature are preserved throughout the metastatic process of breast cancer. Cancer Res. 2005;65:9155–8.

Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447:661–678.

West J, Bianconi G, Severini S, Teschendorff AE. Differential network entropy reveals cancer system hallmarks. Sci Rep. 2012;2:802.

Wu B. Differential gene expression detection using penalized linear regression models: the improved SAM statistics. Bioinformatics. 2005;21:1565–71.

Wu D, Smyth GK. Camera: a competitive gene set test accounting for inter-gene correlation. Nucleic Acids Res. 2012;40:e133.

Wu G, Stein L. A network module-based method for identifying cancer prognostic signatures. Genome Biol. 2012;13:R112.

Wu D, Lim E, Vaillant F, Asselin-Labat ML, Visvader JE, Smyth GK. ROAST: rotation gene set tests for complex microarray experiments. Bioinformatics. 2010a;26:2176–82.

Wu Z, Zhao XM, Chen L. A systems biology approach to identify effective cocktail drugs. BMC Syst Biol. 2010b;4(Suppl 2):S7.

Yaari G, Bolen CR, Thakar J, Kleinstein SH. Quantitative set analysis for gene expression: a method to quantify gene set differential expression including gene–gene correlations. Nucleic Acids Res. 2013;41(18):e170–e170.

Zhang X, Lu X, Shi Q, Xu XQ, Leung HC, Harris LN, Iglehart JD, Miron A, Liu JS, Wong WH. Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. BMC Bioinf. 2006;7:197.

Zhang B, Li H, Riggins RB, Zhan M, Xuan J, Zhang Z, Hoffman EP, Clarke R, Wang Y. Differential dependency network analysis to identify condition-specific topological changes in biological networks. Bioinformatics. 2009;25:526–32.

Zhao XM, Wang RS, Chen L, Aihara K. Automatic modeling of signal pathways by network model. J Bioinform Comput Biol. 2009;7(2):309322.

Zhao XM, Wang RS, Chen L, Aihara K. Uncovering signal transduction networks from high-throughput data by integer linear programming. Nucleic Acids Res. 2008;36:e48.

Zhao J, Zhang S, Wu LY, Zhang XS. Efficient methods for identifying mutated driver pathways in cancer. Bioinformatics. 2012;28:2940–7.

Chapter 3
Identifying Driver Mutations in Cancer

Jack P. Hou and Jian Ma

Abstract A key question in cancer genomics is how to distinguish "driver" mutations, which contribute to tumorigenesis, from functionally neutral "passenger" mutations. Driver mutation is critically important for understanding the molecular mechanisms of cancer development and progression, which will ultimately help tailor more targeted and effective treatments for patients. In this chapter, we introduce recent developments in computational methods for identifying driver mutations. We summarize existing methods into several major categories and discuss challenges in discovering the whole spectrum of driver mutations in cancer for future computational and systems biology studies.

Keywords Cancer · Genomics · Driver mutation · Systems biology

3.1 Introduction

3.1.1 What is Driver Mutation?

Rapid advances in next-generation sequencing technologies have paved the way for comprehensive analysis for large numbers of cancer genomes (Stratton 2013). Through these advances, scientists have uncovered a large number of genetic mutations and other alterations (e.g., copy number changes, epigenetic changes,

J. P. Hou · J. Ma (✉)
Department of Bioengineering, University of Illinois, Urbana–Champaign, IL, USA
e-mail: jianma@illinois.edu

J. P. Hou
Medical Scholars Program, University of Illinois, Urbana–Champaign, IL, USA

J. Ma
Institute for Genomic Biology, University of Illinois, Urbana–Champaign, IL, USA

B. Shen (ed.), *Bioinformatics for Diagnosis, Prognosis and Treatment of Complex Diseases*, Translational Bioinformatics 4, DOI: 10.1007/978-94-007-7975-4_3,
© Springer Science+Business Media Dordrecht 2013

and structural variations) pertaining to cancer (Green et al. 2011). To understand the significant alterations that cause cancer is to discover the source of carcinogenesis—information that we can utilize to improve treatments for patients. However, the complexity of cancer and the tremendous amount of genomic data remain a daunting obstacle for us to fully understand cancer mutations. Cancer cells may often exhibit hundreds upon thousands of different mutations and other alterations in its genome that affect a wide array of genes representing many diverse functions. However, the vast majority of these genes do not have a significant impact on tumorigenesis (Hanahan and Weinberg 2011). A key question in cancer genomics is how to distinguish "driver" mutations, which contribute to tumorigenesis (Greenman et al. 2007), from functionally neutral "passenger" mutations. Such driver mutations (e.g., point mutations or copy number changes) are critically important to elucidate key biological pathways that are perturbed in cells and eventually lead to proliferation, angiogenesis, or metastasis (Hanahan and Weinberg 2011).

Detecting driver mutations is necessary for understanding the molecular mechanisms of carcinogenesis. Determining the driver will also aid in verifying and discovering new prognostic and diagnostic markers in cancer as well as therapeutic targets for potential cancer drugs. Therefore, recently in the field of computational cancer genomics, many researchers have developed computational methods to identify driver mutations (Zhang et al. 2013). Overall, these methods have different underlying principles to achieve similar goals. We can group these different methods that identify driver mutations in cancer into four broad categories:

- Sequence-Based Approaches: methods that assess the functional impact a mutation has on the candidate driver gene and its protein product (Kumar et al. 2009; Adzhubei et al. 2010; Yue et al. 2006; Reva et al. 2011; Gonzalez-Perez et al. 2012; Gonzalez-Perez and Lopez-Bigas 2012) (i.e. MutationAssessor, SIFT, Polyphen2, TransFic, SNPs3D, Oncodrive-FM).
- Machine Learning-Based Approaches: methods that use machine-learning algorithms to model existing knowledge of drivers and passengers to classify driver mutations (Hanahan and Weinberg 2011; Adzhubei et al. 2010; Carter et al. 2009; Bromberg and Rost 2007; Douville et al. 2013) (i.e. CHASM, Polyphen2, SNAP, CRAVAT).
- Frequency-Based Approaches: methods that differentiate drivers and passengers by the number of mutations seen in the candidate driver gene in contrast to the expected number of mutations from functionally neutral passengers (Boca et al. 2010; Dees et al. 2012; Reimand and Bader 2013; Lawrence et al. 2013) (i.e. MutSig, ActiveDriver, MuSiC).
- Pathway-Based Approaches: methods that identify drivers based on the impact a mutated gene would have on gene interactions and biological pathways (Wendl et al. 2011; Ciriello et al. 2012; Vandin et al. 2012; Ng et al. 2012; Bashashati et al. 2012) (i.e. MEMo, Dendrix, DriverNet, PARADIGM-Shift).

The methods described above all excel in explaining some of the biological properties associated with driver mutations (Zhang et al. 2013). Unfortunately, no model exists that can identify all the driver mutations in any given cancer with great accuracy and precision, and many existing models tend to disagree with each other (Zhang et al. 2013). Because of this, there is no computational gold standard for driver mutations in cancer (Tran et al. 2012). In this chapter, we will discuss in detail the methods associated with each of the four broad categories. We will also introduce the strengths and potential limitations of each method.

3.1.2 Properties of Driver Mutations

As stated earlier, driver mutations differ from passenger mutations in that drivers will actively alter a cell's function to display tumorigenic properties, hence "driving" the cancer, whereas passenger mutations simply occur by happenstance. Not providing functions that "drive" the cancer, passenger mutations are simply along for the ride. Drivers can have a wide variety of functions and operate on a variety of mechanisms; however, all drivers provide selective advantage to a mutant cell, allowing it to thrive, grow, and most importantly, divide rapidly to out-compete the non-mutant cells (Bunz 2008). The selective advantage, illustrated in review by Hanahan and Weinberg fall under one of six functions, called "hallmarks" of cancer cells: (1) Sustaining Proliferative Signaling, (2) Evading Growth Suppressors, (3) Resisting Cell Death, (4) Enabling Replicative Immortality, (5) Inducing Angiogenesis, and (6) Invasion and Metastasis (Hanahan and Weinberg 2011).

3.1.3 Evolutionary Model of Cancer

The concept of driver mutations can be best explained by the clonal evolution model of cancer. The clonal evolution model of cancer, as first presented by Peter Nowell in 1976, states that cancer neoplasms originate from a single cell, or clone (Nowell 1976). Over time, the original clone accumulates somatic mutations (Nowell 1976). Although the vast majority of somatic mutations induced this way are functionally neutral or damaging to the clone, in rare instances, a mutation in a hallmark gene will be advantageous to a clone. For this reason, mutated genes with hallmark properties are considered cancer genes (Nowell 1976). The cancer gene, with a hallmark property, will provide the clone with a unique advantage and higher overall fitness that allow it to survive, prosper, and out-compete other cells. This results in an outgrowth of the clone with the new mutation called a neoplasia (Bunz 2010).

A single mutation in a cancer gene is often not enough to trigger cancer (Knudson 1971). The vast majority of neoplasia are not equipped to sustain its expansion and will fail to progress and eventually die, marking the end of the particular clone (Nowell 1976; Bunz 2010). This is due to selective pressures such as the body's immune system response, changes in the cellular microenvironment, or even self-induced pressures such as a shortage of oxygen as a result from its proliferative success (Kim et al. 2009). Just as most somatic mutations will not lead to cancer genes, most neoplasia will not lead to cancer. However, in rare cases, the clone will accumulate new mutations over time, some of which will lead to the formation of new cancer genes that will provide additional growth and fitness advantages for the clone, allowing for the clone to adapt and thrive in the microenvironment and even spread to others (Nowell 1976; Bunz 2010).

The clonal evolution model illustrates many concepts that are required to better understand cancer driver mutations. First, for a mutation to be considered a driver, it must have a significant functional impact on a hallmark gene and/or biological pathway (Hanahan and Weinberg 2011). Second, since a single cancer gene gone awry is not enough to trigger cancer, cancers generally have multiple drivers (Torkamani and Schork 2008). Third, although cancer is driven by multiple drivers with hallmark properties, there are many combinations of different drivers that may lead to the same end result of cancer (Leiserson et al. 2013). Therefore, the drivers within each individual tumor may vary, highlighting the concept of tumor heterogeneity.

3.1.4 Types of Cancer Genes

There are two main types of cancer genes: oncogenes and tumor suppressors. Oncogenes are genes in which a gain of function alteration contributes to the development of cancer (Bunz 2010; Croce 2008). Genes that can become oncogenes are considered proto-oncogenes. Mutations in oncogenes are considered activating mutations as the oncogenic version of these genes present increased activity, thereby being classified as Gain-of-Function mutations. Oncogenes are generally dominant and only one mutated allele of a proto-oncogene is required for the gene to show cancer-like properties. Examples of oncogene functions are involved in functions such as Growth Factors, Receptor and Cytoplasmic Tyrosine Kinases, Serine and Threonine kinases, Regulatory GTPases, and transcription factors. Examples of oncogenes include EGFR, RAS, WNT, MYC, ERK, and TRK (Bunz 2010; Croce 2008).

In contrast, a tumor suppressor is a gene that protects a cell from becoming cancerous. A loss of function of a tumor suppressor through genetic alteration contributes to the development of cancer (Bunz 2010; Sherr 2004). Mutations in tumor suppressors are considered inactivating mutations, resulting in Loss-of-Function mutations. Unlike oncogenes, tumor suppressors are generally recessive, and for that reason, both alleles of a tumor suppressor are required to be inactivated for a functional effect, i.e. the so called "two-hit" model (Knudson 1971). Examples of tumor-suppressor gene functions include repression of genes responsible to continue the cell cycle, triggering apoptosis, blocking contact-inhibition, and repairing DNA. Examples of tumor suppressors include TP53, RB1, PTEN, BRCA1, BRCA2, PIK3CA, AKT, and APC (Bunz 2010; Sherr 2004).

3.1.5 Types of Genetic Alterations in Cancer

There are many different ways a gene can be altered. The question of where and how a gene is altered is very crucial to assessing the impact of a particular mutation. Not all mutations and genetic alterations will have the same impact on the gene (Bunz 2010; Yokota 2000). For example, a mutation in a coding region is more likely to have an impact on a gene's activity than one in a non-coding region (Kryukov et al. 2005). Even though recent studies have shown that alterations in non-coding sequences can be impactful to cancer progression (Vinagre et al. 2013; Landa et al. 2013), most current methods in detecting drivers tend to narrow the scope in coding regions only (Bunz 2010). Nevertheless, even in exonic regions, some types of mutations tend to have more impact on the overall well-being of the cell than others.

The simplest and most intuitive type of genetic mutation is the point mutation. Single base-pair substitutions refer to the replacement of a single nucleotide with another and they can be divided into three groups: silent, missense, and nonsense mutations. Silent mutations occur in the third "wobble" position of a codon (Crick 1966). Due to the redundancy of amino acid codes, silent mutations are substitutions that do not occur in a change in a protein. Silent mutations generally have the least impact, as they do not alter the primary structure of the resulting protein, although they have been shown to have minor effects on the secondary and tertiary structure of the resulting protein. A missense mutation occurs when the single base-pair change results in a single amino acid change. A missense mutation can affect all structures of the resulting protein: primary, secondary, tertiary, and quaternary. The effects of a missense mutation depend both on the similarity of the replacement protein to the original and the position of the mutation. A nonsense mutation is a mutation in which the single base pair substitution transforms an amino acid codon to a stop codon. Nonsense mutations lead to premature truncation of the protein, rendering it non-functional.

In addition to point mutations, small insertions and deletions (indels) can cause frame-shift mutations, resulting in a completely new set of codons as an indel will shift the reading frame. Like nonsense mutations, these proteins are nonfunctional. These faulty proteins are usually degraded and are responsible for the formation of null alleles (Bunz 2010).

Point mutations and indels are not the only form of genetic alterations that can lead to cancer genes. An example of large-scale mutations is copy number variation (CNV). CNVs cause changes of the number of copies of a chromosomal region. CNVs may be either amplifications, presentation of multiple copies of a gene, or deletions, the loss of gene copies. Other examples of large-scale mutations include chromosomal translocations, the interchange of genetic parts from non-homologous chromosomes; chromosomal inversions, reversing sections of a chromosome; and loss of heterozygosity, the deletion of an allele (Bunz 2008). There are other forms of genetic alterations that are epigenetic in nature. Even though these alterations have no effect on the genomic sequence itself (mainly through DNA methylation and histone modification), they can sometimes have profound effects in tumor progression. For example, DNA methyltransferases target CpG islands in the promoter region leading to spontaneous deamination and lowering of gene expression by restricting transcription, effectively silencing the gene. Promoters often are unmethylated in normal cells but hypermethylated in cancer cells.

3.2 Overview of Computational Methods to Identify Driver Mutations

The initial type of methods that identified driver mutations in cancer relied on simple recurrence as a measurement. In simple recurrence, drivers and passengers were classified by the number of times they were observed in patient populations (Jones et al. 2008). Although this method was crucial in identifying *some* common drivers such as TP53 and EGFR (Jones et al. 2008), it soon became clear that based on the biological properties of driver mutations, several difficult challenges need to be overcome in order to determine all of the driver mutations in cancer.

3.2.1 Challenges for Driver Mutation Identification

Many difficulties in identifying driver mutations arise from the concept of tumor heterogeneity, the concept that no two cancer genomes will exhibit the same mutation profiles (Stratton 2013; Pe'er and Hacohen 2011). Therefore, two patients with the same cancer may have vastly different drivers. Additionally, drivers and passengers may switch roles such that a driver in one patient may be a passenger in another patient (Cooke et al. 2010). The advent of cancer subtypes has explained some of the heterogeneity; however, it is at best a compromise. Tumor heterogeneity contributes to the long-tail distribution of the frequency cancer mutations. The long-tail hypothesis states that cancer is driven not only by a few common genes that are mutated in many patients, but also many genes that are not mutated in many patients (i.e. less frequently mutated genes) (Ding et al. 2010). This implies that there will be many rare, yet undiscovered driver mutations that are obscured by tumor heterogeneity.

Another challenge in driver mutation identification is determining what constitutes a mutation. Not all mutations are created equal, some mutations display greater functional impact on a gene in terms of its protein structure and will be more damaging (Kumar et al. 2009). Even genes that have functionally damaging mutations across many patients are not necessarily drivers. Some genes have little functionality in cancer development and progression but are mutated frequently by chance. The most famous example of a highly recurrent passenger gene is the TTN gene. TTN is the largest gene in the human genome, and it functions as a molecular spring for the passive elasticity in muscle cells (Nair and Banerji 2013).

TTN does not have a large impact in many of the flagship cancer pathways (Lawrence et al. 2013). However, due to its large size, it is often mutated in cancer cells due to random chance alone, confounding the results of many methods.

A third challenge is to map the biological function of potential driver mutations. As demonstrated in the TTN example, some genes may present damaging mutations but due to the gene's function being unrelated to cancer pathways, they are most likely to be passengers. Individual driver genes do not operate by themselves; rather they interact with many other genes in complex biological networks (Bashashati et al. 2012). Therefore, driver mutations must be verified by their biological functions. A driver mutation is expected to interact with other genes in various cancer pathways to further promote different hallmarks of cancer (Hanahan and Weinberg 2011; Schwartzentruber et al. 2012).

3.2.2 Resources Available for Driver Mutation Identification

For researchers interested in identifying driver mutations, there exists a wealth of publicly-available data regarding molecular signature data, compendiums on driver mutations, pathway databases, and comparison tools that all can be utilized to achieve a greater understanding of driver mutations in cancer. Perhaps the most comprehensive of these resources is The Cancer Genome Atlas (TCGA), a resource of molecular alterations over large cohorts of patients representing a wide array of cancers (Cancer Genome Atlas Research Network 2008). With regards to curated catalogs of known somatic mutations in cancer, the Sanger Institute's COSMIC and the Cancer Gene Census, maintain a well-defined comprehensive list of common mutations already identified as drivers (Bamford et al. 2004; Futreal et al. 2004). Other tools such as Biocarta (Kim et al. 2012), NCI Pathway interaction Database (PID) (Schaefer et al. 2009), Reactome (Croft et al. 2011), or the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto 2000) all provide valuable information on curated cancer pathways for evaluating potential driver genes.

3.2.3 Summary of Different Algorithms for Driver Mutation Identification

Name	Type	Website
SIFT	Sequence-based	http://sift.jcvi.org/
PolyPhen2	Sequence-based Machine learning-based;	http://genetics.bwh.harvard.edu/pph2/

(continued)

(continued)

Name	Type	Website
MutationAssessor	Sequence-based	http://www.bitnos.com/info/mutation-assessor
TransFic	Sequence-based; aggregate method	http://bg.upf.edu/transfic/help
Oncodrive-FM	Sequence-based; aggregate method	http://bg.upf.edu/group/projects/oncodrive-fm.php
SNPs3D	Sequence-based; aggregate method	http://www.snps3d.org/
CHASM	Machine learning-based	http://wiki.chasmsoftware.org/index.php/Main_Page
CRAVAT	Machine learning-based; aggregate method	http://www.cravat.us/
SNAP	Machine learning-based	https://rostlab.org/services/snap/
MutSig	Frequency-based	http://www.broadinstitute.org/cancer/cga/mutsig
MutSigCV	Frequency-based	http://www.broadinstitute.org/cancer/cga/mutsig
ActiveDriver	Frequency-based	http://www.baderlab.org/Software/ActiveDriver
MuSiC	Frequency-based; Pathway-based	http://gmt.genome.wustl.edu/genome-music/0.2/doc/
MEMo	Pathway-based	http://cbio.mskcc.org/tools/memo/
HotNet	Pathway-based	http://compbio.cs.brown.edu/projects/hotnet/
Dendrix	Pathway-based	http://compbio.cs.brown.edu/projects/dendrix/
DriverNet	Pathway-based	http://www.bioconductor.org/packages/2.12/bioc/html/DriverNet.html
Paradigm-Shift	Pathway-based	http://sysbio.soe.ucsc.edu/paradigm/tutorial/

3.3 Sequence-Based Approaches

The underlying belief in these approaches is that mutations that have functional impact on a gene are more likely to be driver mutations in cancer. These methods assess the functional impact of mutations by predicting the consequences, either through evolutionary impact on conserved regions or changes in the resulting amino acid and potential effects on the protein's secondary and tertiary structure. Examples of these approaches include Separating Tolerant from Intolerant (SIFT) which performs multiple sequence alignments (MSA) to determine the evolutionary impact of altered amino acids in protein homologs to predict functional impacts (Kumar et al. 2009); Polyphen2 combines a multiple sequence alignment to detect mutations with a Naïve Bayes Classifier (NBC) to train the potential functional impact (Adzhubei et al. 2010).

Results from many sequence-based approaches applied to cancer studies have shown that mutations in driver genes tend to have a much higher functional impact to the sequence and resulting protein structure than those of non-driver genes (Reva et al. 2007, 2011; Gonzalez-Perez et al. 2012; Gonzalez-Perez and Lopez-Bigas 2012). These methods also have the advantage of being able to evaluate individual patients' mutations to identify the drivers (Reva et al. 2007, 2011; Bashashati et al. 2012). However, these approaches also present several drawbacks as well. These methods are unable to separate mutations that provide a selective advantage to the overall cell fitness (Zhang et al. 2013). By definition, only mutations that provide a selective advantage to the tumor's growth and survival can be considered driver mutations (Hanahan and Weinberg 2000). Therefore, sequence-based approaches often struggle in separating driver mutations from passenger mutations. This drawback has prompted many groups to look into other methods to detect driver mutations, and for this reason, sequence-based approaches are not commonly used as the sole determinant of novel driver mutations (Zhang et al. 2013; Adzhubei et al. 2010; Yue et al. 2006). Nevertheless, these tools are widely applied as filters, comparison tools, and confirmation for more cancer-specific driver mutation methods.

3.3.1 MutationAssessor

The aforementioned sequence-based methods are generic methods to identify functionally relevant mutations and are not specific to cancer driver mutations. However, some methods have shown to perform well in detecting impactful mutations. One method is MutationAssessor, which predicts the consequence of a mutation using a Functional Impact Score (FIS). The FIS is a metric used to quantify a mutation's impact on a gene by observing the evolutionary conserved patterns from a MSA using combinatorial entropy formalism (Reva et al. 2011).

The FIS of any non-synonymous mutation can be calculated as the average of two conservation scores: the general conservation score S_i^C and the subtype conservation score S_i^S. A mutation in a conserved region is more likely to have a functional impact than a mutation in a non-conserved region (Henikoff and Henikoff 1992). MutationAssessor measures the impact of a mutation from the wild-type amino acid residue α to the mutant β using an entropy score. The general conservation score at position i with respect to the MSA to go from $S_i^C(\alpha \rightarrow \beta)$ therefore is:

$$S_i^C(\alpha \rightarrow \beta) = -\ln\left(\frac{n_i(\beta) + 1}{n_i(\alpha)}\right) \tag{3.1}$$

where $n_i(\alpha)$ is the number of sequences which display the residual α (the wild type) at position i and $n_i(\beta)$ is the number of sequences which display the residual β (the mutant) at position i. This change predicts the functional impact of a protein by determining if a change in the amino acid sequence is highly conserved or not. MutationAssessor takes one step further by assessing the entropy difference of the particular subfamily of the observed difference $S_i^S(\alpha \to \beta)$. The rationale of determining subfamily impact is to model different interaction partners or substrates on the background of similar, conserved biochemical or cellular function (Sarid et al. 1987). To determine subfamilies, a clustering algorithm is used to divide the MSA into subfamilies and the subfamily conservation score $S_i^S(\alpha \to \beta)$ is a measure of the entropy difference between the $\alpha \to \beta$ change with regards to the subfamily that the β residual belongs.

$$S_i^S(\alpha \to \beta) = -\ln\left(\frac{n_i^p(\beta) + 1}{n_i^p(\alpha)}\right) \tag{3.2}$$

where $n_i^p(\beta)$ and $n_i^p(\alpha)$ in equation are the residual counts of α and β with respect to a particular subfamily p. The FIS score for MutationAssessor is simply the average of the two aforementioned conservation scores.

MutationAssessor applied the FIS score for 10,000 mutations cataloged in COSMIC and it was shown that genes with a high FIS score were much more likely to become drivers (Reva et al. 2007).

3.3.2 TransFic

There have been methods that combine the predictive value of several methods to determine the impact of genes in cancer. One example is TransFic, a method that combines the scores from MutationAssessor, SIFT, and Polyphen2, and compares their scores to the distribution of scores of alterations observed in genes with similar functional annotations to select for drivers (Gonzalez-Perez et al. 2012). The use of functional annotations in TransFic was applied to obtain a better grasp on the function of a particular driver in question.

The process of selection is illustrated below:

1. Obtain the Functional Annotations of the gene of interest using four sources: Gene Ontology Biological Process (GOBP) and Molecular Function (GOMF) categories, canonical pathways (CP), and Pfam domain (Dom) (Henikoff and Henikoff 1992; Dejongh et al. 2004; Chagoyen and Pazos 2010; Yu et al. 2012; Punta et al. 2012).
2. Determine the alterations associated with all genes related to the most specific functional term of the original gene of interest. This allows TransFic to not only calculate the impact of an altered gene, but also predict its biological function.

3. If less than 20 alterations are found, the user may choose to add other alterations in genes that have similar functions as the original gene of interest. This allows for an accurate reading of the functional impact score even with less available input.
4. Calculate and normalize the SIFT, Polyphen2 and MutationAssessor scores. The SIFT and Polyphen2 scores first undergo a logit transformation.
5. Calculate the mean, standard deviation and other summary statistics to determine the aggregate FIS score of both the gene and the potential function.

The authors compared their method to each of the individual methods that they aggregated and found that the aggregated results that were more concordant with COSMIC's category of driver mutations. They tested their score with the breast cancer driver PIKC3A and found that the impact of the mutation was more mild than previously thought. Another software developed by the same lab was Oncodrive-FM (Gonzalez-Perez and Lopez-Bigas 2012). Oncodrive-FM uses SIFT and Polyphen2, along with other driver mutation software such as MutSig in order to determine to select driver genes that present accumulated functional impact mutations across a gene (Gonzalez-Perez and Lopez-Bigas 2012).

3.3.3 SNPs3D

SNPs3D is another sequence-based approach that attempts to combine information from many different sources to draw conclusions (Yue et al. 2006). SNPs3D is made up of three gene modules: one concerning the impact a non-synonymous SNP (in our case, a point mutation in a tumor) has on the network, one that connects genes to other related genes based on a PubMed literature search, and a third which provides users with a literature score to measure how likely a gene is related to certain diseases. SNPs3D is unique in that it associates literature scores as a direct measurement to disease association (Yue et al. 2006).

SNPs3D covers the sequence-based data of a driver mutation using two methods: the first determining the amino acid substitution's stability on a proteins folded state (Yue et al. 2005) and the second being a conservation score similar to the one presented in MutationAssessor (Yue and Moult 2006). SNPs3D also links genes together to form gene to gene interactions based on the number of PubMed search results returning the pair of genes. It also counts abstracts from PubMed to link a mutated gene with a disease (Stapley and Benoit 2000). Using this integrated approach, SNPs 3D discovered candidate genes for a long list of diseases, including around 200 potential candidates for Lung Cancer (Yue et al. 2006).

3.4 Machine Learning-Based Methods

Machine-learning approaches operate by training a classifier on a gold standard of driver and passenger mutations to develop a model, which is utilized to determine the drivers and passengers of a new dataset. Generally, these methods train their data from a catalog of missense mutations, and the classifiers themselves range from Naïve Bayes Classifiers to Random Forests to Neural Networks. Machine-learning based approaches have better ability to distinguish drivers from passengers than methods that only consider mutation's functional impact. Once a model is classified, the model can be fitted to any number of patients or groups. However, the machine-learning approaches heavily rely on a gold standard of driver and passenger mutations as a training set, which could be problematic as there currently is no established computational gold standard. Even though COSMIC and the CGC have good compendium for common drivers, they do not take into account rare drivers (Futreal et al. 2004).

3.4.1 CHASM

One example of a machine learning-based method is the Cancer-specific High-throughput Annotation of Somatic Mutations (CHASM). CHASM seeks to identify and rank missense mutations most likely to augment tumor cell proliferation (Carter et al. 2010). CHASM applies a Random Forest Classifier on 49 predictive features including amino acid substitution properties, alignment-based estimates of evolutionary conservation at the mutated position, predicted structural changes at the mutated position and annotations from the UniProtKB feature table. The Random Forest Algorithm is a decision tree classifier that uses a set of random classification trees to vote on a classification of a particular mutation as "driver" and "passenger". Each tree then "votes" for the eventual classification of the alteration (Carter et al. 2009, 2010; Gnad et al. 2013).

The authors selected 2,488 missense mutations breast, colorectal, and pancreatic cancers. The driver mutations selected were from COSMIC and various biological studies in which specific genes were demonstrated to have proliferative roles, and the passenger mutations were computer generated via simulation with an algorithm that recapitulates base substitutions found in brain tumors (Carter et al. 2010). The authors reported higher sensitivities and specificities than traditional sequence-based methods such as SIFT and Polyphen2. Additionally, when training the classifier, the authors reported that many of the variables by themselves only explained a small percentage of the model, which the authors used to justify their rationale behind Random Forests. Random Forests work with each variable jointly rather than as individuals. When applied to a GBM dataset, the authors predicted that 49 of the 607 missense mutations in the GBM dataset, or 8 %, were drivers (Carter et al. 2009).

3.4.2 CRAVAT

A recent machine learning-based method, Cancer-Related Analysis of Variants Toolkit (CRAVAT), seeks to provide predictive scores on the importance of somatic alterations of in cancer genes using a variety of classifier tools (Douville et al. 2013). CRAVAT is unique because it (1) combines the results of multiple classifiers to hone in on both the impact of the driver and the biological function of a somatic alteration; (2) provides a user-friendly workflow where users can submit their jobs to the server and receive both the gene's importance rating and a variety of PubMed literature sources that relate to the important drivers that CRAVAT predicted; and (3) is not limited by the size of the dataset (Douville et al. 2013).

CRAVAT uses three machine learning tools for its workflow: SnvGet, CHASM, and VEST (Carter et al. 2009, 2013; Wong et al. 2011). CRAVAT uses SnvGet to get classifier information for the subsequent CHASM and VEST runs. SnvGet returns 86 pre-computed features for each alteration such as physiochemical properties of amino acid residues; scores derived from multiple sequence alignments of protein or DNA; region-based amino acid sequence composition; predicted properties of local protein structure; and annotations from the UniProtKB feature tables (Wong et al. 2011). The features are then used by CHASM to predict whether or not the alteration in question is a driver, and then VEST (designed by the same authors as CHASM), which also utilizes a Random Forest classifier to determine the function impact of the predicted protein. The *p*-values from both tests are aggregated to return a list of functional driver genes for the user (Douville et al. 2013).

3.4.3 Polyphen2 and SNAP

In addition to CHASM, several other machine-learning approaches have been used to identify driver mutations. Polyphen2, as mentioned earlier as a sequence based method, uses the Naïve Bayes Classifier (NBC) to predict functional impact, improving on the traditional multiple sequence-based approach with knowledge from machine learning (Adzhubei et al. 2010). The alignment output from Polyphen2 is used to select the features for the Naïve Bayes Classifier, which is then used to classify them on function. The NBC works by solving the probability of a sampling belonging to a group c from all groups C using Baye's rule with respect to features $F_1, F_2 \ldots F_n$. The group with the highest probability that a sample could belong is the predicted classifier.

Another method, SNAP, utilizes a neural network to predict the functional effects of non-synonymous SNP, which can be applied to missense mutations to predict drivers (Bromberg and Rost 2007). Both Polyphen2 and SNAP are general functional impact algorithms that can be applied to cancer but are not necessarily created to specifically model cancer mutations.

3.5 Frequency-Based Methods

The third class of driver mutation identification software is methods based on mutation frequency. In the early days of driver mutation identification, simple recurrence was the first method to determine driver mutations. Drivers were defined by the number of times a gene was mutated (Schwartzentruber et al. 2012). Although many common driver mutations were detected using this method, simple recurrence has since fallen out of favor as it does not account for (1) rare mutations in the long tail of driver gene distribution and (2) propensity to select genes that have a high probability due to chance to be mutated by being large or having a high background mutation rate.

Frequency-based methods are among the most powerful methods in classifying common driver genes and passenger genes, and these methods have been some of the most widely-adopted and widely-utilized methods in driver mutation detection (D'Antonio and Ciccarelli 2013). However, one drawback of frequency-based methods is that these methods, like machine-learning based methods, require a large amount of input data from many patients to operate.

3.5.1 MutSig

One of the most-utilized frequency-based methods is MutSig (Banerji et al. 2012). The original MutSig assumes a single average background mutation rate, μ, which can be tailored to be category-specific: μ_c.

Examples of category specific criteria taken from a Lung Carcinoma study were (1) transitions in C's or G's in CpG dinucleotides; (2) transversions in C's or G's in CpG dinucleotides; (3) transitions in other C's or G's; (4) transversions in other C's or G's; (5) transitions at A's or T's; (6) transversions in A's or T's; and (7) small insertions/deletions, nonsense and splice site mutations (Lawrence et al. 2013). Then to calculate a p-value for each gene based on category-specific background rates, a score s is calculated for each gene. The score of each gene's mutation significance s_g is based on the binomial probability distribution given the parameters of the number of mutations in the category n_c, the number of bases covered by those mutations N_c, and that category's background mutation rate: μ_c.

$$s_g = \sum_c -10 \times \text{binomial}(n_c, N_c, \mu_c) \tag{3.3}$$

After calculating the score, the background distributions of all the mutation rates are convoluted and a p-value is calculated by calculating the probability that the convoluted mutation rates can exceed the score s_g. A Benjamin-Hochberg correction is used to correct for multiple testing (Lawrence et al. 2013). The authors of the original MutSig applied the data to a Lung cancer dataset and found 450 candidate drivers that were mutated at a frequency much higher than the expected frequency as assumed from the background mutation rate (Greulich et al. 2012).

3.5.2 MutSigCV

Recently, a newly-published version of MutSig, MutSigCV (Lawrence et al. 2013), has been released. MutSigCV offers additional features to the original MutSig. MutSig corrects for the extensive false positive findings of previous driver mutation identification software by correcting for the heterogeneity of the mutation rates among genes, the mutations rates among patients, and among the mutation types themselves by allowing separate models for multiple types of heterogeneity. MutSigCV also incorporates molecular properties of the gene that may co-vary with the mutation rate of the gene into their model. Examples include gene expression, DNA replication time, open versus closed chromatin status, local GC content, and local gene density (Lawrence et al. 2013).

In MutSigCV, each gene is placed in a high-dimensional covariate space and the gene's nearest neighbors are identified to supplement information to the background mutation rate of the gene in question. The information from the nearest neighbors of the gene, dubbed "Bagel", is combined with the gene's own mutation rates to estimate the background mutation rate. This process, combined with category and patient-specific background mutation rates (calculated via the original MutSig model) provide the mutation rates used to calculate the significance of each gene.

The authors of MutSigCV analyzed 3,083 tumor normal pairs to both look for sources of heterogeneity and for novel driver mutations. The authors found that tissue type mutation rate are highly variable and that lung and skin cancers tend to have high mutation rates although much of the variation can also be attributed to the patients themselves (Lawrence et al. 2013). The authors also studied the type of mutation present for tissue types, and found that lung cancer tended to have more C→T mutations while melanoma patients tended to have more C→A mutations. The regional heterogeneity was one of the most variable, meaning that certain genes are much more likely to mutate by chance than others, and that that mutation rates tended to coincide with gene expression and the time of DNA replication. Taking into account this heterogeneity, the method assigned each gene and tumor type a score, which was used to correct the background rate of mutations in specific genes for specific tumors, and patients. This approach was used to confirm common drivers, eliminate false positive drivers, and suggest possible new drivers (Lawrence et al. 2013).

3.5.3 ActiveDriver and MuSiC

Other recent methods include ActiveDriver and MuSiC (Dees et al. 2012; Reimand and Bader 2013). ActiveDriver is a method developed to discover driver genes in among genes with phosphorylation single nucleotide variants (pSNV). ActiveDriver performs a hypothesis test to determine whether or not the phosphosite-specific mutation rate is the same as the gene-wide mutation rate for particular genes using generalized linear regression tests. The authors of ActiveDriver found that their approach identified many common phospho-specific drivers such as TP53 and EGFR as well as new candidate genes in FLNB, GRM1, POU2F1 (Reimand and Bader 2013).

MuSiC also employs the concept of selecting for genes that tend to be mutated more than a background mutation rate in their novel test, Significantly Mutated Gene (SMG) test. The background rate was a combination of mutated genes in the entire sample set of all patients, mutated genes in the patient, and mutated genes within the subgroup of the gene in question (Dees et al. 2012). MuSiC also supplement their results using pathway analysis through the PathScan algorithm (Wendl et al. 2011), which combines individual selection of driver genes to a multiple-sample value using the Fisher-Lancaster approach (Wendl et al. 2011) to determine the mutated pathway of the driver genes in their analysis.

3.6 Pathway-Based Methods

The most recent type of model to determine driver mutation relies on biological pathways. Pathway-based models have been shown to be effective not only in reliably determining common driver mutations, but also have been able to pinpoint the biological pathways that could be the source of the cancer (Ciriello et al. 2013). As a result, pathway-based methods have a unique advantage over other types of methods in that they take into account gene interactions and potential biological effects rather than simply viewing driver genes individually (Wu et al. 2010). For example, a particular candidate driver gene that shows significantly more mutations in cancer than in normal cells may still not be a true driver gene (Michor and Polyak 2010). If the candidate gene does not affect a cancer pathway or does not interact with many genes that are crucial in cancer-pathways, the candidate gene may have no true biological connection to cancer. Pathway-based approaches allow us to verify functional impactful candidate drivers. These methods are sometimes used to supplement other methods as was demonstrated in the case of ActiveDriver and MuSiC, as measure of the biological significance of their methods (Dees et al. 2012; Reimand and Bader 2013; Wendl et al. 2011).

3.6.1 MEMo

Some pathway-based approaches are not built with specific cancer genes in mind, but rather, these approaches are aimed at discovering driver pathways, groups of genes that may interact together to promote tumorigenesis. Mutual Exclusivity Modules in cancer (MEMo) serves to determine groups of genes that contribute to tumorigenesis (Ciriello et al. 2012). These gene groups, or modules, together are highly recurrent, have similar pathway impact in terms of biological processes, and also are mutually exclusive meaning that only one gene in each gene group is mutated at a time in any given patient. This idea follows the mutual exclusivity rule in cancer pathways, i.e., generally one mutated gene in a pathway is enough to alter the pathway's function. The algorithm for MEMo is described below:

1. Build binary event matrix of significantly altered genes. The binary event matrix (B) is an $n \times m$ matrix where n is the number of genes in the dataset and m is the number of samples (patients) being observed. As a binary event matrix, a cell in the matrix $B_{i,j}$ will be 0 if a gene i is altered in the sample j.
2. Build a gene network to identify gene pair interactions. This step involves the building of a gene network that will gauge the interactions and pathways present in cancer genes. The authors at MEMo built two gene networks: the first being a combination Human Interaction Network based on both curated and non-curated networks, and the second one simply based on manual curation.
3. Extract Cliques: MEMo then finds all cliques in the network. A clique is a fully connected subgraphs such that each subgraph cannot be contained by another fully connected subgraph.
4. Assess each clique for mutual exclusivity. The idea of this step is to determine whether or not the clique has both highly recurrent gene alterations, and also whether or not only one gene in the subgraph is mutated at once. MEMo tests on whether the set of genetic alterations occurs by chance. MEMo builds a null model by randomly permuting the event matrix, and then applies a Markov Chain Monte Carlo method called "permutation switching" to randomly generate networks to find simulated cliques. The cliques are tested for mutual exclusivity under the null model, thus allowing MEMo to determine an empirically derived p-value to gauge the mutual exclusivity of the cliques.

The authors of MEMo discovered several mutually-exclusive modules in GBM such as EGFR, PDGFRA, and NF1 and TP53, CDKN2A, and GLI1. One of the genes in these modules is likely to be altered in any given patient. MEMo is a unique approach at observing cancer as it acknowledges that although patients may have different mutations to drive the cancer, many of those mutations have similar biological effects eventually (Ciriello et al. 2012).

3.6.2 HotNet and Dendrix

In the spirit of finding subnetworks in cancer, Vandin et al. developed two algorithms to determine the impact of mutated genes have on biological pathways: HotNet and Dendrix (Vandin et al. 2011, 2012). HotNet algorithm combines mutation data and protein–protein interaction network information to find subnetworks of genes that are mutated in a significant number of cancer patient (Vandin et al. 2011). Using mutation and gene interaction data on an undirected graph, HotNet uses a heat diffusion algorithm where a mutated sends a "heat" signature based on the number of mutations present in that gene evenly to its neighbors such that genes with lower degrees of connectivity receive a larger proportion of "heat" than those with high connectivity. The idea behind HotNet is that genes with lower connectivity will define the boundaries of the neighborhood (the subnetwork) as they will retain heat better, allowing HotNet to pinpoint subnetworks.

Dendrix, on the other hand, determines driver pathways using two concepts: Mutually Exclusivity (as demonstrated in MEMo) and coverage (recurrence). Modeling a gene interaction network as an adjacency matrix, Dendrix finds the submatrix within the matrix that will maximum coverage, that is, cover the most patients while being mutually exclusive, that is not having any two genes in the submatrix mutated simultaneously within a patient (Vandin et al. 2012). Dendrix uses a greedy MCMC method to do so. After selecting a starter gene, Dendrix selects the neighbor that has the most mutations without any of those mutations being in a patient that already had a mutation in a previously selected gene. One frequently sampled gene set from Dendrix's application to GBM was CDKN2B, RB1, CYP27B1 (Vandin et al. 2012).

3.6.3 DriverNet

One of the most recent pathway-based methods is DriverNet (Bashashati et al. 2012). DriverNet models both gene mutation events and differential expression events of a group of patients into a bipartite graph. The algorithm then applies pathway information to select for mutated genes that are the most well-connected to genes that are differentially expressed. The DriverNet algorithm is a greedy optimization algorithm aimed at determining driver genes as genes with the most pathway impact, which they measure as genes that create the most outlying differentially expressed genes. The greedy optimization algorithm is described below:

1. Create a bipartite graph $B(V^m, V^0, E)$, a graph whose vertices can be divided into two disjoint sets V^m and V^0 such that every edge connects a vertex in V^m to one in V^0. In DriverNet's case, V^m is a mutation matrix built in a similar fashion as MEMo's binary event matrix. V^0 is a binary $n \times m$ matrix where n is the number of genes in the dataset and m is the number of samples (patients) being observed. V^0 is equal to 1 for gene i with respect to patient j if the normalized difference between the tumor and normal expression exceeds a certain threshold. E is an adjacency matrix representing the gene network that connects V^m and V^0 in the bipartite graph. E can be built by similar procedures as MEMo's adjacency matrix.
2. Let Z be the set of all connected outlying events, and z be the set of covered outlying events (initially a null set).

3. Choose the mutated gene that contains the largest number of uncovered outlying expression events. Add that to the driver mutation list. Add the outlying events to z.
4. Remove the mutated gene and its connecting edges from the bipartite graph *B*.
5. Stop when all connected outlying events all covered (when Z = z).

DriverNet combines gene expression, mutation information among groups of patients, and biological pathways (Bashashati et al. 2012). The authors of DriverNet tested their results in Breast Cancer and Glioblastoma datasets and found an abundance of infrequently mutated genes: 22 in the Breast Cancer dataset and 13 in Glioblastoma. The advantage of DriverNet is that it is less dependent on recurrence and therefore can detect rare mutation.

3.6.4 PARADIGM-Shift

PARADIGM-Shift predicts functions of driver genes as gain-of-function or loss-of-function genes in specific cancer pathways (Ng et al. 2012). PARADIGM-Shift has the ability to determine not only if a candidate driver is functionally impactful, but also the type of impact that the driver gene may show. The authors utilized PAthway Recognition Algorithm using Data Integration on Genomic Models (PARADIGM) (Vaske et al. 2010), using gene expression and cy number change signals as inputs to determine the impact of upstream and downstream genes of a candidate driver. The difference activity in upstream and downstream genes of the driver determines a gain-of-function (high downstream, low upstream activity) or loss-of-function (high upstream, low downstream activity).

The activity score was determined by PARADIGM, which uses belief-propagation on a factor graph to compute the log-posterior odds score called inferred pathway levels (IPLs) for each gene, complex, protein family and cellular process using gene expression, copy number and/or genetic interaction. Genes that are more active in a tumor with more activity have positive IPL scores while genes with less activity in the tumor than normal cells have negative IPL scores (Vaske et al. 2009). PARADIGM-Shift runs two iterations of PARADIGM, one with the gene of interest and its upstream genes in the pathways to measure the loss of function score, and one with only the gene of interest and its downstream genes to measure the gain of function score. The PARADIGM-Shift score is the difference of the two paradigm runs. The authors of PARADIGM-Shift applied their approach to both common, TP53, and uncommon, NFE2L2, genes to analyze the impact (Ng et al. 2012).

3.7 Discussion

Each of the four approaches, and the various methodologies associated with each of the approaches, has different advantages and addresses many of the challenges associated with driver mutations. Unfortunately, no method can solve all the challenges, and no perfect model exists that can fully reverse engineer the clonal evolution model of cancer and select only drivers that serve a function relating to the hallmarks of cancer. The task of accounting for tumor heterogeneity, genetic function, and mutation severity is indeed daunting. Many researchers, therefore, have applied multiple methods to determine driver mutations (Adzhubei et al. 2010; Bashashati et al. 2012). The multi-step approach allows for researchers to address multiple challenges in driver mutation identification at the same time.

In addition to the current challenges involved in driver mutation identification, there are also many future avenues of studying driver mutations that have yet to be identified and modeled. Some examples include analyzing the cumulative effects of passenger mutations, accounting for intra-tumor heterogeneity, and predicting the effects of mutations in non-coding regions.

A study from McFarland et al. found that even though a single passenger mutation has a negligible impact on tumorigenesis, the cumulative effect of all passengers may affect a cell's tumor progression model in ways not explainable by widely accepted driver mutation models (McFarland et al. 2013). Much intra-tumor heterogeneity is also ignored by driver mutation methods as most cancer genome sequencing project sequences a bulk tumor tissue from a population of cancer cells. In other words, the sequencing is a simple average of the cells, and no model exists to explain intra-tumor heterogeneity (Michor and Polyak 2010).

The methods described in this chapter are mostly only applicable to point mutations in coding regions of the genome. As described earlier, only a small subset of cancer mutations are point mutations. Detailed impacts of larger scale mutations and structural rearrangements have yet to be described. Additionally, only 2 % of the genome codes for proteins, leaving 98 % of the genome in non-coding regions unexplained. Mutations in non-coding regions can have profound impact on gene regulation related to cancer development and progression. Currently, no driver mutation software can systematically predict the effects of alterations in non-coding sequences. All these challenges need to be addressed by future computational methods.

References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7:248–9.

Bamford S, Dawson E, Forbes S, Clements J, Pettett R, Dogan A, Flanagan A, Teague J, Futreal PA, Stratton MR, Wooster R. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. Br J Cancer. 2004;91:355–8.

Banerji S, Cibulskis K, Rangel-Escareno C, Brown KK, Carter SL, Frederick AM, Lawrence MS, Sivachenko AY, Sougnez C, Zou L, Cortes ML, Fernandez-Lopez JC, Peng S, Ardlie KG, Auclair D, Bautista-Pina V, Duke F, Francis J, Jung J, Maffuz-Aziz A, Onofrio RC, Parkin M, Pho NH, Quintanar-Jurado V, Ramos AH, Rebollar-Vega R, Rodriguez-Cuevas S, Romero-Cordoba SL, Schumacher SE, Stransky N, Thompson KM, Uribe-Figueroa L, Baselga J, Beroukhim R, Polyak K, Sgroi DC, Richardson AL, Jimenez-Sanchez G, Lander ES, Gabriel SB, Garraway LA, Golub TR, Melendez-Zajgla J, Toker A, Getz G, Hidalgo-Miranda A, Meyerson M. Sequence analysis of mutations and translocations across breast cancer subtypes. Nature. 2012;486:405–9.

Bashashati A, Haffari G, Ding J, Ha G, Lui K, Rosner J, Huntsman DG, Caldas C, Aparicio SA, Shah SP. DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. Genome Biol. 2012;13:R124.

Boca SM, Kinzler KW, Velculescu VE, Vogelstein B, Parmigiani G. Patient-oriented gene set analysis for cancer mutation data. Genome Biol. 2010;11:R112.

Bromberg Y, Rost B. SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res. 2007;35:3823–35.

Bunz F. Principles of cancer genetics. Dordrecht: Springer, 2008. p. 325.

Bunz F. Principles of Cancer Genomics. Berlin: Springer; 2010.

Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. 2008;455:1061–8.

Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, Vogelstein B, Karchin R. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. Cancer Res. 2009;69:6660–7.

Carter H, Samayoa J, Hruban RH, Karchin R. Prioritization of driver mutations in pancreatic cancer using cancer-specific high-throughput annotation of somatic mutations (CHASM). Cancer Biol Ther. 2010;10:582–7.

Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics. 2013;14(Suppl 3):S3.

Chagoyen M, Pazos F. Quantifying the biological significance of gene ontology biological processes—implications for the analysis of systems-wide data. Bioinformatics. 2010;26:378–84.

Ciriello G, Cerami E, Aksoy BA, Sander C, Schultz N. Using MEMo to discover mutual exclusivity modules in cancer. *Curr Protoc Bioinformatics*, Chap. 8: Unit 8 17 (2013).

Ciriello G, Cerami E, Sander C, Schultz N. Mutual exclusivity analysis identifies oncogenic network modules. Genome Res. 2012;22:398–406.

Cooke SL, Ng CK, Melnyk N, Garcia MJ, Hardcastle T, Temple J, Langdon S, Huntsman D, Brenton JD. Genomic analysis of genetic heterogeneity and evolution in high-grade serous ovarian carcinoma. Oncogene. 2010;29:4905–13.

Crick FH. Codon–anticodon pairing: the wobble hypothesis. J Mol Biol. 1966;19:548–55.

Croce CM. Oncogenes and cancer. N Engl J Med. 2008;358:502–11.

Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, Caudy M, Garapati P, Gopinath G, Jassal B, Jupe S, Kalatskaya I, Mahajan S, May B, Ndegwa N, Schmidt E, Shamovsky V, Yung C, Birney E, Hermjakob H, D'Eustachio P, Stein L. Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res. 2011;39:D691–7.

D'Antonio M, Ciccarelli FD. Integrated analysis of recurrent properties of cancer genes to identify novel drivers. Genome Biol. 2013;14:R52.

Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, Mooney TB, Callaway MB, Dooling D, Mardis ER, Wilson RK, Ding L. MuSiC: identifying mutational significance in cancer genomes. Genome Res. 2012;22:1589–98.

Dejongh M, Van Dort P, Ramsay B. Linking molecular function and biological process terms in the ontology for gene expression data analysis. Conf Proc IEEE Eng Med Biol Soc. 2004;4:2984–6.

Ding L, Wendl MC, Koboldt DC, Mardis ER. Analysis of next-generation genomic data in cancer: accomplishments and challenges. Hum Mol Genet. 2010;19:R188–96.

Douville C, Carter H, Kim R, Niknafs N, Diekhans M, Stenson PD, Cooper DN, Ryan M, Karchin R. CRAVAT: cancer-related analysis of variants toolkit. Bioinformatics. 2013;29:647–8.

Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. A census of human cancer genes. Nat Rev Cancer. 2004;4:177–83.

Gnad F, Baucom A, Mukhyala K, Manning G, Zhang Z. Assessment of computational methods for predicting the effects of missense mutations in human cancers. BMC Genomics. 2013;14(Suppl 3):S7.

Gonzalez-Perez A, Lopez-Bigas N. Functional impact bias reveals cancer drivers. Nucleic Acids Res. 2012;40:e169.

Gonzalez-Perez A, Deu-Pons J, Lopez-Bigas N. Improving the prediction of the functional impact of cancer mutations by baseline tolerance transformation. Genome Med. 2012;4:89.

Green ED, Guyer MS, National Human Genome Research I. Charting a course for genomic medicine from base pairs to bedside. Nature. 2011;470:204–13.

Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C, Edkins S, O'Meara S, Vastrik I, Schmidt EE, Avis T, Barthorpe S, Bhamra G, Buck G, Choudhury B, Clements J, Cole J, Dicks E, Forbes S, Gray K, Halliday K, Harrison R, Hills K, Hinton J, Jenkinson A, Jones D, Menzies A, Mironenko T, Perry J, Raine K, Richardson D, Shepherd R, Small A, Tofts C, Varian J, Webb T, West S, Widaa S, Yates A, Cahill DP, Louis DN, Goldstraw P, Nicholson AG, Brasseur F, Looijenga L, Weber BL, Chiew YE, DeFazio A, Greaves MF, Green AR, Campbell P, Birney E, Easton DF, Chenevix-Trench G, Tan MH, Khoo SK, Teh BT, Yuen ST, Leung SY, Wooster R, Futreal PA, Stratton MR. Patterns of somatic mutation in human cancer genomes. Nature. 2007;446:153–8.

Greulich H, Kaplan B, Mertins P, Chen TH, Tanaka KE, Yun CH, Zhang X, Lee SH, Cho J, Ambrogio L, Liao R, Imielinski M, Banerji S, Berger AH, Lawrence MS, Zhang J, Pho NH, Walker SR, Winckler W, Getz G, Frank D, Hahn WC, Eck MJ, Mani DR, Jaffe JD, Carr SA, Wong KK, Meyerson M. Functional analysis of receptor tyrosine kinase mutations in lung cancer identifies oncogenic extracellular domain mutations of ERBB2. Proc Natl Acad Sci USA. 2012;109:14476–81.

Hanahan D, Weinberg RA. The hallmarks of cancer. Cell 2000;100:57–70.

Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell 2011;144:646–74.

Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA. 1992;89:10915–9.

Jones S, Zhang X, Parsons DW, Lin JC, Leary RJ, Angenendt P, Mankoo P, Carter H, Kamiyama H, Jimeno A, Hong SM, Fu B, Lin MT, Calhoun ES, Kamiyama M, Walter K, Nikolskaya T, Nikolsky Y, Hartigan J, Smith DR, Hidalgo M, Leach SD, Klein AP, Jaffee EM, Goggins M, Maitra A, Iacobuzio-Donahue C, Eshleman JR, Kern SE, Hruban RH, Karchin R, Papadopoulos N, Parmigiani G, Vogelstein B, Velculescu VE, Kinzler KW. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. Science. 2008;321:1801–6.

Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28:27–30.

Kim Y, Lin Q, Glazer PM, Yun Z. Hypoxic tumor microenvironment and cancer cell differentiation. Curr Mol Med. 2009;9:425–34.

Kim S, Kon M, DeLisi C. Pathway-based classification of cancer subtypes. Biol Direct. 2012;7:21.

Knudson AG Jr. Mutation and cancer: statistical study of retinoblastoma. Proc Natl Acad Sci USA. 1971;68:820–3.

Kryukov GV, Schmidt S, Sunyaev S. Small fitness effect of mutations in highly conserved non-coding regions. Hum Mol Genet. 2005;14:2221–9.

Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4:1073–81.

Landa I, Ganly I, Chan TA, Mitsutake N, Matsuse M, Ibrahimpasic T, Ghossein RA, Fagin JA. Frequent Somatic TERT promoter mutations in thyroid cancer: higher prevalence in advanced forms of the disease. J Clin Endocrinol Metab. 2013;98:E1562–6.

Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, Kiezun A, Hammerman PS, McKenna A, Drier Y, Zou L, Ramos AH, Pugh TJ, Stransky N, Helman E, Kim J, Sougnez C, Ambrogio L, Nickerson E, Shefler E, Cortes ML, Auclair D, Saksena G, Voet D, Noble M, DiCara D, Lin P, Lichtenstein L, Heiman DI, Fennell T, Imielinski M, Hernandez B, Hodis E, Baca S, Dulak AM, Lohr J, Landau DA, Wu CJ, Melendez-Zajgla J, Hidalgo-Miranda A, Koren A, McCarroll SA, Mora J, Lee RS, Crompton B, Onofrio R, Parkin M, Winckler W, Ardlie K, Gabriel SB, Roberts CW, Biegel JA, Stegmaier K, Bass AJ, Garraway LA, Meyerson M, Golub TR, Gordenin DA, Sunyaev S, Lander ES, Getz G. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 2013;499:214–8.

Leiserson MD, Blokh D, Sharan R, Raphael BJ. Simultaneous identification of multiple driver pathways in cancer. PLoS Comput Biol. 2013;9:e1003054.

McFarland CD, Korolev KS, Kryukov GV, Sunyaev SR, Mirny LA. Impact of deleterious passenger mutations on cancer progression. Proc Natl Acad Sci USA. 2013;110:2910–5.

Michor F, Polyak K. The origins and implications of intratumor heterogeneity. Cancer Prev Res (Phila). 2010;3:1361–4.

Ng S, Collisson EA, Sokolov A, Goldstein T, Gonzalez-Perez A, Lopez-Bigas N, Benz C, Haussler D, Stuart JM. PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis. Bioinformatics 2012;28:i640–6.

Nowell PC. The clonal evolution of tumor cell populations. Science 1976;194:23–8.

Pe'er D, Hacohen N. Principles and strategies for developing network models in cancer. Cell. 2011;144:864–73.

Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer EL, Eddy SR, Bateman A, Finn RD. The Pfam protein families database. Nucleic Acids Res. 2012;40:D290–301.

Reimand J, Bader GD. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. Mol Syst Biol. 2013;9:637.

Reva B, Antipin Y, Sander C. Determinants of protein function revealed by combinatorial entropy optimization. Genome Biol. 2007;8:R232.

Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. 2011;39:e118.

Sarid J, Halazonetis TD, Murphy W, Leder P. Evolutionarily conserved regions of the human c-myc protein can be uncoupled from transforming activity. Proc Natl Acad Sci U S A. 1987;84:170–3.

Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH. PID: the Pathway Interaction Database. Nucleic Acids Res. 2009;37:D674–9.

Schwartzentruber J, Korshunov A, Liu XY, Jones DT, Pfaff E, Jacob K, Sturm D, Fontebasso AM, Quang DA, Tonjes M, Hovestadt V, Albrecht S, Kool M, Nantel A, Konermann C, Lindroth A, Jager N, Rausch T, Ryzhova M, Korbel JO, Hielscher T, Hauser P, Garami M, Klekner A, Bognar L, Ebinger M, Schuhmann MU, Scheurlen W, Pekrun A, Fruhwald MC, Roggendorf W, Kramm C, Durken M, Atkinson J, Lepage P, Montpetit A, Zakrzewska M, Zakrzewski K, Liberski PP, Dong Z, Siegel P, Kulozik AE, Zapatka M, Guha A, Malkin D, Felsberg J, Reifenberger G, von Deimling A, Ichimura K, Collins VP, Witt H, Milde T, Witt O, Zhang C, Castelo-Branco P, Lichter P, Faury D, Tabori U, Plass C, Majewski J, Pfister SM, Jabado N. Driver mutations in histone H3.3 and chromatin remodelling genes in paediatric glioblastoma. Nature 2012;482:226–31.

Sherr CJ. Principles of tumor suppression. Cell 2004;116:235–46.

Stapley BJ, Benoit G. Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. Pac Symp Biocomput. 2000:529–540.

Stratton MR. Journeys into the genome of cancer cells. EMBO Mol Med. 2013;5:169–72.

Thasni KAT, Ratheeshkumar T, Rojini G, Sivakumar KC, Nair RS, Srinivas G, Banerji A, Somasundaram V, Srinivas P. Structure activity relationship of plumbagin in BRCA1 related cancer cells. Mol Carcinog. 2013;52:392–403.

Torkamani A, Schork NJ. Prediction of cancer driver mutations in protein kinases. Cancer Res. 2008;68:1675–82.

Tran B, Dancey JE, Kamel-Reid S, McPherson JD, Bedard PL, Brown AM, Zhang T, Shaw P, Onetto N, Stein L, Hudson TJ, Neel BG, Siu LL. Cancer genomics: technology, discovery, and translation. J Clin Oncol. 2012;30:647–60.

Vandin F, Upfal E, Raphael BJ. Algorithms for detecting significantly mutated pathways in cancer. J Comput Biol. 2011;18:507–22.

Vandin F, Upfal E, Raphael BJ. De novo discovery of mutated driver pathways in cancer. Genome Res. 2012;22:375–85.

Vaske CJ, House C, Luu T, Frank B, Yeang CH, Lee NH, Stuart JM. A factor graph nested effects model to identify networks from genetic perturbations. PLoS Comput Biol. 2009;5:e1000274.

Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, Haussler D, Stuart JM. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics. 2010;26:i237–45.

Vinagre J, Almeida A, Populo H, Batista R, Lyra J, Pinto V, Coelho R, Celestino R, Prazeres H, Lima L, Melo M, Rocha AG, Preto A, Castro P, Castro L, Pardal F, Lopes JM, Santos LL, Reis RM, Cameselle-Teijeiro J, Sobrinho-Simoes M, Lima J, Maximo V, Soares P. Frequency of TERT promoter mutations in human cancers. Nat Commun. 2013;4:2185.

Wendl MC, Wallis JW, Lin L, Kandoth C, Mardis ER, Wilson RK, Ding L. PathScan: a tool for discerning mutational significance in groups of putative cancer genes. Bioinformatics. 2011;27:1595–602.

Wong WC, Kim D, Carter H, Diekhans M, Ryan MC, Karchin R. CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer. Bioinformatics. 2011;27:2147–8.

Wu G, Feng X, Stein L. A human functional protein interaction network and its application to cancer data analysis. Genome Biol. 2010;11:R53.

Yokota J. Tumor progression and metastasis. Carcinogenesis. 2000;21:497–503.

Yu N, Seo J, Rho K, Jang Y, Park J, Kim WK, Lee S. hiPathDB: a human-integrated pathway database with facile visualization. Nucleic Acids Res. 2012;40:D797–802.

Yue P, Moult J. Identification and analysis of deleterious human SNPs. J Mol Biol. 2006;356:1263–74.

Yue P, Li Z, Moult J. Loss of protein structure stability as a major causative factor in monogenic disease. J Mol Biol. 2005;353:459–73.

Yue P, Melamud E, Moult J. SNPs3D: candidate gene and SNP selection for association studies. BMC Bioinform. 2006;7:166.

Zhang J, Liu J, Sun J, Chen C, Foltz G, Lin B. Identifying driver mutations from sequencing data of heterogeneous tumors in the era of personalized genome sequencing. Brief Bioinform. 2013.

Chapter 4
Biomarker Discovery with Text Mining and Literature Based Discovery

Fei Zhu and Bairong Shen

Abstract The huge numbers of biomedical publications provide us valuable data for research. However, how to get usable information from these integrated but unstructured biomedical is a difficult problem in front of us, which calls for biomedical text mining techniques aiming at extracting novel knowledge from scientific texts. In this chapter, we will introduce basis of text mining and examine some frequently used algorithms, tools, and data sets. With the development of systems biology, researchers tend to understand complex biomedical systems from a systems biology viewpoint. Thus, the full utilization of text mining to facilitate systems biology research is fast becoming a major concern. To address this issue, we describe the general workflow of text mining in systems biology and each phase of the workflow. Finally, we will discuss the text mining technology for research on biomarkers.

Keywords Text mining · Literature mining · Biomarker · Systems biology · Knowledge discovery

4.1 Introduction to Biomedical Text Mining

Biomedical texts provide abundant knowledge for biomedical research. Text mining can help us from a mountain of text mining the useful information and knowledge, and now is widely used in biomedical research. Since 2000, the number of publications using PubMed "text mining" as key words has a substantial increase. Many researchers have make full use of the advantage of text mining technology to discover novel knowledge to improve the development of

F. Zhu · B. Shen (✉)
Center for Systems Biology, Soochow University, No. 1. Shizi Street, Suzhou, Postbox 206
215006 Jiangsu, China
e-mail: bairong.shen@suda.edu.cn

B. Shen (ed.), *Bioinformatics for Diagnosis, Prognosis and Treatment of Complex Diseases*, Translational Bioinformatics 4, DOI: 10.1007/978-94-007-7975-4_4,
© Springer Science+Business Media Dordrecht 2013

biomedical research, especially those of social category malignant diseases, such as cancer.

Text mining involves multiple several computing domains, such as machine learning, natural language processing, biostatistics, information technology, and pattern recognition. There are also a lot of related applications on biomedical text mining, such as identification malignant tumor related genes, protein (genes, proteins, etc.) in biological medicine, the relationship between the biomedical entities (protein–protein and gene–disease, etc.), and extracting knowledge from text generate hypotheses. In the past few years, many review articles for biomedical text mining have been published. From this aspect, we believe text mining applications in biomedical areas.

4.2 Tasks and Phases of Biomedical Text Mining

The goal of text mining is to achieve the recessive knowledge, which is hidden in the unstructured text and render it in an explicit form. It usually has four steps: information retrieval, information extraction, knowledge discovery, hypothesis generation. Information retrieval system designed to achieve required information for a particular theme; information extraction system is used to extract the scheduled types of information, such as the relationship between extraction; knowledge discovery system help us extracted from the text of novel knowledge; biomedical facts infer unknown hypothesis generation system based on text, as we can see in Fig. 4.1 (Zhu et al. 2013). Therefore, general biomedical text mining tasks including information retrieval, named entity recognition and relationship extraction, knowledge discovery and hypothesis generation.

4.2.1 Information Retrieval

In addition to the traditional information retrieval system, as well as advanced intellectual information retrieval system, different data resources will be integrated into a single system to improve our understanding of complex biological systems. For example, Saliva Ontology from Salivaomics Knowledge Base (Ai et al. 2010) as a term and the vocabulary of relations in facilitate data retrieval and the integration of research together with data analysis and data mining; QuExT (Matos et al. 2010), which can be used to find documents containing concepts related to query words followed a concept-oriented query expansion methodology to find documents containing concepts related to query words. In the genome era, in the progress of biotechnology and high-throughput genetic analysis methods, there will be a text mining, information retrieval tools needs to help researchers to find related articles to help them study work.

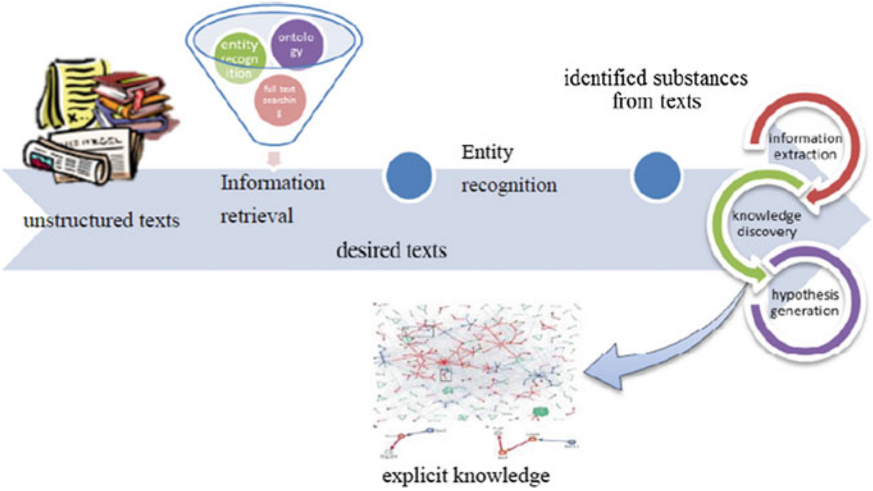


Fig. 4.1 Conventional tasks and phases involved in biomedical text mining, generally including information retrieval, named entity recognition and relationship extraction, knowledge discovery and hypothesis generation

4.2.2 Named Entity Recognition

One of the most important step in the extraction of knowledge is named entity recognition, and its aim is identifying specific terms, for example, biomarker, gene, protein, disease, and drug (Leser and Hakenberg 2005). Biomedical term identification involves several computing technologies. Yet, in fact, there are still factors, such as several different written forms in a biomedical term, which will cause mistakes in automatic identification (Dagar et al. 2011). In addition, as a term can be expressed in different ways, for example, it can be said as a disease, cancer, or astronomy sign, it makes the task ever harder to be solved well. What's more, ambiguity problems can be caused from abbreviations of terms.

Now there are three kinds of techniques in entity recognition: dictionary-based approaches, rule-based approaches and machine learning approaches (Cohen and Hersh 2005; Li et al. 2009). Because there is no complete reference dictionary, dictionary-based approaches is easy to miss unregistered terms (Rebholz-Schuhmann et al. 2011). The rule-bases approaches uses rules to identify terms from texts. However there is no rules that is always effective for call cases (Rebholz-Schuhmann et al. 2011). Most conventional machine learning approaches generally require dataset, which in fact takes tremendous human efforts to build, to learn and construct a model for identifying terms, thereby machine learning approaches are often tend to be data-driven and application domain-oriented. As a result, it is difficult to apply machine learning approaches to broad areas.

Machine learning approaches such as Hidden Markov Models (HMM) (Ephraim and Merhav 2002), Support Vector Machines (SVMs) (Habib and Kalita 2010), Conditional Random Fields (CRFs) (He and Kayaalp 2008), and Maximum Entropy (ME), have been used for named entity recognition. For examples, Zhou and Su (2004) used a system based on HMM with biomedical information as domain knowledge to recognize protein, DNA, RNA, cell-type, and cell-line. Kazama et al. used SVMs to identify protein, DNA, cell-type, cell-line, and lipid, with a 73.6 % F1 rate (Kazama et al. 2002). Tsai et al. (2006) developed a CRFs system to extract protein mentions, achieving a 78.4 % F1 rate. Lin et al. used ME to recognize 23 categories of biological terms with a 72 % F1 rate (Lin 2004). Presently, the best F1 rates for biomedical named entity recognition systems are not as good as the results from general purpose ones (Chang et al. 2010). Researchers have tried their best to improve the performance, by combining different ways and proposing hybrid approaches (Zhu and Shen 2012), conducting post-processing after machine learning conducting post-processing, and take biomedical domain knowledge (Sasaki et al. 2008). In addition, one of the tasks in BioCreative III is focused on gene normalization, which identifies gene mentions and links these genes to standard identifiers (e.g., database identifiers).

4.2.3 Relation Extraction

Relationship extraction in biomedical area is focused on investigating biomedical relation extraction from biomedical terms (Arighi et al. 2011). There is much related work. The system developed by Abacha and Zweigenbaum (2011) could identify the correct semantic relationship between each pair of entities using MetaMap (Aronson and Lang 2010) to identify medical substances. A linguistic patterns approach is used in their system to determine the semantic relationship between each pair. The system developed by Chun et al. (2006) is able to find out gene–disease relations from Medline. Presently, in the current genomic era, many researchers are interested in mining gene–gene interactions, protein–protein interactions, and other interactions in genome-wide associations that provide useful scaffolds for further integrative analysis of gene expression and database annotation (Cohen and Hersh 2005; Wren and Garner 2004; Raychaudhuri et al. 2002; Raychaudhuri and Altman 2003), as well as other extensive relationships (Krallinger et al. 2011). Eskin and Agichtein (2004) applied text mining technology and combined it with sequence analysis to discover protein subcellular localizations, and the results seemed to be highly accurate. Li et al. (2010) took the method based on text mining to determine interaction from the biomedical literature. They used the Bayes method from genome and proteomics data set to validate the results of interaction by integrating heterogeneous types of evidence. The systems developed by Agarwal et al. (2011) can be used to determine an associated with the interaction of protein–protein, as well as the map interaction related articles. Tsai (2011) build a text mining and visualization framework, which was proposed to find the details of the interactions between proteins and identifying the sequence of amino acids to deeply understand the function of the protein in the protein interaction interface.

In addition, researchers are focusing on the more extensive biomedical relationship between genes and other biomedical entities, as well as the relationship between proteins and other biomedical entities, such as gene–disease relationships and protein sub-cellular relationships. For example, the system developed by Garten et al. (2010) can systematically access information to analyze genetic, cellular, and molecular aspects of the plant Arabidopsis thaliana. Trugenberger et al. (2013) studied the relationship between diseases and disease areas. Turenne et al. (2012) constructed a statistical document classifier that was based on MEDLINE citations to determine whether a drug had caused adverse effects. Their systems contributed to current drug safety procedure. Pena-Hernandez et al. implemented an extraction tool to find gene relationship and up-to-date pathways from literature Epstein (2009).

4.2.4 Knowledge Discovery

Facts, information, and description of knowledge, no matter implicit or explicit, refer to theory or practical understanding of a topic or field (Frawley et al. 1992). Knowledge discovery is creation of knowledge from huge amounts of structured or unstructured data. Knowledge discovery is a very important part of text mining. Extracting the knowledge from the biomedical text is a process, the purpose of which is to find out the answer to biomedical problems, such as new drug targets detection and biomarkers identification. The Crab, developed by Korhonen et al. (2012) fully integrated text mining technologies to extract relevant information for cancer risk assessment. Their work indicates that text mining pipeline can contribute to the research of the biomedical and complex task. In addition, Nam and Park (2012) integrated text mining with previous work, finding that there are two pathways involved in predictor gene set indicative of susceptibility to early-onset colorectal cancer overcoming the shortages of genome-wide expression research work of colorectal cancer. Knowledge discovery has the ability to integrate with other sources of data to generate a new explanation context (Mack and Hehenberger 2012). Urzua et al. (2010), for example, through the text mining technology research with microarray data, found that post-transcription control of ovarian process could be responsible for observed tumor and reproductive phenotypes. They also speculated that it is repetitive cycling that represented the actual link between ovarian tumorigenesis and reproductive records.

4.2.5 Hypothesis Generation

Some facts or information could be unexplained well by present knowledge. By scientific hypothesis which is a test to solve the problem rather than a theory, it is possible to put forward some suggestions to further study. Experiments can be used to evaluate the proposed hypothesis before solve this problem. Scientific hypothesis is like a scientific imagination, which is based on the existing evidence and knowledge.

Hypothesis generation from biomedical texts is a method to find new knowledge through the clues hidden in texts. Biomedical literature is used to create potential information to make inferences or biomedical hypothesis as a treasure trove. Hypothesis generation is very important in text mining, for the biomedical researchers to infer unknown biomedical facts, and it can be used to guide experimental design or explain the existing experimental results. It gradually gets more and more attentions. Swanson (1986) used pattern rules to determine a hidden link between of fish oil and Raynaud's syndrome in published literature. Li et al. (2009) constructed Alzheimer's disease-specific drug-protein network from protein interaction networks by using text mining approach. They put forward a new hypothesis that diltiazem and quinidine could be candidate drugs for Alzheimer treatment. Hanisch et al. (2005); Hettne et al. (2007) used an association-based technique and natural language processing tools to generate a sorted list related to disease genes, and extract the relationship between the gene and lipopolysaccharide. Topinka and Shyu (2006) also used text mining-based as well as structure-based protein–protein interaction to predict cancer interaction networks.

4.3 Workflow of Text Mining Based Systems Biology Research

Complex biological systems tend to be understood nowadays from a viewpoint of systems biology (Macilwain 2011), a network based on systems biology can be constructed by aggregating previously reported associations from the literature or various databases. For instance, based on associations reported in the literature, Hayasaka et al. (2011) constructed a network of genes, genetic diseases, and brain areas. Sharma et al. (2006) gathered a serial of genes that known disease-related and used text mining to build an interaction network, confirming that 19 genes were related to prostate cancer after analysis. Therefore, it's no doubt a new hot topic to take full advantage of text mining to facilitate systems biology research. Texts acquisition, bio-entity terms recognition, complex relation extraction, new knowledge discovery, and hypothesis generation in turn gradually become the conventional flow of text mining that based on systems biology research.

From many available sources, firstly we get related biomedical texts in the general phase of text mining of systems biology, such as PubMed. Whereas although it is convenient to obtain packed data download service from plenty of literature bases, there exist several problems, such as timely updating and literature quantity control. We can use some scripts or write programs to automatically download the texts by application programming interface provided by many literature base systems. As an instance, users can obtain up-to-date texts facilely through E-utility of PubMed (McEntyre and Lipman 2001).[1]

Afterwards using named entity recognition tools to extract biomedical mentions, and the most frequently used ones are gene names, protein names, mRNA (message RNA) names, miRNA (micro RNA) names, metabolism related terms, and cell terms. Correcting and normalizing terms are the most prior because the results of NER are the fundamental of the successive steps. Since the terms prove to be correct and normalized, dictionary-based approaches become the usual preference in order to achieve the aim. Besides, we can also use automatized identify approaches and post manual curation which guarantees a pretty high precision, along with many other resources, such as Gene Ontology to get a normalized bio-term.

Then a bio-entity interaction network with the bio-terms can be constructed, such as gene–gene interaction network, metabolism pathways and so on. A number of interaction extraction tools can be utilized to obtain interactions automatically from inputted texts. We can investigate some other recognized biomedical entities, biological entities and bio-factors considered to be related with cancer, and then find out how they work in the network and in what way they affect the network. We pay attention to build certain networks and their variations, such as protein–protein interaction networks (Papp et al. 2011) and variations in metabolism network from texts after focusing on how components and structures change in dynamic contexts in the next stage. Some validation and inference algorithms can be used to correct and optimize the network owing to the high false negative rate in text mining-based networks, along with many resources, such as homology, co-expression data, rich domain data, and co-biological process data, through which to strength some nodes and interactions with strong evidence, then to update or remove a false one. As a result a bio-entity interactome which based on multiple sources of interaction evidence can be developed, such as protein–protein interactome (Li et al. 2010). Ultimately, all the networks and components can be utilized for further studies.

It plays an important part of signaling pathway reconstruction to understand the molecular mechanisms in cancer. Functionality of construct signaling pathway maps offered by some advanced text mining tools are obtained from manual literature search, and evaluated by canonical pathway databases (Alexopoulos et al. 2010). We believe that hypothesis for future work can be proposed through the networks and pathways gained.

An illustration of a text mining-assisted biomedical study workflow from a systems biology viewpoint is showed as Fig. 4.2.

[1] Pubmed. http://www.ncbi.nlm.nih.gov/pubmed/.

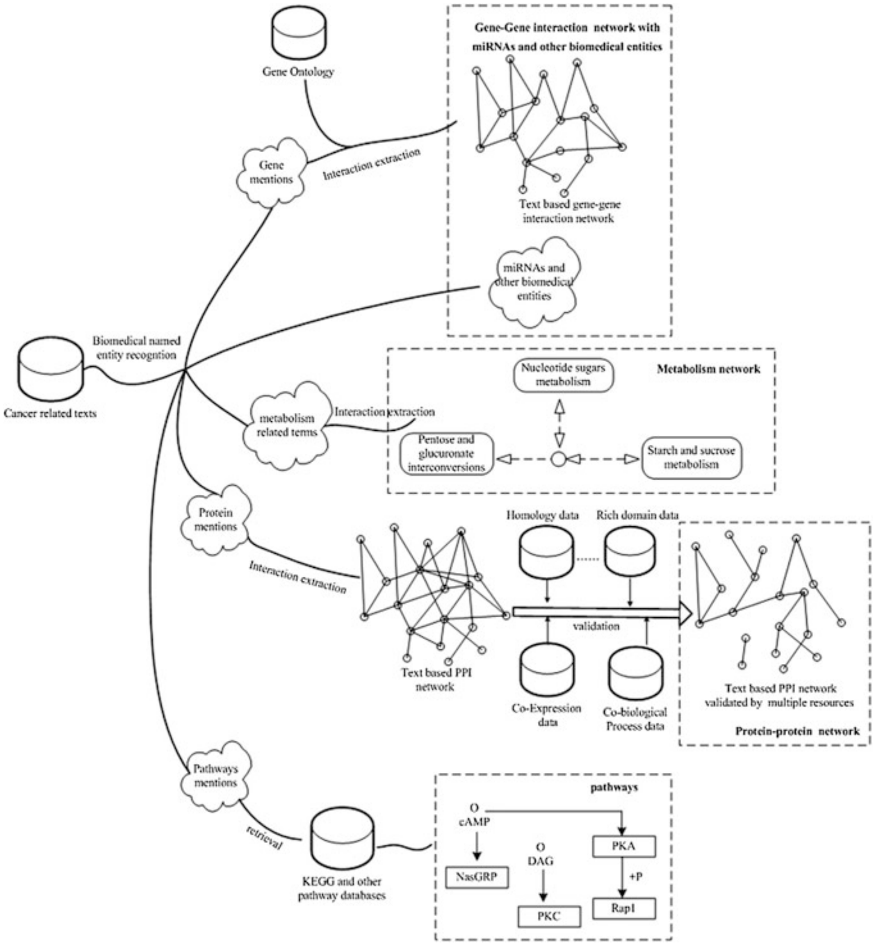

Gene Ontology
Gene-Gene interaction network with miRNAs and other biomedical entities
Gene mentions
Interaction extraction
Text based gene-gene interaction network
miRNAs and other biomedical entities
Biomedical named entity recognition
Cancer related texts
Metabolism network
Nucleotide sugars metabolism
metabolism related terms
Interaction extraction
Pentose and glucuronate interconversions
Starch and sucrose metabolism
Protein mentions
Homology data
Rich domain data
Interaction extraction
validation
Text based PPI network
Co-Expression data
Co-biological Process data
Text based PPI network validated by multiple resources
Protein-protein network
Pathways mentions
retrieval
KEGG and other pathway databases
pathways
cAMP
PKA
DAG
NasGRP
+P
PKC
Rap1

Fig. 4.2 An illustration of a text mining-assisted cancer study workflow from a systems biology viewpoint (adopted from Zhu et al. 2013)

4.4 MicroRNAs Discovery with Biomedical Text Mining

MicroRNAs (miRNAs), which play a key role in diverse biological processes, are small RNA molecules that regulate genes. They are now perceived as a key layer of post-transcriptional control within the networks of gene regulation. In several diseases, such as cancer, miRNAs expression is altered; therefore it is very likely that altering miRNA expression could lead to human diseases. Several evidences suggest that there is functional association between miRNAs and cancer. It is worthy to understand this functional association: firstly, miRNAs can control cell proliferation and apoptosis; secondly, most human miRNAs are located at fragile sites in the genome that are commonly altered in human cancer; thirdly, comparing to normal tissues, miRNAs are widely deregulated. Nowadays, many studies have produced large number of miRNA-disease associations which are more than 70 cases and showed that the mechanisms of miRNAs involved in diseases are very complex. The quantity of microarray data analyzing the gene expression in diseases is exponentially increasing, which leads to disease gene signatures delivered on a regular basis, and it is better to define gene signatures which bear a signature of regulatory activity of miRNAs for diseases than explore the dysregulated biological pathways and cellular processes in diseases.

Lasso regression model was used to predict functional associations between miRNAs and diseases based on gene signatures of each by Qabaja et al. (2013). They evaluated the performance of it as a miRNA enrichment analysis method as a proof of concept, and then evaluated the performance of Lasso regression model on the disease-miRNA interaction networks. They found that gold standard data was biased toward certain diseases that had around hundred associated miRNAs, and there also existed other diseases associated with very few miRNAs. They focused on prostate cancer as a case study to further validate the novel miRNA-disease associations predicted by the model. The results showed a promise of finding underlying functional associations between miRNAs and diseases using regression models for integrating disease and miRNA signatures.

Their work is a key step to understand disease development to decipher miRNA-disease functional association. Integrating disease signature with miRNA target interactions to build miRNA-disease functional association shows promise to decipher significant associations between diseases and miRNAs. It is important for the uncovered interactions to understand diseases and patterns underlying miRNA-disease associations despite the limitations in the current work.

4.5 Biomarker Identification Using Text Mining from PubMed

Identifying molecular biomarkers is an essential task now to assess the different phenotypic states of cells or organisms. The PubMed database offers an enriched source to explore the biomarkers across human disease and to mine the biomarkers related to diseases, meanwhile, text mining has become a critical technique for designing future predictive and personalized medicine. Hereby, integrating text mining has become a fast emerging research area in many specifically biomarker discovery studies. Therefore, efficient text mining tools and developed algorithm are exceedingly needed.

The method proposed by Li and Liu (2012) is based on text mining technique and the PubMed database, accompanied with the full text search-engine technology (Lucence) and a complex network of biological and signaling pathways, which provides a clear text mining to discover biomarkers. First they created a DBXML database from the PubMed database, then constructed a DBXML database, and finally use the finite state machine to conform the disease-related biomarkers. The workflow is shown as Fig. 4.3. They used a finite state machine (FSM) to identify biomarkers which are mined from the PubMed database, pathways, and associated diseases. The association between the biomarkers and the diseases can be output to refine the biomarkers which are accepted by FSM.

Pharmaceutical research is undergoing a great change. Recent trends in the pharmaceutical industry are changing the way of drug discovery, which utilizes computational methods. What's more, drugs are designed to be targeted to different populations and individuals with common biological characteristics, called personalized medicine (Garten et al. 2010). The characteristics are called biomarkers and/or phenotypes. Due to its importance to targeted drug design and personalized medicine, Trugenberger et al. (2013) conducted a pilot experiment to discover potential novel biomarkers and phenotypes for diabetes and obesity by self-organized text mining of about 120,000 PubMed abstracts, public clinical trial summaries, and internal Merck research documents, with appropriate manual curation. Their approach showed benefit to discover biomarker, as well as the future impact on pharmaceutical research, such as shortening time-to-market of novel drugs, and speeding up early recognition of dead ends and adverse reactions.

Turenne et al. (2012) also utilized text mining approaches to extract transcription factors involved in bovine embryo development. They developed a model for the work which integrated information on different mammalian species from different literature and biology databases. They proposed 489 TF as potential participants of embryo proliferation and differentiation, with a recall of 95 % with regard to a biological gold standard defined in 2011 and an extension of more than 3 times the gold standard of TF detected so far in elongating tissues. Their work showed potential in applying to a wide range of biological processes.

The research work of target-specific drugs is less cost-efficient; on the other hand, high-throughput genomic technologies are incapable of deliver novel first-in-class drugs as expected. Some researchers tried to solve the crisis of blockbuster drug development by innovative bioinformatics approaches, such as text mining. The work by Epstein (2009) took advantage of text mining to get an optimal

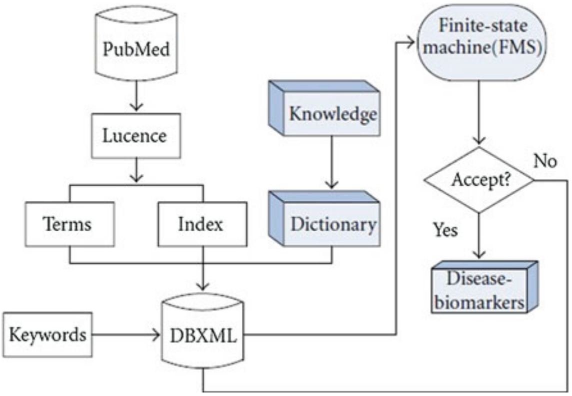


Fig. 4.3 The workflow of biomarker identification using text mining from PubMed (adopted from Li and Liu 2012)

clinical trial for new anticancer drugs. It is shown that text mining can refine therapeutic hypotheses and thus reduce empirical reliance on preclinical model development and early-phase clinical trials. Moreover, as personalised medicine evolves, this approach may inform biomarker-guided phase III trial strategies for noncytotoxic (antimetastatic) drugs that prolong patient survival without necessarily inducing tumor shrinkage.

Target discovery is a most crucial step in the biomarker and drug discovery pipeline to diagnose and fight human diseases, which can be grouped into two categories: a system approach and a molecular approach (Yang et al. 2012). Now that data mining of available biomedical data and information has greatly boosted target discovery in the omics era, it is time to develop efficient data mining methods to fuel target discovery in the post-genomics era. Text mining has been broadly applied to identify disease-associated entities and to understand their roles in diseases in identifying of disease-associated entities. Recently experts are dedicated to develop mining tools for extracting interaction networks related to human diseases from the literature in identifying disease-associated networks.

4.6 Data Sets and Tools for Biomedical Text Mining

4.6.1 Named Entity Recognition

There are many systems or tools for some biomedical named entity recognition, as listed in Table 4.1.

4.6.2 Synonym and Abbreviation Recognition

There are several synonym and abbreviation dictionaries and tools for biomedical entity, synonym and abbreviation recognition, as listed in Table 4.2. Biomedical scientists are able to use them for free.

4.6.3 Relation Extraction

There are many available tools collected from published literatures for biomedical scientists to use, as listed in Table 4.3.

Tables 4.4 and 4.5 list the standard datasets that have been manual/semi-automate annotate and curate. These datasets could be used to either evaluate the performance of named entity recognition system or to develop machine-learning based approach for entity recognition system.

Table 4.1 Some frequently used biomedical named entity recognition systems

System	Brief introduction
ABNER[a] (Settles 2005)	ABNER is a software tool for molecular biology text analysis. It uses linear-chain conditional random fields approach with orthographic and contextual features
GENIATagger[b] (Tsuruoka and Tsujii 2005)	The GENIA tagger is specifically tuned for biomedical text such as MEDLINE abstracts. It is a useful preprocessing tool for information extraction from biomedical documents
LingPipe[c] (Carpenter 2007, 2006)	LingPipe provides three generic, trainable chunkers to carry on named entity recognition. LingPipe can be used to identify biomedical entities such as genes, organisms, malignancies, and chemicals
Yapex[d] (Franzen et al. 2002)	Yapex is a rule-based system named entity recognition system that utilizes lexical and syntactic analysis to identify protein names

[a] ABNER. http://pages.cs.wisc.edu/~bsettles/abner/

[b] GENIATagger. http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi?page=GENIA+Tagger

[c] LingPipe. http://www.alias-i.com/lingpipe/

[d] Yapex. http://www.sics.se/humle/projects/prothalt/

Table 4.2 Standard annotated data sets for biomedical named entity recognition

Corpus name	Brief introduction
Acromine[a] (Okazaki and Ananiadou 2006)	The abbreviation dictionary of Acromine is automatically constructed from the whole MEDLINE. Acromine showed it was quite good then it was applied to the whole MEDLINE
BioLexicon (Thompson et al. 2011)	The BioLexicon brings together terminologies from several large public bioinformatics data resources such as UniProtKb, ChEBI and NCBI. The BioLexicon represents terms in conjunction with lexical and statistical information so as to improve performance of text mining
GENETAG[b] (Tanabe et al. 2005)	GENETAG is one of the most important standardized standard data sets for biomedical named entity recognition testing. It has 20,000 MEDLINE sentences for gene/protein term identification. 15,000 GENETAG sentences were used for the BioCreAtIvE Task 1A Competition
GO[c]	The Gene Ontology (GO) project is a major bioinformatics initiative aiming at standardizing the representation of gene and gene product. GO provides a controlled vocabulary of terms for describing gene product characteristics and gene product annotation data

[a] Acromine. http://www.nactem.ac.uk/software/acromine/

[b] GENETAG. //ftp.ncbi.nlm.nih.gov/pub/tanabe/

[c] GO. http://www.geneontology.org/

Reprinted with permission from Zhu et al. 2013

Table 4.3 Some useful tools for relationship extraction

System	Brief introduction
BCMS[a] (Leitner et al. 2008)	BioCreative MetaServer (BCMS) is a meta-service for information extraction which can generate annotations for PubMed/Medline abstracts, covering gene names, gene IDs, species, and protein–protein interactions
Chilibot[b] (Chen and Sharp 2004)	Chilibot searches PubMed abstracts about specific relationships between proteins, genes, or keywords. The results are returned as a graph
HPID[c] (Han et al. 2004)	The human protein interaction database (HPID) provides human protein interaction information from existing structural and experimental data, and integrated human protein interactions derived from BIND, DIP and HPRD. Users can find potential interaction between with input protein and proteins of the databases. The protein IDs in EMBL, Ensembl, MIM, RefSeq, HPRD and NCBI can be used during interaction search
HPRD[d] (Peri et al. 2003; Prasad 2009)	The Human Protein Reference Database (HPRD) is a platform for human protein interaction networks and disease association. All the information in HPRD has been manually extracted from the literature by experts for each in the proteome. HPRD can visually deploy the results
iHOP[e] (Hoffmann and Valencia 2004, 2005)	Information Hyperlinked over Proteins (iHOP) can generate a network of concurring genes and proteins from millions of PubMed abstracts. iHOP utilizes genes and proteins as hyperlinks between sentences and abstracts; hence the information can be converted into an integrated navigable resource
IntAct[f] (Kerrien et al. 2007)	IntAct provides analysis tools for molecular interaction as well as interaction database of which data were derived from literature curation or user submissions
MedScan[g] (Novichkova et al. 2003)	MedScan collected information and data retrieval from multiple sources of public information, text, journals, and various datasets, and then transformed into biological relationships which could be used for hypothesis generating and verification, disease understanding, drug and patient management
PubGene[h] (Jenssen et al. 2001)	The retrieve names of gene and protein by PubGene are cross-referenced to each other and to relevant terms with goal of understanding biological function, importance in disease and their relationship
Reactome[i] (Vastrik et al. 2007, 2009)	Reactome is an open-source data analysis tools, as well as a manually curated and peer-reviewed database including interaction, reaction and pathway data. Reactome can be used for interaction, reaction and pathway-based analysis

[a] BCMS. http://bcms.bioinfo.cnio.es/
[b] Chilibot. http://www.chilibot.net/
[c] HPID. http://wilab.inha.ac.kr/hpid/
[d] HPRD. http://www.hprd.org/
[e] iHOP. http://www.ihop-net.org/UniPub/iHOP/
[f] IntAct. http://www.ebi.ac.uk/intact/main.xhtml
[g] MedScan. http://www.ariadnegenomics.com/technology-research/medscan/
[h] PubGene. http://www.pubgene.org/
[i] Reactome. http://www.reactome.org/

Table 4.4 Some standard annotated data sets for relation extraction

Data set name	Brief introduction
BioInfer[a] (Pyysalo et al. 2007; Ginter et al. 2007)	BioInfer is a XML-based format corpus protein–protein interaction. The data of BioInfer were from five well-known protein–protein interaction corpora: AIMed, BioInfer, LLL, IEPA, and HPRD50
HIV-1, human PI[b] (Fu et al. 2009; Ptak et al. 2008; Pinney et al. 2009)	HIV-1 corpus contains summary of all known interactions of HIV-1 proteins with host cell proteins, other HIV-1 proteins, or proteins from disease organisms associated with HIV/AIDS
LLL 05[c]	The LLL05 is composed by annotation indicating agent and target of a gene interaction, a dictionary of named entities as well as variants and synonyms, and linguistic information. The LLL05 can be used to evaluate the ability of systems to identify gene/proteins interactions
PICorpus[d] (Johnson et al. 2007)	PICorpus is a protein–protein interaction corpus which was originally created at the PDG. PICorpus can be used for a variety of biomedical text mining tasks, such as named entity extraction, relation identification and relation extraction systems
PDZBase[e] (Beuming et al. 2005)	PDZBase contains 339 PDZ-domain mediated protein–protein interactions, which have been manually extracted. All the interactions are mediated directly by the PDZ-domain, and identified in vivo or in vitro experiments. The information of the binding-sites of interacting proteins are known
STRING[f] (Jensen et al. 2009)	STRING provides known and predicted protein interactions, including physical and functional associations derived from Genomic context, high-throughput experiments, coexpression and previous knowledge

[a] BioInfer. http://mars.cs.utu.fi/BioInfer/
[b] HIV-1ProteinInteraction. http://www.ncbi.nlm.nih.gov/RefSeq/HIVInteractions/index.html
[c] LLL05. http://genome.jouy.inra.fr/texte/LLLchallenge/
[d] PICorpus. http://bionlp-corpora.sourceforge.net/picorpus/index.shtml
[e] PDZBase. http://icb.med.cornell.edu/services/pdz/start
[f] STRING. http://string.embl.de/

4.7 Challenges and Future Work

Nowadays, the life sciences are rapidly revolutionized by highly developed throughput experimental methods, meanwhile the application of text mining technology in the frontier research in life science are accelerated with the widespread of the cloud computing application, therefore it is worth discussing topic

Table 4.5 Some commonly used standard annotated data sets for text mining

Data set name	Brief introduction
BioCreative III[a]	BioCreative III works for evaluating text mining and information extraction systems applied to the biomedical domain. BioCreative III has several data set for three tasks: cross-species gene identification and normalization, protein–protein interactions extraction, and interactive demonstration task for gene indexing and retrieval task
BioInfer[b] (Pyysalo et al. 2007; Ginter et al. 2007)	BioInfer is a XML-based format corpus protein–protein interaction. The data of BioInfer were from five well-known protein–protein interaction corpora: AIMed, BioInfer, LLL, IEPA, and HPRD50
BioText[c] (Rosario and Hearst 2004a,b, 2005a, b; Hearst and Rosario 2001; Schwartz and Hearst 2003)	BioText was initially constructed by 1,000 randomly selected MEDLINE abstracts from the results of a query on the term yeast. The dataset was then manually annotated and further verified. BioText has 954 correct pairs, including abbreviation definitions, protein–protein interaction data, and relations between disease treatment entities
GENIA[d] (Kim et al. 2003)	The GENIA data set is one of the most frequently used dataset for evaluation of biomedical and biological information extraction and text mining systems. The data set contains 1,999 Medline abstracts, selected using a PubMed query for terms human, blood cells, and transcription factors. The GENIA data set has many sub data set, aiming for part-of-Speech annotation, constituency (phrase structure) syntactic annotation, term annotation, event annotation, relation annotation, and coreference annotation
PICorpus[e] (Johnson et al. 2007)	PICorpus is a protein–protein interaction corpus which was originally created at the PDG. PICorpus can be used for a variety of biomedical text mining tasks, such as named entity extraction, relation identification and relation extraction systems

[a] BioCreAtIvE. http://www.pdg.cnb.uam.es/BioLINK/workshop_BioCreative_04/results/

[b] BioInfer. http://mars.cs.utu.fi/BioInfer/

[c] BioText. http://biotext.berkeley.edu/data.html

[d] GENIA. http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/geniaform.cgi

[e] PICorpus. http://bionlp-corpora.sourceforge.net/picorpus/index.shtml

for the work and challenges in the future application of text mining in cancer researches.

The first is the challenge of using biomedical text mining technologies in the personalized medicine development. We know that a complex disease such as cancer has many factors including race, gender, age and environment (Kountourakis et al. 2012; Chandolu and Dass 2012; Chlebowski et al. 2012; Foroughi et al. 2012; Wei and Giovannucci 2012; Hassanein et al. 2012; Hoffe and Balducci 2012). So it has become a trend that with patients' biomedical information collected and analyzed the medicine and the therapies are tailored to individual patients. By text mining technique, Ando et al. (2007) have identified qualitatively the differences in the focus of life review interviews by patients' age, gender, disease age and stage, and Jemal et al. (2011) integrated compound-target relationships which are related with cancer and presented the spectrum of research on personalized medicine and compound-target interactions. During text mining, all these important aspects will be taken into consideration by the personalized medicine (Mattila et al. 2012). One solution of it is before text mining categorizing data first rather than treat them together without any pre-processing, but it is a difficult work to do it. What's more, it's harder for text mining to find a good biomarker for all cases.

The second is that molecular mechanism of complex disease is very sophisticated since different genes or gene sets from the same pathway or network can cause the same phenotype. Hence mining texts from a hierarchical network instead of a single level is needed in order to study the complex mechanisms of cancer. Different levels which contain motif (Wang 2012; Chatterjee and Kumar 2011), pathway (Staiger et al. 2012; Giordano and Sinha 2012; Liu et al. 2012), module (Hjermstad et al. 2012; Chaudhry and Siddiqui 2012; Khoshnevisan et al. 2012) and network (Ramasubbu et al. 2012; Logue and Morrison 2012) are analyzed and studied in systems biomedicine. The resulting hierarchical data offer us valuable materials to conduct text mining. Nevertheless, it's really difficult to categorize texts to hierarchical network correctly, to integrate text mining results from different levels and discover new knowledge with a systems biomedicine view.

The third is to use the text mining techniques in translational medicine research which is an emerging field of biomedicine to involve the transformation of laboratory findings into novel diagnosis and treatment of patients (Azuaje et al. 2012). In order to improve treatment, we can apply the knowledge of pre-clinical in clinic. Translational medicine are used to facilitate the course of diseases predicting, preventing, diagnosing, and treating, in the research of which the calling for bioinformatics to act as a driver rather than a passenger requires text mining to do much more. Nevertheless, due to various stages of information and various sources of evidence taken into consideration and the integration of Omics and clinical data set to find out novel knowledge for both biology and medicine domains, biomedical text mining will face much difficulty. Recently discovered disease genes are confirmed and potential susceptibility genes are identified by numerous this sort of applications, such as the data integration and data mining platform presented by Liekens et al. (2011).

The forth is to understand more about complex disease by integrating at molecule, cell, tissue, organ, individual and even population levels. However, high levels which actually has a close relationship with cancer phenotypes are seldom focused on with most of the current text mining studies reporting at molecular level. Opportunities for successful disease diagnosis and treatments are really provided despite that it's a great challenge for cancer today to mine text at high levels and then integrate the text information of all these levels.

The last is that due to natural language text which contains ambiguities caused by semantics, slang and syntax and also suffered from noisy in texts often inconsistent, text mining can be a big challenge. Consequently, it's not suitable to be used blindly owing to too many errors contained in the mined information. Fortunately there is some solutions. The first is actually pre-processing turns the unstructured texts to structured texts with semantic tags by manually reading and understanding texts, analyzing them, and then adding semantic tags, so it can easily realize the aim with high precision rate. Nevertheless, it's a truly restricted solution for requiring vast human efforts and time-consuming, which only limits mining ability. The second is to provide vast biomedical texts on which carrying on text mining, after that analyzing the candidate results and screening out the final results. In order to enhance mining efficiency and the quality of the mined knowledge, we usually employ domain knowledge during the mining process. Compared with the first solution, this is more powerful on knowledge discovering despite that the mined results may still contain more errors. These two solutions are different on dealing with the correctness of the texts to be mined, for the formal is by carefully manual pre-processing while the latter is by post-processing by experts. And the third is to use some advanced statistical analysis to clean data roughly and after that to conduct mining on them. It is a compromising solution of the advanced two approaches.

4.8 Conclusions

Currently, the existing huge body of biomedical texts and their rapid growth makes it impossible for researchers to process the information manually. Researchers can use biomedical text mining to facilitate their work. We have reviewed the important research issues related to text mining in the biomedical fields. We have also provided a review of the state-of-the-art applications and datasets used for text mining in biomarker discovery, thereby providing researchers with the necessary resources to apply or develop text mining tools in their research. We introduced the general workflow of text mining to support systems biology and we illustrated each phase in detail. We can see that text mining has been used widely. However, to fully utilize text mining, it is still necessary to develop new methods for full text mining and for highly complex texts, as well as platforms for integrating other biomedical knowledge bases.

References

Abacha AB, Zweigenbaum P. Automatic extraction of semantic relations between medical entities: a rule based approach. J Biomed Seman. 2011;2(Suppl 5):S4.

Agarwal S, Liu F, Yu H. Simple and efficient machine learning frameworks for identifying protein–protein interaction relevant articles and experimental methods used to study the interactions. BMC Bioinform. 2011;12(Suppl 8):S10.

Ai J, Smith B, Wong DT. Saliva Ontology: an ontology-based framework for a Salivaomics knowledge base. BMC Bioinform. 2010;11:302.

Alexopoulos LG, et al. Construction of signaling pathways and identification of drug effects on the liver cancer cell HepG2. Conf Proc IEEE Eng Med Biol Soc. 2010;2010:6717–20.

Ando M, Morita T, O'Connor SJ. Primary concerns of advanced cancer patients identified through the structured life review process: a qualitative study using a text mining technique. Palliat Support Care. 2007;5(3):265–71.

Arighi CN, et al. Overview of the BioCreative III workshop. BMC Bioinform. 2011;12(Suppl 8):S1.

Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc. 2010;17(3):229–36.

Azuaje FJ, et al. Bioinformatics as a driver, not a passenger, of translational biomedical research: perspectives from the 6th Benelux bioinformatics conference. J Clin Bioinform. 2012;2:7.

Beuming T, et al. PDZBase: a protein–protein interaction database for PDZ-domains. Bioinformatics. 2005;21(6):827–8.

Carpenter B. Character language models for Chinese word segmentation and named entity recognition. 2006.

Carpenter B. LingPipe for 99.99 % recall of gene mentions. 2007.

Chandolu V, Dass CR. Cell and molecular biology underpinning the effects of PEDF on cancers in general and osteosarcoma in particular. J Biomed Biotechnol. 2012;2012:740295.

Chang Y-C, Tsai RTH, Hsu W-L. New challenges for biological text-mining in the next decade. J Comput Sci Technol. 2010;25:169–79.

Chatterjee S, Kumar D. Unraveling the design principle for motif organization in signaling networks. PLoS ONE. 2011;6(12):e28606.

Chaudhry Z, Siddiqui S. Health related quality of life assessment in Pakistani paediatric cancer patients using PedsQLTM 4.0 generic core scale and PedsQLTM cancer module. Health Qual Life Outcomes. 2012;10(1):52.

Chen H, Sharp BM. Content-rich biological network constructed by mining PubMed abstracts. BMC Bioinform. 2004;5:147.

Chlebowski RT, et al. Diabetes, metformin, and breast cancer in postmenopausal women. J Clin Oncol. 2012.

Chun HW, et al. Extraction of gene–disease relations from Medline using domain dictionaries and machine learning. 2006. (Citeseer).

Cohen AM, Hersh WR. A survey of current work in biomedical text mining. Brief Bioinform. 2005a;6(1):57–71.

Cohen AM, Hersh WR. A survey of current work in biomedical text mining. Brief Bioinform. 2005b;6(1):57–71.

Dagar, A, et al. Epilepsy surgery in a pediatric population: a retrospective study of 129 children from a tertiary care hospital in a developing country along with assessment of quality of life. Pediatr Neurosurg. 2011.

Ephraim Y, Merhav N. Hidden markov processes. IEEE Trans Inform Theory. 2002;48(6): 1518–69.

Epstein RJ. Unblocking blockbusters: using boolean text-mining to optimise clinical trial design and timeline for novel anticancer drugs. Cancer Inform. 2009;7:231–8.

Eskin E, Agichtein E. Combining text mining and sequence analysis to discover protein functional regions. Pac Symp Biocomput. 2004;288–99.

Foroughi F, Saadat N, Salehian MT. Encapsulated insular carcinoma of the thyroid arising in Graves' disease: report of a case and review of the literature. Int J Surg Pathol. 2012.

Franzen K, et al. Protein names and how to find them. Int J Med Inform. 2002;67(1–3):49–61.

Frawley WJ, Piatetsky-Shapiro G, Matheus CJ. Knowledge discovery in databases: an overview. AI Mag. 1992;13:57–70.

Fu W, et al. Human immunodeficiency virus type 1, human protein interaction database at NCBI. Nucleic Acids Res. 2009;37(Database issue):D417–22.

Garten Y, Coulet A, Altman RB. Recent progress in automatically extracting information from the pharmacogenomic literature. Pharmacogenomics. 2010;11(10):1467–89.

Ginter F, et al. BioInfer relationship annotation manual. 2007.

Giordano CN, Sinha AA. Cytokine networks in Pemphigus vulgaris: an integrated viewpoint. Autoimmunity. 2012.

Habib MS, Kalita J. Scalable biomedical named entity recognition: investigation of a database-supported SVM approach. Int J Bioinform Res Appl. 2010;6(2):191–208.

Han K, et al. HPID: the human protein interaction database. Bioinformatics. 2004;20(15):2466–70.

Hanisch D, et al. ProMiner: rule-based protein and gene entity recognition. BMC Bioinform. 2005;6(Suppl 1):S14.

Hassanein M, et al. The state of molecular biomarkers for the early detection of lung cancer. Cancer Prev Res (Phila). 2012.

Hayasaka S, Hugenschmidt CE, Laurienti PJ. A network of genes, genetic disorders, and brain areas. PLoS ONE. 2011;6(6):e20907.

He Y, Kayaalp M. Biological entity recognition with conditional random fields. AMIA Annu Symp Proc. 2008;293–7.

Hearst MA, Rosario B. Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In: Proceedings of 2001 conference on empirical methods in natural language processing (EMNLP 2001). Pittsburgh, PA; 2001.

Hettne KM, et al. Automatic mining of the literature to generate new hypotheses for the possible link between periodontitis and atherosclerosis: lipopolysaccharide as a case study. J Clin Periodontol. 2007;34(12):1016–24.

Hjermstad MJ, et al. The EORTC QLQ-OH17: a supplementary module to the EORTC QLQ-C30 for assessment of oral health and quality of life in cancer patients. Eur J Cancer. 2012.

Hoffe S, Balducci L. Cancer and age: general considerations. Clin Geriatr Med. 2012;28(1):1–18.

Hoffmann R, Valencia A. A gene network for navigating the literature. Nat Genet. 2004;36(7):664.

Hoffmann R, Valencia A. Implementing the iHOP concept for navigation of biomedical literature. Bioinformatics. 2005;21(Suppl 2):ii252–8.

Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA Cancer J Clin. 2011;61(2):69–90.

Jensen LJ, et al. STRING 8—a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res. 2009;37(Database issue):D412–6.

Jenssen TK, et al. A literature network of human genes for high-throughput analysis of gene expression. Nat Genet. 2001;28(1):21–8.

Johnson HL, et al. Corpus refactoring: a feasibility study. J Biomed Discov Collab. 2007;2:4.

Kazama J, Makino T, Ohta Y, Tsujii J. Tuning support vector machines for biomedical named entity recognition. In: Association for computational linguistics. NJ, USA; 2002.

Kerrien S, et al. IntAct—open source resource for molecular interaction data. Nucleic Acids Res. 2007;35(Database issue):D561–5.

Khoshnevisan A, et al. Translation and validation of the EORTC brain cancer module (EORTC QLQ-BN20) for use in Iran. Health Qual Life Outcomes. 2012;10(1):54.

Kim JD, et al. GENIA corpus—semantically annotated corpus for bio-textmining. Bioinformatics. 2003;19(Suppl 1):i180–2.

Korhonen A, et al. Text mining for literature review and knowledge discovery in cancer risk assessment and research. PLoS ONE. 2012;7(4):e33427.

Kountourakis P, et al. Barrett's esophagus: a review of biology and therapeutic approaches. Gastrointest Cancer Res. 2012;5(2):49–57.

Krallinger M, et al. The protein–protein interaction tasks of BioCreative III: classification/ranking of articles and linking bio-ontology concepts to full text. BMC Bioinform. 2011;12(Suppl 8):S3.

Leitner F, et al. Introducing meta-services for biomedical information extraction. Genome Biol. 2008;9(Suppl 2):S6.

Leser U, Hakenberg J. What makes a gene name? Named entity recognition in the biomedical literature. Brief Bioinform. 2005;6(4):357–69.

Li H, Liu C. Biomarker identification using text mining. Comput Math Methods Med. 2012;2012:135780.

Li L, Zhou R, Huang D. Two-phase biomedical named entity recognition using CRFs. Comput Biol Chem. 2009a;33(4):334–8.

Li J, Zhu X, Chen JY. Building disease-specific drug-protein connectivity maps from molecular interaction networks and PubMed abstracts. PLoS Comput Biol. 2009b;5(7):e1000450.

Li X, et al. A mouse protein interactome through combined literature mining with multiple sources of interaction evidence. Amino Acids. 2010;38(4):1237–52.

Liekens AM, et al. BioGraph: unsupervised biomedical knowledge discovery via automated hypothesis generation. Genome Biol. 2011;12(6):R57.

Lin YF. BIOKDD04: 4th workshop on data mining in bioinformatics (with SIGKDD conference). In: A maximum entropy approach to biomedical named entity recognition; 2004.

Liu KQ, et al. Identifying dysregulated pathways in cancers from pathway interaction networks. BMC Bioinform. 2012;13(1):126.

Logue JS, Morrison DK. Complexity in the signaling network: insights from the use of targeted inhibitors in cancer therapy. Genes Dev. 2012;26(7):641–50.

Macilwain C. Systems biology: evolving into the mainstream. Cell. 2011;144(6):839–41.

Mack R, Hehenberger M. Text-based knowledge discovery: search and mining of life-sciences documents. Drug Discov Today. 2012;7:89–98.

Matos S, et al. Concept-based query expansion for retrieving gene related publications from MEDLINE. BMC Bioinform. 2010;11:212.

Mattila J, et al. Design and application of a generic clinical decision support system for multiscale data. IEEE Trans Biomed Eng. 2012;59(1):234–40.

McEntyre J, Lipman D. PubMed: bridging the information gap. CMAJ. 2001;164(9):1317–9.

Nam S, Park T. Pathway-based evaluation in early onset colorectal cancer suggests focal adhesion and immunosuppression along with epithelial-mesenchymal transition. PLoS One. 2012;7.

Novichkova S, Egorov S, Daraselia N. MedScan, a natural language processing engine for MEDLINE abstracts. Bioinformatics. 2003;19(13):1699–706.

Okazaki N, Ananiadou S. Building an abbreviation dictionary using a term recognition approach. Bioinformatics. 2006;22(24):3089–95.

Papp B, Notebaart RA, Pal C. Systems-biology approaches for predicting genomic evolution. Nat Rev Genet. 2011;12(9):591–602.

Peri S, et al. Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res. 2003;13(10):2363–71.

Pinney JW, et al. HIV-host interactions: a map of viral perturbation of the host system. AIDS. 2009;23(5):549–54.

Prasad TSK, et al. Human protein reference database—2009 update. Nucleic Acids Res. 2009;37(Database issue):D767–72.

Ptak RG, et al. Cataloguing the HIV type 1 human protein interaction network. AIDS Res Hum Retroviruses. 2008;24(12):1497–502.

Pyysalo S, et al. BioInfer: a corpus for information extraction in the biomedical domain. BMC Bioinform. 2007;8:50.

Qabaja A, Alshalalfa M, Bismar TA, Alhajj R. Protein network-based Lasso regression model for the construction of disease-miRNA functional interactions. EURASIP J Bioinform Syst Biol. 2013;1:3.

Ramasubbu R, et al. The Canadian network for mood and anxiety treatments (CANMAT) task force recommendations for the management of patients with mood disorders and select comorbid medical conditions. Ann Clin Psychiatry. 2012;24(1):91–109.

Raychaudhuri S, Altman RB. A literature-based method for assessing the functional coherence of a gene group. Bioinformatics. 2003;19(3):396–401.

Raychaudhuri S, Schutze H, Altman RB. Using text analysis to identify functionally coherent gene groups. Genome Res. 2002;12(10):1582–90.

Rebholz-Schuhmann D, et al. Assessment of NER solutions against the first and second CALBC silver standard corpus. J Biomed Seman. 2011;2(Suppl 5):S11.

Rosario B, Hearst MA. Multi-way relation classification: application to protein–protein interactions. 2005.

Rosario B, Hearst MA. Classifying semantic relations in bioscience text. In: Proceedings of the 42nd annual meeting of the association for computational linguistics (ACL 2004). Barcelona; 2004.

Rosario B, Hearst MA. Classifying semantic relations in bioscience texts. 2004.

Rosario B, Hearst MA. Multi-way relation classification: application to protein–protein interaction. In: HLT-NAACL'05. Vancouver; 2005.

Sasaki Y, et al. How to make the most of NE dictionaries in statistical NER. BMC Bioinform. 2008;9(Suppl 11):S5.

Schwartz AS, Hearst MA. A simple algorithm for identifying abbreviation definitions in biomedical text. Pac Symp Biocomput. 2003;451–62.

Settles B. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. Bioinformatics. 2005;21(14):3191–2.

Sharma P, et al. Mining literature for a comprehensive pathway analysis: a case study for retrieval of homocysteine related genes for genetic and epigenetic studies. Lipids Health Dis. 2006;5:1.

Staiger C, et al. A critical evaluation of network and pathway-based classifiers for outcome prediction in breast cancer. PLoS ONE. 2012;7(4):e34796.

Swanson DR. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspect Biol Med. 1986;30:7–18.

Tanabe L, et al. GENETAG: a tagged corpus for gene/protein named entity recognition. BMC Bioinform. 2005;6(Suppl 1):S3.

Thompson P, et al. The BioLexicon: a large-scale terminological resource for biomedical text mining. BMC Bioinform. 2011;12:397.

Topinka CM, Shyu C. Predicting cancer interaction networks using text-mining and structure understanding. In: AMIA annual symposium proceeding. 2006.

Trugenberger CA, et al. Discovery of novel biomarkers and phenotypes by semantic technologies. BMC Bioinform. 2013;14(51):51.

Tsai FS. Text mining and visualisation of protein–protein interactions. Int J Comput Biol Drug Des. 2011;4(3):239–44.

Tsai T, et al. Integrating linguistic knowledge into a conditional random fieldframework to identify biomedical named entities. Expert Syst Appl. 2006;30(1):117–28.

Tsuruoka Y, Tsujii J. Bidirectional inference with the easiest-first strategy for tagging sequence data. In: Association for computational linguistics Morristown, NJ, USA; 2005.

Turenne N, Tiys E, Ivanisenko V, Yudin N, Ignatieva E, Valour D, Degrelle SA, Hue I. Finding biomarkers in non-model species: literature mining of transcription factors involved in bovine embryo development. BioData Min. 2012;5(12):1–12.

Urzua U, Owens G, Zhang GM, Cherry JM, Sharp JJ. Tumor and reproductive traits are linked by RNA metabolism genes in the mouse ovary: a transcriptome-phenotype association analysis. BMC Genomics. 2010;11.

Vastrik I, et al. Reactome: a knowledge base of biologic pathways and processes. Genome Biol. 2007;8(3):R39.

Vastrik I, et al. Correction: Reactome: a knowledge base of biologic pathways and processes. Genome Biol. 2009;10(2):402.

Wang B. BRCA1 tumor suppressor network: focusing on its tail. Cell Biosci. 2012;2(1):6.

Wei MY, Giovannucci EL. Lycopene, tomato products, and prostate cancer incidence: a review and reassessment in the PSA screening era. J Oncol. 2012;2012:271063.

Wren JD, Garner HR. Shared relationship analysis: ranking set cohesion and commonalities within a literature-derived relationship network. Bioinformatics. 2004;20(2):191–8.

Yang Y, Adelstein S, Kassis AI. Target discovery from data mining approaches. Drug Discov Today. 2012;17.

Zhou GD, Su J. Exploring deep knowledge resources in biomedical name recognition. In: JNLPBA; 2004.

Zhu F, Shen B. Combined SVM-CRFs for biological named entity recognition with maximal bidirectional squeezing. PLoS ONE. 2012;7(8):1–8.

Zhu F, Patumcharoenpol P, Zhang C, Yang Y, Chan J, Meechai A, Vongsangnak W, Shen B. Biomedical text mining and its applications in cancer research. J Biomed Inform. 2013;46(2):200–11.

Chapter 5
Protein Binding Interfaces and Their Binding Hot Spot Prediction: A Survey

Qian Liu and Jinyan Li

Abstract In living organisms, genes are the blueprints or library, specifying instructions for building proteins. Proteins constitute the bulk of cells. Proteins mutual binding and interactions play a vital role in numerous functions and activities, such as signal transduction, enzymatic reactions, immunoreactions and inter-cellular communications. This survey provides basic knowledge of proteins and protein binding. First, we describe proteins' fundamental elements, structures and functions. In Sect. 5.2, we present concepts related to protein binding and interactions. In Sect. 5.3, we explain why protein binding interfaces have a uneven distribution of binding free energy. In the Sects. 5.4 and 5.5, we explain why protein interfaces are complicated and how the current studies deal with this difficult problem. In Sect. 5.6, we present an overview on methods to model and predict binding free energy of protein interactions. Section 5.7 concludes this survey with a summary.

Keywords Protein structures · Protein–protein binding · Binding interface · Binding hot spots

5.1 Proteins: An Elementary Introduction

The basic building blocks of every protein are named amino acids. There are 20 types of amino acids. Each amino acid consists of two parts: a backbone and a side chain. The backbone consists of three groups: the amino group (NH_2), a central carbon (the alpha carbon or CA) and a carboxyl group (COOH). The backbone is the same for all of the 20 types of amino acids. Side chains have different combinations of heavy atoms, each corresponding to one of the 20 types of amino

Q. Liu · J. Li (✉)
Advanced Analytics Institute and Center for Health Technologies,
University of Technology Sydney, Broadway NSW 2007, Australia
e-mail: jinyan.li@uts.edu.au

B. Shen (ed.), *Bioinformatics for Diagnosis, Prognosis and Treatment of Complex Diseases*, Translational Bioinformatics 4, DOI: 10.1007/978-94-007-7975-4_5,
© Springer Science+Business Media Dordrecht 2013

acids. Heavy atoms are those atoms such as Carbon (C), Nitrogen (N), Oxygen (O) and Sulfur (S).

The twenty amino acids are shown in Fig. 5.1. Their 3-letter (1-letter) codes are: Ile(I), Val(V), Leu(L), Phe(F), Cys(C), Met(M), Ala(A), Gly(G), Thr(T), Ser(S), Trp(W), Tyr(Y), Pro(P), His(H), Glu(E), Gln(Q), Asp(D), Asn(N), Lys(K), and Arg(R) in the order with I as the most hydrophobic and R as the least hydrophobic with respect to their hydrophobicity (Kyte and Doolittle 1982).

Gly is the smallest amino acid without a side chain of any heavy atom. Ala is the second smallest amino acid and has only one heavy carbon atom CB (or the beta carbon) in its side chain; CB is the first heavy atom of side chains which has a covalent bond with CA. The side chains of the other amino acids have more number of heavy atoms. Please refer to Fig. 5.1 for more details.

Because of different side chains, the twenty amino acids possess various physicochemical properties. For example, the side chains from Arg, His and Lys are positively charged, while the side chains in Asp and Glu are negatively charged; Ser, Thr, Asn and Gln have polar side chains, but Ala, Ile, Leu, Met, Phe, Trp, Tyr and Val have hydrophobic side chains and especially the side chains in Phe, Trp and Tyr contain aromatic rings. Hydrophobic amino acids generally tend not to contact with water molecules, while hydrophilic (polar and charged) amino acids prefer a high affinity for water.

5.1.1 A Definition for Proteins

In a real-cell environment, the amino NH_2 group of an amino acid can react with the carboxyl COOH group of another amino acid, forming an amino covalent bond and resulting in the release of a water molecule; this covalent bond is also called peptide bond in biology. After this condensation reaction, the involving amino acids are also referred to as (amino acid) residues. A set of residues create a linear-sequence polymer, called a peptide for a shorter or a protein for a longer polymer (the left-side of Fig. 5.2).

In vivo with a three-dimensional (3D) space, a residue of a protein can interact with some other residues under physical, chemical and biological rules, producing a 3D structure. One example of protein 3D structures is shown in the right-side sub-figure in Fig. 5.2, whose protein sequences are in the left-side sub-figure. Generally, a protein sequence determines a unique, stable, intended and correct 3D structure (also called native conformation). It is also widely accepted that similar sequences have similar 3D structures, but similar 3D structures may have different protein sequences.

Proteins, structures can be characterized from the following four aspects: protein primary structures, secondary structures, tertiary structures and quaternary structures. Primary structures are protein sequences themselves, while the others represent different levels of protein 3D structures. In protein 3D structures, some segments have such favored regular shapes as alpha helix or beta sheet stabilized by hydrogen

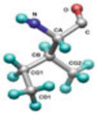
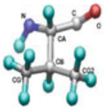
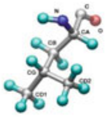
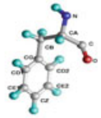
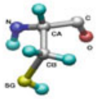
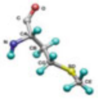
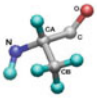
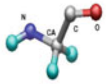
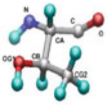
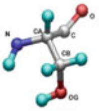
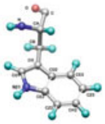
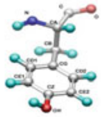
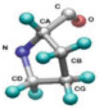
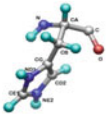
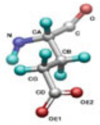
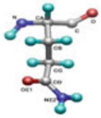
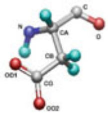
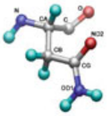
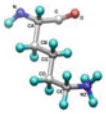
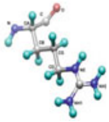
	Amino Acids			
Full name	Isoleucine	Valine	Leucine	Phenylalanine
3-letter name	Ile	Val	Leu	Phe
1-letter name	I	V	L	F
Hydrophobicity	4.5	4.2	3.8	2.8
Structure				
Full name	Cysteine	Methionine	Alanine	Glycine
3-letter name	Cys	Met	Ala	Gly
1-letter name	C	M	A	G
Hydrophobicity	2.5	1.9	1.8	-0.4
Structure				
Full name	Threonine	Serine	Tryptophan	Tyrosine
3-letter name	Thr	Ser	Trp	Tyr
1-letter name	T	S	W	Y
Hydrophobicity	-0.7	-0.8	-0.9	-1.3
Structure				
Full name	Proline	Histidine	Glutamic acid	Glutamine
3-letter name	Pro	His	Glu	Gln
1-letter name	P	H	E	Q
Hydrophobicity	-1.6	-3.2	-3.5	-3.5
Structure				
Full name	Aspartic acid	Asparagine	Lysine	Arginine
3-letter name	Asp	Asn	Lys	Arg
1-letter name	D	N	K	R
Hydrophobicity	-3.5	-3.5	-3.9	-4.5
Structure				

Fig. 5.1 Twenty standard amino acids, and their structures and hydrophobicity. Here, the structures are from http://share.chuagh.net/science/Share_AminoAcids.php?structure=1 and the hydrophobicity values are from http://www.cgl.ucsf.edu/chimera/docs/UsersGuide/midas/hydrophob.html

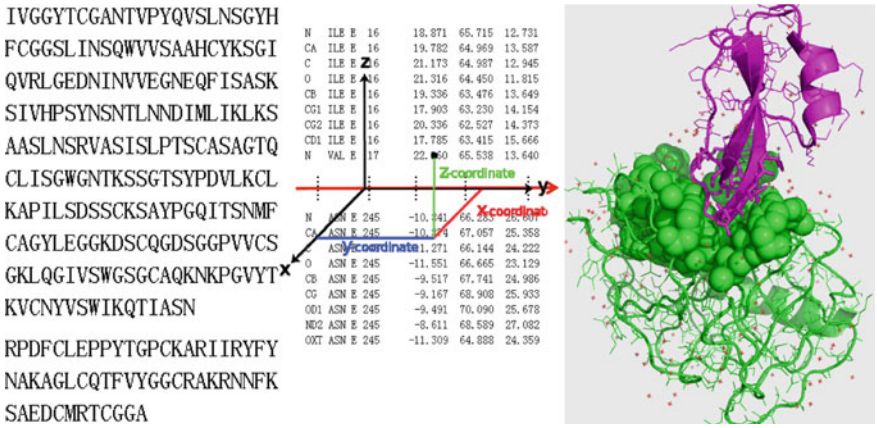

IVGGYTCGANTVPYQVSLNSGYH
FCGGSLINSQWVVSAAHCYKSGI
QVRLGEDNINVVEGNEQFISASK
SIVHPSYNSNTLNNDIMLIKLKS
AASLNSRVASISLPTSCASAGTQ
CLISGWGNTKSSGTSYPDVLKCL
KAPILSDSSCKSAYPGQITSNMF
CAGYLEGGKDSCQGDSGGPVVCS
GKLQGIVSWGSGCAQKNKPGVYT
KVCNYVSWIKQTIASN
RPDFCLEPPYTGPCKARIIRYFY
NAKAGLCQTFVYGGCRAKRNNFK
SAEDCMRTCGGA
N ILE E 16 18.871 65.715 12.731
CA ILE E 16 19.782 64.969 13.587
C ILE E 16 21.173 64.987 12.945
O ILE E 16 21.316 64.450 11.815
CB ILE E 16 19.336 63.476 13.649
CG1 ILE E 16 17.903 63.230 14.154
CG2 ILE E 16 20.336 62.527 14.373
CD1 ILE E 16 17.785 63.415 15.666
N VAL E 17 22. 60 65.538 13.640
z
y
x
Z-coordinate
X-coordinat
Y-coordinate
N ASN E 245 -10. 41 66.
CA ASN E 245 -10. 4 67.057 25.358
ASN 1.271 66.144 24.222
O ASN E 245 -11.551 66.665 23.129
CB ASN E 245 -9.517 67.741 24.986
CG ASN E 245 -9.167 68.908 25.933
OD1 ASN E 245 -9.491 70.090 25.678
ND2 ASN E 245 -8.611 68.589 27.082
OXT ASN E 245 -11.309 64.888 24.359

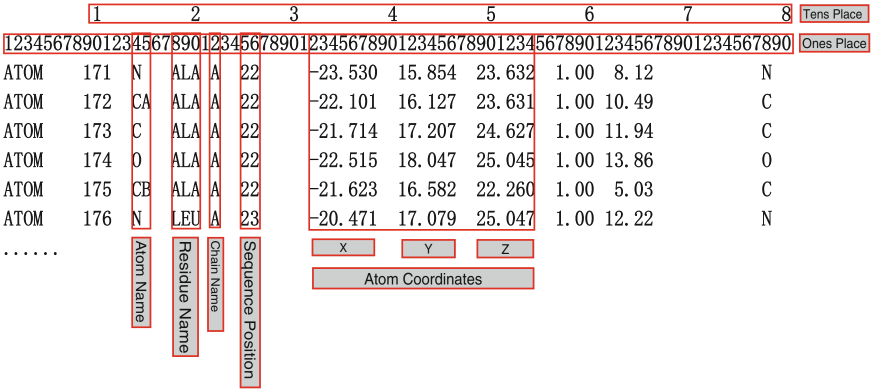
Fig. 5.2 An example of protein sequences and their 3D structures. The *middle* sub-figure represents 3D coordinates of each atom in proteins. In the *right-side* sub-figure, one protein is in *magenta* while the other is in *green*; the binding interface is in a *sphere* view for the *green* protein and in a *stick* view for the *magenta*

bonds, while others shapes are irregular, specially termed as loop/turn/random coil. These local shapes are known as secondary structures of proteins. Further, a protein 3D conformation is defined as a tertiary structure if a single protein is involved, and as a quaternary structure when the conformation is about a protein complex. A protein complex is the binding or contact of two or more proteins by non-covalent bonds. Each protein in its quaternary structures is also referred to as a protein chain or a proteomer. Please refer to the right-side sub-figure in Fig. 5.2 for an example of protein secondary structures, tertiary structures or quaternary structures.

The native 3D conformations of proteins determine their functions. Failure to fold into intended shapes usually produces inactive proteins with different properties, such as toxic prions (please refer to http://en.wikipedia.org/wiki/Protein_folding). This failure may result in many diseases, including Creutzfeldt-Jakob disease, bovine spongiform encephalopathy (mad cow disease), amyloid-related illnesses such as Alzheimer's Disease and familial amyloid cardiomyopathy or polyneuropathy, as well as intracytoplasmic aggregation diseases such as Huntington's and Parkinson's disease (please refer to http://en.wikipedia.org/wiki/Protein_folding). Therefore, it is vital to determine and understand protein 3D structures.

5.1.2 Database of Protein 3D Structural Data

Many experimental methods have been developed to determine protein 3D conformations, including X-ray crystallography, electron microscopy and Nuclear magnetic resonance (NMR) spectroscopy. Among these techniques, X-ray crystallography is the most popular and prolific technique. X-ray crystallography



```
         1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
ATOM    171  N   ALA A  22     -23.530  15.854  23.632  1.00  8.12           N
ATOM    172  CA  ALA A  22     -22.101  16.127  23.631  1.00 10.49           C
ATOM    173  C   ALA A  22     -21.714  17.207  24.627  1.00 11.94           C
ATOM    174  O   ALA A  22     -22.515  18.047  25.045  1.00 13.86           O
ATOM    175  CB  ALA A  22     -21.623  16.582  22.260  1.00  5.03           C
ATOM    176  N   LEU A  23     -20.471  17.079  25.047  1.00 12.22           N
......
```

Fig. 5.3 An example of atomic coordinates in PDB

(please refer to http://en.wikipedia.org/wiki/X-ray_crystallography) uses the following way to determine the atomic arrangement of proteins within a crystal. Firstly, according to elastic scattering, X-ray crystallography employs X-ray to produce a diffraction pattern of regularly spaced spots called a reflection. Then, several reflections are obtained through rotating the crystal. Finally, the mathematical method of Fourier transforms is used to convert the two-dimensional reflection images taken at different rotations into a three-dimensional model of the density of electrons within the crystal. According to electron density, atomic mean positions in a crystal can be determined. X-ray crystallography is relatively affordable, and can obtain high-resolution information of atoms in proteins. So far, it has resolved 66,989 crystal structures of biological molecules according to the statistics of Protein Data Bank (PDB).

PDB is a key resource to store 3D structural data of those biological molecules, such as proteins and nucleic acids. PDB can be accessed by URL http://www.rcsb.org/. In each PDB entry of proteins, heavy atoms of each residue are stored with their own 3D coordinates as shown in Fig. 5.3. Based on these atomic coordinates, protein 3D conformations can be visualized. For example, the right-side sub-figure in Fig. 5.2 is produced according to the coordinates in the middle sub-figure of Fig. 5.2. Therefore, PDB structures are primary information to investigate proteins and their structures; several important databases already derive and store the classification of PDB structural data, such as SCOP (Murzin et al. 1995) and CATH (Orengo et al.1997).

5.2 Protein Binding Interfaces

A living organism is a dynamic integral world. Its components contact each other for performing biological functions. As a fundamental component of living organisms, proteins rarely function alone; most proteins must work in collaboration with other macromolecules, such as other proteins or DNA. Without doubt, close interactions of proteins are indispensable to fulfill molecular functions and biological processes. A good example of this is antigen–antibody interactions. When foreign objects antigens intrude, antibodies can specially identify and neutralize antigens so that immune systems can well defend the body against the attack from antigens. In this process, the antigen–antibody close interactions are essential to immune systems.

Meanwhile, protein interactions are very specific: on one hand, not all segments of proteins are able to bind to other molecules, although a protein can have more than one binding segment; on the other hand, a binding segment of a protein should only contact with specific binding segments in some other proteins instead of with all binding segments or all other proteins—these specific binding segments should have a certain corresponding characteristic.

Close and specific protein interactions play a vital and fundamental role in molecular functions and biological processes, but why binding happens like this and what are binding segments are unknown yet. Thus, discovering the principle of protein interactions and the specificity of protein binding is a fundamental and challenging problem in proteomics, resulting in a lot of useful applications, such as drug design and protein engineering.

5.2.1 Diversity of Protein Interactions

Protein interactions are diverse according to various criteria. For example, some interactions are permanent and usually very stable. Thus once forming protein complexes, they only exist in their complex form in all of the lifetime of the complexes (Nooren and Thornton 2003). In contrast, others associate to accomplish a particular function upon a molecular stimulus and dissociate after that (Nooren and Thornton 2003). This kind of interactions are termed as transient interactions in comparison to permanent interactions.

Biological interactions can also be grouped depending on whether their protomers of interactions can be found or not as stable tertiary structures on their own in vivo (Irene 2003). On certain physiological conditions and environments, if protomers of interactions cannot be found as stable tertiary structures on their own in vivo (Irene 2003), this kind of interactions are referred to as obligate interactions; otherwise, those interactions are non-obligate interactions. Similarly, two-state folding complexes and three-state complexes were also used to describe obligate and non-obligate interactions; this criterion depends on the different transition processes in protein folding and binding (Tsai and Nussinov 1998): (1) in non-obligate interactions, two protomers of interactions fold separately, and then recognize each other to form the complexes. Thus, this process is three-state; (2) in contrast, two protomers fold and bind together without the intermediate states in obligate interactions, and these interactions are two-state complexes (Tsai and Nussinov 1998) or obligomers (Ofran and Rost 2003).

The criterion for obligate/two-state and non-obligate/three-state complexes is similar to and confused with, but really different from, the criterion for permanent and transient interactions: (1) obligate interactions or two-state complexes are stable, and their protein chains function only in the complex form (Irene 2003). So, this kind of interactions are permanent interactions. (2) Correspondingly, protomers in transient interactions fold separately, and they have stable tertiary structures; thus, transient interactions are non-obligate. (3) Clearly, there exist non-obligate permanent complexes, such as antigen–antibody interactions, and the complexes of thrombin and rodniin inhibitor (Irene 2003). In these complexes, each protomer fold into a stable tertiary structure, and then the protomers come together to form protein complexes which do not dissociate again. However, the stability of complexes much depends on physiological conditions or environments: a continuum exists between obligate and non-obligate interactions, while an interaction may be mainly transient in vivo but become permanent under certain cellular conditions (Irene 2003). Thus, these two criteria are mostly combined in literature works, although they focus on different aspects of protein complexes.

Meanwhile, protein interactions can happen between two identical protein chains and called homodimers (Irene 2003); otherwise, the interactions are called heterodimers (Irene 2003). Homodimers are generally obligate or permanent interactions, while most of heterodimers are non-obligate or transient interactions. Furthermore in a more complicated way, protein interactions can be categorized into six subtypes (Ofran and Rost 2003): intra-domain, domain–domain, homo-complexes (interfaces of transient interactions between identical protein chains), homo-obligomers (interfaces of permanent interactions between identical protein chains), hetero-obligomers (interfaces of permanent interactions between different protein chains) and hetero-complexes (interfaces of transient interactions between different protein chains). Here, a domain is a consecutive segment of protein sequences with a particular and repeatable three-dimensional structure; it may evolve, function, and exist independently of the rest of protein chains.

It is clear that these different types of biological interactions possess their unique binding behaviors. This diversity of protein interactions necessitates individual investigation for each different type of protein interactions based on their quaternary structures.

5.2.2 Binding Interfaces in Protein Interactions

In quaternary structures of protein interactions, physicochemical properties are not uniform everywhere, such as evolutionary conservation, and solvent accessible surface area (ASA). According to the difference of solvent accessible surface area, protein quaternary structures can be divided into several segments, such as interfaces, surfaces, and intraproteins/interior. Intraproteins can further be subdivided into domain interfaces and intradomains according to evolutionary/structural conservation or functional properties. Among these segments of protein quaternary structures, interfaces are significantly important.

In a protein complex, the region where protein chains come into contact is a binding site; or for both sides, an interface (Tuncbag et al. 2009). Protein binding interfaces specify how proteins involve in various activities in living cells. Thus, interface analysis becomes a key point to reveal the rules governing protein interactions and their cellular processes involved.

A protein binding interface is generally considered to be composed of two relatively large, spatially close protein surfaces of atoms/residues with good geometric shape and chemical complementarity. It is also believed that the formation of protein chain interfaces is driven by various natural weak forces such as hydrogen bonds, electrostatic interactions, van der Waals forces, salt bridges and hydrophobic attractions. Thus, protein interfaces have complicated physicochemical properties.

In folding and binding process of proteins, a certain solvent environment is a crucial factor. Thus, it is helpful to measure how much molecules (e.g., residues and atoms) may be exposed to water/solvent environments in their protein 3D structures. To do that, solvent accessible surface area was first described by Lee and Richards in 1971. In a protein 3D structure, each heavy atom is represented by a sphere defined by its van der Waals radius; then, ASA is calculated by rolling a ball of a particular radius to probe the surface of atoms (Hubbard and Thornton 1993). The typical radius of the rolling ball is 1.4 Å. An example of ASA for several atoms is shown in Fig. 5.4. After that, the ASA sum of the certain group of atoms in a specific residue is ASA of the residue; the aggregate ASA of residues is protein ASA. Several softwares are developed to calculate ASA, and NACCESS is the well-known one (Hubbard and Thornton 1993).

It is clear that ASA is not even everywhere in protein quaternary structures: some residues/atoms are completely buried with zero ASA, while the other may be greatly exposed to the solvent; protein surfaces are heavily exposed to the solvent with larger ASA, while intraproteins/interior is little exposed to the solvent with smaller ASA; protein interfaces should be exposed to the solvent in unbound states, and little exposed in bound states. It seems that ASA is one of physical features of protein binding interfaces. It can be used to measure interface size and contact area. This measure is about the change of ASA (ΔASA) upon the formation of protein complexes, which is considered to provide a measure of binding strength (Jones and

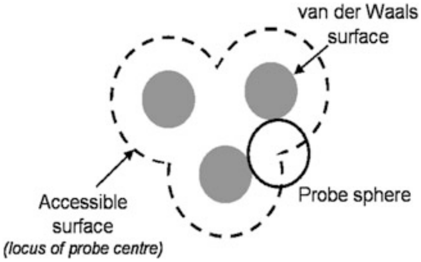


Fig. 5.4 An example of ASA for atoms. This example is from http://www.ccp4.ac.uk/dist/html/areaimol.html

Thornton 1996). Computationally for two protein chains in a protein complex, there is a difference between the sum of their ASAs each in their own tertiary structures (i.e., ASA_1, and ASA_2) and ASA of their quaternary structures (ASA_c) upon the formation of the complex. The difference, $\Delta ASA = (ASA_1 + ASA_2 - ASA_c)/2$, is generally referred to as interface size or contact area for the interactions. In detail, some atoms/residues may also undergo a similar change of their ASAs upon the formation of protein complexes; these atoms/residues are considered to be more likely involved in protein interfaces.

Besides ASA, other outstanding chemical, physical and geometric properties (Neuvirth et al. 2004; William et al. 2001) as follows can also be used to characterize protein 3D structures especially protein binding interfaces.

- Compositional features consisting of residue composition (Zhu et al. 2006; De et al. 2005; Rudra et al. 2005; Lukman et al. 2008) and propensity (Bahadur et al. 2004), residue pairs (Glaser et al. 2001; Miyazawa and Jernigan 1996; Moont et al. 1999) and atomic pairs (Mintseris and Weng 2003; Ponstingl et al. 2000, 2003; Zhang et al.1997). These compositional features are good descriptors for different segments of proteins (e.g., interfaces, surfaces and intraproteins/interior) in protein 3D structures. Protein binding interfaces tend to have a core with hydrophobic residues.
- Geometric features including planarity, shape complementarity, circularity (Jones and Thornton 1996, 1997) and secondary structures (Neuvirth et al. 2004). In a protein quaternary structure, there exists a gap between the tertiary structures of its protein chains. This kind of gaps can be measured by gap index, planarity and shape complementarity.
- Chemical features that contain hydrophobicity and polarity (Jones and Thornton 1996; Bahadur et al. 2003, 2004; Young et al. 1994).
- There are also other features, such as evolutionary residue conservation (William et al. 2001; Zhu et al. 2006). A new protein sequence usually is not created from scratch; it is generated from existing sequences by deletion, insert and mutations of residues, and shifting of parts of old sequences. Thus, some sequences can have similar or same residues in specific corresponding positions; these residues are more evolutionarily conserved. Interfacial residues are considered to be more conserved than surface residues.

Many of these features have been involved in important findings. For example, interface area of biological interactions is found to be much larger than that in non-biological interactions (William et al. 2001; Zhu et al. 2006; De et al. 2005; Bahadur et al. 2003, 2004; Janin and Rodier 1995; Janin 1997; Carugo and Argos 1997). These above features also suggest that protein interfaces are very complicated. Here, biological interactions are those molecular binding which form in solution or in their physiological states to perform biological processes and molecular functions, while non-biological interactions are produced manually or their monomers are randomly in close vicinity without biological significance. Good examples of non-biological binding are decoys in protein docking or crystal packing in PDB—the former are artificial binding produced by computational docking algorithms, while the latter are enforced by the crystallographic packing environment and formed during the crystallization process (Tuncbag et al. 2009), but both of them do not occur in solution or in their physiological states.

5.3 Uneven Distribution of Binding Free Energy and Binding Hot Spots in Protein Interfaces

Assume that there is free energy G with proteins in an unbound state; when proteins contact each other upon the formation of protein complexes, the free energy is $G' \cdot \Delta G = G - G'$ is the change of the free energy of protein binding, or ΔG is the required energy to disassemble proteins in the bound state into the unbound state. For protein interfaces, ΔG is binding free energy.

Binding free energy is not evenly distributed in protein interfaces (Bogan and Thorn 1998). As shown in Fig. 5.5, the red part is considered to contribute most to the binding free energy, and the green part least. This uneven energetic distribution can be probed by experimental approaches. A widely-used experimental approach is alanine scanning mutagenesis (Wells 1991; Clackson and Wells 1995). This approach selectively mutates an individual side chain of a residue from interfaces

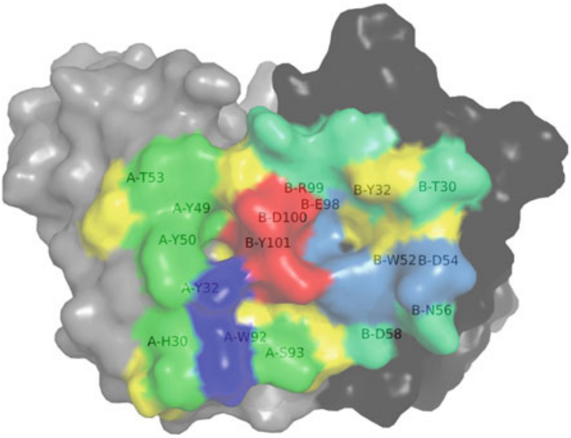


Fig. 5.5 An example of the uneven distribution of the binding free energy in the antibody protein interface in the PDB entry 1VFB. In this figure, the parts in *black*, *red*, *skyblue* and *limegreen* are from Chain B, and the parts in *gray*, *blue* and *green* are from Chain A. Those interfacial residues in *yellow* are not probed by the experimental alanine mutation method. $\Delta\Delta G$ of the red part is ≥ 2 kcal/mol; $\Delta\Delta G$ of the *blue* and *skyblue* part is <2 kcal/mol and ≥ 1 kcal/mol; $\Delta\Delta G$ of the *green* and *limegreen* part is <1 kcal/mol. The letter *strings* represent the specific residues in the interface: the first letter is the chain name; the letters following '-' are one-letter residue types and their positions in protein sequences

to Ala mostly by eliminating the side chain beyond CB. Why Ala is the substitution residue of choice is that it yet does not alter backbone conformations (as Gly or Pro can) nor does it impose extreme electrostatic or steric effects (Cunningham and Wells 1989; Lefevre et al. 1997). Then, kinetic and thermodynamic measurements are employed to determine ΔG^{mut} of binding after mutations. Finally, the energetic contribution of individual side chains to protein binding (Bogan and Thorn 1998) is calculated by using $\Delta\Delta G = \Delta\ G - \Delta\ G^{mut}\cdot$ $\Delta\Delta G$ is the change of binding free energy after alanine mutations. Through this way, it has been proved that there exist a small fraction of residues in protein interfaces, contributing most to binding stability and binding free energy. These residues are called 'binding hot spots' (Clackson and Wells 1995). In literature works, residues are considered as hot spot residues if their $\Delta\Delta G$s are not less than a threshold, such as $\geq$1.0 kcal/mol (Grosdidier and Recio 2008), or $\geq$2.0 kcal/mol (Bogan and Thorn 1998).

Binding hot spots are considered to be very important to binding affinity and strength of protein interactions. For instance, in the complex of bovine pancreatic trypsin inhibitor (BPTI) and β-Trypsin, the mutation of Lys-15 in BPTI caused a pronounced drop in the binding energy, a 230-fold decrease in the association rate constant for the trypsin/K15A complex (Maria Jose et al. 1996).

Binding hot spots defined by $\Delta\Delta G$ does not take the contribution from backbone atoms into account, but they provide a new opportunity to understand complicated protein binding and to pinpoint the governing principles under specific protein binding. Thus, identification of binding hot spots is particularly advantageous for many predictions in structural analysis of proteins, such as docking algorithms to construct protein quaternary structures (Moont et al. 1999), identification of protein binding sites (Neuvirth et al. 2004; Bradford and Westhead 2005) and reliable predictions on interaction types for new protein complexes (Zhu et al. 2006; Mintseris and Weng 2003; Bernauer et al. 2005). In practical applications, binding hot spots imply target candidates of drug design. For example, compared to 2007 H1N1, the mutations of 2009 H1N1 result in several new binding hot spots which increase the fatal property of 2009 H1N1; the interfacial mutation E104D causes human triosephosphate isomerase deficiency (Daar et al. 1986). Hence, mutations are pathogenic if they remove hot spots in expected sites or add hot spots in unexpected sites, and detecting hot spots can help design drugs against diseases.

5.4 Definition of Protein Interfaces: An Intractable but Fundamental Research Problem

In protein complexes, protein binding interfaces are an essential segment. Protein binding interfaces under a perfect definition should only contain those residues which have large positive contribution to binding free energy from both backbone atoms and atoms of side chains. However, $\Delta\Delta G$ is only the contribution from atoms of side chains, while the contribution from backbone atoms is hard, if not impossible, to experimentally measure. Also, it is expensive and time-consuming to biologically identify binding free energy of interfacial residues. So to computationally analyze protein quaternary structures, definitions of protein binding interfaces and their atomic contacts should be first given based on atomic 3D coordinates of proteins.

Generally, a protein interface is defined as two parts: contact residues and neighbor residues of the contact residues as contact scaffold (Tsai et al. 1996, 1997; Halperin et al. 2004; Li et al. 2004; Keskin et al. 2004, 2005; Mintz et al. 2005). Neighbor residues are those residues which are from the same chain as contact residues and also near contact residues. The 'near' is mostly measured based on the spatial distance between CA atoms of contact residues and of neighbor residues (Tsai et al. 1996). Contact residues are widely defined by the following several criteria.

One definition of contact residues/atoms uses atomic distance (Ofran and Rost 2003; Tsai et al. 1996; Gong et al. 2005; Lawrence and Colman 1993; Larsen et al. 1998; Preissner et al. 1998; Korkin et al. 2005; Davis and Sali 2005). The measure under this definition evaluates how two residues/atoms are spatially close from interacting proteins. However, there is no gold standard to define 'close' contacts between two residues/atoms. Some works consider two residues close enough to contact if the spatial distance of their CA, or CB atoms, or the residue's center of mass is less than a threshold (Glaser et al. 2001; Fischer et al. 2006), e.g., 9 Å (Tsai et al. 1996), while others take all heavy atoms of residues into account (Mintseris and Weng 2003; Halperin et al. 2004; Li et al. 2004). To define contacts between heavy atoms, two ways are commonly used in the literature (Glaser et al. 2001; Tsai et al. 1996): one is to take an absolute value, e.g., 6 Å (Mintseris and Weng 2003), as a threshold for atomic spatial distance; and the other is to take into account the van der Waals radii of atoms and use the sum of the corresponding atomic radii plus a value as a threshold—this value can be 0.5 Å (Halperin et al. 2004; Li et al. 2004; Keskin et al. 2004; Guney et al. 2008), 1.5 Å (Liu and Li 2009), 2.75 Å (Li and Liu 2009; Liu and Li 2010; Li and Li 2010) or 3.0 Å (Larsen et al. 1998).

Another definition considers the change of solvent accessible surface area (ΔASA) upon the formation of protein complexes (Zhu et al. 2006; Bahadur et al. 2003, 2004; Glaser et al. 2001; Jones and Thornton 1997; Gong et al. 2005; Chakrabarti and Janin 2002). Under this definition, contact residues/atoms are those residues/atoms whose ΔASA is > 0.1 Å^2 (Bahadur et al. 2003, 2004) or > 1.0 Å^2 (Zhu et al. 2006; Jones and Thornton 1997; Cho et al. 2009) upon the formation of complexes.

A more complicated definition of contacts is based on Voronoi diagrams of protein complexes (Li and Li 2010; Headd et al. 2007; Cazals et al. 2006; Bouvier et al. 2009; Bernauer et al. 2008; Poupon 2004). Given a set of points p, its general Voronoi diagram, $VD(p)$, is defined by Voronoi cells of $s \in p$; a Voronoi cell of an s is the region comprising all points closer to s than to any other points in p. The

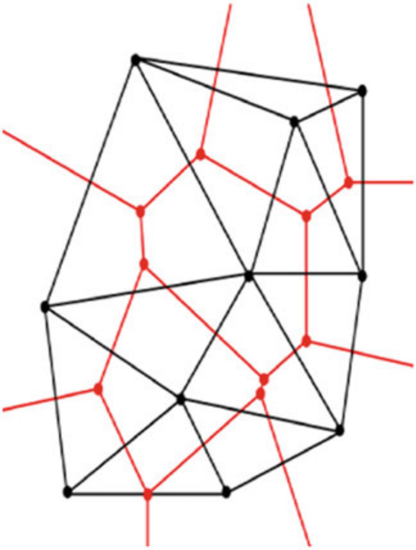


Fig. 5.6 Voronoi diagram (*red*) and Delaunay triangulation (*black*). This figure is from http://en.wikipedia.org/wiki/Delaunay_triangulation

Delaunay triangulation of p is a graph where each node is an s and two nodes s_i and s_j, $i \neq j$, have an edge if they share a Voronoi facet. So, a Delaunay triangulation is the dual graph of a Voronoi diagram (please refer to http://en.wikipedia.org/wiki/Delaunay_triangulation). Figure 5.6 shows an example of a Voronoi diagram and its dual graph Delaunay triangulation. In 3D structures of protein complexes, each atom/residue can be considered as an s. Two atoms are considered to contact with each other if they share a facet in their Voronoi diagram or an edge in their Delaunay triangulation.

Meanwhile from an energetic perspective, van der Waals energy can also be used to define contact residues (Tsai et al. 1996).

These various definitions above suggest that it is very complicated to accurately provide a computational definition for protein binding interfaces based on essential known quaternary structures of proteins. Also, the above definitions cannot capture the specific property of protein binding, because they will define large interfaces of non-biological interactions for any monomers which are in close vicinity; instead in living organisms, protein binding only happen when certain chemical, physical and biological rules are satisfied. These rules are so intractable that they are beyond the simple measures, distance and ASA in the above definitions. Hence, the above definitions are almost impossible to identify computational interfaces of protein binding which are quite similar to, if not exactly same as, what interfaces should be under the perfect definition (a definition according to the contribution of binding free energy to complex formation). This results in that the principles governing protein binding are elusive (Janin 1995), although they are necessary to understand molecular functions and biological processes in living organisms.

5.5 Current Research on Binding Interfaces

One ultimate goal in bioinformatics is to unveil the governing principles behind protein interfaces; these principles can considerably benefit structural analysis of proteins and drug design. Biological experiments are the best way of the principle discovery, since wet-lab experiments can exactly tell where binding sites and interfaces are and which residues are binding hot spots. However, protein interfaces, especially binding hot spots, intricately imply the governing principles but don't directly pinpoint.

In other words, after determining binding interfaces and their binding hot spots, discovery of the governing principles is still a challenging problem. What is worse is that, due to the structural and functional diversity of protein interactions and the large dataset of protein structures, it is an arduous and exhausting task, if not impossible, for biologists to extract the rules governing protein interactions without computational aid. Therefore, many endeavors have been tried to understand specific characteristics of protein interfaces through computational methods, since the pioneer works on protein recognition in 1975 (Chothia and Janin 1975).

5.5.1 Computational Structural Analysis of Protein Interfaces

In the light of the structural and functional diversity of protein interfaces, computational interfacial investigation can fall into several typical problems as follows but not limited to:

1. **Modeling and prediction of binding hot spots** is a key problem of analyzing binding interfaces and even of structural analysis in computational biology. As discussed before, protein binding interfaces are too complicated like fog covering the governing principles. Binding hot spots occur like several lights interspersed in the fog to help us uncover what is behind the fog.
2. **Statistical analysis of physicochemical properties for interfaces or for binding hot spots** aims at using statistical methods to find out what are specific properties of binding interfaces and binding hot spots, and why and how they can be interfaces and binding hot spots.
3. **Prediction of protein interaction types** wants to know which binding is permanent or obligate, which is transient or non-obligate, and which is non-biological. Different types of protein interactions generally possess various binding behaviors. This classification can make interfacial analysis more effective.
4. **Identification of binding sites** tries to pinpoint those atomic/residue clusters on protein surfaces which can bind to specific surfaces on other proteins for forming binding interfaces and fulfilling biological functions. The knowledge

from binding hot spots and true binding interfaces can provide some hints for this identification.

5. **Protein docking** is a process of predicting quaternary structures based on given protein tertiary structures. Protein docking first searches potential binding interfaces based on shape complementarity with the rigid docking assumption, and then scores each potential interfaces to identify interfaces which are likeliest to be true interfaces. In this process, accurate binding site prediction and precise hot spot identification will reduce the huge search space in the first step, while both binding hot spots and the knowledge from true binding interfaces, such as interfacial pairwise potentials, can help to produce functional scores in the second step. Furthermore, the task in the second step is exactly the problem of binding type prediction: distinguishing true binding interfaces from false decoys.
6. Other analyses include predictions of protein secondary and of tertiary structures, and investigation of protein–RNA/DNA binding and of protein-small molecule binding. These problems are not closely related to this survey.

These problems are closely related to each other. For example, identification of binding hot spots can specialize physicochemical analysis of protein interfaces, much improve predictions of protein interaction types and of binding sites, and also benefit docking algorithms. Correspondingly, interfacial physicochemical analysis can help to understand other problems; and types of protein interactions can specialize and simplify analysis of binding hot spots and interfacial physicochemical properties.

5.5.2 Limitations in the Current Approaches to Protein Interfaces Analysis

In computational structural biology, there are a lot of interesting problems as discussed above. In this survey, we mainly focus on problems (1), (2) and (3). We found that there are many limitations in these three directions.

(a) **The organizational topology inside O-ring**: Many works endeavored to dissect hot spots in protein interfaces. Binding hot spots were first characterized by the profound and influential O-ring theory based on the topological shape of their surrounding residues (Bogan and Thorn 1998). The O-ring theory confirms that the residues of the O-ring likely function as a role to occlude bulk water molecules from the hot spots inside the O-ring. However, the organizational topology of the ring-inside, energetically more important hot spot residues is not specified by the O-ring theory.

(b) **Prediction of binding hot spots**: Considerable computational methods have been designed to predict $\Delta\Delta G$ and binding hot spots including FoldX (Guerois et al. 2002; Schymkowitz et al. 2005), Robetta (Kortemme and Baker 2002;

Kortemme et al. 2004), PP_SITE (Gao et al. 2004), KFC (Steven et al. 2008), MINERVA (Cho et al. 2009), APIS (Xia et al. 2010), ISIS (Ofran and Rost 2007) and a web server HOTPOINT (Tuncbag et al. 2009). However, none of these approaches can characterize and identify binding hot spots with high prediction performance. A clear picture of binding hot spots still remains uncovered. Without doubt, binding hot spots are proving their own complexity. They cannot be accurately characterized by sole physicochemical property or even by most of existing physicochemical properties (Cho et al. 2009). For example, although conservation is considered to be closely related to binding hot spots (Keskin et al. 2005), some works observed that conservation less help the prediction of binding hot spots (Cho et al. 2009). Meanwhile, on the one hand, Keskin's study (Tuncbag et al. 2009) suggested that statistical pairwise potentials of interface residues can improve classification performance of binding hot spots; on the other hand, Huang's study (Xia et al. 2010) reported that residue pairwise potentials do not perform well in their hot spot prediction.

(c) **Contradictory conclusions in structural analysis of binding interfaces**: In computational structural analysis of protein interactions, accurate knowledge from binding interfaces is a starting point. Although binding interfaces are intensively investigated in the literature, some conclusions about interfaces are contradictory. For example, some works reported that there was no significant change in going from interfaces to other segments of proteins, such as intra-proteins (Glaser et al. 2001), while another works found protein interfaces were unique and different from other segments (Ofran and Rost 2003; Jones and Thornton 1997; Lo Conte et al. 1999). Furthermore, some works found that hydrophobic clusters play a determining role in protein–protein interactions (Young et al. 1994; Yan et al. 2008); in contrast, another works argued that it is the hydrophilic rather than the hydrophobic effect that is dominant in biochemical processes including protein–protein association (Ben-Naim 2006).

(d) **Prediction of binding types**: Different types of protein interactions, for example permanent versus transient complexes, have various binding behaviors. When different types of protein interactions are considered on a large dataset, specific analysis of interfaces can facilitate understanding of the principles of protein binding. So, several good classification methods have been proposed for predicting different types of protein interactions, such as ACV (Mintseris and Weng 2003), NOXclass (Zhu et al. 2006) and DiMoVo (Bernauer et al. 2008). But they mostly employ physicochemical properties of interface residues/atoms, e.g., residues with ASA change >0.1 Å^2 (Bahadur et al. 2003, 2004) or >1.0 Å^2 (Zhu et al. 2006; Jones and Thornton 1997) upon the formation of complexes. It is well-known that binding hot spots are significantly important in the stability of proteins and hot spot residues can capture more specificity of various binding behaviors of proteins. But it is not clear how this knowledge can help the prediction of binding types.

(e) **Computational definition of binding interfaces and their contacts** is a fundamental but intractable problem in previous works and in computational structural biology. In the literature, although several computational definitions for binding interfaces were proposed as discussed in Sect. 5.4, none of them exactly follows the biological meaning of protein interfaces. An evidence is crystal packing in PDB or decoys in protein docking. These computational definitions will detect a larger 'artifact interfaces' in crystal packing, no matter the definition is based on the change of atom/residue ASA, or on atomic distance, or on Voronoi diagrams of protein complexes. What is even worse is that it is very hard to distinguish these crystal-packing 'binding interfaces' from true ones. Under these computational definitions, crystal packing contacts may also make interfacial analysis full of noises. Anyway, crystal-packing interfaces clearly and intuitively expose the disadvantages of these computational definitions for protein interfaces.

5.6 Modeling and Prediction of Binding Hot Spots

Protein interfaces are very complicated to analyze; they can be dissected from several perspectives, as shown in Sect. 5.1. In this section, we present a survey on the models and prediction methods for binding hot spots.

Literature works dissecting binding hot spots can fall into three groups: (1) some studies tried to characterize topological properties of binding hot spots, e.g., the O-ring theory; (2) many other works endeavored to estimate binding free energies or to predict binding hot spots qualitatively, for example, Robetta (Kortemme and Baker 2002), FoldX (Schymkowitz et al. 2005) and KFC (Steven et al. 2008); and (3) several other authors built databases/servers based on wet-lab methods or computational methods to assist the analysis of binding hot spots, such as ASEdb and Hotsprint. Since 2000, several surveys (Tuncbag et al. 2009; Irina 2007; Fernandez-Recio et al. 2011) summarized computational analysis of binding hot spots, aiming at facilitating development of more effective prediction methods for binding free energy and binding hot spots.

5.6.1 Modeling Binding Hot Spots

Hot spots of binding free energy were initially conceptualized by Clackson and Wells (1995), supported by an outstanding finding that a "functional epitope" (hot spot) between human growth hormone and its receptor accounts for more than three-quarters of the binding free energy (Clackson and Wells 1995). This hot spot was found to be a central hydrophobic region dominated by two Trps and also geometrically surrounded by energetically less important contact residues that are generally hydrophilic and partially hydrated (Clackson and Wells 1995).

On the basis of the pioneering observations and studies of Clackson and Wells (1995), Bogan and Thorn (1998) investigated 402 alanine mutations and was the first to formalize a hypothesis named the O-ring theory to more intuitively characterize the topological shape of the surrounding residues of binding hot spots (Bogan and Thorn 1998). The O-ring theory confirmed that occlusion of solvent is a necessary condition for highly energetic interactions (hot spots) and the residues on O-ring likely function to occlude bulk water molecules from binding hot spots.

Following this step, after observing high correlation between structurally conserved interface residues and the experimental enrichment energies (Zengjian et al. 2000; Ma et al. 2003), Nussinov and her colleges suggested two specific organizations of hot spots in binding interfaces (Halperin et al. 2004; Keskin et al. 2005): on one hand, both experimental hot spots and conserved residues tend to couple a two-chain interface with higher local packing density (Halperin et al. 2004); on the other hand, binding sites (one side of the interfaces) might have several ‘hot regions’, which are locally tightly packed regions containing the clustered, networked, structurally conserved residues (Keskin et al. 2005).

Although the above works provided some nice topology organizations of binding hot spots, none of them demonstrate what is the topology between hot spot residues and how two hot spots ‘couple’ across binding interfaces. This problem was discussed in (a) of Sect. 5.2.

5.6.2 *Estimating* $\Delta\Delta$G *of Binding Hot Spots*

With the above topological organization and understanding of binding hot spots, prediction methods were designed to calculate $\Delta\Delta G$ of residue mutations or qualitatively identify binding hot spots. The knowledge used in these prediction methods can be categorized into three kinds (Guerois et al. 2002): (1) one is from physical effective energy functions; (2) another is statistical potentials of residue or atom contacts based on protein databases; and (3) the other is empirical information obtained from protein engineering experiments. Based on these kinds of knowledge, atomistic simulations were used to calculate $\Delta\Delta G$ of residue mutations. Atomistic simulations apply such strategies as the rigorous free energy perturbation (Kollman 1993), thermodynamic integration (Gouda et al. 2003), and the approximate approaches (e.g., MM-PBSA) (Gouda et al. 2003; Huo et al. 2001).

However, atomistic simulations are time-consuming. The most rapid approaches to the estimation of $\Delta\Delta G$ of residue mutations are the empirical or knowledge-based (statistical) approaches in conjunction with simple physical models (Kortemme and Baker 2002). The well-known examples include FoldX and Robetta. Robetta is a simple physical model for estimating binding energy of hot spots (Kortemme and Baker 2002; Kortemme et al. 2004). This method uses all heavy atoms and polar hydrogens to represent proteins. Then, it computationally mutated a residue into alanine (a residue type Ala), and then repacked a local sphere of 5 Å radius of the site of the mutation. Binding free energy changes upon mutations were calculated by an energy function which linearly combines such terms as Lennard-Jones potentials, an orientation-dependent hydrogen bond potential, Coulomb electrostatics, and an implicit solvation model.

Similarly, FoldX (Guerois et al. 2002; Schymkowitz et al. 2005) uses a linear combination of empirical terms to calculate binding free energy. It uses several empirical terms similar to Ro-betta, such as hydrogen bonds (water-intermediate hydrogen bonds included) and Coulomb electrostatics; other empirical terms used by FoldX are hydrophobic and polar solvation, the Van der Waals terms and so on. The weights of these empirical terms were optimized by empirical mutant data of protein experiments.

Recently, Benedix et al. proposed a structure-based method, CC/PBSA for fast estimation of the effect of mutations for protein folding and binding (Benedix et al. 2009). Given a protein structure, no matter it has a wild type or a mutant, CC/PBSA uses the program Concoord to generate alternative protein conformations. Concoord (de Groot et al. 1997) first finds pairwise interatomic distance bounds, and then starts from random coordinates and iteratively corrects the coordinates until satisfying all distance constraints. After that, CC/PBSA considers the following terms to design its energy function: Coulomb electrostatics, polar solvation free energy, a Lennard-Jones potential for solute–solute interactions and protein–protein interaction surface. Several of these terms are similar to those used in Robetta and FoldX.

All of these computational methods achieved good prediction performance based on experimental mutations. For example, the overall correlation between the observed and Robetta-calculated changes in binding free energy had an average unsigned error of 1.06 kcal/mol for interface mutations (Kortemme and Baker 2002); and on 367 mutant of 9 protein complexes, CC/PBSA obtained the correlation with $R = 0.79$ and $\delta = 1.19$ kcal/mol.

5.6.3 *Identifying Binding Hot Spots Qualitatively*

However, predicted energies by the above methods still have a large discrepancy from experimentally measured energy changes (Cho et al. 2009). Thus, computational methods were designed to qualitatively identify binding hot spots using those generated knowledge from protein structures, such as atomic contacts or evolutionary/structural conservation; the used protein structures can be quaternary structures, tertiary structures, or even primary structures with orderly increasing difficulty to accurately predict binding hot spots.

5.6.3.1 Hot Spot Identification Based on Quaternary Structures

Most of previous prediction methods of binding hot spots were designed based on known protein quaternary structures. Gao et al. (2004) developed PP_SITE using hydrogen bonds, hydrophobic and van der Waals interactions to qualitatively estimate individual contribution of each interfacial residue to protein binding. This method correctly predicted 75 hot spot residues with 88 % success rate (recall). Cho et al. (2009) combined 54 multifaceted features to develop their predictive model, MINERVA, for interaction hot spots; these features are composed of different levels of protein information, including structure, sequence and molecular interactions. They then used a decision tree to select the best subset of features, and treated them as input of SVM to build their classifier MINERVA. They claimed that MINERVA is better than Robetta, FoldEF and KFC. In their analysis, they found that weighted atomic packing density, relative surface area burial and weighted hydrophobicity are the top 3 features for identifying hot spots; π-related interactions, hydrogen bonds, and salt bridges are observed to be closely related to binding hot spots as expected.

Hot Spot Identification based on Simple rules KFC (Steven et al. 2008; Steven 2007) uses a decision-tree model to produce some rules for classifying binding hot spots. KFC comprises K-FADE (based on shape specificity features calculated by the Fast Atomic Density Evaluator) and K-CON (based on biochemical contact features). K-FADE and K-CON are two decision tree classifiers to improve the ability of hot spot prediction. These two decision trees employ the following features to represent a residue: physical and chemical features, shape specificity, and biochemical contacts such as atomic contacts, hydrogen bonds and salt bridges. Similarly based on simple rules, Hotsprint (Guney et al. 2008) detects hot spots using some thresholds of conservation, ASA, and residue propensity, while HotPoint (Tuncbag et al. 2009) combines conservation, ASA and residue pairwise potentials to produce hot spot rules.

Machine-learning Algorithms to Identify Hot Spots Since 2009, machine-learning algorithms, such as SVM and Bayesian networks, have been actively applied to improve prediction performance of binding hot spots. Lise et al. (2009) considered energetic terms such as van der Waals potentials, solvation energy, hydrogen bonds and Coulomb electrostatics, and took them as input features of SVM to classify binding hot spots. They found that transductive SVM can achieve the best performance with precision 56 % and recall 65 % for binding hot spots with $\Delta\Delta G \geq 2$ kcal/mol. Further, they treated predictions involving either an arginine or a glutamic residue separately, and improved the classification performance with precision 61 % and recall 69 % (Stefano Lise et al. 2011). It seems that different types of residues may have their unique way of contributing to protein binding. It should be better to construct individual hot spot classifiers for each type of residues when enough mutations are available.

Xia et al. (2010) introduced an ensemble classifier of nine SVM classifiers to predict binding hot spots. This method investigated a wide variety of 62 features from a combination of protein sequences and structure information, and used F-score to select non-redundant and relevant features for SVM classifiers. In particular, this method combined a new feature, protrusion index with solvent accessibility, to significantly improve the prediction performance.

Assi et al. used Bayesian Networks to design a novel probabilistic method (Salam et al. 2010) for binding hot spots. This method combines three main sources of information. One source is energetic term of computational alanine scanning by FoldX, another source is evolutionary determinants of mutated residues and of their contact residues, while the other source is structural environment such as proportion of side chain atoms involving contacts and the ratio of the number of structurally neighbor residues over the average number of neighbor residues of the type of mutated residues. This method achieved 0.71 F1 score on a BID dataset.

Network-based Approaches Different from these feature-based methods above, graphs and networks of residues are also used to investigate the topological organization of binding hot spots. Similar to O-ring, Li and Li proposed a novel descriptor of atoms and residues, called burial level, to enhance hot spot prediction performance (Li and Li 2010). By this method, they built an atomic contact graph for a protein complex, where vertices are atoms and edges are atom contacts. They defined the burial level of an atom in this graph as the length of the shortest path from this atom to its nearest exposed atom to the bulk solvent. The burial level of an atom or a residue indicates the extent it is buried inside the complex. As hot spot residues are protected by O-rings (Bogan and Thorn 1998; Li and Liu 2009), hot spot residues always have low ASA and high burial levels. But Li and Li claimed that a high burial level seems to be more sufficient than ASA: there are very few highly buried interfacial residues that are not hot spot residues. Based on this concept, a GCR model is built for binding hot spot prediction (Li and Li 2010) and has achieved good performance.

Similarly, network-based approach (Del 2005) is also used to identify the properties of key interfacial residues. In this approach, protein complexes are represented by graphs with residues as vertexes and residue contacts as edges. On a dataset with 48 dimer complexes, these protein complex graphs were proved to exhibit characteristics that resemble a small-world network; in these networks, 83 % of predicted highly central residues were found to correspond to or directly contact an experimentally determined hot spot. Also, Tuncbag et al. used graphs to visualize residue contact networks of protein interfaces. Edges in this graph are weighted according to an energy function derived from knowledge-based potentials. Then, they constructed min-cut tree to simplify and summarize contact graphs (Tuncbag 2010). They observed that binding hot spots are the highest degree node and also in a few paths in the min-cut tree.

5.6.3.2 Hot Spot Identification Based on Tertiary or Primary Structures

All of the above analyses require protein quaternary structures. Grosdidier and Recio (2008) tried to identify binding hot spots from protein tertiary structures. They used computational docking to produce docking solutions for protein tertiary structures; then they estimated normalized interfacial propensity (NIP) for each residue according to averaged ASA of 100 lowest-energy docking structures. Those residues with higher NIP were considered to be predicted hot spots. This method was reported to have comparative prediction performance to those methods based on protein quaternary structures.

Further, another work ISIS (Ofran and Rost 2007) endeavored to predict binding hot spots from amino acid sequences, without determined 3D structures of proteins. ISIS employed features are those kinds of information generated from amino acid sequences, such as sequence environment of residues, evolutionary profile, predicted secondary structures and solvent accessibility, and conservation. In this method, almost all binding residues predicted by ISIS were found to experimentally have significant effect on protein binding, while more than 90 % of the negative predictions contribute little to protein binding.

5.6.4 Databases for Binding Hot Spots

To facilitate the analysis of binding hot spots, several databases of binding hot spots were built. Alanine Scanning Energetics Database (ASEdb) (Thorn and Bogan 2001) and Binding Interface Database (BID) (Fischer et al. 2003) store binding hot spots determined by wet-lab experiments. These two databases were widely used to verify the effectiveness of hot spot prediction and estimation. The difference of these two databases is that hot spot residues in ASEdb have quantitative $\Delta\Delta G$, while binding strength in BID is qualitative measures such as "Strong", 'Intermediate', 'Weak' or 'Insignificant'.

However, the size of binding hot spots in these two databases is very limited, since wet-lab experiments are expensive and time-consuming. Thus, several servers or databases of computational hot spots were also constructed, for example, KFC server (Steven et al. 2008) and Hotsprint database (Guney et al. 2008). HotSprint (Guney et al. 2008) stores those computational hot spots for 35,776 protein interfaces among 49,512 protein interfaces extracted from the multi-chain structures in PDB (as of February 2006); those computational hot spots were derived based on residues conservation score, propensity, and ASA, and they are highly correlated with the experimental hot spots with a sensitivity of 76 %.

As a summary, these literature works are grouped as Table 5.1 for a quick review. Table 5.1 also suggests that although there have been a lot of works on the analysis of binding hot spots, the principles underlying binding hot spots are not yet clearly elaborated. In particular, the prediction performance is still low, even when almost of all available sequence, structural and molecular-contact information and features are used, as reviewed in Sects. 6.2 and 6.3. Meanwhile, there are conflicting observations for some features in terms of whether a feature can help uncover the puzzles behind binding hot spots. These problems were also discussed in (b) of Sect. 5.2.

Table 5.1 Summary of main previous works in dissection of binding hot spots

Objective	Previous works	Findings and properties
	(Clackson and Wells 1995)	The pioneering works of binding hot spots
Characteristic of hot spots	(Bogan and Thorn 1998)	The O-ring theory
	(Zengjian et al. 2000; Ma et al. 2003)	High correlation of structurally conserved residues and the experimental enrichment
	(Halperin et al. 2004)	Hot spots tend to couple binding interfaces
	(Keskin et al. 2005)	Hot regions of binding hot spots
Estimating $\Delta\Delta G$	Robetta (Kortemme and Baker 2002)	Decomposing $\Delta\Delta G$ into several terms such as hydrogen bonds, Coulomb electrostatics, Lennard-Jones potential
	FoldX (Guerois et al. 2002; Schymkowitz et al. 2005)	
	CC/PBSA (Benedix et al. 2009)	
Identify hot spots	PP_SITE (Gao et al. 2004)	
	MINERVA (Cho et al. 2009)	Quantifying different levels of protein, information, including structure, sequence and molecular interactions
	KFC (Steven et al. 2008; Steven 2007)	Using simple rules
	Hotsprint (Guney et al. 2008)	
	HotPoint (Tuncbag et al. 2009)	
	HSPred (Lise et al. 2009, 2011)	Machine-learning approaches
	APIS (Xia et al. 2010)	
	PCRPi (Salam et al. 2010)	
	GCR (Li and Li 2010)	Graph-based approaches
	(Del 2005)	
	(Tuncbag 2010)	
	(Grosdidier and Recio 2008)	Based on tertiary structures
	ISIS (Ofran and Rost 2007)	Based on primary structures
Database of hot spots	ASEdb (Thorn and Bogan 2001)	Each mutation has a quantitative $\Delta\Delta G$
	BID (Fischer et al. 2003)	Each mutation has a qualitative measure: 'Strong', 'Intermediate', 'Weak' or 'Insignificant'

5.7 Conclusion

Specific protein–protein interactions play a vital and fundamental role in molecular functions and biological processes. It has been attracting intensive investigations. However, protein interactions have high structural and functional diversity, and the existing investigations are still far away from the discovery of the principles governing protein–protein binding. In protein complexes, binding hot spots contribute to the binding free energy remarkably. This survey provides a comprehensive overview of protein binding interfaces and especially binding hot spots besides an elementary description of proteins and protein interactions' fundamentals. The survey lists the advancement areas for the dissection studies on protein binding, and also particularly demonstrates the difficulties in the research problems of computational modeling of protein binding interfaces and in the prediction of binding hot spots. These problems can be future directions to uncovering the mystery of protein–protein interactions.

Acknowledgments This work is partly funded by an Australia Research Council Discovery Project (DP130102124).

References

Assi SA, Tanaka T, Rabbitts TH, Fernandez-Fuentes N. PCRPi: presaging critical residues in protein interfaces, a new computational tool to chart hot spots in protein interfaces. Nucl Acids Res. 2010;38(6).

Bahadur RP, Chakrabarti P, Rodier F, Janin J. Dissecting subunit interfaces in homodimeric proteins. Proteins. 2003;53(3):708–19.

Bahadur RP, Chakrabarti P, Rodier F, Janin J. A dissection of specific and non-specific protein–protein interfaces. J Mol Biol. 2004;336(4):943–55.

Benedix A, Becker CM, de Groot BL, Caflisch A, Bockmann RA. Predicting free energy changes using structural ensembles. Nat Methods. 2009;6(1):3–4.

Ben-Naim A. On the driving forces for protein–protein association. J Chem Phys. 2006;125(2):24901.

Bernauer J, Bahadur RP, Rodier F, Janin J, Poupon A. DiMoVo: a Voronoi tessellation-based method for discriminating crystallographic and biological protein–protein interactions. Bioinformatics. 2008;24:652–8.

Bernauer J, Poupon A, Aze J, Janin J. A docking analysis of the statistical physics of protein–protein recognition. Phys Biol. 2005;2:1–2.

Bogan Andrew A, Thorn Kurt S. Anatomy of hot spots in protein interfaces. J Mol Biol. 1998;280(1):1–9.

Bouvier B, Griinberg R, Nilges M, Cazals F. Shelling the Voronoi interface of protein–protein complexes reveals patterns of residue conservation, dynamics, and composition. Proteins. 2009;76(3):677–92.

Bradford James R, Westhead David R. Improved prediction of protein–protein binding sites using a support vector machines approach. Bioinformatics. 2005;21(8):1487–94.

Carugo O, Argos P. Protein–protein crystal-packing contacts. Protein Sci. 1997;6(10):2261–3.

Castro MJM, Anderson S. Alanine point-mutations in the reactive region of bovine pancreatic trypsin inhibitor: effects on the kinetics and thermodynamics of binding to β-trypsin and a-chymotrypsin. Biochemistry. 1996; 11435–11446.

Cazals F, Proust F, Bahadur RP, Janin J. Revisiting the Voronoi description of protein–protein interfaces. Protein Sci. 2006;15(9):2082–92.

Chakrabarti P, Janin J. Dissecting protein–protein recognition sites. Proteins. 2002;47(3):334–43.

Cho K, Kim D, Lee D. A feature-based approach to modeling protein–protein interaction hot spots. Nucl Acids Res. 2009;37(8):2672–87.

Chothia C, Janin J. Principles of protein–protein recognition. Nature. 1975;256:705–8.

Clackson T, Wells J. A hot spot of binding energy in a hormone-receptor interface. Science. 1995;267:383–6.

Conte LL, Chothia C, Janin J. The atomic structure of protein–protein recognition sites. J Mol Biol. 1999;285(5):2177–98.

Cunningham BC, Wells JA. High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. Science. 1989;244(4908):1081–5.

Daar IO, Artymiuk PJ, Phillips DC, Maquat LE. Human triose- phosphate isomerase deficiency: a single amino acid substitution results in a thermolabile enzyme. Proc Natl Acad Sci USA. 1986;83(20):7903–7.

Darnell SJ, Legault L, Mitchell JC. KFC server: interactive forecasting of protein interaction hot spots. Nucl Acids Res. 2008.

Darnell SJ, Page D, Mitchell JC. An automated decision-tree approach to predicting protein interaction hot spots. Proteins. 2007;36:W265–9.

Davis FP, Sali A. PIBASE: a comprehensive database of structurally defined protein interfaces. Bioinformatics. 2005;21(9):1901–7.

De S, Krishnadev O, Srinivasan N, Rekha N. Interaction preferences across protein–protein interfaces of obligatory and non-obligatory components are different. BMC Struct Biol. 2005;5:15.

Del Sol A, O'Meara P. Small-world network approach to identify key residues in protein–protein interaction. Proteins. 2005;58:672–82.

Fernandez-Recio J. Prediction of protein binding sites and hot spots. Wiley Interdisc Rev Comput Mol Sci. 2011, 1–19.

Fischer TB, Arunachalam KV, Bailey D, Mangual V, Bakhru S, Russo R, Huang D, Paczkowski M, Lalchandani V, Ramachandra C, Ellison B, Galer S, Shapley J, Fuentes E, Tsai J. The binding interface database (BID): a compilation of amino acid hot spots in protein interfaces. Bioinformatics. 2003;19(11):1453–4.

Fischer TB, Holmes JB, Miller IR, Parsons JR, Tung L, Hu JC, Tsai J. Assessing methods for identifying pair-wise atomic contacts across binding interfaces. J Struct Biol. 2006;153(2):103–12.

Gao Y, Wang R, Lai L. Structure-based method for analyzing protein–protein interfaces. J Mol Model. 2004;10:44–54.

Glaser F, Steinberg DM, Vakser IA, Ben-Tal N. Residue frequencies and pairing preferences at protein–protein interfaces. Proteins. 2001;43(2):89–102.

Gong S, Park C, Choi H, Ko J, Jang I, Lee J, Bolser DM, Donghoon O, Kim DS, Bhak J. A protein domain interaction interface database: InterPare. BMC Bioinform. 2005;6:207.

Gouda H, Kuntz ID, Case DA, Kollman PA. Free energy calculations for theophylline binding to an RNA aptamer: Comparison of MM-PBSA and thermodynamic integration methods. Biopolymers. 2003;68(1):16–34.

de Groot B, van Aalten D, Scheek R, Amadei A, Vriend G, Berendsen H. Prediction of protein conformational freedom from distance constraints. Proteins. 1997;29(2):240–51.

Grosdidier S, Recio JF. Identification of hot-spot residues in protein–protein interactions by computational docking. BMC Bioinformatics. 2008;9(1):447.

Guerois R, Nielsen JE, Serrano L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1,000 mutations. J Mol Biol. 2002;320(2):369–87.

Guney E, Tuncbag N, Keskin O, Grsoy A. HotSprint: database of computational hot spots in protein interfaces. Nucl Acids Res. 2008;36:662–6.

Halperin I, Wolfson H, Nussinov R. Protein–protein interactions: coupling of structurally conserved residues and of hot spots across interfaces. Implications for docking. Structure. 2004;12(6):1027–38.

Headd JJ, Ban YEA, Brown P, Edelsbrunner H, Vaidya M, Rudolph J. Protein–protein interfaces: properties, preferences, and projections. J Proteome Res. 2007;6(7):2576–86.

Hubbard SJ, Thornton JM. 'NACCESS', computer program. Technical report, Department of Biochemistry Molecular Biology, University College London, 1993.

Huo S, Massova I, Kollman PA. Computational alanine scanning of the 1:1 human growth hormone-receptor complex. J Comput Chem. 2001;23:15–27.

Janin J. Elusive affinities. Proteins. 1995; 30–39.

Janin J. Specific versus non-specific contacts in protein crystals. Nat Struct Biol. 1997;4:973–4.

Janin J, Rodier F. Protein-protein interaction at crystal contacts. Proteins. 1995;23(4):580–7.

Jones S, Thornton JM. Principles of protein–protein interactions. Proc Natl Acad Sci USA. 1996;93(1):13–20.

Jones S, Thornton JM. Analysis of protein–protein interaction sites using surface patches. J Mol Biol. 1997;272(1):121–32.

Keskin O, Ma B, Nussinov R. Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues. J Mol Biol. 2005;345(5):1281–94.

Keskin O, Tsai CJ, Wolfson H, Nussinov R. A new, structurally nonredundant, diverse data set of protein–protein interfaces and its implications. Protein Sci. 2004;13(4):1043–55.

Kollman P. Free energy calculations: applications to chemical and biochemical phenomena. Chem Rev. 1993;93(7):2395–417.

Korkin D, Davis FP, Sali A. Localization of protein-binding sites within families of proteins. Protein Sci. 2005;14(9):2350–60.

Kortemme T, Baker D. A simple physical model for binding energy hot spots in protein–protein complexes. Proc Natl Acad Sci USA. 2002;99(22):14116–21.

Kortemme T, Kim DE, Baker D. Computational alanine scanning of protein–protein interfaces. Sci STKE. 2004;2004(219).

Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol. 1982;157(1):105–32.

Larsen TA, Olson AJ, Goodsell DS. Morphology of protein–protein interfaces. Structure. 1998;6(4):421–7.

Lawrence MC, Colman PM. Shape complementarity at protein/protein interfaces. J Mol Biol. 1993;234(4):946–50.

Lefevre F, Remy MH, Masson JM. Alanine-stretch scanning mutagenesis: a simple and efficient method to probe protein structure and function. Nucl Acids Res. 1997;25(2):447–8.

Li J, Liu Q. 'Double water exclusion': a hypothesis refining the O-ring theory for the hot spots at protein interfaces. Bioinformatics. 2009;25(6):743–50.

Li X, Keskin O, Ma B, Nussinov R, Liang J. Protein–protein interactions: hot spots and structurally conserved residues often locate in complemented pockets that pre-organized in the unbound states: implications for docking. J Mol Biol. 2004;344(3):781–95.

Li Z, Li J. Geometrically centered region: a "wet" model of protein binding hot spots not excluding water molecules. Proteins. 2010;78(16):3304–16.

Lise S, Archambeau C, Pontil M, Jones DT. Prediction of hot spot residues at protein–protein interfaces by combining machine learning and energy-based methods. BMC Bioinform. 2009;10(1):365.

Lise S, Buchan D, Pontil M, Jones DT. Predictions of hot spot residues at protein–protein interfaces using support vector machines. PLoS ONE. 2011;6(2):e16774.

Liu Q, Li J. Propensity vectors of low-ASA residue pairs in the distinction of protein interactions. Proteins. 2009;78(3):589–602.

Liu Q, Li J. Protein binding hot spots and the residue–residue pairing preference: a water exclusion perspective. BMC Bioinform. 2010;11(1):244.

Lukman S, Sim K, Li J, Chen YPP. Interacting amino acid preferences of 3D pattern pairs at the binding sites of transient and obligate protein complexes. In *APBC,* p.69–78 (2008).

Ma B, Elkayam T, Wolfson H, Nussinov R. Protein–protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. Proc Natl Acad Sci. 2003;100(10):5772–7.

Mintseris J, Weng Z. Atomic contact vectors in protein–protein recognition. Proteins. 2003;53(3):629–39.

Mintz S, Peleg AS, Wolfson HJ, Nussinov R. Generation and analysis of a protein–protein interface data set with similar chemical and spatial patterns of interactions. Proteins. 2005;61(1):6–20.

Miyazawa Sanzo, Jernigan Robert L. Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. J Mol Biol. 1996;256:623–44.

Moont G, Gabb HA, Sternberg MJ. Use of pair potentials across protein interfaces in screening predicted docked complexes. Proteins. 1999;35(3):364–73.

Moreira IS, Fernandes PA, Ramos MJ. Hot spots—a review of the protein–protein interface determinant amino-acid residues. Proteins. 2007;68(4):803–12.

Murzin A, Brenner S, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol. 1995;247:536–40.

Neuvirth H, Raz R, Schreiber G. ProMate: a structure based prediction program to identify the location of protein–protein binding sites. J Mol Biol. 2004;338(1):181–99.

Nooren IMA, Thornton JM. Diversity of protein–protein interactions. EMBO J. 2003;22(14):3486–92.

Ofran Y, Rost B. Analysing six types of protein–protein interfaces. J Mol Buol. 2003;325(2):377–87.

Ofran Y, Rost B. Protein-protein interaction hotspots carved into sequences. PLoS Comput Bio. 2007;3(7).

Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH: a hierarchic classification of protein domain structures. Structure (London, England: 1993). 1997;5(8):1093–108.

Ponstingl H, Henrick K, Thornton JM. Discriminating between homodimeric and monomeric proteins in the crystalline state. Proteins. 2000;41(1):47–57.

Ponstingl H, Kabir T, Thornton JM. Automatic inference of protein quaternary structure from crystals. J Appl Crystallogr. 2003;36(5):1116–22.

Poupon A. Voronoi and Voronoi-related tessellations in studies of protein structure and interaction. Curr Opin Struct Biol. 2004;14(2):233–41.

Preissner R, Goede A, Frommel C. Dictionary of interfaces in proteins (DIP) data bank of complementary molecular surface patches. J Mol Biol. 1998;280:535–50.

Saha RP, Bahadur RP, Chakrabarti P. In terresidue contacts in proteins and protein–protein interfaces and their use in characterizing the homodimeric interface. J Proteome Res. 2005; 4:1600–1609.

Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX web server: an online force field. Nucl Acids Res. 2005, 33(Web Server issue).

Thorn KS, Bogan AA. ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. Bioinformatics. 2001;17(3):284–5.

Tsai CJ, Lin SL, Wolfson HJ, Nussinov R. A dataset of protein–protein interfaces generated with a sequence-order-independent comparison technique. J Mol Biol. 1996;260(4):604–20.

Tsai CJ, Lin SL, Wolfson HJ, Nussinov R. Studies of protein–protein interfaces: a statistical analysis of the hydrophobic effect. Protein Sci. 1997;6(1):53–64.

Tsai CJ, Xu D, Nussinov R. Protein folding via binding and vice versa. Fold Des. 1998;3(4):R71–80.

Tuncbag N, Gursoy A, Keskin O. Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy. Bioinformatics. 2009;25(12):1513–20.

Tuncbag N, Kar G, Keskin O, Gursoy A, Nussinov R. A survey of available tools and web servers for analysis of protein–protein interactions and interfaces. Brief Bioinform. 2009;10:217–32.

Tuncbag N, Salman FS, Keskin O, Gursoy A. Analysis and network representation of hotspots in protein interfaces using minimum cut trees. Proteins: Struct, Funct, Bioinf. 2010;78(10):2283–94.

Valdar WSJ, Thornton JM. Conservation helps to identify biologically relevant crystal contacts. J Mol Biol. 2001; 313(2):399–416.

Wells JA. Systematic mutational analyses of protein–protein interfaces. Methods Enzymol. 1991;202:390–411.

Xia JF, Zhao XM, Song J, Huang DS. APIS: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility. BMC Bioinform. 2010;11(1):174.

Yan C, Wu F, Jernigan RL, Dobbs D, Honavar V. Characterization of protein–protein interfaces. Protein J. 2008;27(1):59–70.

Young L, Jernigan RL, Covell DG. A role for surface hydrophobicity in protein–protein recognition. Protein Sci. 1994;3(5):717–29.

Zengjian H, Ma B, Wolfson H, Nussinov R. Conservation of polar residues as hot spots at protein interfaces. Proteins: Struct, Funct, Bioinf. 2000;39(4):331–42.

Zhang C, Vasmatzis G, Cornette JL, DeLisi C. Determination of atomic desolvation energies from the structures of crystallized proteins. J Mol Biol. 1997;267(20):707–26.

Zhu H, Domingues FS, Sommer I, Lengauer T. NOXclass: prediction of protein–protein interaction types. BMC Bioinformatics. 2006;7:27.

Part II
Network Based Diagnosis of Complex Diseases

Chapter 6
Systems Biology Studies of Gene Network and Cell Signaling Pathway in Cancer Research

Junbai Wang, Ben Davidson and Tianhai Tian

Abstract This chapter intends to review the most recent development in computational investigation of regulatory networks. It covers both top-down systems biology approaches (i.e. data mining methods for analyzing large amount of omics datasets) and bottom-up systems biology methods (i.e. mathematical modeling using differential equations or chemical reaction systems) for reconstructing cancer-related biological networks in general. Particularly, two case studies are provided to illustrate the usage of these approaches for developing genetic regulatory networks and cell signaling pathways using microarray and proteomics datasets, respectively. A future outlook of this research field is also discussed.

Keywords Cancer · Microarray · Proteomics · Genetic regulatory networks · Cell signalling pathways

6.1 Introduction

Microarray technology revolutionized many biomedical research fields. In particular, the advancement of DNA microarray technologies has enabled researchers to measure a large number of gene expression activities simultaneously.

J. Wang (✉) · B. Davidson
Pathology Department, Oslo University Hospital—Norwegian Radium Hospital, Montebello
0310 Oslo, Norway
e-mail: junbai.wang@rr-research.no

B. Davidson
University of Oslo, Institute of Clinical Medicine, 0316 Oslo, Norway
e-mail: ben.davidson@medisin.uio.no

T. Tian (✉)
School of Mathematical Sciences, Monash University, Melbourne, VIC 3800, Australia
e-mail: tianhai.tian@monash.edu

B. Shen (ed.), *Bioinformatics for Diagnosis, Prognosis and Treatment of Complex Diseases*, Translational Bioinformatics 4, DOI: 10.1007/978-94-007-7975-4_6,
© Springer Science+Business Media Dordrecht 2013

The genome-wide expression measurements provide opportunities to investigate global gene regulations under various external conditions. Thus, many researchers are attracted by a problem of reverse engineering of gene regulatory networks from microarray expression data. It has inspired development of many computational algorithms to address the network issue (Wang 2008; de Jong 2002). However, there are limitations in reconstructing gene regulatory networks by using microarray gene expression data only. For real biological networks, it needs to consider diverse information (i.e. gene expression at the mRNA level, the protein level, and the metabolite level) to accurately describe the gene regulatory system. For this reason, many computational approaches have been designed to incorporate multiple molecular biology information into a uniform framework for inferring gene regulatory networks (Wang 2007; Bar-Joseph et al. 2003).

In the post-genomic era, proteomics is considered as the next crucial step to study biological systems because it allows large-scale determination of genetic and cellular functions at the protein level (Aebersold and Mann 2003; Hummon et al. 2006). The proteome is the entire complement of proteins, including the post-translational modifications (PTMs) that are made to a particular set of proteins. The purpose of proteomics research is to determine the relative or absolute amount of a biological sample. In recent years, the advanced proteomic technologies have provided powerful methods for analyzing protein samples, emerging as a potent tool for rapidly identifying proteins from complex biological samples, and for characterizing protein post-translational modifications and protein–protein interactions (Cox and Mann 2011; Cravatt et al. 2007). An important application of MS-based proteomics is to study cell signaling cascades that involve the binding of extracellular signaling molecules to cell-surface receptors triggering events inside the cell (Choudhary and Mann 2010). In this process, phosphorylation, a key reversible PTM, plays a key role in regulating protein function and localization in cell signaling networks. Phosphoproteomics is a branch of proteomics that identifies and characterizes proteins containing a phosphate group as a PTM (Choudhary and Mann 2010). In recent years phosphoproteome studies have provided a global and integrative description of cellular signaling networks (Gilchrist et al. 2006; Olsen et al. 2006). However, the complex nature of the cell signaling pathways remains to be completely understood as to how they are exactly regulated in vivo and what are the important parameters that determine their dynamics (Heinrich et al. 2002). To improve our understanding of signaling pathways, mathematical modeling allows us to make testable predictions and validate biological hypotheses regarding the signal transduction mechanisms regulating various cellular functions (Bourret 2008). The advances in proteomics technologies offer an unprecedented opportunity to understand how living organisms execute necessary functions at systems levels. From a systems biology perspective, the highly accurate temporal dynamic data generated by phosphoprotemics are valuable resources to infer unknown model parameters and to accurately model complex cell signaling networks.

Both gene network and signaling pathways have been investigated many years by various methods. The present work aims to divide the methods into two big categories, “top-down approach” and “bottom-up method”. Then the evolution of each category will be described carefully and case studies of typical methods will be provided. Top-down approach is defined as methods that apply advanced mathematical models to mine or explore massive experimental datasets, in order to acquire meaningful biological networks (i.e. gene–gene or protein-gene interactions) from measured evidence. It often obtains a large network. Bottom-up method usually tries to describe a small-scale gene or signaling network by detailed mathematical modeling, which may reproduce or predict known or unknown biological interaction (i.e. gene–gene association) based on experimental evidence. Both approaches are important in the bioinformatics research, especially in human cancer research which needs constant development in order to meet the challenge of diagnosis or prognosis. This chapter first gives a brief review on various types of methods in systems biology, and then uses two case studies to demonstrate the application of these methods in developing genetic regulatory networks and cell signaling pathways.

6.2 Top–Down Approaches

For biological network studies, Boolean network (Liang et al. 1998) was very popular at the initial stage of the field because of its simplicity. Then a pioneer paper published in 2000 set up a long-standing method—Bayesian Networks—for gene regulatory network studies (Friedman et al. 2000). In addition, at microarray related fields, Bayesian Networks is frequently used to infer gene–gene interactions or to predict signaling pathways based on gene expression profiles (Wang 2008; Pe’er 2005). Nevertheless, in the category of top-down approach, there are mainly four types of methods (i.e. probabilistic graphical models, regression methods, information theory, and network topology) that are often used or refined by researchers to explore the biological networks. Due to the limit of space, other methods that are not included in the above-mentioned four categories will only be briefly mentioned in this work.

6.2.1 Probabilistic Graphical Models

Probabilistic graphical models are widely used in bioinformatics field, though there are many variations: for example, Bayesian networks (BN), Gaussian Graphical models (GGM), Hidden Markov model (HMM), and Factor Graph (FG). Among them, GGM may be the simplest method that can be used to infer gene networks by gene expression profiles. It is suited for analyzing continuous data that summarize pair-wise interactions in a correlation matrix, which is also known as covariance selection model. For analyzing discrete data such as classified binary gene expression profiles (i.e. low or high), Graphical Log-linear models (GLM) is an option.

If the measurements contain both continuous and discrete data then Mixed Graphical models (MGM) is a good choice to explore the interactions. More detailed description of GGM, GLM and MGM algorithms, as well as their real applications in biological networks are available in papers (Wang 2007, 2011; Wang et al. 2003a, 2005).

However, there are several limitations in GGM types of probabilistic graphical models. For example, GGM is not able to identify non-linear interactions, not able to infer dynamical gene regulations and cannot integrate multiple data sources to predict gene–gene interaction. Especially, the model requires that the number of conditions should be far greater than the number of genes if researchers want to obtain an accurate biological network. For that reason, Bayesian Networks is the first choice for many researchers to infer gene–gene interactions because of the flexibility of the algorithm that can be modified to fit different demanding (Penfold and Wild 2011). Additionally, recent publications had shown good results in inferring gene–gene interaction by Bayesian network integration of multiple sources (Yu et al. 2008). Many variations of Bayesian networks are also widely used in gene regulation: for example, Factor graph was applied on cancer genomic data (Vaske et al. 2010) to infer patient-specific pathway activities, HMM was used to identify human epigenetic regulation patterns in breast cancer (Bonneville and Jin 2013), and Dynamic Bayesian networks were used to model peptide fragmentation from tandem mass spectrometry (Klammer et al. 2008). All in all, probabilistic graphical models are an active research field that not only keeps developing new algorithms, but also extends its application to a wide area.

6.2.2 Regression Method

Application of regression methods in gene network studies has not been noticed before a publication (Rogers and Girolami 2005) of using Bayesian regression method to predict gene regulatory networks from microarray gene expression profiles. The method does not need a threshold value to select possible interactions as required by many probabilistic graphical models including Bayesian networks. In addition, it is suited to study large gene–gene interactions when compared to massive computational resources that are needed for Bayesian networks. Interestingly, one published high-profile paper (Pujana et al. 2007) had applied the regression type of methods to omic data (i.e. combined gene expression profiling with functional genomic and proteomic data from various species) to infer large gene networks in human breast cancer. Since then, further studies have been carried out to develop new regression type of methods and apply them to various biological networks: for example, regression with mixed effects for identifying anti-HIV therapies according to genomic and clinical factors (Rosen-Zvi et al. 2008), and several novel regression algorithms designed to estimate time-varying interactions between genes (Song et al. 2009; Ahmed and Xing 2009) or association analysis of quantitative trait network (Kim et al. 2009). More recently, new regression methods for interaction studies have been developed: for instance, spare multitask regression for detecting common mechanism of response to therapeutic targets (Zhang and Gray 2010), identifying gene regulatory networks by incorporating diverse information in cancer (Li et al. 2012), estimating gene–gene interactions from biological lineages in an application to breast cancer data (Parikh et al. 2011). The regression type of methods for biological network study is an attractive research topic, which motivates the development of novel algorithms to tackle various problems related to inference of associations. Gene network study may be benefited from the computational efficiency of regression methods in data mining or exploring large biological datasets.

6.2.3 Information Theory

Information theory is another widely used method for inferring gene–gene interaction. It is able to predict both linear and non-linear associations. One of the well-known algorithms that based on information theory is ARACNE (Margolin et al. 2006), which identifies gene–gene interaction by estimating pair-wise gene expression profile mutual information. This method was successfully applied to a number of lymphoma-related cancer studies such as reverse engineering of gene networks in human B cells (Basso et al. 2005) and identifying BCL6 target genes that control multiple pathways in normal germinal center B cells (Basso et al. 2010). Recently, several new algorithms based on mutual information were developed to reconstructing dynamic gene regulatory networks (Wang et al. 2013a). This suggests that the prediction of gene–gene interactions based on information theory is worthy for further development.

6.2.4 Network Topology

Network topology study includes network structure, network robustness and network motifs etc. It is an important way to interpret biological networks, especially for large gene–gene interaction networks. For example, after applying a top-town approach to a set of biological data such as microarray gene expression profiles, gene–gene interaction networks can be reconstructed from either time-series measurements or condition-specific experiments. Usually, the predicted networks are large and may contain hundred or even several thousands of genes. Thus, how to interpret such massive networks becomes a challenging task. Many interesting network structure studies have been published by Barabasi AA, one of the essential ones is human disease network, where bipartite network was constructed to identify disease-associated genes (Goh et al. 2007). Other works that use network topology to investigate biological networks include protein interaction networks (Evlampiev and Isambert 2008), network robustness in gene regulatory networks (Ciliberti et al. 2007), and dynamical structure of the human protein interaction networks that predicts breast cancer outcome (Taylor et al. 2009). Network topology was also used to correlate molecular signaling pathway with cancer patient survivability (Breitkreutz et al. 2012), and new network topology algorithm was developed to distinguish direct target in gene expression regulatory networks (Feizi et al. 2013). More detailed description of application of network motifs in gene interaction studies, and how to interpret biological networks with the help of visualization and analysis patterns can be found in Merico et al. (2009). In summary, network topology is a very useful tool to interpret and visualize the biological networks.

6.2.5 Other Methods

Other computational approaches that can be used to study biological networks include linear programming (Wang et al. 2007), combinatorial association logic networks (de Ridder et al. 2010), dependence Trees (Costa et al. 2008), structural equation models (Cai et al. 2013), the ReliefF algorithm (Wu et al. 2013), and Tensor computation (Li et al. 2011). However, the significance of those new methods in gene network studies remains unclear, and more time and further research is needed to verify them (Rockman 2008).

6.3 Bottom–Up Approaches

The bottom-up systems biology is aimed at investigating the functional property of small subsystems of biological networks based on the high level of mechanistic details in the molecular levels. It starts from the bottom of biological systems, namely molecular molecules, by formulating the interactive behavior of each component of a manageable part of the system. Simulations of the model make testable predictions of the system dynamics that will be validated by further experimental studies. As research and understanding progresses, the developed models will typically be enlarged by the inclusion of more molecular components and/or processes at a higher level of mechanistic details. Based on the experimental studies that determine the biological properties of components and regulations between these components, the bottom-up approaches will select an appropriate mathematical model, infer unknown parameters of the model based on experimental data and simulate the dynamics of each component under various experimental conditions and perturbations (Schlitt and Brazma 2007; Bruggeman and Westerhoff 2007).

6.3.1 Boolean Network Model

Boolean networks are based on the assumption that the state of a component at any particular time point is binary (0/1), and thus the state of the network with

N components is defined as a vector of N-elements of 0s and 1s. A Boolean network model provides the rules of the on/off switches functioning of all components simultaneously in a series of discrete time steps. Due to its simplicity, the Boolean network has the capacity to deal with large network with a large number of molecular components (Paul et al. 2006). To realize the potential difference between the states in different cells, randomly generated networks are used to study the dynamics of complex systems (Paul et al. 2006). Stochastic extensions to the deterministic Boolean networks were proposed: they are the so-called noisy networks or Probabilistic Boolean Networks (Shmulevich et al. 2002). Furthermore, a generalized model was proposed by the introduction of the notion of gene state and image, the latter representing the substance produced by the respective gene (Thomas et al. 1995). A time delay was included in the simulation between the change of the state of the gene and that of the image. Another major extension of the Boolean network is the finite state model, in which the state of a component is more than two (Mao and Resat 2004). This approach locates somewhere between the Boolean network and continuous dynamic model, depending on the number of potential states of a component. It combines the advantages of Boolean networks such as simplicity and low computational cost, and the advantages of continuous models, such as more presentations of concentrations and time.

6.3.2 Deterministic Differential Equations

Although the Boolean networks can be used to study large-scale regulatory networks, the assumption of binary states is too simple to study network with complex dynamics. To describe biological systems in detail, differential equations models are widely used to study genetic regulatory networks, cell signaling pathways and metabolic networks. In the differential equations, each item represents the synthesis, degradation, trans-location or form trans-formation (binding, activation) of a molecular species. When the differential equations reach a steady state, namely the differentiation of each equation is zero, the system will be reduced substantially, which has been used to study the metabolic networks. The differential equation model can be developed by detailed chemical reactions directly, or by using more sophisticated functions such as the Michaelis-Menten or Hill function to design reduced models (Endy and Brent 2001). In addition, other modeling techniques, such as time delay and memory, have been used to reduce the complexity of models (Monk 2003).

For genetic regulatory networks, it is widely accepted that the stochastic models should be used because certain species (such as DNA and mRNA) may have small copy numbers (it will be discussed in the next subsection). However, the ODE systems have been widely used to model cell signaling pathways and metabolic systems because of the large number of protein kinases (Janes and Lauffenburger 2013). In addition, multi-scale models have been developed for systems that have species of both small molecular copy number and very large copy number as well

(Tian et al. 2007). Although it has been recognized that the spatial effects inside the cell are important for cellular processes (Kholodenko et al. 2010), it is difficult to use partial differential equations (PDE) to study cellular processes at the current stage because of the huge computing time and lack of information for protein spatial distribution. So far only a limited research work used PDE to model cell signaling pathway (Kholodenko 2003), or alternatively the component model has been proposed to model the spatial effects in a practical way (Schoeberl et al. 2002).

The cost of the detailed dynamic modeling is the requirement of rate constants and initial molecular concentrations. Due to the sparse of experimental data, the inference of unknown model parameters is still a challenging issue in systems biology. A common approach is to collect quantitative information from literature that may be obtained from different cell types and based on different experimental conditions. The advances in high-throughout technologies have generated huge amount of data that may make it possible to infer parameters based on datasets obtained in a single experimental condition, though the analysis of the omics datasets is still a challenge in bioinformatics.

6.3.3 Discrete Stochastic Models

Since the pioneered research work on stochastic modeling of the regulatory network of λ-phage (Arkin et al. 1998), there have been an increasing number of studies in the last decade investigating the origins of noise in biological networks and its crucial role in determining the key properties of biological networks (Kaern et al. 2005). It has been proposed that noise in the form of random fluctuations arises in biological networks in one of two ways: internal (intrinsic) noise or external (extrinsic) noise (Elowitz et al. 2002; Ozbudak et al. 2004). Empirical discoveries have stimulated explosive research interests in developing stochastic models for a wide range of biological systems, including gene regulatory networks, cell signaling pathways, and metabolic pathways (Raj and van Oudenaarden 2008; Wilkinson 2009; Tian et al. 2007; Kar et al. 2009).

For cellular processes associated with small numbers of certain key molecules, the standard chemical framework described by systems of ODEs breaks down. The stochastic simulation algorithm (SSA) represents a discrete modeling approach and an essentially exact procedure for numerically simulating the time evolution of a well-stirred reaction system (Gillespie 1977). Since the SSA can be very computationally inefficient, Gillespie (2001) proposed the t-leap methods in order to improve the efficiency of the SSA while maintaining acceptable losses in accuracy. We have proposed the binomial tau-leap method to avoid the possible negative numbers generated in the Poisson tau-leap method (Tian and Burrage 2004). These effective simulation methods in return provided innovative methodologies for designing stochastic models of biological systems (Tian and Burrage 2006).

To deal with the intrinsic noise in reactions with time delay, the delay stochastic simulation algorithm (DSSA) was designed by introducing time delay into the SSA (Barrio et al. 2006). Unlike the SSA, the DSSA characterizes chemical systems that contain both fast and slow reactions. The DSSA was also extended to describe chemical events that have multiple delays or stochastic delay that follows a given probabilistic distribution. In recent years, the DSSA has been widely used to model and simulate the dynamics of genetic regulatory networks and cell signalling pathways (Marquez-Lago et al. 2010). In addition, a number of effective simulation methods have been proposed to reduce the huge computing load of the DSSA (Leier et al. 2008). Other modelling techniques proposed recently include the slow-scale linear noise approximation and stochastic quasi-steady-state assumption. Most recently a new modelling approach has been proposed to simulate chemical reaction systems with memory reactions (Tian 2013).

6.3.4 Continuous Stochastic Models

The widely used continuous stochastic models are stochastic differential equations (SDE). There are two major approaches to introduce noise into deterministic models. The first one is the Langevin approach that studies the intrinsic noise of the system due to species of small copy numbers (Gillespie 2000). This method can be regarded as using a Gaussian random variable to approximate the Poisson random variable in the Poisson tau-leap method. Another approach is to use the Wiener process to represent external noise representing environmental fluctuations (Hasty et al. 2000). Both the additive noise and multiplicative noise have been used to describe the effect if random perturbations to the basal production rate and noise source that alters the transcription rate, respectively. An interesting approach is to use the multiplicative noise to represent intrinsic noise by using a threshold value in the rate constant (Tian and Burrage 2004). In this way the effect of intrinsic noise is significant when the copy number of a species is low. In addition, stochastic models can be developed based on the information regarding the noise in experimental data. When the noise in microarray gen e expression data is represented by the Poisson noise and multiplicative noise, the SDE model also includes both Poisson and Gaussian random variables as well (Tian 2010).

An important question in stochastic modeling is the development of a framework that includes both intrinsic and external noise simultaneously. Lei used the transcriptional system of a single gene to derive the analytic expression about the interacting effects of external and internal noise (Lei 2009). A similar question is how to distinguish effects of intrinsic noise and extrinsic noise in experimental observations. To tackle the challenge, a number of theoretical studies have been carried out to derive the analytical expressions of the mean and variance of system dynamics via a simple mathematical model (Pedraza and Paulsson 2008).

6.3.5 Reverse-Engineering of Dynamic Models

Two major inference methods, namely the optimization methods and Bayesian inference methods, are commonly used for estimating unknown models.

Optimization methods start with an initial guess, and then search exhaustively within the parameter space, aiming at minimizing an objective function (Chou et al. 2006; Gonzalez et al. 2007). The objective function represents the fitness of the model, and is usually defined as the error between the output of the model and a set of experimental data (Lillacci and Khammash 2010). With these two basic approaches of the gradient-based nonlinear optimization method and evolutionary based method, many researches attempted to build various techniques such as linear and nonlinear least-squares fitting, simulated annealing, genetic algorithms, and evolutionary computation (Mendes and Kell 1998; Kirkpatrick et al. 1983; Ashyraliyev et al. 2008). Although optimization methods have been successfully applied for biological systems, there are still some limitations using these methods, especially using the local optimization methods. To address these issues, the use of several state-of-the-art deterministic and stochastic global optimization methods has been explored (Moles et al. 2003).

The Bayesian inference methods is able to extract useful information from noisy or uncertain data (Wilkinson 2007). Different from the optimization methods, the main advantage of these methods is their ability to infer the whole probability distributions of the parameters, rather than just a point estimate. Also, handling estimations for stochastic systems using these methods is more robust as for deterministic systems (Toni et al. 2009). Meanwhile, computational time is the main obstacle for this approach as analytical approaches are not feasible for non-trivial problems and mostly numerical solutions are hard to achieve, as we need to solve for high-dimensional integration problems. Nonetheless, some developments have taken place during the last 20 years and the most recent advancements in Bayesian computation include the Markov chain Monte Carlo (MCMC) techniques, ensemble methods, and sequential Monte Carlo (SMC) methods that do not require likelihoods (Wang 2011; Penfold and Wild 2011; Sisson et al. 2007; Battogtokh et al. 2002). All these techniques have been successfully applied to biological systems.

6.4 Case Study of Top–Down Approaches

To illustrate the application of top-town approaches in investigating gene regulatory networks in cancer research, a case study by microarray gene expression profiles in ovarian cancer is provided. The aim of this study is to identify genes that are strongly associated with patient overall survival time. Here a total of 60 microarray experiments (60 ovarian cancer patients) were performed under Affymetrix chip (HG-U133_plus_2, ~54,675 probes per array) at Pathology of Department, Oslo University Hospital. Preprocessing and normalization of raw Affymetrix measurements were carried out by R package (Bioconductor). Probes with weak quality (e.g. probes have more than 80 % absent calls), and low variations across the experiments (e.g. ratios of probe maximum intensity to the probe minimum intensity is lesser than 2) were removed. After the pre-processing of microarray data, 28,091 probes were obtained for further data analysis where the probe intensity values were log2 transformed and converted to Z-scores.

First, differentially expressed genes within a predefined clinical group (e.g. good overall survival ≥36 months vs. poor overall survival <36 months) were identified by using pair-wise Fisher's linear discriminant (Wang et al. 2003b). The top 1 % (~ 280 genes) of the most differentially expressed genes was selected for further study. These putative differential expressed genes were clustered in ten groups since similar gene expression pattern may represent similar gene function in a regulatory network (Wang 2007; Wang et al. 2002). Clustering results are available at (http://folk.uio.no/junbaiw/ben/top1_overall_clusters/overall.html).

Then, the centers of ten gene expression clusters and the corresponding two categories of overall survival for 60 samples (i.e. good vs. poor survival) were combined together. The data matrix contains both continues and discrete data, which is suited for mixed graphical model (Wang 2007) to infer associations between gene expression clusters and the overall survival. The predicted gene-survival network (P value < 0.01) is displayed by Cytoscape in Fig. 6.1, where only gene clusters 3 and 4 are directly correlated to patient overall survival. In addition, both clusters 3 and 4 are interacting with cluster 5. Subsequently, a detailed gene functional study of clusters 3, 4 and 5 were carried out by DAVID tool. The results please refer to Table 6.1, in which tissue expression enrichment test indicates that genes of cluster 3 are associated with normal breast tissue from a breast cancer patient, vascular hemangioma and invasive ductal carcinoma; cluster 4 only linked to normal tissues such as brain and white blood cells; but cluster 5 is correlated with several cancers (i.e. breast cancer, vascular high grade comedo DCIS endothelium, and ovary serous adenocarcinoma). In KEGG pathway analysis, only cluster 5 is associated with some disease-related pathways such as autoimmune thyroid disease. A GO study of the clusters reveals that genes of cluster 5 are strongly linked to immune response and positive regulation of response to stimulus. Thus, cluster 5 may be the most important gene group that affects the overall survival rate of ovarian cancer patients according to the network study of integrated gene expression profiles and patient overall survival.

From the above case study, it shows that gene–gene interaction networks can be correlated to diverse information such as patient overall survival or even protein binding motifs and 3-D chromosomal structures (Wang 2007; Morigen 2009; Wang et al. 2013b) if a proper top-down network inference approach is used. Such integrated analysis of gene regulatory networks will help researchers tremendously to understand and to explore the complex human genome regulation.

6.5 Case Study of Bottom–Up Methods

The mitogen-activated protein (MAP) kinase cascade communicates signal from the growth factor receptors on the cell surface to effector molecules located in the cytoplasm and nucleus. This pathway comprises a set of three protein kinases,

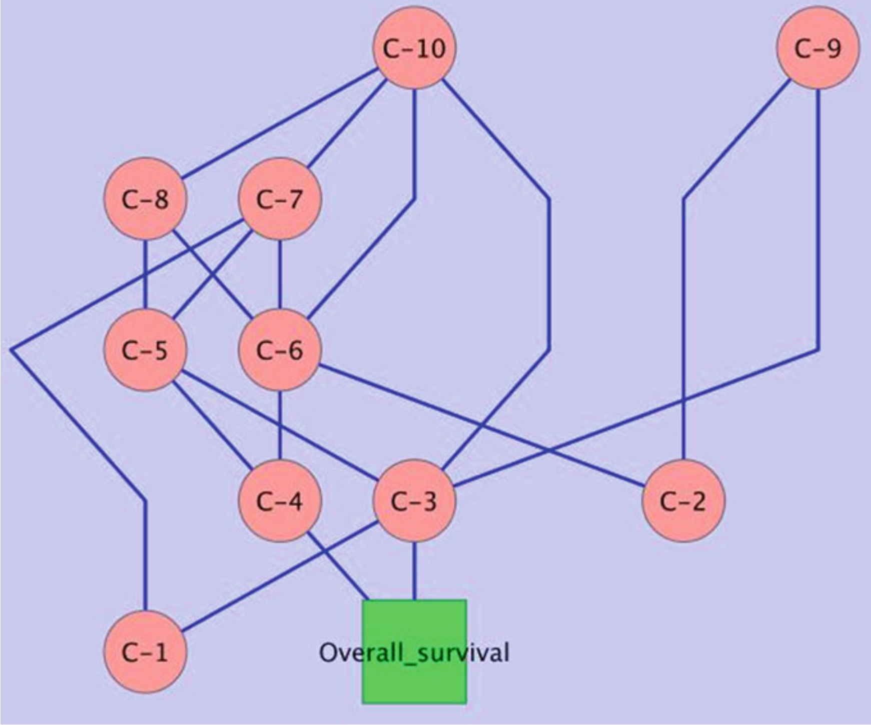


Fig. 6.1 Predicted gene-patient survival network by applying mixed graphical models on integrated gene expression clusters and patient *overall survival*

namely Raf, MEK and ERK, with a highly conserved molecular architecture that acts sequentially (Thomas and Huganir 2004). Activated MAP kinase phosphorylates multiple substrates, including transcription factors, protein kinases, phospholipases and cytoskeletal proteins, as well as regulates a wide range of physiological responses, such as cell proliferation, differentiation, apoptosis, and tissue development. Over the last decade, the MAP kinase pathway has been used repeatedly as a testable paradigm for pioneering computational systems biology. By focusing on Ras-dependent activation of the MAP kinase module, Huang and Ferrell developed the first mathematical model that predicted highly ultra-sensitive responses of the MAP kinase cascade, which were then confirmed by experimentation (Huang and Ferrell 1996). The success of this work stimulated a great deal of interest in designing kinetic models that provided testable predictions and novel insights into signaling events (Tian et al. 2007; Schoeberl et al. 2002; Bhalla et al. 2002; Chen et al. 2009).

Using the MAP kinase pathway as the test system, we designed a novel computational framework for developing mathematical models of cell signaling pathway based on the available proteomic data (Tian and Song 2012). The proposed mathematical model for the system in Fig. 6.2 includes 33 differential

Table 6.1 Top 5 results in functional study of gene clusters 3, 4 and 5 by DAVID tool

	Tissue expression	KEGG pathway	GO
C3	vascular_hemangioma_3rd	Vascular smooth muscle contraction	Vasodilation
	mammary gland_normal breast tissue from a breast cancer patient (corresponding to IDC7)_3rd		Protein ubiquitination
	brain_GBM_3rd	Ubiquitin mediated proteolysis	Protein modification by small protein conjugation
			Ubiquitin-protein ligase activity
	mammary gland_normal breast tissue from a breast cancer patient (corresponding to IDC7)_3rd		Small conjugating protein ligase activity
	mammary gland_Grade I, ER+, PR+, Her2-invasive ductal carcinoma_3rd		
C4	brain_null_3rd	NA	Ubiquitin thiolesterase activity
	retina_retinal pigment epithelium (RPE) and choroid_3rd		Peptidase activity, acting on L-amino acid peptides
	white blood cells_normal leukocytes_3rd		
	retina_normal macula_3rd		
	retina_normal macula_3rd		Peptidase activity
			Thiolester hydrolase activity
C5	vascular_high-grade comedo DCIS endothelium_3rd	Allograft rejection	Cellular monovalent inorganic cation homeostasis
	spinal cord_normal spinal cord_3rd	Graft-versus-host disease	Monovalent inorganic cation homeostasis
	mammary gland_breast myoepithelium_3rd	Type I diabetes mellitus	
	mammary gland_ER+, PR+, HER2-, grade II_3rd	Autoimmune thyroid disease	Immune response
	ovary_serous adenocarcinoma_3rd	Asthma	Positive regulation of response to stimulus
			Cellular sodium ion homeostasis

NA means not available

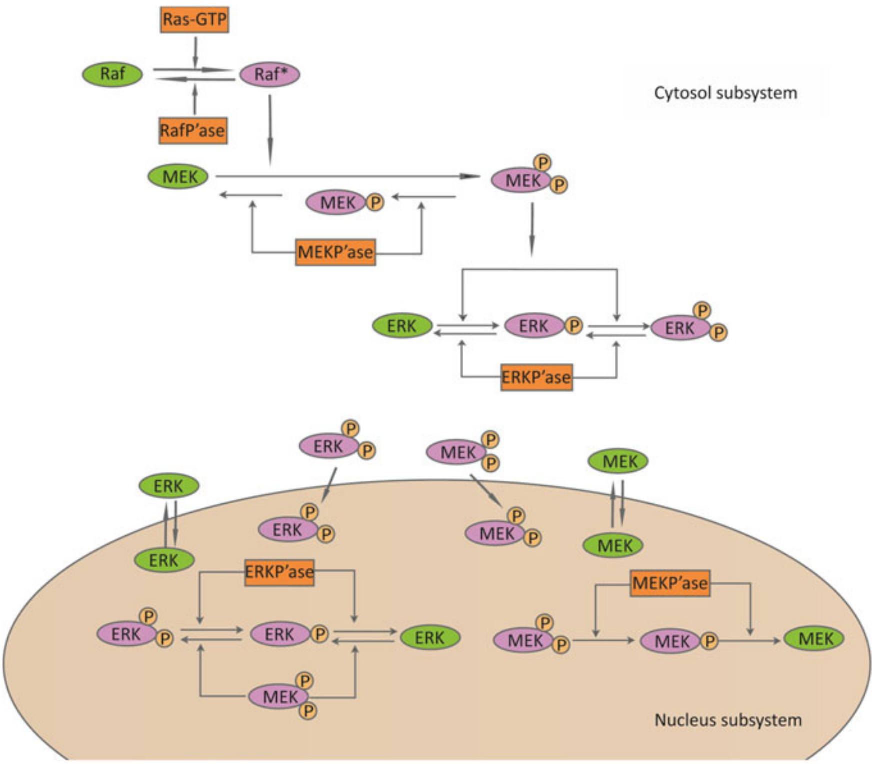


Fig. 6.2 Schematic representation of the MAK kinase pathway (Tian and Song 2012). This pathway comprises a cytosolic subsystem and a nuclear subsystem. In cytosolic *Ras-GTP*, the signal input activates *Raf* molecules in a single step. This activation is followed by activation of *MEK* kinase by activated *Raf** in a single-step processive module. The activated *MEKpp* in turn activates *ERK* in a two-step distributive module. Both the activated and un-activated *MEK* and *ERK* kinases diffuse between the cytosol and nucleus freely. In the nucleus, the activated *MEKpp* further activates *ERK* kinase via the distributive module. In addition, phosphatases, termed as *Raf-P'ase*, *MEK-P'ase* and *ERK-P'ase*, deactivate the activated *Raf**, *MEKpp* and *ERKpp* kinases, respectively, at different subcellular locations

equations representing the dynamics of 33 variables in the system. The number of the unknown rate constants in the proposed model is 57. We first used the genetic algorithm to infer the model kinetic rates based on the proteomic dataset (Olsen et al. 2006). Since Ras activity was not available in this dataset, we used the Ras activity monitored in vivo by FRET imaging as the signal input of the MAP kinase module (Fujioka et al. 2006). Since the kinase activities in the proteomic dataset were available at most five time points, we used the linear interpolation to generate kinase activities at other 16 time points during the time interval [0, 20] (min). To be consistent with the normalized kinase activities in the proteomic dataset (Olsen et al. 2006), the simulated activity of each kinase was also normalized by its activity at 5 min. The parameter set that produced smaller simulation error with respect to the proteomics data was selected as the estimated model rate constants. Because of the local maximal issue of the genetic algorithm, we implemented the genetic algorithm with different random seeds that led to different estimates of the model kinetic rates. We obtained 20 sets of estimated rate constants and selected the top 10 estimates with smaller simulation errors to the proteomic data for further analysis.

We then used the robustness property of the model as an additional criterion to select the optimal rate constants. We first used the estimated kinetic rates without any perturbation to produce a simulation that was used as the standard kinase activity. Then for each set of model rate constants, we perturbed the value of each parameter by using the generated random numbers. New simulations were obtained by using the perturbed rate constants, and we compared the new simulations with the standard simulation. The system with a particular set of rate constants is more stable if the difference between the new simulations and standard simulation is smaller. For each set of estimated rate constants, we generated 10,000 sets of perturbed rate constants by using the uniformly distributed random variable. The kinase activities at different subcellular locations together with the total activities of each kinase were collected at 20 min and we calculated the mean and variance of each kinase activity. Based on Kitano's definition of robustness (Kitano 2007), we proposed to use the average behavior, which is the sum of all the means of each kinase activity, and the nominal behavior, which is the sum of all the variances of each kinase activity, as the measure of the robustness property (Tian and Song 2012).

Figure 6.3 shows simulation results of the MAP kinase pathway using the model that has both small estimation error and good robustness property. To compare with the proteomic data, simulations were also normalized by the simulated kinase activity at 5 min. Simulations showed that the simulated kinase activities matched the Raf* activities in the cytosol (Fig. 6.3b) and ERKpp activities in both the cytosol and nucleus (Fig. 6.3f) quite well. In fact, the proteomic data of the normalized ERK activity in the cytosol are very close to those in the nucleus (Fig. 6.3f). However, there is a large difference between the simulated MEK activities and proteomic data in Fig. 6.3c. Note that there is a significant difference between the MEK kinase proteomic data in the cytosol and nucleus. The simulated MEK activities in the nucleus match the proteomic data very well.

To demonstrate the feasibility of our modeling approach, we compared our simulated kinase activities in Fig. 6.3 with the kinase activities measured in vivo by Western blotting that were taken from Fig. 7 in Fujioka et al. (2006). It shows that our computer simulation matched the Raf activity (Fig. 6.3b) and ERK activity (Fig. 6.3d) very well. However, the measured MEK activity in Fig. 6.3c is different from the proteomic data, and interestingly, the simulated MEK activity locates in the middle of the proteomic data and Western blotting data. Note that the simulated MEK activity is smaller, rather than being larger than the proteomic data, when time increases. The reason may be that, in order to match the ERK kinase activity that

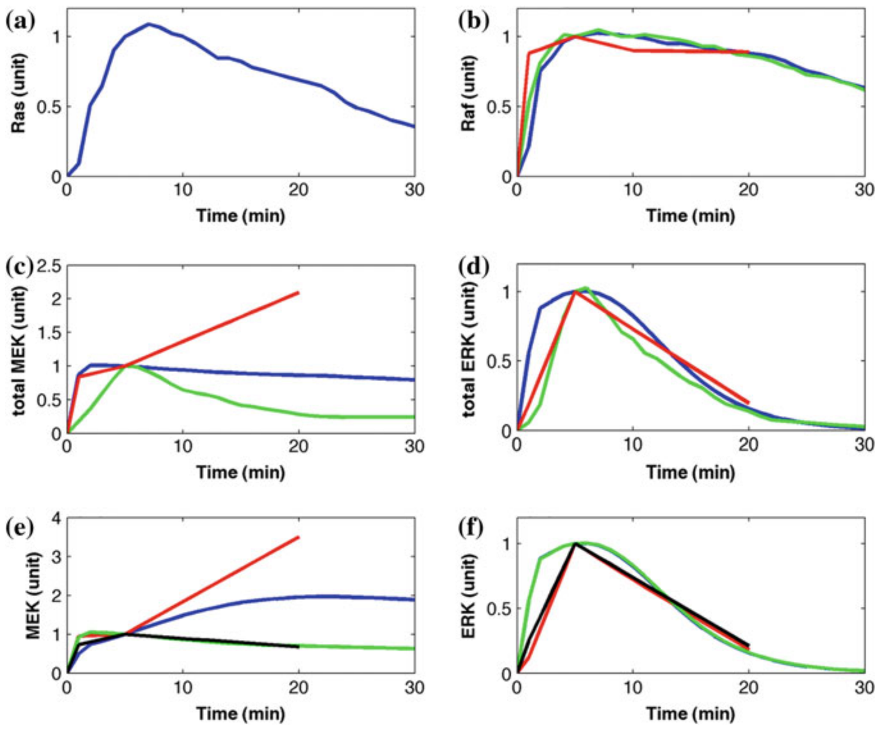


Fig. 6.3 Simulations of the normalized kinase activities (Tian and Song 2012). **a** Normalized Ras activity as the signal input from Fujioka et al. (2006). **b** Raf activity. **c** Total MEK activity. **d** Total ERK activity [*blue-line* simulation, *green-line* normalized Western blotting data (Fujioka et al. 2006), *red-line* proteomic data (Olsen et al. 2006)]. **e** MEK activity. **f** ERK activity at different locations (*blue-line* simulation in the cytosol, *red-line* proteomic data in the cytosol, *green-line* simulation in the nucleus, *black-line* proteomic data in the nucleus)

decreases significantly from 10 to 20 min, MEK kinase activity should be smaller and smaller in this time period. This observation suggests that in the cell signaling cascade, the downstream signal activity may be used to calibrate the measurement errors of the upstream signals that are present in the proteomic datasets.

6.6 Conclusion

Top-down and bottom-up methods for reverse engineering biological networks are complementary to each other. The former one allows researcher to explore or mining massive genomic high-throughput experiment datasets, to recover very large network or interaction then to discover hypothesis for further more detailed research investigation. However, the recovered networks may not reflect the dynamical activity of realistic biological systems. This weakness may be tackled by a bottom-up approach that is able to design a detailed mathematical model to describe dynamical activity of a realistic biological system, though the method only suits for a small scaled network. In the future, the connection between the two types of approaches needs to be strengthened. Novel algorithms that combine both approaches are needed in order to understand large-scale gene regulatory network and cell signaling pathway in human cancer research.

Acknowledgments JBW thanks the support from Norwegian Cancer Society (project number-DNK 2192630-2012-33376) and Norwegian Research Council (NOTUR project number nn4605k).

References

Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature. 2003;422:198–207.

Ahmed A, Xing EP. Recovering time-varying networks of dependencies in social and biological studies. Proc Natl Acad Sci USA. 2009;106:11878–83.

Arkin A, Ross J, McAdams HH. Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected *Escherichia coli* cells. Genetics. 1998;149:1633–48.

Ashyraliyev M, Jaeger J, Blom JG. Parameter estimation and determinability analysis applied to Drosophila gap gene circuits. BMC Syst Biol. 2008;2:83.

Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, et al. Computational discovery of gene modules and regulatory networks. Nat Biotechnol. 2003;21:1337–42.

Barrio M, Burrage K, Leier A, Tian T. Oscillatory regulation of Hes1: discrete stochastic delay modelling and simulation. PLoS Comput Biol. 2006;2:e117.

Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, et al. Reverse engineering of regulatory networks in human B cells. Nat Genet. 2005;37:382–90.

Basso K, Saito M, Sumazin P, Margolin AA, Wang K, et al. Integrated biochemical and computational approach identifies BCL6 direct target genes controlling multiple pathways in normal germinal center B cells. Blood. 2010;115:975–84.

Battogtokh D, Asch DK, Case ME, Arnold J, Schuttler HB. An ensemble method for identifying regulatory circuits with special reference to the qa gene cluster of Neurospora crassa. Proc Natl Acad Sci USA. 2002;99:16904–9.

Bhalla US, Ram PT, Iyengar R. MAP kinase phosphatase as a locus of flexibility in a mitogen-activated protein kinase signaling network. Science. 2002;297:1018–23.

Bonneville R, Jin VX. A hidden Markov model to identify combinatorial epigenetic regulation patterns for estrogen receptor alpha target genes. Bioinformatics. 2013;29:22–8.

Bourret RB. Signal transduction meets systems biology: deciphering specificity determinants for protein–protein interactions. Mol Microbiol. 2008;69:1336–40.

Breitkreutz D, Hlatky L, Rietman E, Tuszynski JA. Molecular signaling network complexity is correlated with cancer patient survivability. Proc Natl Acad Sci USA. 2012;109:9209–12.

Bruggeman FJ, Westerhoff HV. The nature of systems biology. Trends Microbiol. 2007;15:45–50.

Cai X, Bazerque JA, Giannakis GB. Inference of gene regulatory networks with sparse structural equation models exploiting genetic perturbations. PLoS Comput Biol. 2013;9:e1003068.

Chen WW, Schoeberl B, Jasper PJ, Niepel M, Nielsen UB, et al. Input-output behavior of ErbB signaling pathways as revealed by a mass action model trained against dynamic data. Mol Syst Biol. 2009;5:239.

Chou IC, Martens H, Voit EO. Parameter estimation in biochemical systems models with alternating regression. Theoret Biol Med Model. 2006;3:25.

Choudhary C, Mann M. Decoding signalling networks by mass spectrometry-based proteomics. Nat Rev Mol Cell Biol. 2010;11:427–39.

Ciliberti S, Martin OC, Wagner A. Innovation and robustness in complex regulatory gene networks. Proc Natl Acad Sci USA. 2007;104:13591–6.

Costa IG, Roepcke S, Hafemeister C, Schliep A. Inferring differentiation pathways from gene expression. Bioinformatics. 2008;24:i156–64.

Cox J, Mann M. Quantitative, high-resolution proteomics for data-driven systems biology. Annu Rev Biochem. 2011;80:273–99.

Cravatt BF, Simon GM, Yates JR 3rd. The biological impact of mass-spectrometry-based proteomics. Nature. 2007;450:991–1000.

de Jong H. Modeling and simulation of genetic regulatory systems: a literature review. J Comput Biol. 2002;9:67–103.

de Ridder J, Gerrits A, Bot J, de Haan G, Reinders M, et al. Inferring combinatorial association logic networks in multimodal genome-wide screens. Bioinformatics. 2010;26:i149–57.

Elowitz MB, Levine AJ, Siggia ED, Swain PS. Stochastic gene expression in a single cell. Science. 2002;297:1183–6.

Endy D, Brent R. Modelling cellular behaviour. Nature. 2001;409:391–5.

Evlampiev K, Isambert H. Conservation and topology of protein interaction networks under duplication-divergence evolution. Proc Natl Acad Sci USA. 2008;105:9863–8.

Feizi S, Marbach D, Medard M, Kellis M. Network deconvolution as a general method to distinguish direct dependencies in networks. Nat Biotechnol. 2013;31:726–33.

Friedman N, Linial M, Nachman I, Pe’er D. Using Bayesian networks to analyze expression data. J Comput Biol. 2000;7:601–20.

Fujioka A, Terai K, Itoh RE, Aoki K, Nakamura T, et al. Dynamics of the Ras/ERK MAPK cascade as monitored by fluorescent probes. J Biol Chem. 2006;281:8917–26.

Gilchrist A, Au CE, Hiding J, Bell AW, Fernandez-Rodriguez J, et al. Quantitative proteomics analysis of the secretory pathway. Cell. 2006;127:1265–81.

Gillespie DT. Exact stochastic simulation of coupled chemical-reactions. J Phys Chem. 1977;81:2340–61.

Gillespie DT. The chemical Langevin equation. J Chem Phys. 2000;113:297–306.

Gillespie DT. Approximate accelerated stochastic simulation of chemically reacting systems. J Chem Phys. 2001;115:1716–33.

Goh KI, Cusick ME, Valle D, Childs B, Vidal M, et al. The human disease network. Proc Natl Acad Sci USA. 2007;104:8685–90.

Gonzalez OR, Kuper C, Jung K, Naval PC Jr, Mendoza E. Parameter estimation using Simulated Annealing for S-system models of biochemical networks. Bioinformatics. 2007;23:480–6.

Hasty J, Pradines J, Dolnik M, Collins JJ. Noise-based switches and amplifiers for gene expression. Proc Natl Acad Sci USA. 2000;97:2075–80.

Heinrich R, Neel BG, Rapoport TA. Mathematical models of protein kinase signal transduction. Mol Cell. 2002;9:957–70.

Huang CY, Ferrell JE Jr. Ultrasensitivity in the mitogen-activated protein kinase cascade. Proc Natl Acad Sci USA. 1996;93:10078–83.

Hummon AB, Richmond TA, Verleyen P, Baggerman G, Huybrechts J, et al. From the genome to the proteome: uncovering peptides in the Apis brain. Science. 2006;314:647–9.

Janes KA, Lauffenburger DA. Models of signalling networks—what cell biologists can gain from them and give to them. J Cell Sci. 2013;126:1913–21.

Kaern M, Elston TC, Blake WJ, Collins JJ. Stochasticity in gene expression: from theories to phenotypes. Nat Rev Genet. 2005;6:451–64.

Kar S, Baumann WT, Paul MR, Tyson JJ. Exploring the roles of noise in the eukaryotic cell cycle. Proc Natl Acad Sci USA. 2009;106:6471–6.

Kholodenko BN. Four-dimensional organization of protein kinase signaling cascades: the roles of diffusion, endocytosis and molecular motors. J Exp Biol. 2003;206:2073–82.

Kholodenko BN, Hancock JF, Kolch W. Signalling ballet in space and time. Nat Rev Mol Cell Biol. 2010;11:414–26.

Kim S, Sohn KA, Xing EP. A multivariate regression approach to association analysis of a quantitative trait network. Bioinformatics. 2009;25:i204–12.

Kirkpatrick S, Gelatt CD Jr, Vecchi MP. Optimization by simulated annealing. Science. 1983;220:671–80.

Kitano H. Towards a theory of biological robustness. Mol Syst Biol. 2007;3:137.

Klammer AA, Reynolds SM, Bilmes JA, MacCoss MJ, Noble WS. Modeling peptide fragmentation with dynamic Bayesian networks for peptide identification. Bioinformatics. 2008;24:i348–56.

Lei J. Stochasticity in single gene expression with both intrinsic noise and fluctuation in kinetic parameters. J Theor Biol. 2009;256:485–92.

Leier A, Marquez-Lago TT, Burrage K. Generalized binomial tau-leap method for biochemical kinetics incorporating both delay and intrinsic noise. J Chem Phys. 2008;128:205107.

Li W, Liu CC, Zhang T, Li H, Waterman MS, et al. Integrative analysis of many weighted co-expression networks using tensor computation. PLoS Comput Biol. 2011;7:e1001106.

Li W, Zhang S, Liu CC, Zhou XJ. Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. Bioinformatics. 2012;28:2458–66.

Liang S, Fuhrman S, Somogyi R. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. Pacific symposium on biocomputing. 1998. p. 18–29.

Lillacci G, Khammash M. Parameter estimation and model selection in computational biology. PLoS Comput Biol. 2010;6:e1000696.

Mao LY, Resat H. Probabilistic representation of gene regulatory networks. Bioinformatics. 2004;20:2258–69.

Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics. 2006;7(1):S7.

Marquez-Lago TT, Leier A, Burrage K. Probability distributed time delays: integrating spatial effects into temporal models. BMC Syst Biol. 2010;4:19.

Mendes P, Kell D. Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. Bioinformatics. 1998;14:869–83.

Merico D, Gfeller D, Bader GD. How to visually interpret biological data using networks. Nat Biotechnol. 2009;27:921–4.

Moles CG, Mendes P, Banga JR. Parameter estimation in biochemical pathways: a comparison of global optimization methods. Genome Res. 2003;13:2467–74.

Monk NA. Oscillatory expression of Hes1, p53, and NF-kappaB driven by transcriptional time delays. Curr Biol: CB. 2003;13:1409–13.

Morigen WJ. BayesPI—a new model to study protein-DNA interactions: a case study of condition-specific protein binding parameters for Yeast transcription factors. BMC Bioinformatics. 2009;10:345.

Olsen JV, Blagoev B, Gnad F, Macek B, Kumar C, et al. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. Cell. 2006;127:635–48.

Ozbudak EM, Thattai M, Lim HN, Shraiman BI, van Oudenaarden A. Multistability in the lactose utilization network of *Escherichia coli*. Nature. 2004;427:737–40.

Parikh AP, Wu W, Curtis RE. Xing EP TREEGL: reverse engineering tree-evolving gene networks underlying developing biological lineages. Bioinformatics. 2011;27:i196–204.

Paul U, Kaufman V, Drossel B. Properties of attractors of canalyzing random Boolean networks. Phys Rev E: Stat, Nonlin, Soft Matter Phys. 2006;73:026118.

Pedraza JM, Paulsson J. Effects of molecular memory and bursting on fluctuations in gene expression. Science. 2008;319:339–43.

Pe'er D. Bayesian network analysis of signaling networks: a primer. Sci STKE. 2005; pl4.

Penfold CA, Wild DL. How to infer gene networks from expression profiles, revisited. Interface Focus. 2011;1:857–70.

Pujana MA, Han JD, Starita LM, Stevens KN, Tewari M, et al. Network modeling links breast cancer susceptibility and centrosome dysfunction. Nat Genet. 2007;39:1338–49.

Raj A, van Oudenaarden A. Nature, nurture, or chance: stochastic gene expression and its consequences. Cell. 2008;135:216–26.

Rockman MV. Reverse engineering the genotype-phenotype map with natural genetic variation. Nature. 2008;456:738–44.

Rogers S, Girolami M. A Bayesian regression approach to the inference of regulatory networks from gene expression data. Bioinformatics. 2005;21:3131–7.

Rosen-Zvi M, Altmann A, Prosperi M, Aharoni E, Neuvirth H, et al. Selecting anti-HIV therapies based on a variety of genomic and clinical factors. Bioinformatics. 2008;24:i399–406.

Schlitt T, Brazma A. Current approaches to gene regulatory network modelling. BMC Bioinformatics. 2007;8(6):S9.

Schoeberl B, Eichler-Jonsson C, Gilles ED, Muller G. Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. Nat Biotechnol. 2002;20:370–5.

Shmulevich I, Dougherty ER, Kim S, Zhang W. Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. Bioinformatics. 2002;18:261–74.

Sisson SA, Fan Y, Tanaka MM. Sequential Monte Carlo without likelihoods. Proc Natl Acad Sci USA. 2007;104:1760–5.

Song L, Kolar M, Xing EP. KELLER: estimating time-varying interactions between genes. Bioinformatics. 2009;25:i128–36.

Taylor IW, Linding R, Warde-Farley D, Liu Y, Pesquita C, et al. Dynamic modularity in protein interaction networks predicts breast cancer outcome. Nat Biotechnol. 2009;27:199–204.

Thomas GM, Huganir RL. MAPK cascade signalling and synaptic plasticity. Nat Rev Neurosci. 2004;5:173–83.

Thomas R, Thieffry D, Kaufman M. Dynamical behavior of biological regulatory networks. 1. Biological role of feedback loops and practical use of the concept of the loop-characteristic state. Bull Math Biol. 1995;57:247–76.

Tian T. Stochastic models for inferring genetic regulation from microarray gene expression data. Bio Systems. 2010;99:192–200.

Tian T. Chemical memory reactions induced bursting dynamics in gene expression. PLoS ONE. 2013;8:e52029.

Tian T, Burrage K. Binomial leap methods for simulating stochastic chemical kinetics. J Chem Phys. 2004a;121:10356–64.

Tian T, Burrage K. Bistability and switching in the lysis/lysogeny genetic regulatory network of bacteriophage lambda. J Theor Biol. 2004b;227:229–37.

Tian T, Burrage K. Stochastic models for regulatory networks of the genetic toggle switch. Proc Natl Acad Sci USA. 2006;103:8372–7.

Tian T, Song J. Mathematical modelling of the MAP kinase pathway using proteomic datasets. PLoS ONE. 2012;7:e42230.

Tian T, Harding A, Inder K, Plowman S, Parton RG, et al. Plasma membrane nanoswitches generate high-fidelity Ras signal transduction. Nat Cell Biol. 2007;9:905–14.

Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf MP. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. J R Soc Interface. 2009;6:187–202.

Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics. 2010;26:i237–45.

Wang J. A new framework for identifying combinatorial regulation of transcription factors: a case study of the yeast cell cycle. J Biomed Inform. 2007;40:707–25.

Wang J. Computational biology of genome expression and regulation—a review of microarray bioinformatics. J Environ Pathol Toxicol Oncol. 2008;27:157–79.

Wang J. Computational study of associations between histone modification and protein-DNA binding in yeast genome by integrating diverse information. BMC Genomics. 2011;12:172.

Wang J, Delabie J, Aasheim H, Smeland E, Myklebost O. Clustering of the SOM easily reveals distinct gene expression patterns: results of a reanalysis of lymphoma study. BMC Bioinformatics. 2002;3:36.

Wang J, Myklebost O, Hovig E. MGraph: graphical models for microarray data analysis. Bioinformatics. 2003a;19:2210–1.

Wang J, Bo TH, Jonassen I, Myklebost O, Hovig E. Tumor classification and marker gene prediction by feature selection and fuzzy c-means clustering using microarray data. BMC Bioinformatics. 2003b;4:60.

Wang J, Cheung LW, Delabie J. New probabilistic graphical models for genetic regulatory networks studies. J Biomed Inform. 2005;38:443–55.

Wang RS, Wang Y, Zhang XS, Chen L. Inferring transcriptional regulatory networks from high-throughput data. Bioinformatics. 2007;23:3056–64.

Wang J, Chen B, Wang Y, Wang N, Garbey M, et al. Reconstructing regulatory networks from the dynamic plasticity of gene expression by mutual information. Nucleic Acids Res. 2013a;41:e97.

Wang J, Lan X, Hsu PY, Hsu HK, Huang K, et al. Genome-wide analysis uncovers high frequency, strong differential chromosomal interactions and their associated epigenetic patterns in E2-mediated gene regulation. BMC Genomics. 2013b;14:70.

Wilkinson DJ. Bayesian methods in bioinformatics and computational systems biology. Briefings Bioinform. 2007;8:109–16.

Wilkinson DJ. Stochastic modelling for quantitative description of heterogeneous biological systems. Nat Rev Genet. 2009;10:122–33.

Wu M, Liu L, Hijazi H, Chan CA. A multi-layer inference approach to reconstruct condition-specific genes and their regulation. Bioinformatics. 2013;29:1541–52.

Yu H, Zhu S, Zhou B, Xue H, Han JD. Inferring causal relationships among different histone modifications and gene expression. Genome Res. 2008;18:1314–24.

Zhang K, Gray JW, Parvin B. Sparse multitask regression for identifying common mechanism of response to therapeutic targets. Bioinformatics. 2010;26:i97–105.

Chapter 7
A Network Systems Approach to Identify Functional Epigenetic Drivers in Cancer

Andrew E. Teschendorff and Martin Widschwendter

Abstract Aberrant epigenetic regulation is a key cancer hallmark. Epigenetic changes observed in pre-neoplastic lesions and cancer also provides a promising avenue for the discovery of novel biomarkers for early detection, diagnosis and prognosis, as well as offering novel therapeutic opportunities. However, the biological interpretation and functional significance of the epigenetic changes in cancer is still unclear. This chapter describes an emerging computational systems framework for elucidating the observed epigenetic deregulation in cancer and other complex diseases. As we shall see, the novel graph-theoretical approach presented here provides a powerful framework for the identification of epigenetic biomarkers associated with common phenotypes. Moreover, it provides a convenient platform in which to perform integrative multi-dimensional analysis, allowing functional epigenetic modules driving disease to be identified. We illustrate the computational method with applications to ageing and the early detection of endometrial cancer. The methods and data presented here provide a concrete example of systems-medicine: the application of a systems-approach to identify a biomarker with great potential for clinical application.

Keywords Epigenetic regulation · Biomarkers · Cancer

A. E. Teschendorff (✉)
Statistical Cancer Genomics, UCL Cancer Institute, 72 Huntley Street,
London WC1E 6BT, UK
CAS-MPG Partner Institute for Computational Biology, Chinese Academy of Sciences,
Shanghai Institute for Biological Sciences, 320 Yue Yang Road, 200031 Shanghai, China
e-mail: a.teschendorff@ucl.ac.uk

M. Widschwendter
Department of Women's Cancer, University College London,
London WC1E 6BT, UK
e-mail: m.widschwendter@ucl.ac.uk

B. Shen (ed.), *Bioinformatics for Diagnosis, Prognosis and Treatment of Complex Diseases*, Translational Bioinformatics 4, DOI: 10.1007/978-94-007-7975-4_7,
© Springer Science+Business Media Dordrecht 2013

7.1 Introduction

7.1.1 General Background, Aims and Chapter Organisation

Complex genetic diseases like cancer represent an enormous economic and social cost to modern society (Brawley 2011). This burden is set to double by 2030 in response to a rapidly ageing population and emerging epidemics such as those associated with obesity or alcohol intake (Brawley 2011). Thus, there is an urgent need to identify novel biomarkers across every stage of disease progression, from risk prediction and early detection, to diagnosis and prognosis, and finally to drug response (Sawyers 2008). However, identifying such biomarkers is itself costly and subject to numerous challenges (Sawyers 2008). One key challenge is of a statistical and computational nature: identifying robust biomarkers from very large and complex data sets is notoriously hard, because data is often noisy, incomplete and the number of biological samples is low relative to the number of molecular features being measured. Furthermore, the functional and biological significance of the candidate biomarkers in the context of the given disease is often unclear. Thus, there is an urgent to develop novel statistical and computational methods that can tame the underlying complexity of cancer genomic data, allowing more robust and biologically significant biomarkers to be identified. To achieve this, we advocate a systems-level approach that can identify more robust and biologically meaningful biomarkers.

In this chapter we present a relatively novel systems approach for identifying epigenetic biomarkers associated with a phenotype of interest. We show how this systems approach can be naturally extended to integrate multi-dimensional genomic data (e.g. gene expression, copy number and DNA methylation), allowing, for instance, functional epigenetic driver modules to be identified. We illustrate the method by identifying robust network biomarkers associated with ageing and endometrial cancer. Our main aim with this chapter is to demonstrate how a systems approach can identify a robust biomarker with potential clinical application, as we do in the context of early detection of endometrial cancer.

Briefly, we have organized this chapter as follows. In Sect. 7.1.2 we provide a brief overview of epigenomics in ageing and cancer, justifying why we focus on the epigenome in our search for cancer biomarkers. We follow this with a brief description and justification for adopting a systems-network perspective for the identification of biomarkers. In Sect. 7.2 we provide the rationale for using a systems-network approach to analyze epigenomic, and in particular, DNA methylation data. In Sect. 7.3 we describe the systems method in detail and illustrate, as proof of principle, its application to identify differential methylation interactome hotspots associated with ageing. In Sect. 7.4 we describe the extension of the algorithm to include gene expression data, and apply it to endometrial cancer to identify a functional epigenetic driver of this cancer. We discuss the biological and clinical significance of this finding. The final section Sect. 7.5 presents our conclusions.

7.1.2 Epigenomics Links Ageing, Stem Cell Biology and Cancer

Although epigenetic changes have traditionally not featured as a cancer hallmark (Hanahan and Weinberg 2011), its importance in cancer has risen prominently over the last few years (Baylin and Ohm 2006; Feinberg et al. 2006; Feinberg and Vogelstein 1983; Irizarry et al. 2009; Jones and Baylin 2007), emerging as one of the novel key cancer hallmarks (Hanahan 2012). Indeed, not only can epigenetic aberrations, such as epigenetic silencing of tumour suppressor genes, represent driver events (Vogelstein et al. 2013), but many of the recently discovered cancer driver mutations affect genes encoding epigenetic enzymes (Shen and Laird 2013). Thus, modulation of the epigenome may play an important mediating, if not causal, role in carcinogenesis (Shen and Laird 2013).

One of the most important epigenetic aberrations seen in cancer involves changes in DNA methylation (DNAm). DNAm is a covalent modification of DNA, of regulatory potential, targeting cytosines in a CpG context (Deaton and Bird 2011), although cytosine methylation in a non-CpG context has also been observed (Lister et al. 2009). By comparing the DNA methylomes of cancer and normal cells, two key features of the cancer methylome landscape have emerged: (1) high-CpG dense promoters are often hypermethylated in cancer, and (2) these methylated islands are often immersed in relatively large blocks ($\sim$1–3 Mb) characterised by a global hypomethylation (Berman et al. 2011; Hansen et al. 2011; Wen et al. 2012). Thus, while promoters located within CpG islands are normally unmethylated, they often become methylated in cancer, a mark which is generally associated with gene silencing (Deaton and Bird 2011; Feinberg et al. 2006). On the other hand, most of the genome, which is depleted of CpGs and is normally methylated, incurs widespread methylation loss in cancer, potentially resulting in overexpression of oncogenes and genomic instability (Berman et al. 2011; Hansen et al. 2011; Wen et al. 2012). A number of further observations have been made which support the view that DNA methylation changes may play a key role in carcinogenesis. First, is the observation that cancer differentially methylated regions (cDMRs) overlap significantly with tissue-specific DMRs (Irizarry et al. 2009). This is an important observation since it is consistent with the fact that cancer cells represent a highly undifferentiated state. Second, a number of studies have shown that promoter hypermethylation in cancer preferentially targets genes with key roles in stem cell biology (Ohm et al. 2007; Schlesinger et al. 2007; Widschwendter et al. 2007), notably genes carrying the bivalent activation (H3K4me3) and repression (H3K27me3) marks in human embryonic stem cells (hESCs) (Bernstein et al. 2006). These genes overlap strongly with those marked by the PolyComb Repressive Complex (PRC2) (PolyComb Group Targets—PCGTs) in hESCs (Lee et al. 2006), many of which encode transcription factors which are necessary for the differentiation of stem cells. Thus, in hESCs these genes are kept at a low, basal level of expression, poised for immediate activation if the cell receives a cue to differentiate. The finding that these genes may become irreversibly silenced in cancer through promoter DNA methylation has been invoked as a possible mechanism supporting an epigenetic progenitor origin of cancer, since silencing of key differentiation genes could “lock” stem or progenitor cells in a state of permanent self-renewal, thus suppressing differentiation and promoting a state where further mutations (genetic or epigenetic) can accumulate (Baylin and Ohm 2006; Feinberg et al. 2006; Widschwendter et al. 2007).

Further support for an epigenetic progenitor origin of cancer has come from studies that have analysed DNA methylation changes in pre-neoplastic cells (Teschendorff et al. 2010, 2012), normal cells exposed to risk factors (Issa 2011; Teschendorff et al. 2012), as well as in relation to ageing (Maegawa et al. 2010; Rakyan et al. 2010; Teschendorff et al. 2010), the major risk factor for most cancers (Fraga et al. 2007; Fraga and Esteller 2007). What these studies have demonstrated is that the epigenetic changes one often observes in cancer, are already seen to accumulate with age in normal tissue (Issa 2011; Teschendorff et al. 2010), and that they are also present in cytologically normal cells predisposed to future morphological transformation (Teschendorff et al. 2012). Moreover, age-associated changes in DNAm also overlap significantly with those DNAm changes associated with cancer risk factors, independently of age. For instance, this is the case for smoking (Selamat et al. 2012), inflammation (Issa et al. 2001, 2011; Suzuki et al. 2009), obesity (Xu et al. 2013) and viral infections (Lechner et al. 2013; Teschendorff et al. 2012). Thus, the epigenome, and the DNA methylome in particular, is a plastic entity, capable of recording the exposure of cells to environmental risk factors. It follows that epigenetic marks offer the potential to provide biomarkers for early detection or risk prediction, and indeed, a number of studies have provided preliminary evidence that this may be possible in epithelial (Teschendorff et al. 2012) as well as non-epithelial tissue (Brennan et al. 2012; Xu et al. 2013).

In addition to early detection and risk prediction, epigenetic biomarkers also offer great promise in diagnosis (deVos et al. 2009; Gruetzmann et al. 2008; Lofton-Day et al. 2008), prognosis (Heyn and Esteller 2012; Zhuang et al. 2012) and prediction (Amatu et al. 2013; Cancer Genome Atlas Research Network 2008; Heyn and Esteller 2012). For instance, DNAm of SEPT9 measured from serum DNA has been proposed as a diagnostic marker for colorectal cancer (Gruetzmann et al. 2008). The prognostic potential of DNA methylation changes was demonstrated in Zhuang et al. (2012). Specifically, there it was shown that whereas hypermethylation of PCGTs was an early event in carcinogenesis, that hypomethylation of specific CpG sites that are normally found methylated in hESCs (termed “MESCs”-Methylated in hESCs) carried prognostic significance. Indeed, Zhuang et al. (2012) derived a hypomethylation signature which was prognostic across four different gynaecological cancers and which was further aggravated in metastatic lesions.

7.1.3 Network Biomarkers

Identification of epigenetic biomarkers from large-scale omic studies is subject to the same difficulties one is faced with in the gene expression context, including the presence of numerous false positives, potentially small effect sizes and signal-to-noise ratios, unwanted variation caused by known or unknown confounding factors, and ultimately, also the interpretation of the selected biomarkers.

To address these difficulties, Chuang et al. (2007) proposed integrating gene expression data with a protein–protein interaction (PPI) network and to perform statistical inferences on the integrated network. See Chen et al. (2013) for a recent review of methods integrating gene expression with PPI networks. In Chuang et al. (2007), the authors assigned statistics of differential expression to connected subnetworks, and devised a greedy search algorithm to identify interactome hotspots of differential expression. Similar algorithms were also developed by other authors [see e.g. (Beisser et al. 2010; Dittrich et al. 2008; Ulitsky et al. 2010; Ulitsky and Shamir 2007)]. In principle, the integration of gene expression data with a PPI network offers significant advantages over methods that don't use a PPI. First, since the PPI links genes at a functional level, identifying PPI hotspots of differential expression offers an improved framework for biological interpretation. Thus, specific deregulated mechanisms or pathways can be readily identified. Second, while individual differential gene expression changes may be small, their coordinated changes within a closely connected PPI module or pathway may be highly significant when considered at the level of the whole module or pathway. Consequently, differential expression hotspots are less likely to be false positives. Third, since there is no reason to expect that confounding factors (e.g. chip/batch effects) would target specific PPI modules or pathways, performing the inference at the systems-level should therefore be more robust to such confounders. Indeed, the PPI can be viewed as providing a scaffold on which it is then possible to filter out or deconvolve the effects of technical artefacts and noise.

Although it could be argued that the integration of gene expression data with a PPI network is based on the premise that genes whose coding proteins interact, are more likely to be correlated at the level of gene expression, this assumption is not really needed. It is certainly the case, as demonstrated by many studies that neighbors in the PPI do indeed show, on average, much stronger gene expression correlations across samples than non-neighbors (Bhardwaj and Lu 2005, 2009; Taylor et al. 2009; West et al. 2012). Thus, gene expression changes related to a phenotype of interest will also likely be more correlated locally between neighbors in the PPI than for non-neighbors. However, it could also be the case that two neighboring genes in the PPI, which are both overexpressed in a given phenotype, may be overexpressed due to a different subset of samples, hence their correlation could be weak. Indeed, this is what happens in the case of genomic aberrations, which have been shown to target the same pathways in all individuals, but where the specific aberrations within the pathway often differ between individuals. Thus, there is a key distinction to be made between integrating molecular profiles with the PPI compared to integrating the statistics of differential change with the same PPI. The latter approach does not in principle require molecular profiles to be correlated, and can also lead to the identification of differential expression hotspots. In fact, this latter approach can be seen as a form of Functional Supervised Analysis (FSA), in which univariate statistics of differential expression are used in the context of the PPI, to identify connected subnetworks where an overall statistic is maximised. While many of the interactions in the PPI may bear no relevance to the biological context of the study, it can still provide a powerful means of identifying functional modules. In this regard, we should recall that a popular non-network based alternative, Gene Set Enrichment Analysis (GSEA) (Subramanian et al. 2005), is based on assessing the enrichment of a univariately ranked list of differentially expressed genes (DEGs) against an independent database of biological terms and pathways [e.g. the Molecular Signatures Database (Subramanian et al. 2005)], most of which may also be completely unrelated to the biological context of the study.

The considerations above lead to two distinct notions of interactome "modularity": one type of modularity is at the level of correlations of the molecular profiles and would correspond to an interactome (i.e. a PPI) where the correlations of neighbors is significantly larger than for non-neighbors. This type of modularity, refered to as "correlation modularity", may be unrelated to changes associated with a phenotype, since correlations would reflect common variations across samples within the same phenotype. The second type of interactome modularity is associated with a phenotype of interest, and describes an interactome that contains hotspots of association, i.e. connected subnetworks where a significant number of members have profiles that are significantly associated with the phenotype of interest. We now turn to investigate these different notions of modularity in the context of epigenomic data, specifically DNA methylation.

7.2 A Systems Approach to Epigenomics

7.2.1 The Human Interactome Exhibits DNA Methylation Correlation Modularity

That the human interactome exhibits correlation modularity at the level of gene expression has been observed in many studies (Bhardwaj and Lu 2005, 2009; Taylor et al. 2009; West et al. 2012), and is a reflection of the fact that neighboring genes in the PPI are often part of the same physical complex or molecular pathway, thus requiring that they be co-expressed under the same conditions. It is therefore natural to ask if such correlation modularity is also present at the level of DNA methylation. Intuitively, since DNAm is one of the marks influencing gene expression levels, one would expect that DNAm levels would also exhibit some level of correlation modularity. As demonstrated by West et al. (2013), using Illumina Infinium 27 k data (Bibikova and Fan 2010; Bibikova et al. 2009) and using as representative CpG the one closest to the transcription start site of genes, DNAm exhibits a relatively strong level of correlation modularity, of a magnitude similar to that seen in gene expression. In Fig. 7.1a, we reproduce this result for an Illumina 27 k data set consisting of 24 liquid based cytology normal cervical smear samples (Teschendorff and Widschwendter 2012).

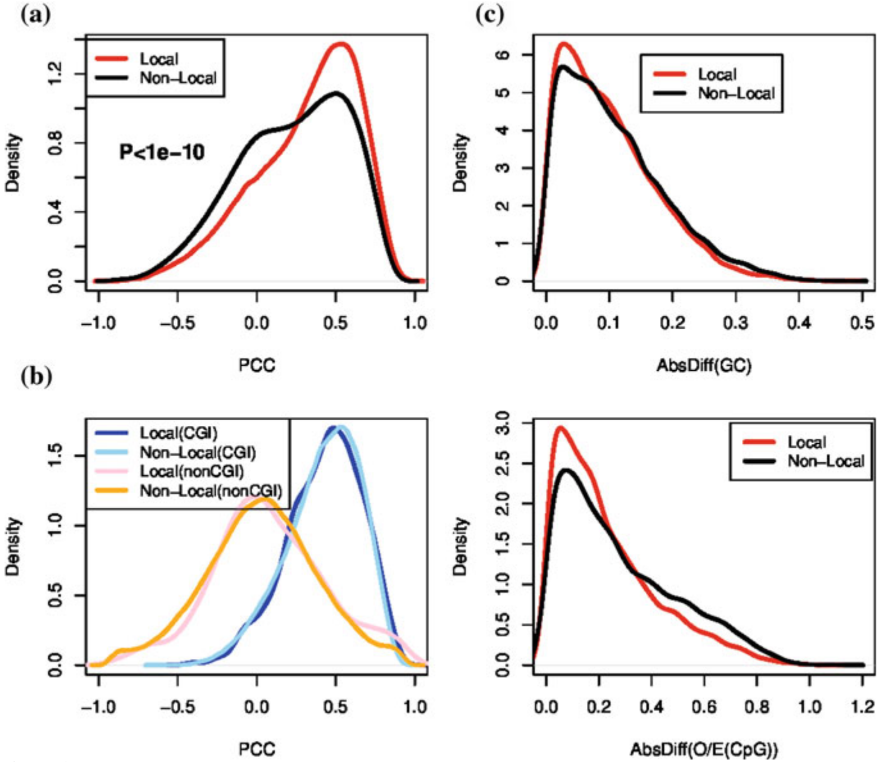


Fig. 7.1 **a** Density distribution of Pearson correlation coefficients (*PCC*) for DNA methylation profiles of genes that interact in the human interactome (*Local*) versus genes that have not been reported to interact (*Non-local*). PCCs were estimated by taking the CpG closest to the transcription start site (from Illumina 27 k data) and computed across 24 cytologically normal liquid-based cytology samples. **b** As (**a**), but now restricting independently to edges where both gene promoters are of high CpG density (*blue* and *skyblue curves*), or where both gene promoters are of low CpG density (*magenta* and *orange curves*). **c** Density distributions of the absolute differences in fractional GC content (fGC, *upper* panel) and observed to expected CpG density ratio O/E(CpG) for *local* and *non-local* edges in the *PPI*

It is well known that in the normal physiological state, CpG methylation levels are largely determined by the surrounding CpG density (Deaton and Bird 2011). Thus, it is of interest to consider if the correlation distributions are dependent on CpG density. Remarkably, for the same data set, we observe that correlations in DNAm are generally positive for CpGs located within CpG islands (CGI), whilst the correlations are symmetricly distributed around 0 for CpGs in non CpG dense regions (Fig. 7.1b). Moreover, when adjusted for CpG density, the marked difference between local and non-local correlations disappears (Fig. 7.1b). Thus, the stronger local correlations observed in Fig. 7.1a are driven by underlying differences in CpG density patterns. Indeed, Fig. 7.1c demonstrates that gene promoters of interacting proteins are more likely to share a similar CpG density than gene promoters of non-neighbors. This is particularly true for the ratio of observed to expected CpG density (Fig. 7.1c). We can conclude therefore that the DNA methylation correlation modularity exhibited by the human interactome can be attributed mostly to sequence specific features of the gene promoters of interacting proteins, and is thus a property that is hardwired in the structure of the interactome.

7.3 Differential Methylation Interactome Hotspots

7.3.1 General Considerations

Next, we turn to the question of whether DNA methylation exhibits interactome modularity in relation to a specific phenotype of interest. In other words, do differential methylation changes associated with a phenotype occur randomly in the context of the human interactome, or do they target specific gene modules or pathways?

As before, we first need to assign a DNA methylation profile to every gene in the interactome. In the case of Illumina Infinium 27 k data, where on average there are 2 CpGs per gene (located within the gene promoters), we pick the CpG closest to the transcription start site (TSS), since this one is more likely to be of functional significance. In the case of Illumina 450 k data, one may consider taking an average over probes within 200 bp of the TSS. In this way, every gene in the human interactome and represented in the DNAm assay can be assigned a statistic of differential methylation assessing its strength of association with a phenotype of interest. Which statistical test is used depends on the nature of the phenotype considered, but crucially, every gene must be subjected to the same statistical test.

Next, we describe how to infer modules, defined as connected subnetworks for which an abnormally high proportion of their constituent members are significantly associated with the phenotype of interest. In doing so, it is key to compare to an appropriate null distribution. As we shall see, our strategy will be to use two different null distributions, one to assess the statistical significance of the modularity score relative to the network as a whole. This test thus takes the topology of the network into account and is part of the inference algorithm itself. The second hypothesis test evaluates the significance of the modularity scores for the same network but with the DNA methylation profiles randomly reassigned in the interactome, thus allowing the significance of the inferred modules to be assessed solely in relation to the DNA methylation profiles. This second test thus ensures that the inferred modules are not biased by the underlying network topology.

7.3.2 Module Detection Using a Spin-Glass Algorithm

To detect the modules, we phrase the problem as that of identifying subnetworks of high edge weight density, where the weights are a direct function of the statistics of differential methylation (West et al. 2013). To clarify this, let t_i denote the statistic of differential methylation of gene i and suppose genes i and j are neighbors. We then define the weight w_{ij} between them as

$$w_{ij} = \frac{|t_i| + |t_j|}{2|t|_{\max}} \tag{7.1}$$

where |.| denotes the absolute value and $|t|_{\max}$ denotes the maximum value over all genes (nodes) in the network. This normalisation ensures that the weights are positive and bounded between 0 and 1. Other possible definitions of the weights are possible, for instance, taking the absolute value of the average of the two statistics, which would then favour modules where neighbors show the same directional changes in DNAm. We also note that $w_{ij} = 0$ if genes i and j do not interact at the protein level.

The purpose of encoding the differential methylation statistics into the edge weights (and not as node attributes) is done mainly out of mathematical convenience, since a number of module detection algorithms are best formulated in terms of edge densities. One such algorithm is a spin-glass (SPG) module detection method presented in (Reichardt and Bornholdt 2006), and which was extensively applied and validated in (West et al. 2013). Briefly, the spin-glass algorithm formulates the problem of community/module detection as that of finding the ground state of an infinite ranged spin glass Potts model (Reichardt and Bornholdt 2006). Specifically, modules are found via a Hamiltonian

$$\mathrm{H}(\{\sigma\}) = -\sum_{i \neq j} (w_{ij} - p_{ij})(\sigma_i, \sigma_j) \tag{7.2}$$

where σ_i denotes the module that node i belongs to, w_{ij} is the weighted adjacency matrix of the network, and where p_{ij} describes the probability of an edge between nodes i and j according to some appropriate null model. In the above expression, δ (σ_i, σ_j) is the Kronecker delta and $\gamma > 0$ is a tunable parameter of the algorithm. It can be shown that this Hamiltonian rewards internal edges (i.e. those within an inferred subnetwork, or equivalently within the same spin state) as well as non-edges between inferred subnetworks, while also penalizing internal non-edges, and edges between different subnetworks (Reichardt and Bornholdt 2006). The choice of parameter γ controls the relative energy contributions of edges and non-edges occuring both internally and externally of the inferred subnetworks (Reichardt and Bornholdt 2006).

The spin-glass algorithm has a number of key attractive features. First, there is only one main tunable parameter γ ($0 < \gamma < 1$), and crucially, this parameter controls the size of the inferred modules. As we shall see later, module size is of key importance since very large modules will exhibit large overlaps and thus exhibit redundancy, while very small modules will be harder to interpret biologically and are more likely to represent false positives. Thus, there is an optimal module size range, which from independent biological considerations is on the order of 10–200 genes (West et al. 2013). The tuning of the parameter γ is described in detail in West et al. (2013). Importantly, the optimal values of this parameter (normally in the range $\gamma \sim 0.5$–0.6) are fairly insensitive to the data and phenotypes considered (West et al. 2013). A second nice feature of the spinglass algorithm is that it admits a greedy search implementation (Reichardt and Bornholdt 2006), allowing modules to be identified which are proximal to certain genes of interest (which we shall term “seeds”). A greedy approach is specially attractive because it offers the needed scalability and computational efficiency (Newman 2006). Although inferred modules may not be stable, this can always be tested a posteriori using validation/test data sets. The seeds themselves are typically chosen from a top ranked list of features associated with the phenotype of interest (West et al. 2013). For each seed we thus obtain a module minimising the Hamiltonian using a simulated annealing procedure as implemented in the *spinglass.community* function of the *igraph* R package. We note however that the existence of a module associated with a given seed is not automatic since growing a module from a given seed may not result in reductions of the Hamiltonian, as for instance in the case of genes that represent isolated nodes of association. In typical applications one finds that approximately 50 % of seeds are not associated with any module (West et al. 2013). This is an important point, because, as shown in West et al. (2013), seeds that do not generate modules are less likely to validate in independent data sets and are thus more likely to represent false positives.

It is important to note that the modules inferred using the SPG algorithm describe interactome hotspots of differential methylation, where the hotspot nature is assessed relative to the network as a whole. Hence, this takes the topological edge density of the network into account and the inference of modules could be overly biased to the more highly connected and clustered subnetworks. Thus, the significance of the inferred modularities must also be assessed in a manner which does not depend on the topology of the inferred modules. This assessment is achieved using a Monte-Carlo randomisation approach in which the inferred modules are kept fixed, but where the DNAm profiles are randomised across the network. Performing a large number of Monte-Carlo runs (>1000) thus allows a null distribution of modularity values to be derived for each module independently. Thus, a significance *P* value can be assigned to each module (West et al. 2013).

7.3.3 The Importance of Module Size

The importance of module size for the detection of significant differential methylation hotspots is illustrated in Table 7.1. This table shows, for four different Illumina 27 k DNA methylation data sets, how the average module size affects its significance P-value, as assessed using Monte-Carlo randomisations of the DNAm profiles over the network (see Sect. 7.3.2). To provide the spin-glass (SPG) algorithm with a benchmark, we compare its performance to that of two other module detection algorithms: an agglomerative fast greedy (FG) non-local algorithm (Clauset et al. 2004) and a non-greedy non-local spectral decomposition (SD) algorithm (Newman 2006). These other algorithms attempt to maximise a modularity score similar to that of the spin-glass algorithm (Reichardt and Bornholdt 2006), but differ substantially in the inference procedure, allowing us to assess both the impact of "greediness" and locality.

We can see, across all four data sets, that the SPG algorithm identifies a substantially higher fraction of modules which are statistically significant under the Monte-Carlo randomisation scheme (Table 7.1). This is also reflected by a higher average modularity of the inferred modules (Table 7.1). Moreover, we can see how the improved statistical significance can be attributed to the SPG algorithm inferring on average smaller sized modules. As explained further in West et al., the optimal size of biological modules should be on the order of 50–200 genes, and the SPG algorithm tuned with a $\gamma \sim 0.5$–0.6 generally infers modules in this desired size range.

Table 7.1 For four different data sets (LBC1, LBC2, CVX and EC), we show the number of inferred modules (nMod) among the top 100 seeds, their average size (AvSize), the fraction of these that are significant under the Monte-Carlo randomisation scheme (f($P < 0.05$)) and the average modularity (AvMod) of these significant modules

	nMod	AvSize	f($P < 0.05$)	AvMod
LBC1				
SPG	21	208	**0.52**	1.58
SD	32	211	0.25	1.27
FG	14	502	0.36	1.18
LBC2				
SPG	23	108	**0.22**	1.57
SD	29	262	0.1	1.43
FG	13	593	0.08	1.44
CVX				
SPG	24	44	**0.5**	3.51
SD	56	129	0.09	2.2
FG	17	434	0.29	2.6
EC				
SPG	30	99	**0.63**	3.06
SD	24	301	0.33	1.81
FG	9	809	0.11	1.57

Table is reproduced from West et al. (2013). Data sets are described in detail in West et al. (2013). Briefly, LBC1 and LBC2 are two DNAm data sets of liquid based cytology cervical smear samples containing both normal and neoplasia specimens. CVX is a cervical normal/cancer data set. EC is an endometrial normal/cancer data set. In all cases, the modules represent hotspots of differential methylation between normal and cancer phenotypes

Bold values indicates the maximum value attained by any of the three methods

7.3.4 *Differential Methylation Hotspots Associated with Ageing*

To illustrate and validate the algorithm we consider the case of ageing. Although widespread DNAm changes associated with ageing have been reported [see e.g. (Maegawa et al. 2010; Rakyan et al. 2010; Teschendorff et al. 2010)], it is unknown whether specific molecular pathways are targeted. To address this question, West et al. used the spin-glass module detection algorithm described in Sect. 7.3.3 and applied it to a large Illumina 27 k dataset [abbreviated UKOPS, (Teschendorff et al. 2009)] of whole blood samples from 261 postmenopausal women (age range 50–80 years) to identify age-associated differential methylation hotspots (West et al. 2013). As shown in West et al., a number of hotspots were identified which validated across multiple data sets including different normal tissue types. Three of the most consistent hotspots are illustrated in Fig. 7.2a, two of which (the SOX8 and FZD2/WNT hotspots) were found to target stem-cell differentiation pathways. Importantly, the relevance of the WNT signalling pathway in ageing has been documented before (Brack et al. 2007; Brack and Rando 2007; Maiese et al. 2008). Thus, the observation that many members of this pathway are epigenetically deregulated could underpin its increased functional activity with age, suppressing differentiation and promoting self-renewal (Brack et al. 2007; Brack and Rando 2007; Maiese et al. 2008), which in turn could lead to cancer predisposition (Baylin and Ohm 2006).

To further demonstrate the robustness of these hotspots, we considered an additional whole blood data set (Hannum et al. 2013) where the samples were profiled using the more comprehensive Illumina 450 k arrays (Sandoval et al. 2011). We used the same CpGs as those in the original UKOPS 27 k set, except for those not represented on the 450 k array, in which case we picked the 450 k probe closest to the TSS. The modularity of the three hotspots was calculated and compared to that expected under random permutation of the statistics over the network. Remarkably, for all three modules, the observed modularities were significantly higher than those expected by random chance, thus validating their hotspot nature (Fig. 7.2b). Moreover, the consistency of the directional changes in DNAm across the two cohorts was good (Fig. 7.2c). Thus, these data demonstrate that the algorithm provides a powerful means of performing a functional supervised analysis, and that differential methylation hotspots associated with ageing exist.

7.4 Functional Epigenetic Driver Modules

The statistical framework presented to identify interactome hotspots associated with a phenotype of interest can be extended to include multi-dimensional data. In the context of cancer genomics, such multi-dimensional data are being routinely generated as part of international consortia [e.g. The Cancer Genome Atlas-TCGA

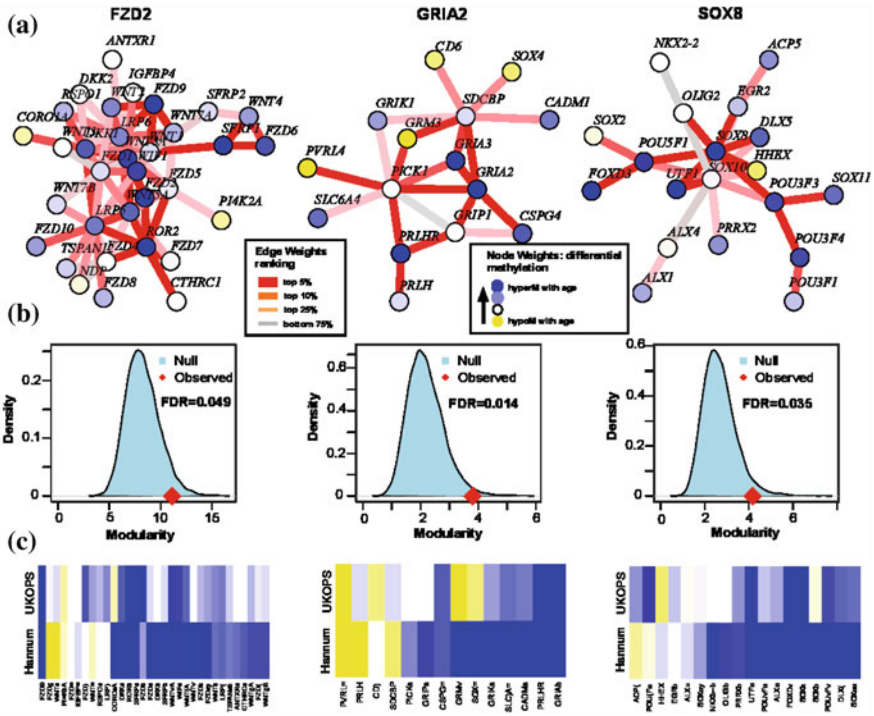


Fig. 7.2 **a** Three hotspots of age-associated differential methylation with seed genes SOX8, GRIA2 and FZD2 as inferred from the UKOPS whole blood data set (Teschendorff et al. 2009) and which were shown to be tissue independent (West et al. 2013). **b** Validation of their hotspot nature in an independent whole blood data set (Hannum et al. 2013) generated with Illumina 450 k arrays. **c** Consistency of the directional changes in DNAm with age between the two cohorts

(Cancer Genome Atlas Research Network 2008)]. In the cancer epigenomics context it may thus be of interest to identify hotspots of differential methylation, but where simultaneously there are also coordinated changes in gene expression. Such hotspots may therefore represent candidate functional epigenetic driver modules. In extending the algorithm to include gene expression data, a number of possibilities exist, depending on whether the expression data is from the same tumour samples (i.e. a matched setting as is the case for many TCGA samples) or from an independent but otherwise equivalent cohort of samples (i.e. unmatched setting). Since most data in the public domain is for the unmatched setting, we henceforth consider the more general scenario where the gene expression data is from an independent set of samples, although importantly, we assume that the independent cohort consists of cancer samples that are comparable in terms of histology and type.

7.4.1 The Functional Epigenetic Module Algorithm

The basic concept of the Functional Epigenetic Module (FEM) algorithm is illustrated in Fig. 7.3. In the case where DNAm and mRNA expression data are available, we can attribute two statistics to each node (gene) i in the network. One of these statistics, $t_i^{(D)}$, describes the association of the gene's DNAm profile with the phenotype of interest, while the other, $t_i^{(R)}$, describes the corresponding association at the gene expression level. Since the DNAm profile is that of the CpG closest to the TSS, it is sensible to assume that hypermethylation at this site is associated with gene suppression. Although the expected anti-correlation between DNAm at the TSS and downstream gene expression is generally not strong, it remains of high statistical significance [see e.g. (Lechner et al. 2013)], and hence this constitutes a sensible way forward. We should point out however, that the algorithm can be easily generalised so as to avoid making this assumption. Thus, focusing on genes where there is the expected anticorrelation between DNAm and mRNA expression, an overall statistic of association can be built by taking the absolute difference of the two statistics, i.e. $\left|t_i^{(D)} - t_i^{(R)}\right|$. If one desires to find modules where hypermethylation leads to gene silencing, one can construct the edge weights in the network as follows (see Fig. 7.3):

$$w_{ij} = \frac{1}{2}\left(H\left(t_i^{(D)}\right)H\left(-t_i^{(R)}\right)\left|t_i^{(D)} - t_i^{(R)}\right| + H\left(t_j^{(D)}\right)H\left(-t_j^{(R)}\right)\left|t_j^{(D)} - t_j^{(R)}\right|\right) \quad (7.3)$$

where $H(x)$ is the Heaviside function defined by $H(x) = 1$ if $x > 0$ and $H(x) = 0$ if $x < 0$. Thus, in the above equation, edge weights would be zero if both genes show overexpression and/or hypomethylation. Alternatively, if one desires to find modules of hypomethylation with concomitant overexpression, then the weight definition above would need to be modified by switching the signs within the Heaviside functions.

With the weights defined as above, modules of significant hypermethylation and concomitant underexpression are then identified using the same spin-glass algorithm as described in the previous sections.

7.4.2 Application to Endometrial Cancer: The HAND2 Module

To illustrate the power of the FEM algorithm, we consider the application to identify epigenetic driver modules in endometrial cancer.

Given that endometrial carcinoma risk is largely determined by non-hereditary factors (Lichtenstein et al. 2000; Schouten et al. 2004), including age, obesity and reproductive factors, it constitutes an ideal system in which to search for epigenetic mechanisms underlying cancer initiation and progression. While estrogen

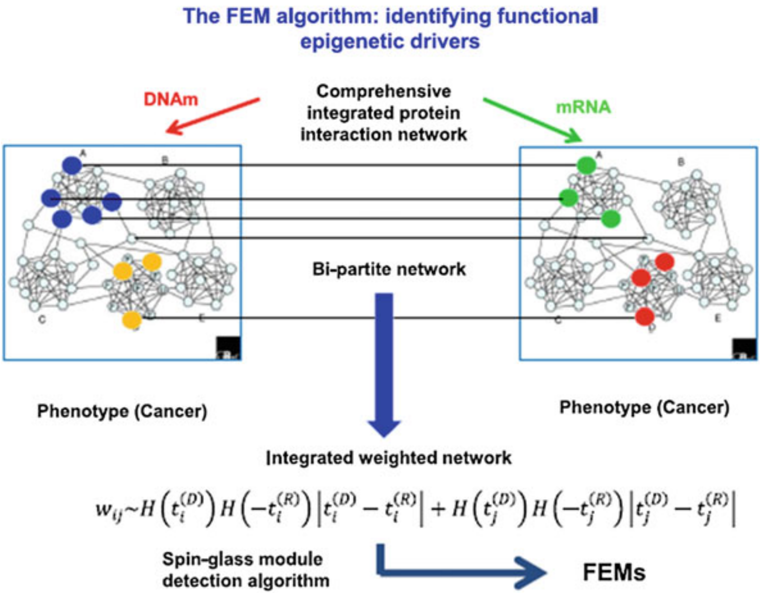

The FEM algorithm: identifying functional epigenetic drivers
DNAm
Comprehensive integrated protein interaction network
mRNA
A
B
C
E
A
B
C
D
E
Bi-partite network
Phenotype (Cancer)
Phenotype (Cancer)
Integrated weighted network
$w_{ij} \sim H\left(t_i^{(D)}\right) H\left(-t_i^{(R)}\right) \left|t_i^{(D)} - t_i^{(R)}\right| + H\left(t_j^{(D)}\right) H\left(-t_j^{(R)}\right) \left|t_j^{(D)} - t_j^{(R)}\right|$
Spin-glass module detection algorithm
FEMs

Fig. 7.3 Illustration of the *FEM* algorithm: nodes of the same *PPI* network are assigned two different statistics according to their association of DNAm (*D*) and mRNA expression (*R*) with a phenotype of interest (e.g. cancer). *Blue* denotes hypermethylated genes, *yellow* hypomethylated. *Red* denotes overexpression, *green* underexpression. Since the DNAm profiles are for CpGs closest to the TSS, one seeks modules where there is significant hypermethylation and corresponding underexpression or vice versa, modules of significant hypomethylation and corresponding overexpression. To find these modules, an integrated weighted network can be constructed with the weights constructed using the formula as shown, where *H* denotes the Heaviside function. The formula shows the case where one seeks modules of hypermethylation and concomitant underexpression. Modules are inferred using the same spin-glass algoritm used earlier

drives cell proliferation, progesterone inhibits proliferation of the endometrium and causes cell differentiation. Thus, conditions which are associated with a functional dominance of estrogen over progesterone (obesity, polycystic ovary syndrome, nulliparity, long-term exposure to unopposed estrogens) are associated with an increased risk for endometrial cancer (Amant et al. 2005). Although it is well established that the tumor-protective and anti-proliferative effect of progesterone on the endometrial epithelium (Yang et al. 2011) is mediated via progesterone receptor (PR) activity in the endometrial stroma and not directly via the epithelial PR (Kurita et al. 1998), very little is known about early molecular changes which contribute to the development of this disease. Thus, we applied the FEM algorithm to an Illumina 27 k DNAm data set consisting of 64 endometroid endometrial cancers and 23 normal endometrial samples from cancer-free women (Jones et al. 2013), and an unmatched Affymetrix (Human Genome 133 Plus 2.0) gene expression data set consisting of 79 endometrioid endometrial cancers and 12 normal samples from the atrophic endometrium [GSE17025, (Day et al. 2011)]. Statistics of differential methylation and differential expression were derived using a common moderated t-statistic framework (Smyth 2004). The distribution of differential expression t-statistics was then scaled to ensure that the statistics of differential methylation and differential expression had similar variance. This was done to ensure that one data type would not overly bias the inference of modules. One hundred seeds were then selected as those with the highest ranked combined t-statistic

$$t_i = H\left(t_i^{(D)}\right)H\left(-t_i^{(R)}\right)\left|t_i^{(D)} - t_i^{(R)}\right| \tag{7.4}$$

where due to the design of the Illumina 27 k array (overrepresented for promoter CpGs), we focused solely on seeds which were significantly hypermethylated and underexpressed in endometrial cancer. Application of FEMto the integrated data set led to the identification of a small number of FEMs, the most significant of which are shown in Table 7.2.

Of the 3 significant FEMs, the *HAND2* module is of special interest. First, *HAND2* itself emerges as a top ranked gene, significantly hypermethylated and underexpressed in endometrial cancer (Fig. 7.4a, b), a result which has been further validated in independent samples [see (Jones et al. 2013)]. Moreover, as shown in Jones et al. (2013), hypermethylation of *HAND2's* promoter is observed already in atypical hyper-plasias, a pre-cancerous lesion. Thus, *HAND2* methylation is an early event. Hence this epigenetic mark offers the potential to provide a non-invasive test for the early detection of endometrial cancer. This was assessed in (Jones et al. 2013) by means of DNA collected from vaginal swabs from women with endometrial cancer, resulting in an AUC of over 0.9, thus providing a test of high sensitivity and specificity (Jones et al. 2013).

Besides the clinical significance of *HAND2*, there is also mounting evidence for its biological significance in endometrial carcinogenesis. First, *HAND2* is a basic helix-loop-helix transcription factor and developmental regulator, as well as a stem cell PCGT (Lee et al. 2006; Srivastava et al. 1997). Thus, the observed hypermethylation in pre-cancerous lesions fits in with the hypotheses formulated earlier that epigenetic aberrations of PCGTs represent early oncogenic events. Second, *HAND2* is expressed in the normal endometrial stroma, with its key physiological function to suppress the production of fibroblast growth factors (FGF) that mediate

Table 7.2 The top three functional epigenetic modules identified in endometrial cancer

CpG(Seed)	EntrezID	Symbol	Size	Mod	P	Members
cg02622316	9753	ZNF96	20	1.42	0.01	ZNF96 OGT RBAK RREB1 PARP12
cg01580681	9464	HAND2	30	1.69	0.003	HAND2 HEY2 PHOX2A GATA4
cg05902852	9863	MAGI2	16	2.65	0.002	MAGI2 PTEN DLL1 CTNND2 TGFA

We list the CpG ID of the seed gene used, the corresponding Entrez gene ID and gene symbol, the size of resulting FEMmodule, its overall modularity (average weight density), its significance *P*-value (as estimated from 1000 Monte-Carlo runs), and some of the genes in the FEM

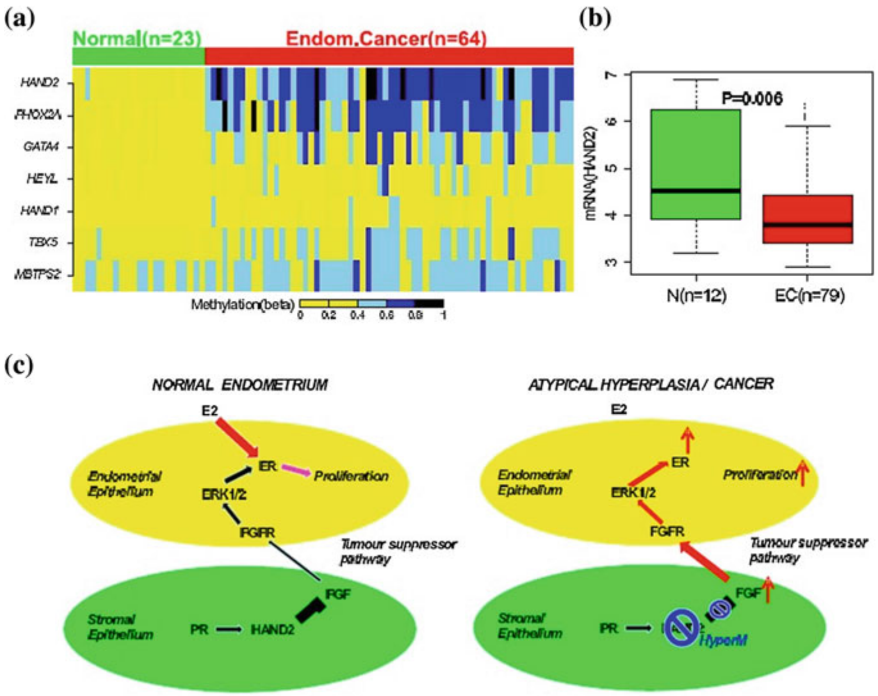


Fig. 7.4 **a** Heatmap of DNAm levels of the significantly associated members of the *HAND2* functional epigenetic module. **b** The concomitant underexpression observed for *HAND2* identifying it as the key driver gene. **c** Hypothesized role of *HAND2* in mediating the tumour suppressive effect of stromal progesterone receptor (*PR*) on the endometrial epithelial cells, which, in atypical hyperplasias/endometrial cancer, becomes disrupted through DNA hypermethylation of *HAND2's* promoter leading to *HAND2* silencing

the paracrine mitogenic effects of estrogen on the endometrial epithelium (Fig. 7.4c) (Li et al. 2011). Finally, *HAND2* is regulated by progesterone and is integral for the progesterone-mediated suppression of estrogen-induced pathways (Fig. 7.4c) (Bagchi et al. 2005; Dassen et al. 2007). Thus, it is plausible that hypermethylation induced silencing of *HAND2* prevents the suppression of FGFs by PR, thus resulting in increased FGF mediated paracrine signaling between the endometrial stromal and epithelial cells. This, in turn, could render the epithelial cells hypersensitive to estrogen exposure through overexpression of the estrogen receptor (Fig. 7.4c). In other words, *HAND2* methylation could shut down the tumour suppressive effect of PR, thus allowing overactivation of the oncogenic estrogen pathway. Since *HAND2* methylation is an early event in endometrial carcinogenesis, the associated silencing could thus underpin an increased susceptibility and risk to endometrial cancer.

7.5 Conclusions

In this chapter we have advocated a systems-epigenomic approach for identifying biomarkers associated with common phenotypes. We have provided a rationale for integrating DNA methylation data with a human interactome network, and presented substantial evidence that such a systems approach can provide novel insights in the epigenomic context. Indeed, using ageing as a proof of principle, we have demonstrated the existence of differential methylation hotspots associated with age, and which target key stem cell differentiation pathways (West et al. 2013). Thus, the epigenetic deregulation of differentiation pathways could result in impaired differentiation of stem and progenitor cells, in line with recent observations (Beerman et al. 2010, 2013; Brack and Rando 2007).

Importantly, we also applied our systems-epigenomic approach to endometrial cancer, a cancer strongly associated with environmental (non-genetic) risk factors, in order to identify functional epigenetic drivers. Remarkably, our approach revealed the existence of epigenetically deregulated functional hotspots, implicating *HAND2* as a key tumour suppressor in endometrial carcinogenesis. As we have seen, its putative tumour-suppressive role in endometrial cancer is entirely consistent with its role of mediating the tumour suppressive effects of PR on the oncogenic estrogen receptor pathways. *HAND2* methylation was shown to be an early event and provided a test of high sensitivity and specificity for the detection of early stage endometrial cancer (Jones et al. 2013). Thus, this chapter provides an example of "systems-medicine", whereby application of a computational systems method has enabled the identification and development of an early detection tool for endometrial cancer. We envisage that the specific FEM algorithmic framework presented here will be of interest to the disease epigenomics community at large. The statistical framework used is flexible, and will allow further and more complex integrative analysis of multi-dimensional cancer genomic data.

Acknowledgments AET was supported by a Heller Research Fellowship. The authors wish to thank everyone involved in the endometrial cancer project, specially Allison Jones and Helga Salvesen for collaborative work on this project.

References

Amant F, Moerman P, Neven P, Timmerman D, Van Limbergen E, Vergote I. Endometrial cancer. Lancet. 2005;366(9484):491–505.

Amatu A, Sartore-Bianchi A, Moutinho C, Belotti A, Bencardino K, Chirico G, Cassingena A, Rusconi F, Esposito A, Nichelatti M, Esteller M, Siena S. Promoter CpG island hypermethylation of the DNA repair enzyme mgmt predicts clinical response to dacarbazine in a phase ii study for metastatic colorectal cancer. Clin Cancer Res. 2013;19(8):2265–72.

Bagchi IC, Li Q, Cheon YP, Mantena SR, Kannan A, Bagchi MK. Use of the progesterone receptor antagonist ru 486 to identify novel progesterone receptor-regulated pathways in implantation. Semin Reprod Med. 2005;23(1):38–45.

Baylin SB, Ohm JE. Epigenetic gene silencing in cancer—a mechanism for early oncogenic pathway addiction? Nat Rev Cancer. 2006;6(2):107–16.

Beerman I, Bhattacharya D, Zandi S, Sigvardsson M, Weissman IL, Bryder D, Rossi DJ. Functionally distinct hematopoietic stem cells modulate hematopoietic lineage potential during aging by a mechanism of clonal expansion. Proc Natl Acad Sci USA. 2010;107(12):5465–70.

Beerman I, Bock C, Garrison BS, Smith ZD, Gu H, Meissner A, Rossi DJ. Proliferation-dependent alterations of the DNA methylation landscape underlie hematopoietic stem cell aging. Cell Stem Cell. 2013;12(4):413–25.

Beisser D, Klau GW, Dandekar T, Mller T, Dittrich MT. Bionet: an r-package for the functional analysis of biological networks. Bioinformatics. 2010;26(8):1129–30.

Berman BP, Weisenberger DJ, Aman JF, Hinoue T, Ramjan Z, Liu Y, Noushmehr H, Lange CP, van Dijk CM, Tollenaar RA, Van Den Berg D, Laird PW. Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. Nat Genet. 2011;44(1):40–6.

Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, Jaenisch R, Wagschal A, Feil R, Schreiber SL, Lander ES. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell. 2006;125(2):315–26.

Bhardwaj N, Lu H. Correlation between gene expression profiles and protein–protein interactions within and across genomes. Bioinformatics. 2005;21(11):2730–8.

Bhardwaj N, Lu H. Co-expression among constituents of a motif in the protein–protein interaction network. J Bioinform Comput Biol. 2009;7(1):1–17.

Bibikova M, Fan JB. Genome-wide DNA methylation profiling. Wiley Interdiscip Rev Syst Biol Med. 2010;2(2):210–23.

Bibikova M, Le J, Barnes B, Saedinia-Melnyk S, Zhou L, Shen R, Gunderson KL. Genome-wide DNA methylation profiling using Infinium$^{®}$ assay. Epigenomics. 2009;1(1):177–200.

Brack AS, Conboy MJ, Roy S, Lee M, Kuo CJ, Keller C, Rando TA. Increased Wnt signaling during aging alters muscle stem cell fate and increases fibrosis. Science. 2007;317(5839): 807–10.

Brack AS, Rando TA. Intrinsic changes and extrinsic influences of myogenic stem cell function during aging. Stem Cell Rev. 2007;3(3):226–37.

Brawley OW. Avoidable cancer deaths globally. CA Cancer J Clin. 2011;61(2):67–8.

Brennan K, Garcia-Closas M, Orr N, Fletcher O, Jones M, Ashworth A, Swerdlow A, Thorne H, Riboli E, Vineis P, Dorronsoro M, Clavel-Chapelon F, Panico S, Onland-Moret NC, Trichopoulos D, Kaaks R, Khaw KT, Brown R, Flanagan JM. Intragenic ATM methylation in peripheral blood DNA as a biomarker of breast cancer risk. Cancer Res. 2012;72(9):2304–13.

Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008;455(7216):1061–8.

Chen J, Chen L, Shen B. Identification of network biomarkers for cancer diagnosis. In: Wang X, editor. Bioinformatics of human proteomics, vol. Translational Bioinformatics vol. 3. Netherlands: Springer; 2013. p. 257–75.

Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. Mol Syst Biol. 2007;3:140.

Clauset A, Newman MEJ, Moore C. Finding community structure in very large networks. Phys Rev E 2004;70(6):066111. doi:10.1103/PhysRevE.70.066111.

Dassen H, Punyadeera C, Kamps R, Klomp J, Dunselman G, Dijcks F, de Goeij A, Ederveen A, Groothuis P. Progesterone regulation of implantation-related genes: new insights into the role of oestrogen. Cell Mol Life Sci. 2007;64(7–8):1009–32.

Day RS, McDade KK, Chandran UR, Lisovich A, Conrads TP, Hood BL, Kolli VS, Kirchner D, Litzi T, Maxwell GL. Identifier mapping performance for integrating transcriptomics and proteomics experimental results. BMC Bioinform. 2011;12:213.

Deaton AM, Bird A. CpG islands and the regulation of transcription. Genes Dev. 2011;25(10): 1010–22.

deVos T, Tetzner R, Model F, Weiss G, Schuster M, Distler J, Steiger KV, Grtzmann R, Pilarsky C, Habermann JK, Fleshner PR, Oubre BM, Day R, Sledziewski AZ, Lofton-Day C. Circulating methylated SEPT9 DNA in plasma is a biomarker for colorectal cancer. Clin Chem. 2009;55(7):1337–46.

Dittrich MT, Klau GW, Rosenwald A, Dandekar T, Mller T. Identifying functional modules in protein–protein interaction networks: an integrated exact approach. Bioinformatics. 2008;24(13):i223–31.

Feinberg AP, Ohlsson R, Henikoff S. The epigenetic progenitor origin of human cancer. Nat Rev Genet. 2006;7(1):21–33.

Feinberg AP, Vogelstein B. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. Nature. 1983;301(5895):89–92.

Fraga MF, Agrelo R, Esteller M. Cross-talk between aging and cancer: the epigenetic language. Ann N Y Acad Sci. 2007;1100:60–74.

Fraga MF, Esteller M. Epigenetics and aging: the targets and the marks. Trends Genet. 2007;23(8):413–8.

Gruetzmann R, Molnar B, Pilarsky C, Habermann JK, Schlag PM, Saeger HD, Miehlke S, Stolz T, Model F, Roblick UJ, Bruch HP, Koch R, Liebenberg V, Devos T, Song X, Day RH, Sledziewski AZ, Lofton-Day C. Sensitive detection of colorectal cancer in peripheral blood by septin 9 DNA methylation assay. PLoS ONE. 2008;3(11):e3759.

Hanahan D. The hallmarks of cancer revisited. Ann Oncol. 2012;23(9).

Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646–74.

Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S, Klotzle B, Bibikova M, Fan JB, Gao Y, Deconde R, Chen M, Rajapakse I, Friend S, Ideker T, Zhang K. Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol Cell. 2013;49(2):359–67.

Hansen KD, Timp W, Bravo HC, Sabunciyan S, Langmead B, McDonald OG, Wen B, Wu H, Liu Y, Diep D, Briem E, Zhang K, Irizarry RA, Feinberg AP. Increased methylation variation in epigenetic domains across cancer types. Nat Genet. 2011;43(8):768–75.

Heyn H, Esteller M. DNA methylation profiling in the clinic: applications and challenges. Nat Rev Genet. 2012;13(10):679–92.

Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, Cui H, Gabo K, Rongione M, Webster M, Ji H, Potash JB, Sabunciyan S, Feinberg AP. The human colon cancer methylome shows similar hypo-and hypermethylation at conserved tissue-specific CpG island shores. Nat Genet. 2009;41(2):178–86.

Issa JP. Epigenetic variation and cellular Darwinism. Nat Genet. 2011;43(8):724–6.

Issa JP, Ahuja N, Toyota M, Bronner MP, Brentnall TA. Accelerated age-related CpG island methylation in ulcerative colitis. Cancer Res. 2001;61(9):3573–7.

Jones A, Teschendorff AE, Li Q, Hayward JD, Kannan A, Mould T, West J, Zikan M, Cibula D, Fiegl H, Lee SH, Wik E, Hadwin R, Arora R, Lemech C, Turunen H, Pakarinen P, Jacobs IJ, Salvesen HB, Bagchi MK, Bagchi IC, Widschwendter M. Role of DNA methylation and epigenetic silencing of hand 2 in endometrial cancer development. PLoS Med. 2013;10(11): e1001551. doi:10.1371/journal.pmed.1001551.

Jones PA, Baylin SB. The epigenomics of cancer. Cell. 2007;128(4):683–92.

Kurita T, Young P, Brody JR, Lydon JP, O'Malley BW, Cunha GR. Stromal progesterone receptors mediate the inhibitory effects of progesterone on estrogen-induced uterine epithelial cell deoxyribonucleic acid synthesis. Endocrinology. 1998;139(11):4708–13.

Lechner M, Fenton T, West J, Wilson G, Feber A, Henderson S, Thirlwell C, Di-bra HK, Jay A, Butcher L, Chakravarthy AR, Gratrix F, Patel N, Vaz F, O'Flynn P, Kalavrezos N, Teschendorff AE, Boshoff C, Beck S. Identification and functional validation of HPV-mediated hypermethylation in head and neck squamous cell carcinoma. Genome Med. 2013;5(2):15.

Lee TI, Jenner RG, Boyer LA, Guenther MG, Levine SS, Kumar RM, Cheva-lier B, Johnstone SE, Cole MF, Isono K, Koseki H, Fuchikami T, Abe K, Murray HL, Zucker JP, Yuan B, Bell GW, Herbolsheimer E, Hannett NM, Sun K, Odom DT, Otte AP, Volkert TL, Bartel DP, Melton DA, Gifford DK, Jaenisch R, Young RA. Control of developmental regulators by polycomb in human embryonic stem cells. Cell. 2006;125(2):301–13.

Li Q, Kannan A, DeMayo FJ, Lydon JP, Cooke PS, Yamagishi H, Srivastava D, Bagchi MK, Bagchi IC. The antiproliferative action of progesterone in uterine epithelium is mediated by hand2. Science. 2011;331(6019):912–6.

Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, Pukkala E, Skytthe A, Hemminki K. Environmental and heritable factors in the causation of cancer—analyses of cohorts of twins from Sweden, Denmark, and Finland. N Engl J Med. 2000;343(2):78–85.

Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar AH, Thomson JA, Ren B, Ecker JR. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009;462(7271):315–22.

Lofton-Day C, Model F, Devos T, Tetzner R, Distler J, Schuster M, Song X, Lesche R, Liebenberg V, Ebert M, Molnar B, Grtzmann R, Pilarsky C, Sledziewski A. DNA methylation biomarkers for blood-based colorectal cancer screening. Clin Chem. 2008;54(2):414–23.

Maegawa S, Hinkal G, Kim HS, Shen L, Zhang L, Zhang J, Zhang N, Liang S, Donehower LA, Issa JP. Widespread and tissue specific age-related DNA methylation changes in mice. Genome Res. 2010;20(3):332–40.

Maiese K, Li F, Chong ZZ, Shang YC. The Wnt signaling pathway: aging gracefully as a protectionist? Pharmacol Ther. 2008;118(1):58–81.

Newman ME. Modularity and community structure in networks. Proc Natl Acad Sci USA. 2006;103(23):8577–82.

Ohm JE, McGarvey KM, Yu X, Cheng L, Schuebel KE, Cope L, Mohammad HP, Chen W, Daniel VC, Yu W, Berman DM, Jenuwein T, Pruitt K, Sharkis SJ, Watkins DN, Herman JG, Baylin SB. A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. Nat Genet. 2007;39(2):237–42.

Rakyan VK, Down TA, Maslau S, Andrew T, Yang TP, Beyan H, Whittaker P, Mc-Cann OT, Finer S, Valdes AM, Leslie RD, Deloukas P, Spector TD. Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. Genome Res. 2010;20(4):434–9.

Reichardt J, Bornholdt S. Statistical mechanics of community detection. Phys Rev E. 2006;74:016110. doi:10.1103/PhysRevE.74.016110.

Sandoval J, Heyn H, Moran S, Serra-Musach J, Pujana MA, Bibikova M, Esteller M. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. Epigenetics. 2011;6(6):692–702.

Sawyers CL. The cancer biomarker problem. Nature. 2008;452(7187):548–52.

Schlesinger Y, Straussman R, Keshet I, Farkash S, Hecht M, Zimmerman J, Eden E, Yakhini Z, Ben-Shushan E, Reubinoff BE, Bergman Y, Simon I, Cedar H. Polycombmediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. Nat Genet. 2007;39(2):232–6.

Schouten LJ, Goldbohm RA, van den Brandt PA. Anthropometry, physical activity, and endometrial cancer risk: results from the Netherlands cohort study. J Natl Cancer Inst. 2004;96(21):1635–8.

Selamat SA, Chung BS, Girard L, Zhang W, Zhang Y, Campan M, Siegmund KD, Koss MN, Hagen JA, Lam WL, Lam S, Gazdar AF, Laird-Offringa IA. Genome-scale analysis of DNA methylation in lung adenocarcinoma and integration with mRNA expression. Genome Res. 2012;22(7):1197–211.

Shen H, Laird PW. Interplay between the cancer genome and epigenome. Cell. 2013;153(1):38–55.

Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. 2004;**3**(Article3).

Srivastava D, Thomas T, Lin Q, Kirby ML, Brown D, Olson EN. Regulation of cardiac mesodermal and neural crest development by the bHLH transcription factor, dHAND. Nat Genet. 1997;16(2):154–60.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA. 2005;102(43):15545–50.

Suzuki H, Toyota M, Kondo Y, Shinomura Y. Inflammation-related aberrant patterns of DNA methylation: detection and role in epigenetic deregulation of cancer cell transcriptome. Methods Mol Biol. 2009;512:55–69.

Taylor IW, Linding R, Warde-Farley D, Liu Y, Pesquita C, Faria D, Bull S, Paw-son T, Morris Q, Wrana JL. Dynamic modularity in protein interaction networks predicts breast cancer outcome. Nat Biotechnol. 2009;27(2):199–204.

Teschendorff AE, Jones A, Fiegl H, Sargent A, Zhuang JJ, Kitchener HC, Wid-schwendter M. Epigenetic variability in cells of normal cytology is associated with the risk of future morphological transformation. Genome Med. 2012;4(3):24.

Teschendorff AE, Menon U, Gentry-Maharaj A, Ramus SJ, Gayther SA, Apostolidou S, Jones A, Lechner M, Beck S, Jacobs IJ, Widschwendter M. An epigenetic signature in peripheral blood predicts active ovarian cancer. PLoS ONE. 2009;4(12):e8274.

Teschendorff AE, Menon U, Gentry-Maharaj A, Ramus SJ, Weisenberger DJ, Shen H, Campan M, Noushmehr H, Bell CG, Maxwell AP, Savage DA, Mueller-Holzner E, Marth C, Kocjan G, Gayther SA, Jones A, Beck S, Wagner W, Laird PW, Jacobs IJ, Widschwendter M. Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. Genome Res. 2010;20(4):440–6.

Teschendorff AE, Widschwendter M. Differential variability improves the identification of cancer risk markers in DNA methylation studies profiling precursor cancer lesions. Bioinformatics. 2012;28(11):1487–94.

Ulitsky I, Krishnamurthy A, Karp RM, Shamir R. DEGAS: de novo discovery of dysregulated pathways in human diseases. PLoS One. 2010;5(10):e13367.

Ulitsky I, Shamir R. Identification of functional modules using network topology and high-throughput data. BMC Syst Biol. 2007;1:8.

Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. Cancer genome landscapes. Science. 2013;339(6127):1546–58.

Wen B, Wu H, Loh YH, Briem E, Daley GQ, Feinberg AP. Euchromatin islands in large heterochromatin domains are enriched for CTCF binding and differentially DNA-methylated regions. BMC Genomics. 2012;13:566.

West J, Beck S, Wang X, Teschendorff AE. An integrative network algorithm identifies age-associated differential methylation interactome hotspots targeting stem-cell differentiation pathways. Sci Rep. 2013;3:1630.

West J, Bianconi G, Severini S, Teschendorff AE. Differential network entropy reveals cancer system hallmarks. Sci Rep. 2012;2:802.

Widschwendter M, Fiegl H, Egle D, Mueller-Holzner E, Spizzo G, Marth C, Weisen-berger DJ, Campan M, Young J, Jacobs I, Laird PW. Epigenetic stem cell signature in cancer. Nat Genet. 2007;39(2):157–8.

Xu X, Su S, Barnes VA, De Miguel C, Pollock J, Ownby D, Shi H, Zhu H, Snieder H, Wang X. A genome-wide methylation study on obesity: differential variability and differential methylation. Epigenetics. 2013;8(5).

Xu Z, Bolick SC, Deroo LA, Weinberg CR, Sandler DP, Taylor JA. Epigenome-wide association study of breast cancer using prospectively collected sister study samples. J Natl Cancer Inst. 2013.

Yang S, Thiel KW, Leslie KK. Progesterone: the ultimate endometrial tumor suppressor. Trends Endocrinol Metab. 2011;22(4):145–52.

Zhuang J, Jones A, Lee SH, Ng E, Fiegl H, Zikan M, Cibula D, Sargent A, Salvesen HB, Jacobs IJ, Kitchener HC, Teschendorff AE, Widschwendter M. The dynamics and prognostic potential of DNA methylation changes at stem cell gene loci in women's cancer. PLoS Genet. 2012;8(2):e1002517.

Chapter 8
Identification of Cancer MicroRNA Biomarkers Based on miRNA–mRNA Network

Wenyu Zhang and Bairong Shen

Abstract It has been previously reported that miRNA regulations were involved in various biological processes. The deregulation activities of microRNA regulators potentially contribute to the pathopoiesis of various kinds of human cancers, and are candidate biomarkers for cancer diagnosis and prognosis. Until now, enormous studies have been conducted to explore potential miRNA biomarkers for different types of cancers. In this chapter, we will first provide a brief introduction about miRNAs biogenesis and their involvement in cancer pathopoiesis, and then reviewed the advances on current available miRNA profiling technologies. Then concise text will be exploited to describe the traditional experiment-dominate approaches for miRNA biomarker discovery. In the next part, intensive efforts are made on the review and summarization of miRNA–mRNA network based computational methods for the discovery of potential miRNA biomarkers. Afterwards, collect and list exsiting online databases relating to cancer miRNA biomarker discovery. Finally, we propose the perspective directions on this research area, and conclude the main context in this chapter.

Keywords Cancer · miRNA biomarker · miRNA–mRNA network

W. Zhang (✉) · B. Shen
Center for Systems Biology, Soochow University, P.O. Box 206 No.1 Shizi Street,
Suzhou 215006 Jiangsu, China
e-mail: michael_0214@126.com

B. Shen
e-mail: bairong.shen@suda.edu.cn

B. Shen (ed.), *Bioinformatics for Diagnosis, Prognosis and Treatment of Complex Diseases*, Translational Bioinformatics 4, DOI: 10.1007/978-94-007-7975-4_8,
© Springer Science+Business Media Dordrecht 2013

8.1 Introduction

MicroRNAs (miRNAs) are a class of small non-coding RNAs, approximately 22 nucleotides in length, which regulate gene expression at the post-transcriptional level through translation inhibition or mRNA cleavage (Bartel 2004). Since the discovery of the first miRNA (i.e., lin-4) in *C.elegans* (Lee et al. 1993; Wightman et al. 1993), enormous studies have been conducted to explore latent miRNA structures in a series of organisms, from viruses to advanced mammalians. Currently, within the most popular miRNA repository—miRBase (Release 20, June 2013) (Griffiths-Jones 2004), there have been 24,521 entries representing hairpin precursor miRNAs, expressing 30,424 mature miRNA products, among 206 distinct species. It is reported that most miRNAs are independently encoded in intergenic regions (Lee et al. 2004) or co-encoded within intron regions of other "host" protein-coding genes (Lin et al. 2008). Also, recent studies implicated the transfer RNAs (tRNAs) might be another origin for miRNA biogenesis (Schopman et al. 2010; Maute et al. 2013). The cardinal mechanism for the biological functions of miRNAs is to bind the 3'UTR (un-translated region) of their target genes through imperfect base pairing in animals (Reinhart et al. 2000), or perfect base pairing in plants (Dugas and Bartel 2004). Generally, the miRNA binding to its target genes will induce the mRNA degradation or protein translation inhibition, albeit several studies have shown its effect on the stabilization of target transcripts (Place et al. 2008).

Until now, there are more than 2,500 mature miRNAs identified in humans, which have potential to approximately regulate 33 % of human protein-coding genes (Lewis et al. 2005). It has been recently shown that miRNA regulations are involved in a wide variety of cellular processes, from cell proliferation, differentiation, development, to apoptosis (Ambros 2004; Bartel 2004). The alterations in miRNA expression have been associated with the pathogenesis and procession of various kinds of diseases, especially cancers (Jay et al. 2007). The abnormal miRNAs are reported to be capable as classifiers to distinguish tumour samples from Normal tissues (Raponi et al. 2009).

As implicated in the previous study (Bielekova and Martin 2004), some features were required to be as a disease biomarker: biological rational, clinical relevance, practicality and correlation with disease activity, etc. For miRNAs, they are involved in various biological processes and clinical related to the disease pathogenesis, as discussed in the above context. Besides, other advantageous features of miRNAs, such as manageable, durability and easily detected, make them more potential to be cancer biomarkers (Guerau-de-Arellano et al. 2012). In fact, enormous attempts have been made to explore candidate miRNA biomarkers in a series of cancers, such as breast cancer (Heneghan et al. 2010; Ramshankar and Krishnamurthy 2013), lung cancer (Gao et al. 2012), and gastric cancer (Li et al. 2012, 2013).

8.2 Advances of miRNA Expression Profiling Methods

The most straightforward approach to explore candidate cancer miRNA biomarkers should be the screening of abnormally expressed miRNAs between cancer tissues and normal tissues, via contemporary miRNA expression profiling detection platforms. Generally, there are two main experimental categories for these miRNA expression profiling techniques: high-throughput screening approaches and low-throughput detection methods. The first category consists of nucleic acid hybridization-based array technologies and cloning-sequencing based approaches, which can be used for simultaneous detection of many miRNAs in an individual experiment. The latter one includes small-scale detection methods, such as Northern Blot, Real time quantitative PCR (RT-qPCR), and in situ hybridization (ISH). The comparison information about these technologies is presented in Table 8.1.

8.2.1 High-Throughput Screening Approaches

Currently, microarray (also called Gene chip) technology may be the most widely used in the filed of transcript expression profiling detection. The implementation of this method is based on the hybridization of slide-localized miRNA-specific

Table 8.1 Current available miRNA expression profiling approaches

	Technical description	Throughput	Identification of novel miRNAs	Cost (per miRNA)
Microarray	Hybridization of fluorescent dye-labeled miRNAs on slide array with localized probes	High	No	Low
Beadarray	Flow cytometry detection of color beads coated with probes that bound to biotynilated miRNA sample	High	No	Low
Deep sequencing	Sequencing and quantification of all the miRNAs after the reverse transcription and PCR amplification	High	Yes	Low
Northern blot	Hybridization of labeled miRNAs on slide with localized probes	Low	No	High
Taqman-based assays	MicroRNAs are first reverse transcribed to cDNA, and then amplified and quantified	Low	No	High
ISH	Localization and quantification of a specific miRNA with a labeled probe in a portion of tissue	Low	No	High

Note This table was collected and modified from Guerau-de-Arellano et al. (2012)

probes (Babak et al. 2004). MiRNAs in the sample are first fluorescent dye-labeled, and then hybridized on slide array with glass-printed probes. After eluting unstable connections, the miRNA abundance is measured according its fluorescent luminance. The main advantage of this approach is parallelized detection of hundreds of miRNAs in an individual experiment.

Another large-scale miRNA expression profiling method is Beadarray technology. Unlike the classical microarray approach, it uses magnetic microspheres tagged with unique DNA sequence to identify the bead. This bead can specifically bind to a chimeric probe, which is utilized to recruit a specific miRNA into a complex. The fluorescent labeling of the complex is implemented with the biotin on the probe. Therefore, the abundance of miRNA expression can be quantified according to the amount of fluorescence on the microspheres through flow instruments.

The aforementioned array-based technologies can merely exploited for the expression profiling of known miRNAs. Deep sequencing techniques are novel miRNA profiling approaches that could detect the expression of all the miRNAs expressed in the target sample, even for miRNAs that are never previously reported. The detailed sequencing procedures may be diverse for different platforms. Generally, the first step for all these techniques is the generation of a miRNA library. During this process, miRNAs are ligated to 5′ and 3′ adaptors for reverse transcription and PCR amplification to generate this library. After that, the miRNAs in the library are further simultaneously sequenced and eventually abundance quantified. Due to their higher specificity and accuracy, the deep sequencing technologies are very promising to take over the dominating position of microarray technique in the area of transcript expression profiling.

8.2.2 Low-Throughput Detection Methods

Even with low flux, low-throughput miRNA expression profiling methods are considered to be more reliable than high-throughput techniques. This is a trade-off problem. More specifically, northern blot method is declaimed as the “gold standard” for characterizing miRNA expression (Ahmed 2007), due to its high specificity. The method is more like a mini-version of microarray. The miRNA abundance in the sample is determined by the hybridization signal from the binding complex of query miRNA and pre-set probes. Except for its low throughput, another defect of this method is low sensitivity for low-abundance samples.

Different from northern blot technique, RT-qPCR is a relative high sensitive approach for miRNA expression characterization. The basic idea is to make miRNA molecules amplifiable, through adding adapters to their fragment ends. After that, the amount of miRNA in the sample is relatively quantified. The PCR amplification process makes it accessible for low-abundance specimen measurement. This technology can be used for quantification of both miRNA precursors and mature miRNAs.

Another small-scale miRNA profiling method is in situ hybridization, which was invented by American cell biologist Joseph Grafton Gall in 1969 (Gall and Pardue 1969). This technology can be used to measure and localize miRNAs within tissue sections and cells.

With the recent progress in miRNA expression profiling detection technologies, there have been growing intense interests in cancer miRNA biomarker discovery studies, as documented by the numbers of related published literatures from NCBI pubmed searching engine in the last 10 years (Fig. 8.1). In the following parts, much more context will be spent on the descriptive review on these cancer miRNA biomarker discovery studies.

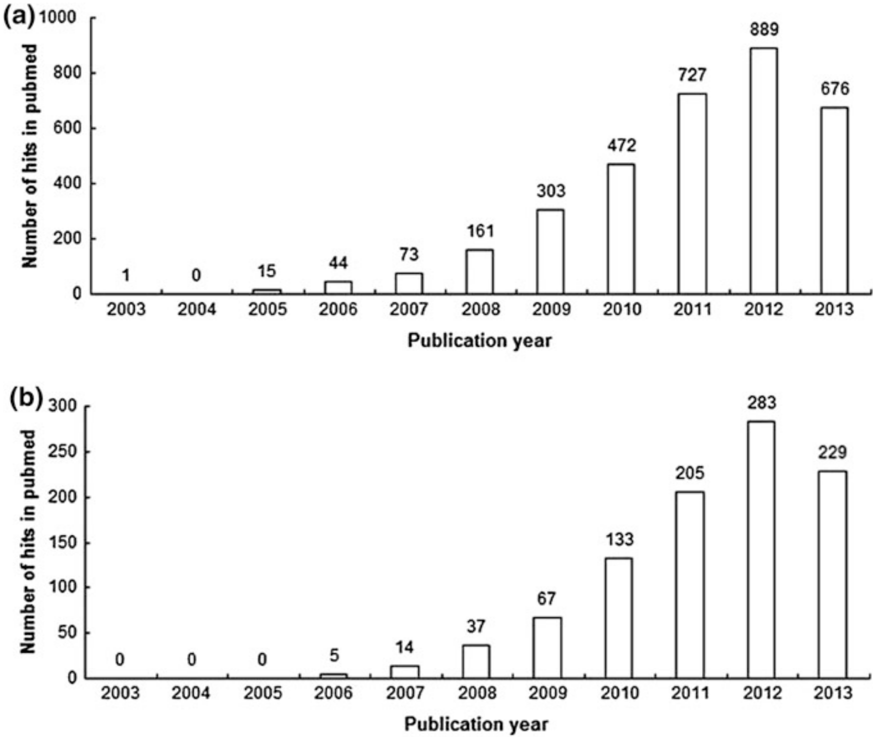


Fig. 8.1 Number of publications related to cancer miRNA biomarker discovery studies during the past decade. *Bars* represent the number of NCBI pubmed hits for query "(cancer[ti] OR carcinoma[ti] OR tumor[ti]) AND (miRNA*[ti] OR microRNA*[ti])" (*Pane a*) and "(cancer[ti] OR carcinoma[ti] OR tumor[ti]) AND (miRNA*[ti] OR microRNA*[ti]) AND (biomarker*[tiab] OR marker*[tiab])" (*Pane b*). The numbers for year 2013 are enumerated until 18th July

8.3 Traditional Approaches for Cancer miRNA Biomarker Discovery

The general procedures of traditional approaches for novel cancer miRNA biomarker discovery can be divided into three steps: (a) detection of differentially expressed miRNAs with high-throughput method (e.g., microarray) between cancer samples and control groups; (b) low-throughput technology (e.g., RT-qPCR) validation of outlier miRNAs detected above; (c) further confirmation of potential miRNA biomarkers on large-scale of case and control specimen via low-throughput experiments. The simplest bioinformatics tool for outlier miRNA identification in the first step is fold-change filtering (usually 2-fold). Following, many well-established statistical tools are employed for this issue, such as z-score, t-test, and Mann–Whitney test. Nevertheless, all these approaches do not take the heterogeneity of cancer samples into account. MacDonald et al. proposed a novel bioinformatics tool to infer chromosomal translocations only existing in the subset of disease samples (MacDonald and Ghosh 2006). Afterwards, this concept was implemented to generate several other outlier gene detection algorithms (Tibshirani and Hastie 2007; Wu 2007; Lian 2008). These available methods could be applied for outlier miRNA screening from high-throughput experiments.

Now, let us exemplify the general routine of traditional approaches for cancer miRNA biomarker discovery. Through the aforementioned procedures, the study of (Li et al. 2012) revealed that miR-199a-3p in plasma as a potential diagnostic biomarker for gastric cancer. Another research group conducted genome-wide miRNA expression profiles detection followed with Real-Time quantitative RT-PCR (qRT-PCR) assays on gastric cancer samples and normal samples, and discovered three elevated expressed miRNA (miR-187(*), miR-371-5p and miR-378) in gastric cancer. Further validation study showed that miR-378 alone could produce 87.5 % sensitivity and 70.73 % specificity in discriminating gastric cancer patients from healthy controls, thus this miRNA was a potential diagnosis biomarker for gastric cancer.

8.4 The Reconstruction of miRNA-mRNA Network

As the expression profiling data from high-throughput technologies (e.g., microarray) is always of high false positive rate, integrative analysis on high-throughput expression profiling data and miRNA-mRNA target network information may be a plausible approach for novel cancer miRNA biomarker discovery (Xu et al. 2011). The miRNA–mRNA network is more exactly a unidirectional graph, reflecting the regulation relationships from miRNAs to their target genes, as presented in Fig. 8.2. Due to the limit of current experimentally validated miRNA–mRNA target pairs (Sethupathy et al. 2006; Xiao et al. 2009; Hsu et al. 2011), the main resources for miRNA–mRNA network reconstruction are from computational

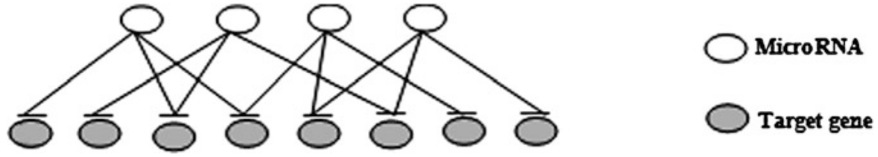


Fig. 8.2 Schematic graph of miRNA–mRNA target network

prediction. Even with potential false positive or negative cases, many of the predicted miRNA–mRNA interactions are confirmed to be credible (Arora and Simpson 2008). Usually, the main resources for miRNA–mRNA network reconstruction are the common miRNA–mRNA target pairs shared by the prediction results from multiple computational approaches, such as PicTar (Krek et al. 2005), miRanda (Enright et al. 2003), and TargetScan (Lewis et al. 2003, 2005). Besides, the negative expression correlations between miRNAs and their putative targets, computed from matched miRNA and mRNA expression profiles can also be applied for the refinement of miRNA–mRNA network reconstruction (Tran et al. 2008; Zhang et al. 2013).

8.5 Computational Based Approaches for Individual miRNA Biomarker Discovery

8.5.1 Based on Gene Expression Data and miRNA–mRNA Network Information

As the general miRNA functions are to regulate the gene expression at mRNA level, it is rational to infer miRNA deregulation from its target genes' expression level changes. Indeed, a number of studies have reported the reverse correlations between expressions of miRNAs and their target genes (Krutzfeldt et al. 2005; Wang and Wang 2006). Base on this theory, Cheng et al. proposed an algorithm to infer microRNA activities by combining gene expression data with miRNA–mRNA network information (Cheng and Li 2008). The basic idea of this algorithm is to analyze the expression changes of target genes for miRNAs. The activity of a miRNA will be inferred to elevated, if the expressions of its target genes tend to be down-regulated, and vice versa. Applying this approach, the cancer miRNA expression patterns can be deduced according to the gene expression profiles between cancer and normal samples.

Similarly, another approach referred as Co-inertia analysis (CIA) was proposed for this issue. CIA is a multivariate coupling approach, which was initially introduced for ecological research (Doledec and Chessel 1994; Dray et al. 2003). It was used to explore the correlation of two sets of variables from two linked data tables. Stephen et al. (Madden et al. 2010) applied this method to detect miRNA activity in different biological conditions. In this case, the two linked tables were gene microarray expression data, and a miRNA frequency table on the same set of genes. The two linked tables were performed two simultaneous non-symmetric correspondence analyses (NSCs), which reduced each data table in a low dimensional space, by projecting each variable on to axes which best discriminate the coordinates of the projected points. Then these two reduced tables were linked to associate the miRNA activity with biological samples. This methodology can be used to identify miRNA deregulation patterns that distinguish disease and normal groups.

8.5.2 Based on miRNA and Gene Expression Profiles and miRNA–mRNA Network Information

As the miRNA–mRNA target relationships are presented a simplified network style, the topological features of this network may be helpful for the identification of candidate cancer miRNA biomarkers. (Xu et al. 2011). introduced an approach based on the miRNA-target dysregulated network (MTDN) to prioritize candidate disease miRNAs, and applied this method to predict novel miRNA biomarkers in prostate cancer. In this methodology, miRNA expression and mRNA expression data, and miRNA–mRNA interaction data were combined to construct MTDN in tumor and non-tumor conditions. Then a support vector machine (SVM) was trained with considering the expression fold change and network topological features of known prostate cancer miRNAs and non-prostate cancer miRNAs in MTDN. Finally, the novel prostate cancer miRNA biomarkers were prioritized with this SVM and in vitro experimentally validated in prostate cancer cell lines. This study also showed the function synergism of miRNAs that were involved in the specific disease or biological process.

In contrary to the functional cooperation, (Zhang et al. 2013). declaimed another bizarre characteristic of cancer miRNAs—strong independent regulation power, which denoted the number of exclusively regulated genes for an individual miRNA. This research group also proposed a novel pipeline to infer candidate cancer miRNA biomarkers. The negative correlations from paired miRNA and mRNA expression profiles, along with computational prediction miRNA-mRNA target pairs were combined to generate a reliable miRNA–mRNA network. This network was further reduced to a sub-network, which only consisted of miRNA nodes exhibiting deregulation patterns from the miRNA expression profiles. In this sub-network, the independent regulation power was calculated for each miRNA. Ultimately, miRNAs with significant great independent regulation power were predicted as potential cancer miRNA biomarkers. The afterwards in vitro experiment validation and systematic analysis confirmed the accuracy of this approach.

8.6 Computational Based Approaches for miRNA Network Biomarker Discovery

8.6.1 The Discovery of Cancer-Related miRNA–mRNA Regulatory Modules

The concept of miRNA regulatory modules (mRMs) was first proposed by Yoon and De Micheli (2005a, b) to indicate groups of miRNAs and target genes that were completely connected in the sub-groups and functionally corporate in specific biological processes. This notion was afterwards applied for cancer studies.

Through integrative analysis on matched miRNA expression and mRNA expression profiles, two cancer mRMs discovery algorithms were proposed based on fuzzy decision tree model (Bonnet et al. 2010) and correspondence latent dirichlet allocation (Liu et al. 2010). Afterwards, (Jayaswal et al. 2011) introduced a clustering method to infer miRNA regulatory modules involved in cancers, through deducing miRNA activities from microRNA gene expression data and computational miRNA–mRNA target information. Considering the miRNA function patterns, there should be negative expression correlations for an individual miRNA and its target genes. Based on this theory, Joung et al. and Tran et al. raised two approaches to discover functional mRMs by combining paired miRNA and mRNA expression data and miRNA–mRNA binding information, through Population-based probabilistic learning method (Joung et al. 2007) and rule induction method (Tran et al. 2008), respectively. Finally, a computational framework for the discovery of cancer related miRNA-gene modules was proposed by simultaneous integration of multiple types of genomic data, including matched miRNA and mRNA expression profile, computational miRNA–mRNA target information, and gene–gene interaction network data, which was generated by integrating protein–protein interaction data with DNA–protein interaction data (Zhang et al. 2011). The brief summary about the aforementioned approaches about cancer miRNA–mRNA regulatory modules discovery is presented in Table 8.2.

8.6.2 The Discovery of Cancer-Related MicroRNA Network Biomarkers

The regulations of miRNAs on mRNAs are only a miniature for the whole biological regulatory network. The incorporation of other information, such as transcriptional factor (TF) regulations may be used for better understanding of specific biological process. Lu et al. designed a computational approach for identification of potential microRNA network biomarkers for the progression stages of gastric cancer (Lu et al. 2011). Within this approach, computational miRNA–mRNA

Table 8.2 Computational approaches on cancer mRMs discovery

Author	Input data	Main algorithm	Publication date	Availability
Yoon and De Micheli	miRNA–mRNA binding data	Identification of miRNA–mRNA complete sub-graph	2005	No
Bonnet et al.	Matched miRNA and mRNA expression data	Co-expression network analysis and a fuzzy decision tree model	2010	Yes
Liu et al.	Expression profiles of miRNAs and mRNAs; with (without) miRNA–mRNA target information	Correspondence latent dirichlet allocation	2010	No
Jayaswal et al.	Gene expression profiles; miRNA–mRNA binding information	Clustering method	2011	No
Joung et al.	miRNA–mRNA binding information; miRNA and mRNA expression profiles	Population-based probabilistic learning	2007	No
Tran et al.	miRNA and gene expression data; miRNA–target binding information	Rule induction	2008	No
Zhang et al.	Matched miRNA and mRNA expression profile; miRNA–mRNA target prediction result; gene–gene interaction network (integration of protein–protein interaction data and DNA–protein interaction data)	SNMNMF algorithm	2011	Yes

target information and TF-miRNA regulation data were combined to generate a novel miRNA network for each individual miRNA. The significance of each miRNA network was evaluated according to its GSEA score, and miRNA networks with higher GSEA scores than pre-set threshold were declaimed as potential gastric cancer miRNA network biomarkers.

Except for miRNA regulations, the ultimate expression levels of miRNAs are determined by many factors, e.g., TF regulations. Through integrating analysis on miRNA-gene binding information and TF-gene binding information, Tran et al. introduced a novel way to discover miR-TF regulatory modules in human genome. In this study, many identified modules have been previously reported to be involved in cancer genesis and development (Tran et al. 2010).

8.7 Databases on Potential Cancer miRNA Biomarkers

With the data accumulation from cancer–miRNA association studies, there have been a couple of online databases that collected the cancer–miRNA associations via text-mining approaches on previous publications. The brief summary about these databases can be referred in Table 8.3.

MiR2Disease is a manually curated database collecting the miRNA deregulation patterns in various human diseases, including cancers (Jiang et al. 2009). It provides the detailed information about miRNA–disease relationships, experimentally

Table 8.3 Databases on cancer–miRNA association

	Brief description	Statistics	Access link
miR2Disease	A manually curated database to provide comprehensive resource of miRNA deregulation in various human diseases, including cancers	349 miRNAs; 163 diseases	http://www.mir2disease.org/
dbDEMC	A publicly available database to collect differentially expressed MiRNAs in various human cancers from previous studies	607 miRNAs; 14 cancers	http://159.226.118.44/dbDEMC/index.html
PhenomiR	A manually curated database to provide differentially regulated miRNA expression information in various human diseases, including cancers	675 miRNAs; 145 diseases	http://mips.helmholtz-muenchen.de/phenomir/index.gsp
miRCancer	A comprehensive collection of miRNA expression profiles in various human cancers which are automatically extracted from published literatures	1300 miRNA; 151 cancers	http://mircancer.ecu.edu/
S-MED	A repository that describes the patterns of miRNA expression found in various human sarcoma tumor types	>700 miRNAs	http://www.oncomir.umn.edu/SMED/index.php
CC-MED	A repository that describes the patterns of miRNA expression found in human colon cancer	39 miRNAs	http://www.oncomir.umn.edu/colon/basic_search.php

validated miRNA targets, and corresponding literature references. Similarly, PhenomiR is another comprehensive repository of deregulation miRNA profiling data for different human diseases and biological processes (Ruepp et al. 2010, 2012). Based on self-defined text-mining rules, miRCancer (Xie et al. 2013) and dbDEMC (Yang et al. 2010) specially focus on the collection of cancer-related differentially expressed miRNAs information. More specifically, there have been also some established databases that merely provide miRNA expression profiling data on certain tumor type, such as S-MED (Sarver et al. 2010), CC-MED (Sarver et al. 2009), and others.

8.8 Future Directions

Although the current cancer miRNA studies have shed some light on our understanding of cancer genesis and development mechanisms, there is still a long way ahead in this new emerging research area. The heterogeneity of cancer is our main concern. Except the compensation of future more advanced miRNA expression detection technologies, another two research directions might also be potential solutions for this issue.

Recently, a new concept referred as personalized medicine (PM) has been proposed to indicate the customization of healthcare. The importance and urgency about PM has already been emphasized years ago (Long 2007). Therefore, the specific information about patients, such as race, genetic makeup, should also be integrated for the future discovery of cancer miRNA biomarkers.

Currently, network view shows that the disorder conditions (diseases, including cancers) may attribute to the deregulation of specific biological process, not simply the alteration of an individual biological molecule. Network biomarkers should be a better choice for cancer diagnosis. As reviewed above, there have been a couple of studies conducted for the identification of cancer miRNA network biomarkers. In the future, the integration of multi-layer information, such as genomic information, epigenomic information, and clinical information, is need for the discovery of miRNA network biomarkers.

8.9 Conclusions

This chapter summarizes current advances of miRNA expression detection technologies, the traditional approaches on cancer miRNA biomarker discovery based on these techniques. Computational-based methods on the identification of individual miRNA biomarker and miRNA network biomarker for cancer diagnosis and prognosis are also reviewed herein. Although studies of cancer miRNA biomarkers are still in their infancy, the evolving miRNA profiling measurement technologies, miRNA network information, and computational algorithms offer new insights on cancer mechanism investigation. We can expect the clinical application of miRNA biomarkers for the diagnosis, staging, and prognosis of cancers in the near future.

References

Ahmed FE. Role of miRNA in carcinogenesis and biomarker selection: a methodological view. Expert Rev Mol Diagn. 2007;7(5):569–603.

Ambros V. The functions of animal microRNAs. Nature. 2004;431(7006):350–5.

Arora A, Simpson DA. Individual mRNA expression profiles reveal the effects of specific microRNAs. Genome Biol. 2008;9(5):R82.

Babak T, Zhang W, Morris Q, Blencowe BJ, Hughes TR. Probing microRNAs with microarrays: tissue specificity and functional inference. RNA. 2004;10(11):1813–9.

Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell. 2004;116(2):281–97.

Bielekova B, Martin R. Development of biomarkers in multiple sclerosis. Brain J Neurol. 2004;127(Pt 7):1463–78.

Bonnet E, Tatari M, Joshi A, Michoel T, Marchal K, Berx G, Van de Peer Y. Module network inference from a cancer gene expression data set identifies microRNA regulated modules. PLoS One. 2010;5(4):e10162.

Cheng C, Li LM. Inferring microRNA activities by combining gene expression with microRNA target prediction. PLoS One. 2008;3(4):e1989.

Doledec S, Chessel D. Co-Inertia Analysis: an alternative method for studying species environment relationships. Freshw Biol. 1994;31(3):277–94.

Dray S, Chessel D, Thioulouse J. Co-inertia analysis and the linking of ecological data tables. Ecology. 2003;84(11):3078–89.

Dugas DV, Bartel B. MicroRNA regulation of gene expression in plants. Curr Opin Plant Biol. 2004;7(5):512–20.

Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS. MicroRNA targets in Drosophila. Genome Biol. 2003;5(1):R1.

Gall JG, Pardue ML. Formation and detection of RNA-DNA hybrid molecules in cytological preparations. Proc Natl Acad Sci USA. 1969;63(2):378–83.

Gao W, Lu X, Liu L, Xu J, Feng D, Shu Y. MiRNA-21: a biomarker predictive for platinum-based adjuvant chemotherapy response in patients with non-small cell lung cancer. Cancer Biol Ther. 2012;13(5):330–40.

Griffiths-Jones S. The microRNA registry. Nucleic acids Res. 2004;32(Database issue):D109–D111.

Guerau-de-Arellano M, Alder H, Ozer HG, Lovett-Racke A, Racke MK. miRNA profiling for biomarker discovery in multiple sclerosis: from microarray to deep sequencing. J Neuroimmunol. 2012;248(1–2):32–9.

Heneghan HM, Miller N, Kelly R, Newell J, Kerin MJ. Systemic miRNA-195 differentiates breast cancer from other malignancies and is a potential biomarker for detecting noninvasive and early stage disease. Oncologist. 2010;15(7):673–82.

Hsu SD, Lin FM, Wu WY, Liang C, Huang WC, Chan WL, Tsai WT, Chen GZ, Lee CJ, Chiu CM et al. miRTarBase: a database curates experimentally validated microRNA-target interactions. Nucleic Acids Res. 2011;39(Database issue):D163–D169.

Jay C, Nemunaitis J, Chen P, Fulgham P, Tong AW. miRNA profiling for diagnosis and prognosis of human cancer. DNA Cell Biol. 2007;26(5):293–300.

Jayaswal V, Lutherborrow M, Ma DD, Yang YH. Identification of microRNA-mRNA modules using microarray data. BMC genomics. 2011;12:138.

Jiang Q, Wang Y, Hao Y, Juan L, Teng M, Zhang X, Li M, Wang G, Liu Y. miR2Disease: a manually curated database for microRNA deregulation in human disease. Nucleic Acids Res. 2009;37(Database issue): D98–D104.

Joung JG, Hwang KB, Nam JW, Kim SJ, Zhang BT. Discovery of microRNA-mRNA modules via population-based probabilistic learning. Bioinformatics. 2007;23(9):1141–7.

Krek A, Grun D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, MacMenamin P, da Piedade I, Gunsalus KC, Stoffel M, et al. Combinatorial microRNA target predictions. Nat Genet. 2005;37(5):495–500.

Krutzfeldt J, Rajewsky N, Braich R, Rajeev KG, Tuschl T, Manoharan M, Stoffel M. Silencing of microRNAs in vivo with 'antagomirs'. Nature. 2005;438(7068):685–9.

Lee RC, Feinbaum RL, Ambros V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell. 1993;75(5):843–54.

Lee Y, Kim M, Han J, Yeom KH, Lee S, Baek SH, Kim VN. MicroRNA genes are transcribed by RNA polymerase II. EMBO J. 2004;23(20):4051–60.

Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell. 2005;120(1):15–20.

Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB. Prediction of mammalian microRNA targets. Cell. 2003;115(7):787–98.

Li C, Li JF, Cai Q, Qiu QQ, Yan M, Liu BY, Zhu ZG. MiRNA-199a-3p in plasma as a potential diagnostic biomarker for gastric cancer. Annals Surg Oncol. 2012.

Li C, Li JF, Cai Q, Qiu QQ, Yan M, Liu BY, Zhu ZG. MiRNA-199a-3p: a potential circulating diagnostic biomarker for early gastric cancer. J Surg Oncol. 2013;108(2):89–92.

Lian H. MOST: detecting cancer differential gene expression. Biostatistics. 2008;9(3):411–8.

Lin SL, Kim H, Ying SY. Intron-mediated RNA interference and microRNA (miRNA). Front Biosci : J Virtual Libr. 2008;13:2216–30.

Liu B, Liu L, Tsykin A, Goodall GJ, Green JE, Zhu M, Kim CH, Li J. Identifying functional miRNA-mRNA regulatory modules with correspondence latent dirichlet allocation. Bioinformatics. 2010;26(24):3105–11.

Long M. Side effects of Tamiflu: clues from an Asian single nucleotide polymorphism. Cell Res. 2007;17(4):309–10.

Lu L, Li Y, Li S. Computational identification of potential microRNA network biomarkers for the progression stages of gastric cancer. Int J Data Min Bioinform. 2011;5(5):519–31.

MacDonald JW, Ghosh D. COPA–cancer outlier profile analysis. Bioinformatics. 2006;22(23):2950–1.

Madden SF, Carpenter SB, Jeffery IB, Bjorkbacka H, Fitzgerald KA, O'Neill LA, Higgins DG. Detecting microRNA activity from gene expression data. BMC Bioinform. 2010;11:257.

Maute RL, Schneider C, Sumazin P, Holmes A, Califano A, Basso K, Dalla-Favera R. tRNA-derived microRNA modulates proliferation and the DNA damage response and is down-regulated in B cell lymphoma. Proc Natl Acad Sci USA. 2013;110(4):1404–9.

Place RF, Li LC, Pookot D, Noonan EJ, Dahiya R. MicroRNA-373 induces expression of genes with complementary promoter sequences. Proc Natl Acad Sci USA. 2008;105(5):1608–13.

Ramshankar V, Krishnamurthy A. Lung cancer detection by screening—presenting circulating miRNAs as a promising next generation biomarker breakthrough. Asian Pac J Cancer Prev : APJCP. 2013;14(4):2167–72.

Raponi M, Dossey L, Jatkoe T, Wu X, Chen G, Fan H, Beer DG. MicroRNA classifiers for predicting prognosis of squamous cell lung cancer. Cancer Res. 2009;69(14):5776–83.

Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE, Horvitz HR, Ruvkun G. The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. Nature. 2000;403(6772):901–6.

Ruepp A, Kowarsch A, Schmidl D, Buggenthin F, Brauner B, Dunger I, Fobo G, Frishman G, Montrone C, Theis FJ. PhenomiR: a knowledgebase for microRNA expression in diseases and biological processes. Genome Biol. 2010;11(1):R6.

Ruepp A, Kowarsch A, Theis F. PhenomiR: microRNAs in human diseases and biological processes. Methods Mol Biol. 2012;822:249–60.

Sarver AL, French AJ, Borralho PM, Thayanithy V, Oberg AL, Silverstein KA, Morlan BW, Riska SM, Boardman LA, Cunningham JM, et al. Human colon cancer profiles show differential microRNA expression depending on mismatch repair status and are characteristic of undifferentiated proliferative states. BMC Cancer. 2009;9:401.

Sarver AL, Phalak R, Thayanithy V, Subramanian S. S-MED: sarcoma microRNA expression database. Lab Inv; J Tech Methods Pathol. 2010;90(5):753–61.

Schopman NC, Heynen S, Haasnoot J, Berkhout B. A miRNA-tRNA mix-up: tRNA origin of proposed miRNA. RNA Biol. 2010;7(5):573–6.

Sethupathy P, Corda B, Hatzigeorgiou AG. TarBase: a comprehensive database of experimentally supported animal microRNA targets. RNA. 2006;12(2):192–7.

Tibshirani R, Hastie T. Outlier sums for differential gene expression analysis. Biostatistics. 2007;8(1):2–8.

Tran DH, Satou K, Ho TB. Finding microRNA regulatory modules in human genome using rule induction. BMC Bioinform. 2008;9(Suppl 12):S5.

Tran DH, Satou K, Ho TB, Pham TH. Computational discovery of miR-TF regulatory modules in human genome. Bioinformation. 2010;4(8):371–7.

Wang X, Wang X. Systematic identification of microRNA functions by combining target prediction and expression profiling. Nucleic Acids Res. 2006;34(5):1646–52.

Wightman B, Ha I, Ruvkun G. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. Cell. 1993;75(5):855–62.

Wu B. Cancer outlier differential gene expression detection. Biostatistics. 2007;8(3):566–75.

Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T. miRecords: an integrated resource for microRNA-target interactions. Nucleic Acids Res. 2009;37(Database issue):D105–D110.

Xie B, Ding Q, Han H, Wu D. miRCancer: a microRNA-cancer association database constructed by text mining on literature. Bioinformatics. 2013;29(5):638–44.

Xu J, Li CX, Lv JY, Li YS, Xiao Y, Shao TT, Huo X, Li X, Zou Y, Han QL, et al. Prioritizing candidate disease miRNAs by topological features in the miRNA target-dysregulated network: case study of prostate cancer. Mol Cancer Ther. 2011;10(10):1857–66.

Yang Z, Ren F, Liu C, He S, Sun G, Gao Q, Yao L, Zhang Y, Miao R, Cao Y, et al. dbDEMC: a database of differentially expressed miRNAs in human cancers. BMC Genomics. 2010;11(Suppl 4):S5.

Yoon S, De Micheli G. Prediction and analysis of human microRNA regulatory modules. In: Conference proceedings: annual international conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society conference; 2005a, vol 5, p. 4799–802.

Yoon S, De Micheli G . Prediction of regulatory modules comprising microRNAs and target genes. Bioinformatics. 2005b;21 Suppl 2:ii93–ii100.

Zhang S, Li Q, Liu J, Zhou XJ. A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules. Bioinformatics. 2011;27(13):i401–9.

Zhang WY, Zang J, Jing XH, Sun ZD, Yang DR, Guo F, Shen BR. Identification of candidate cancer miRNA biomarkers from miRNA regulatory network: with application to prostate cancer. RNA. 2013 (submitted).

Part III
Applications in Detection and Treatment of Complex Diseases

Chapter 9
Ubiquitin and Ubiquitin-Like Conjugations in Complex Diseases: A Computational Perspective

Tianshun Gao, Zexian Liu, Yongbo Wang and Yu Xue

Abstract As one class of most essential and common post-translational modifications (PTMs), ubiquitin and ubiquitin-like (Ub/UBL) conjugations play an important role in almost all aspects of biological processes, and aberrances in the conjugation systems are highly involved in numerous complex diseases. Identification of the Ub/UBL-associated enzymes, substrates and sites is fundamental for understanding the molecular mechanisms of Ub/UBL conjugations, and provides a potential reservoir for discovering disease biomarkers and drug targets. Besides experimental identifications, computational analysis of Ub/UBL conjugations has also emerged as an attractive field. In this chapter, we first summarized the cutting-edge experimental techniques in the large-scale identification of Ub/UBL conjugation substrates, and further emphasized the importance of computational efforts by introducing online databases and predictors for Ub/UBL conjugations. Although computational analysis of Ub/UBL conjugations is still immature, we believe more and more efforts will be paid in the near future.

Keywords Ubiquitin and ubiquitin-like conjugation · Ubiquitination · Sumoylation · Proteomics · Small cell lung cancer

T. Gao · Z. Liu · Y. Wang · Y. Xue (✉)
Department of Biomedical Engineering, College of Life Science and Technology, Huazhong University of Science and Technology, Luoyu road, 1037 Wuhan 430074, Hubei, China
e-mail: xueyu@hust.edu.cn

T. Gao
e-mail: gts.hust@gmail.com

Z. Liu
e-mail: lzx@mail.ustc.edu.cn

Y. Wang
e-mail: hust.wangyb@gmail.com

B. Shen (ed.), *Bioinformatics for Diagnosis, Prognosis and Treatment of Complex Diseases*, Translational Bioinformatics 4, DOI: 10.1007/978-94-007-7975-4_9,
© Springer Science+Business Media Dordrecht 2013

9.1 Introduction

During the past three decades, the ubiquitin-proteasome system (UPS) has been demonstrated to be critical for protein degradation in most cellular processes (Ciechanover 1994; Bedford et al. 2011; Geng et al. 2012). Ubiquitin (Ub) is a small 76aa protein that binds to target proteins and takes them for destruction through Ubiquitination (Ciechanover 1994), which labels mono- or poly-ubiquitin proteins to substrates via an E1 (Ub-activating enzyme)-E2 (Ub-conjugating enzyme)-E3 (Ub-protein ligase) cascade mechanism (Fig. 9.1a). Recently, more than ten Ub-like modifiers (UBLs) have also been identified, such as SUMO, NEDD8, ISG15, Apg8/12, FAT10, Urm1, UFM1 and Hub1 in eukaryotes, prokaryotic Ub-like protein (Pup) and archaeal SAMPs (Hochstrasser 2009; van der Veen and Ploegh 2012). The prokaryotic homologs of Ub, ThiS and MoaD, are potential antecedents of all Ub/UBL modifiers in eukaryotes (Iyer et al. 2006; van der Veen and Ploegh 2012). Analogous to Ub, most UBLs share a β-grasp fold and a C-terminal diglycine motif, and their conjugation processes, such as sumoylation (Fig. 9.1b) and pupylation (Fig. 9.1c), have a conserved enzyme cascade

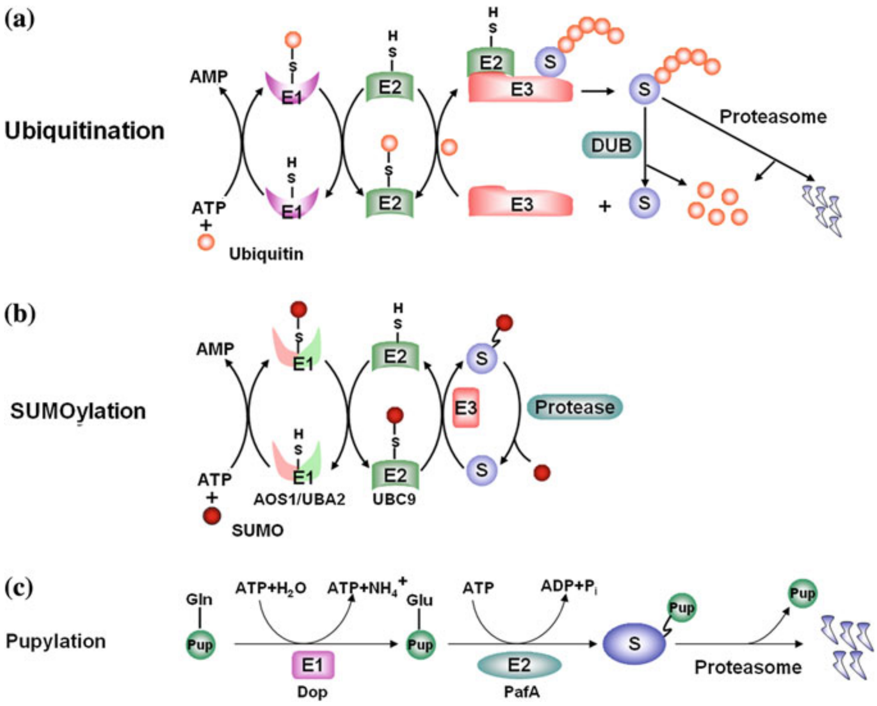


Fig. 9.1 The conjugation processes for **a** ubiquitination, **b** sumoylation, and **c** Pupylation. For ubiquitination, an E1-E2-E3 enzyme cascade mechanism was characterized, and Ub E3 ligases provide the major specificity for substrate recognition. However, SUMO E3 ligases are only cofactors that facilitate the sumoylation, while pupylation doesn't have E3s

mechanism (van der Veen and Ploegh 2012). Ub E3 ligases confer the major specificity of ubiquitination for recognizing substrates (Deshaies and Joazeiro 2009). However, SUMO E3 ligases are only cofactors that facilitate the conjugation of SUMO (Yunus and Lima 2006), and pupylation has only an E1–E2 cascade without any E3 ligases (Striebel et al. 2009). Substrates in the UPS pathway are ubiquitinated through three forms, mono-, multi- and poly-ubiquitination (Sadowski and Sarcevic 2010), while several UBL conjugations, such as SUMO, NEDD8 and SAMP, can also adopt analogical forms for targeting proteins (Ulrich 2008; Ohki et al. 2009; Humbard et al. 2010). Different forms can lead to different fates on substrates (Sadowski and Sarcevic 2010). Mono-ubiquitination affects the activity and location of substrates to be involved in histone regulation, endocytosis and membrane transport (Hicke 2001), while multi- and poly-ubiquitination mainly induce protein degradation as well as non-proteolytic functions (Ciechanover 1994; Rape et al. 2006; Chen and Sun 2009).

Ub and UBL conjugation pathways are implicated in diverse but essential biological functions. Cells usually use these pathways to select specific proteins for destruction, activation or other functions and ensure the fidelity of cellular processes (Ciechanover 1994; Chen and Sun 2009). Thus, aberrances in Ub/UBL conjugation pathways have been identified to be involved in numerous complex diseases (Dahlmann 2007; Bedford et al. 2011), including inflammation (Hochrainer and Lipp 2007; Coornaert et al. 2009), viral infection (Bogunovic et al. 2013), neurodegenerative disease (Hegde and Upadhya 2007; Lehman 2009; Mandel et al. 2009; Deng et al. 2013), cardiac disease (Sohns et al. 2010; Wang 2011), von Hippel-Lindau disease (Kaelin 2007) and several types of cancers (Bonacci et al. 2010; Irminger-Finger 2010; Linehan et al. 2010; Conrad et al. 2011; Escobar et al. 2011; Duncan et al. 2012). However, compared to phosphorylation, in which protein kinases occupied ~30 % of the drug discovery programs in pharmaceutical research and development, ubiquitination owned less than 1 % of drug design (Cohen and Tcherpakov 2010), and only one proteasome Inhibitor Bortezomib was approved currently (Chen et al. 2011). To target complex diseases, theoretically, any components of the UPS and UBL conjugation pathways, including E1s, E2s, E3s, DUBs and proteasomes, can be selected for targeting by small-molecule inhibitors. For example, RING E3s including BARD1 and SIAH (Chasapis and Spyroulias 2009; Irminger-Finger 2010; Wong and Moller 2013), HECT E3 s such as ITCH and SMURF1 (Scheffner and Staub 2007; Melino et al. 2008; Lin et al. 2013), DUBs such as A20 and UCHL1 (Singhal et al. 2008; Coornaert et al. 2009; Day and Thompson 2010), and proteasome subunits such as PSMA7 (Du et al. 2009), had been identified as potential biomarkers of complex diseases. More, inhibitors of several SCF E3 complexes, such as SCF$^{\mathrm{skp2}}$, SCF$^{\beta-\mathrm{TrCP1}}$, SCF$^{\mathrm{CDC4}}$, SCF$^{\mathrm{Met30}}$, have also been identified (Chen et al. 2008; Nakajima et al. 2008; Aghajan et al. 2010; Orlicky et al. 2010). The rapid progresses suggested that Ub/UBL conjugation pathways can be a great reservoir for discovering potential biomarkers and drug targets (Cohen and Tcherpakov 2010).

9.2 Advances in High-Throughput Proteomic Analysis of Ub/UBL Conjugations

Because Ub E3 ligases bind substrates at distinct regions and modify specific lysine residues (Bustos et al. 2012), the Ub-mediated proteasomal substrates can be detected by mutating lysines for poly-ubiquitin chain (Chau et al. 1989), substituting E3-substrate binding site (House et al. 2006) or eliminating all lysines of substrate can disrupt the ubiquitination (Bourgeois-Daigneault and Thibodeau 2012). Since high-affinity Ub antibody, linkage specific antibodies and Ub epitope-tags were developed, further studies were focused on the detection of Ub-conjugated substrates (Muller et al. 1988; Newton et al. 2008). For a substrate containing only one ubiquitinated lysine, a single K to R mutation is enough for identifying the ubiquitination site (Flick et al. 2004). However, for multi-ubiquitinated substrates, accurate identification of all ubiquitination sites needs both individual and combinatorial mutations (Zhong et al. 2005). Reintroducing lysine residues one by one into the lysineless mutant (K0) is also an alternative method for identifying multiple ubiquitination sites (Rufini et al. 2011). However, any attempts based on the mutagenesis can only identify one substrate and several ubiquitination sites at most in a single study (Flick et al. 2004; Zhong et al. 2005; Rufini et al. 2011).

In contrast with conventional studies, high-throughput characterization of ubiquitinated substrates provides a more comprehensive understanding of the ubiquitination dynamics and potential relationships between ubiquitinaton and other important cellular processes. Recently, the technologies of mass spectrometry-based proteomics have a significant improvement for the identification of ubiquitination sites (Jeram et al. 2009; Bustos et al. 2012). In the presence of trypsin, Ub-conjugated substrates can be cleaved into K-GG modified peptides (Fig. 9.2a), which can be regarded as ubiquitination signatures (Denis et al. 2007). Thus, the liquid chromatography-mass spectrometry (LC/MS) analysis can detect a mass shift of 114.043 Da, which represents the diglycine (GG) remnant of Ub (Shi et al. 2011) (Fig. 9.2a).

Analogous to Ub, NEDD8, ISG15 and Pup can also produce K-GG remnants with their C-terminal (K/R) GG sequences by the trypsin cleavage, whereas SUMO can’t because of the absence of a basic residue adjacent to the C-terminal GG motif (Kang and Yi 2011; Osula et al. 2012). Since the LC/MS identification can’t distinguish among K-GGs of Ub, NEDD8 and ISG15, adding MLN4924 but not interferon can effectively block NEDD8ylation and ISG15ylation for exclusively identifying ubiquitinated substrates (Kim et al. 2011; Zhao et al. 2013). However, if Ub was not tagged, only one or several ubiquitination sites of one purified substrate can be identified in vitro (Wang et al. 2005). Thus, with the improvement of Ub epitope-tagging strategies, large-scale analysis of K-GG peptides can be available by the trypsin digestion of hundreds of epitope-tagging Ub-conjugated substrates after in vivo enrichment and purification of Ub-conjugated substrates (Peng et al. 2003; Maor et al. 2007; Danielsen et al. 2011; Kim et al. 2011; Lee et al. 2011; Shi et al. 2011; Oshikawa et al. 2012; Osula et al. 2012; Starita et al. 2012).

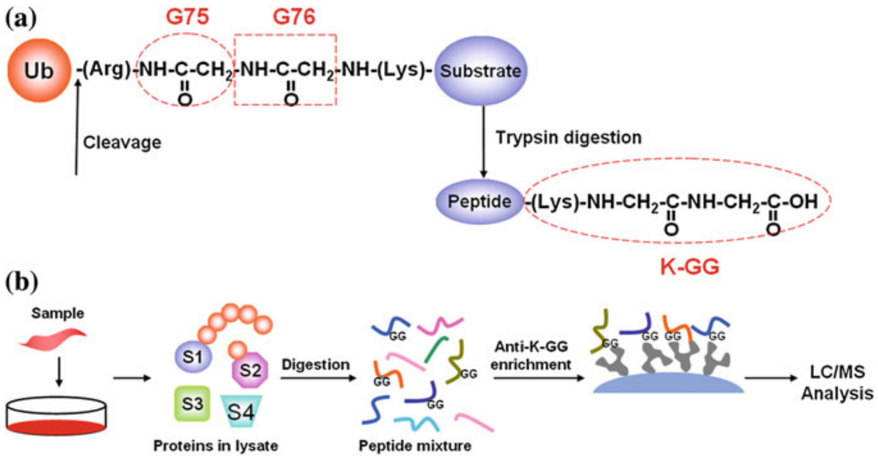


Fig. 9.2 Proteomic analysis of Ub/UBL conjugation substrates. **a** Ub-conjugated substrates can be cleaved into K-GG modified peptides by trypsin. **b** The direct enrichment of in vivo K-GG Peptides from samples has been an efficient approach for the large-scale identification of Ub/UBL conjugation sites

For example, Peng et al. (2003) identified 110 ubiquitination sites and 1,075 ubiquitinated substrates from yeast cells by using His$_6$-tagged Ub. Also, Maor et al. (2007) detected 85 ubiquitination sites and 294 Ub substrates from Arabidopsis cells with GST-tagged Ub. Furthermore, Meierhofer et al. (2008) characterized 44 ubiquitin acceptor sites and 669 ubiquitinated proteins in HeLa cells, using hexahistidine-biotin (HB)-fused Ub. In particular, Oshikawa et al. (2012) identified 1392 ubiquitination sites of 794 proteins in HEK293T cells, with His$_6$-tagged K0-Ub. Additionally, this strategy was also adopted for analyzing other UBL conjugations, such as pupylation, which can also generate -GG remnants for the high-throughput identification (Kang and Yi 2011). In fact, Festa et al. (2010) identified 55 pupylation sites from a single sample in *Mycobacterium tuberculosis* (Mtb). As the further improvement of MS techniques, the higher-throughput identification of K-GG peptides was achieved by the direct enrichment of K-GG Peptides in vivo from cells or tissues (Wagner et al. 2011, 2012; Udeshi et al. 2012, 2013) (Fig. 9.2b). For example, Wagner et al. characterized >20,000 ubiquitination sites of >5,200 proteins in murine tissues. In this regard, direct enrichment of K-GG peptides has attracted more attention for further large-scale assays.

9.3 Data Resources for Ub/UBL Conjugations

Currently, there are 13 databases available for Ub/UBL conjugations (Table 9.1). To circumvent competitions, most databases were focused on certain aspects. For example, Lee et al. (2008) developed a budding yeast-specific database SCUD, including 1 E1, 11 E2s, 42 E3s, 20 DUBs and 940 ubiquitinated substrates.

Table 9.1 A summary of Ub/UBL-related databases

Databases	Main propose	Species	Method[a]	Reference[b]	Number		
					Enzyme[c]	Substrate[d]	Sites
SCUD	Ubiquitin-associated enzymes and ubiquitinated substrates	*S. cerevisiae*	TO	Y	74	940	–
PlantsUPS	Ubiquitin-associated enzymes	Plants	TO	N	8,165	–	–
PlantsUBQ	Ubiquitin-associated enzymes	*A. thaliana*	SL, FA	Y	1,416	–	–
hUbiquitome	Ubiquitin-associated enzymes and ubiquitinated substrates	*H. sapiens*	SL	Y	168	279	36
E3Net	Ubiquitin E3 ligases and substrates	427 species	SL, TO	Y	2,201	4,896	–
UUCD	Ubiquitin-associated enzymes	70 eukaryotes	SL, FA	Y	56,949	–	–
DUDE-db	Ubiquitin-associated enzymes	50 eukaryotes	TO, FA	N	35,228	–	–
UbiProt	Ubiquitylated proteins	9 species	SL, TO	Y	–	1,104	222
UniProt	Ubiquitinated and sumoylated substrates	General	SL, OS	Y	–	–	2,502
SysPTM1.1	~50 PTMs including ubiquitinated substrates	General	MS, SL, TO	Y	–	699	1,164
dbPTM3.0	18 PTMs including ubiquitinated and sumoylated substrates	General	MS, SL, OS, FA	Y	–	–	48,781
mUbiSiDa	Mammalian ubiquitination sites	Mammalians	MS, SL	Y	–	27,272	79,425
PupDB	Pupylated proteins	Prokaryotes	SL, FA	Y	–	1305	215

[a] Method, methods used in collecting the data. *TO* taken from other databases or websites; *SL* manually curated from scientific literature; *PS* further computational analysis; *OS* orthologous sites of experimentally verified Ub/UBL conjugation sites; *MS* mass spectrometry-derived data

[b] Reference, whether the information provided in the databases is traceable to original publications

[c] Enzyme, ubiquitin and ubiquitin-like conjugation enzymes, including E1s, E2s, E3s and DUBs

[d] Substrate, substrates of ubiquitin and ubiquitin-like conjugations

Later, Du et al. (2009) constructed a ubiquitination-associated enzyme database plantsUPS, which contains 24 E1, 417 E2s and 7624 E3s from plants. Also, a similar database of PlantsUBQ was developed for plant Ub enzymes, with 2 E1s, 37 E2s, 1,326 E3s and 51 DUBs (http://plantsubq.genomics.purdue.edu/). Furthermore, the hUbiquitome was released for human ubiquitination, with 1 E1, 12 E2s, 138 E3s, 17 DUBs, 279 substrates and 36 ubiquitination sites (Du et al. 2011). In addition, by constructing the E3-mediated regulatory networks, Han et al. (2012) collected 2,201 E3s and 4,896 substrates. The above databases only contains enzyme information for Ub, while UBL conjugations were not included. Recently, we developed a comprehensive database Ubiquitin and Ubiquitin-like Conjugation Database (UUCD) that contains 738 E1s, 2,937 E2s and 46,631 E3s and 6,647 DUBs in 70 eukaryotic species (Gao et al. 2013). Later, Hutchins et al. (2013) also released a similar database DUDE-db for Ub/UBL conjugations, but only with 267 E1s, 2,095 E2s, 28,985 E3s and 3881 substrates in 50 eukaryotic species.

Additionally, several databases have developed exclusively for Ub/UBL conjugation substrates and sites (Table 9.1). The fist database only containing ubiquitinated substrates and sites was UbiProt, which collected 1,104 substrates and 222 ubiquitination sites (Chernorudskiy et al. 2007). The UniProt also contained substrates and sites for multiple post-translational modifications (PTMs), such as ubiquitination and sumoylation (Magrane and Consortium 2011). Since rapid progresses in MS-based proteomics have generated a large number of Ub/UBL conjugation substrates and sites, collection and integration these data sets will provide useful resources for further analysis. For example, Li et al. (2009) created SysPTM that contained modification information for nearly 50 types of PTMs, including 1,164 ubiquitination sites in 699 substrates. DbPTM 3.0, another PTM resource, contains 48,781 ubiquitination and sumoylation sites (Lu et al. 2013). Recently, Hui et al. provided a comprehensive database, including 79,425 mammalian ubiquitination sites of 27,272 proteins (http://222.193.31.35:8000/mUbiSiDa.php). In particular, a UBL conjugation database of PupDB was developed with 1,305 substrates and 215 pupylation sites (Tung 2012).

9.4 Prediction of Ub/UBL Conjugation Sites

Although more and more Ub/UBL conjugation substrates have been identified, accurate prediction of conjugation sites is still a great challenge. To date, although over 20 approaches have been developed for predicting Ub/UBL conjugation sites, only 13 applicable tools can be accessed (Table 9.2). In Tung and Ho (2008) used 531 physicochemical features and the support vector machines (SVMs) algorithm to develop the first predictor of UbiPred, with a training data set of 157 known ubiquitination sites. Using 442 positive sites, Lee et al. (2011) developed UbSite, which adopted a number of sequence features and the radial basis function networks (RBFNs) algorithm for training. Since different organisms may have different features in proteins selected for ubiquitination, the prediction accuracy might

Table 9.2 Predictors for non- or organism-specific Ub/UBL conjugation substrates and sites

Predictors	Training data set[a]	Specificity[b]	Method[c]
Ubiquitination			
UbiPred	157 ubiquitination sites	General	SVMs
UbSite	442 ubiquitination sites	General	RBFNs
UbPred	265 ubiquitination sites in S. cerevisiae	*S. cerevisiae*	RF
CKSAAP_UbSite	263 ubiquitination sites in S. cerevisiae	*S. cerevisiae*	SVMs
hCKSAAP_UbSite	6118 K sites in human	*H. sapiens*	SVMs
UbiProber	25,194 ubiquitination sites in *H. sapiens*, 5348 in *M. musculus* and 175 in *S. cerevisiae*	General and organism-specific	SVMs
GPS-ARM	74 D-box and 42 KEN-box motifs	General	GPS
Sumoylation			
SUMOplot	N/A	General	HS
SUMOsp1.0	239 sumoylation sites	General	GPS
SUMOpre	268 sumoylation sites	General	SM
SUMOsp2.0	279 sumoylation sites	General	GPS
seeSUMO	425 sumoylation sites	General	RF, SVMs
Pupylation			
GPS-PUP	127 pupylation sites	Prokaryotes	GPS

SVMs support vector machines, *RBFNs* radial basis function networks, *RF* random forest, *GPS* group-based prediction system, *HS* hydrophobic similarity; *SM* statistical method

[a] Training Data Set, the experimentally verified Ub/UBL sites were taken as the positive training data set

[b] Specificity, for general propose or organism-specific prediction

[c] Method, the computational methods used for training

be improved in organism-specific manner. For example, Radivojac et al. (2010) collected 265 yeast ubiquitination sites and developed the first organism-specific predictor of UbPred, with the random forest (RF) algorithm. Also, Chen et al. (2011) adopted the composition of *k*-spaced amino acid pairs (CKSAAPs) of lysine-centered peptides and SVMs algorithm to designed a yeast-specific predictor of CKSAAP_UbSite, with a training data set of 263 known ubiquitination sites. Later, they further constructed a human-specific predictor of hCKSAAP_UbSite with the same approaches (Chen et al. 2013). Recently, Chen et al. (2013) adopted a number of sequence features and used the SVMs algorithm to develop UbiProber, which can predict general or organism-specific ubiquitination sites. With the group-based prediction system (GPS) algorithm, we also developed GPS-ARM for the prediction of anaphase-promoting complex/cyclosome (APC/C) recognition motifs including D-box and KEN-box, which can be recognized by Cdh1 or Cdc20 for the protein degradation (Liu et al. 2012). Thus, the GPS-ARM predicts ubiquitinated substrates but not exact sites (Liu et al. 2012).

Beyond ubiquitination, there have been a considerable number of efforts taken for other UBL conjugations, such as sumoylation and pupylation. Because ~77 % of total sumoylation sites follow a canonical motif of Ψ-K-X-D/E (Ψ is a hydrophobic residue, X is any amino acid) (Xue et al. 2006), the first predictor SUMOplot was developed by evaluating the hydrophobic similarity between given proteins and known sumoylation sites (http://www.abgent.com/sumoplot). Later, using 239 known sumoylation sites as positive samples, we developed SUMOsp1.0 with the GPS algorithm (Xue et al. 2006). With a statistical method, Xu et al. (2008) developed the SUMOpre, which was trained with 268 known sumoylation sites. In 2009, we greatly improved the GPS algorithm and released the SUMOsp 2.0 software package, with a superior performance than other existing tools (Ren et al. 2009). Recently, Teng et al. (2012) used RF and SVMs algorithms to develop the seeSUMO for predicting sumoylation sites. In addition, we also developed an accurate tool of GPS-PUP for the prediction of pupylation sites in prokaryotes (Liu et al. 2011). Due to the page limitation, the computational predictions of Ub/UBL conjugation sites without available programs were not summarized.

9.5 Computational Analysis of Disease-Associated Ub/UBL Conjugations Provides Potential Biomarkers and Drug Targets

To evaluate the importance of Ub/UBL conjugations in diseases and drug targets, we mapped Ub/UBL conjugation enzymes to other databases. First, we obtained 874 human Ub/UBL conjugation enzymes from the UUCD database (Gao et al. 2013), 474 known cancer genes from Cancer Gene Census (Forbes et al. 2011) and 4,096 well-characterized drug targets from Drugbank database (Knox et al. 2011). We mapped cancer genes and drug targets to the human proteomes and got 464 and 2,071 unique sequences, respectively. Also, we mapped all human Ub/UBL conjugation enzymes to the two data sets, and only identified 27 cancer genes and 16 drug targets. The statistical analyses with a hypergeometric distribution demonstrated that both known cancer genes and drug targets were not significantly enriched in Ub/UBL conjugation enzymes (p-value > 0.05). However, we further mapped all enzymes to the KEGG pathways (Kanehisa et al. 2012), and observed that Ub/UBL conjugations are significantly involved in a number of essential pathways (p-value $< 10^{-4}$), such as ubiquitin mediated proteolysis (hsa04120), protein processing in endoplasmic reticulum (hsa04141) and cell cycle (hsa04110) (Table 9.3). In particular, we revealed that Ub/UBL conjugation enzymes are over-represented in the pathway of small cell lung cancer (SCLC, hsa05222) (Table 9.3). Based on the results and KEGG annotations, we illustrated the pathway, and totally detected 12 E3s, 2 E3 complexes and 4 ubiquitinated substrates (Fig. 9.3). The results also demonstrated that ubiquitination plays an important role in SCLC-related PI3 K-Akt signaling, cell cycle, apoptosis and p53 signaling pathways (Fig. 9.3). In this regard, Ub/UBL conjugation enzymes and substrates can be a useful reservoir for further identifying potential biomarkers and drug targets.

Table 9.3 The enrichment analysis of KEGG pathways for 874 human Ub/UBL conjugation enzymes from the UUCD database (Gao et al. 2013) (the hypergeometric distribution, p-value $<10^{-4}$)

KEGG ID	Description	UUCD[a]		Proteome		E-ratio[d]	p-value
		Number[b]	Percentage[c]	Number	Percentage		
The most over-represent KEGG Pathway							
hsa04120	Ubiquitin mediated proteolysis	135	55.10	137	2.21	24.94	8.63E−206
hsa04141	Protein processing in endoplasmic reticulum	32	13.06	165	2.66	4.91	2.32E−14
hsa04110	Cell cycle	22	8.98	125	2.02	4.46	2.35E−09
hsa04114	Oocyte meiosis	20	8.16	110	1.77	4.60	7.24E−09
hsa05222	Small cell lung cancer	13	5.31	84	1.35	3.92	2.12E−05
hsa04914	Progesterone-mediated oocyte maturation	13	5.31	86	1.39	3.83	2.75E−05
hsa04330	Notch signaling pathway	9	3.67	47	0.76	4.85	7.39E−05
The most under-represent KEGG Pathway							
hsa01100	Metabolic pathways	2	0.82	1156	18.64	0.04	6.47E−20

[a] UUCD, proteins in the UUCD database

[b] Number, the number of proteins annotated with the KEGG ID

[c] Percentage the proportion of proteins annotated with the KEGG ID

[d] E-ratio, the enrichment ratio as the proportion of enzymes in UUCD divided by that in the proteome

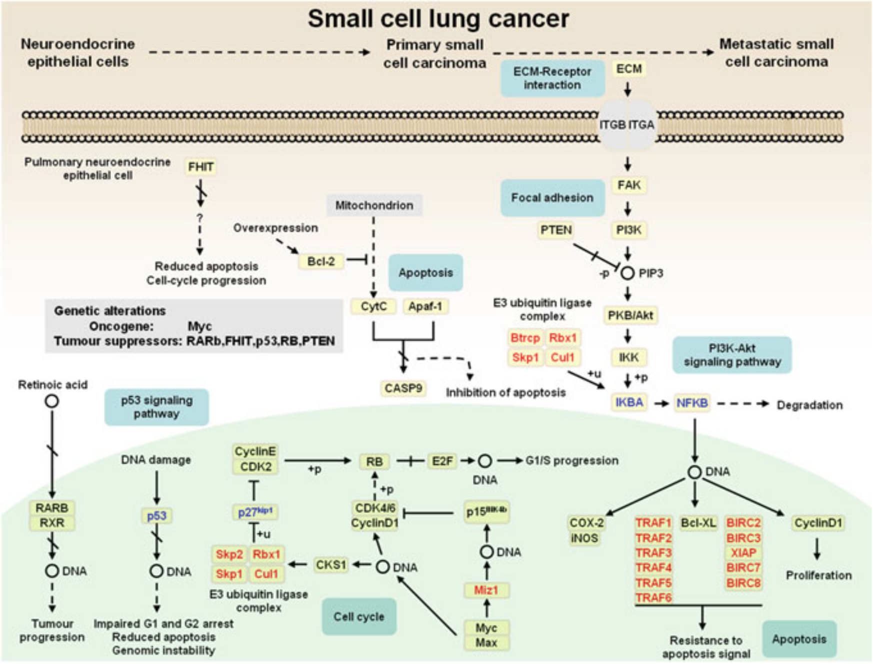


Fig. 9.3 The small cell lung cancer pathway (SCLC, hsa05222) adapted from the KEGG database. The known E3s were shown in *red*, whereas experimentally identified ubiquitinated substrates were shown in *blue*

9.6 Personal Perspectives on Further Computational Analysis of Ub/UBL Conjugations

In this chapter, we presented a brief summarization of current progresses especially computational efforts in Ub/UBL conjugations. Totally, there have been 13 online databases and 13 applicable predictors released for Ub/UBL conjugations. As more and more conjugation substrates and sites have been identified, we believed that more and more databases and tools will be developed in the near future. For further computational studies, we provided several personal perspectives as below:

1. Prediction of conjugation sites for more UBLs. Currently, most computational predictions were focused on ubiquitination and sumoylation, or in a less extent, pupylation. However, over ten UBLs have been characterized, while a number of proteomic analyses of substrates for these UBLs, such as Nedd8-mediated neddylation (Jones et al. 2008) and ISG15-mediated ISGylation (Giannakopoulos et al. 2005). The development of efficient algorithms and predictors can generate useful information for further experimental considerations.
2. Prediction of ubiquitinated substrates and sites in an E3-specific mode. For ubiquitination, the E3 ligases determined the specificity for substrate

recognition. Analogous to phosphorylation which can be catalyzed by ~520 kinases, there were 874 human Ub/UBL conjugation enzymes collected in the UUCD database (Gao et al. 2013). Because different kinases recognize different motifs for modification, we developed a kinase-specific predictor of GPS for the phosphorylation (Xue et al. 2005, 2008). Again, because different E3 ligases exhibited dramatically different sequence or structure profiles, it can be expected that different E3s can recognize distinct motifs for conjugations. In this regard, prediction of E3-specific ubiquitinated substrates and sites will achieve much better performance.

3. Re-construction of Ub/UBL-associated networks. Protein substrates can be modified by E1s, E2s, and E3s and de-modified by DUBs. Thus, the complex relations among Ub/UBL conjugation enzymes and substrates constitute the Ub/UBL-associated networks, which are fundamental for systematically understanding the molecular mechanisms and regulatory roles of Ub/UBL conjugations. Also, how to retrieve useful information from the networks will be a great challenge.

9.7 Conclusion

As a class of important and ubiquitous PTMs, Ub/UBL conjugations has attracted more and more attention to be potential biomarkers or drug targets. Besides both small- or large-scale experimental identifications, computational analysis of Ub/UBL conjugations has also emerged to a promising topic. However, the number of either databases or predictors for Ub/UBL conjugations is still limited, and more efforts should be paid in this field. We believed a better study will generate a deeper understanding on Ub/UBL conjugations and provide useful information for biomedical design.

Acknowledgments This work was supported by grants from the National Basic Research Program (973 project) (2012CB910101, and 2013CB933903), Natural Science Foundation of China (31171263 and 81272578), and International Science and Technology Cooperation Program of China (0S2013ZR0003).

References

Aghajan M, Jonai N, Flick K, Fu F, Luo M, Cai X, Ouni I, Pierce N, Tang X, Lomenick B, et al. Chemical genetics screen for enhancers of rapamycin identifies a specific inhibitor of an SCF family E3 ubiquitin ligase. Nat Biotechnol. 2010;28:738–42.

Bedford L, Lowe J, Dick LR, Mayer RJ, Brownell JE. Ubiquitin-like protein conjugation and the ubiquitin-proteasome system as drug targets. Nat Rev Drug Discov. 2011;10:29–46.

Bogunovic D, Boisson-Dupuis S, Casanova JL. ISG15: leading a double life as a secreted molecule. Exp Mol Med. 2013;45:e18.

Bonacci T, Roignot J, Soubeyran P. Protein ubiquitylation in pancreatic cancer. ScientificWorldJournal. 2010;10:1462–72.

Bourgeois-Daigneault MC, Thibodeau J. Autoregulation of MARCH1 expression by dimerization and autoubiquitination. J Immunol. 2012;188:4959–70.

Bustos D, Bakalarski CE, Yang Y, Peng J, Kirkpatrick DS. Characterizing ubiquitination sites by peptide based immunoaffinity enrichment. Mol Cell Proteomics. 2012.

Chasapis CT, Spyroulias GA. RING finger E(3) ubiquitin ligases: structure and drug discovery. Curr Pharm Des. 2009;15:3716–31.

Chau V, Tobias JW, Bachmair A, Marriott D, Ecker DJ, Gonda DK, Varshavsky A. A multiubiquitin chain is confined to specific lysine in a targeted short-lived protein. Science. 1989;243:1576–83.

Chen D, Frezza M, Schmitt S, Kanwar J, Dou QP. Bortezomib as the first proteasome inhibitor anticancer drug: current status and future perspectives. Curr Cancer Drug Targets. 2011a;11:239–53.

Chen Q, Xie W, Kuhn DJ, Voorhees PM, Lopez-Girona A, Mendy D, Corral LG, Krenitsky VP, Xu W. Moutouh-de Parseval L et al. Targeting the p27 E3 ligase SCF(Skp2) results in p27- and Skp2-mediated cell-cycle arrest and activation of autophagy. Blood. 2008;111:4690–9.

Chen X, Qiu JD, Shi SP, Suo SB, Huang SY, Liang RP. Incorporating key position and amino acid residue features to identify general and species-specific ubiquitin conjugation sites. Bioinformatics. 2013.

Chen Z, Chen YZ, Wang XF, Wang C, Yan RX, Zhang Z. Prediction of ubiquitination sites by using the composition of k-spaced amino acid pairs. PLoS ONE. 2011b;6:e22930.

Chen Z, Zhou Y, Song J, Zhang Z. hCKSAAP_UbSite: Improved prediction of human ubiquitination sites by exploiting amino acid pattern and properties. Biochim Biophys Acta. 2013.

Chen ZJ, Sun LJ. Nonproteolytic functions of ubiquitin in cell signaling. Mol Cell. 2009;33:275–86.

Chernorudskiy AL, Garcia A, Eremin EV, Shorina AS, Kondratieva EV, Gainullin MR. UbiProt: a database of ubiquitylated proteins. BMC Bioinformatics. 2007;8:126.

Ciechanover A. The ubiquitin-proteasome proteolytic pathway. Cell. 1994;79:13–21.

Cohen P, Tcherpakov M. Will the ubiquitin system furnish as many drug targets as protein kinases? Cell. 2010;143:686–93.

Conrad C, Podolsky MJ, Cusack JC. Antiproteasomal agents in rectal cancer. Anticancer Drugs. 2011;22:341–50.

Coornaert B, Carpentier I, Beyaert R. A20: central gatekeeper in inflammation and immunity. J Biol Chem. 2009;284:8217–21.

Dahlmann B. Role of proteasomes in disease. BMC Biochem. 2007;8 Suppl 1:S3.

Danielsen JM, Sylvestersen KB, Bekker-Jensen S, Szklarczyk D, Poulsen JW, Horn H, Jensen LJ, Mailand N, Nielsen ML. Mass spectrometric analysis of lysine ubiquitylation reveals promiscuity at site level. Mol Cell Proteomics. 2011;10:M110 003590.

Day IN, Thompson RJ. UCHL1 (PGP 9.5): neuronal biomarker and ubiquitin system protein. Prog Neurobiol. 2010;90:327–62.

Deng H, Liang H, Jankovic J. F-box only protein 7 gene in parkinsonian-pyramidal disease. JAMA Neurol. 2013;70:20–4.

Denis NJ, Vasilescu J, Lambert JP, Smith JC, Figeys D. Tryptic digestion of ubiquitin standards reveals an improved strategy for identifying ubiquitinated proteins by mass spectrometry. Proteomics. 2007;7:868–74.

Deshaies RJ, Joazeiro CA. RING domain E3 ubiquitin ligases. Annu Rev Biochem. 2009;78:399–434.

Du H, Huang X, Wang S, Wu Y, Xu W, Li M. PSMA7, a potential biomarker of diseases. Protein Pept Lett. 2009a;16:486–9.

Du Y, Xu N, Lu M, Li T. hUbiquitome: a database of experimentally verified ubiquitination cascades in humans. Database (Oxford). 2011;2011:bar055.

Du Z, Zhou X, Li L, Su Z. plantsUPS: a database of plants' ubiquitin proteasome system. BMC Genomics. 2009b;10:227.

Duncan K, Schafer G, Vava A, Parker MI, Zerbini LF. Targeting neddylation in cancer therapy. Future Oncol. 2012;8:1461–70.

Escobar M, Velez M, Belalcazar A, Santos ES, Raez LE. The role of proteasome inhibition in nonsmall cell lung cancer. J Biomed Biotechnol. 2011;2011:806506.

Festa RA, McAllister F, Pearce MJ, Mintseris J, Burns KE, Gygi SP, Darwin KH. Prokaryotic ubiquitin-like protein (Pup) proteome of Mycobacterium tuberculosis [corrected]. PLoS ONE. 2010;5:e8589.

Flick K, Ouni I, Wohlschlegel JA, Capati C, McDonald WH, Yates JR, Kaiser P. Proteolysis-independent regulation of the transcription factor Met4 by a single Lys 48-linked ubiquitin chain. Nat Cell Biol. 2004;6:634–41.

Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, Jia M, Shepherd R, Leung K, Menzies A, et al. COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in cancer. Nucleic Acids Res. 2011;39:D945–50.

Gao T, Liu Z, Wang Y, Cheng H, Yang Q, Guo A, Ren J, Xue Y. UUCD: a family-based database of ubiquitin and ubiquitin-like conjugation. Nucleic Acids Res. 2013;41:D445–51.

Geng F, Wenzel S, Tansey WP. Ubiquitin and proteasomes in transcription. Annu Rev Biochem. 2012;81:177–201.

Giannakopoulos NV, Luo JK, Papov V, Zou W, Lenschow DJ, Jacobs BS, Borden EC, Li J, Virgin HW, Zhang DE. Proteomic identification of proteins conjugated to ISG15 in mouse and human cells. Biochem Biophys Res Commun. 2005;336:496–506.

Han Y, Lee H, Park JC, Yi GS. E3Net: a system for exploring E3-mediated regulatory networks of cellular functions. Mol Cell Proteomics. 2012;11:O111 014076.

Hegde AN, Upadhya SC. The ubiquitin-proteasome pathway in health and disease of the nervous system. Trends Neurosci. 2007;30:587–95.

Hicke L. Protein regulation by monoubiquitin. Nat Rev Mol Cell Biol. 2001;2:195–201.

Hochrainer K, Lipp J. Ubiquitylation within signaling pathways in- and outside of inflammation. Thromb Haemost. 2007;97:370–7.

Hochstrasser M. Origin and function of ubiquitin-like proteins. Nature. 2009;458:422–9.

House CM, Hancock NC, Moller A, Cromer BA, Fedorov V, Bowtell DD, Parker MW, Polekhina G. Elucidation of the substrate binding site of Siah ubiquitin ligase. Structure. 2006;14:695–701.

Humbard MA, Miranda HV, Lim JM, Krause DJ, Pritz JR, Zhou G, Chen S, Wells L, Maupin-Furlow JA. Ubiquitin-like small archaeal modifier proteins (SAMPs) in Haloferax volcanii. Nature. 2010;463:54–60.

Hutchins AP, Liu S, Diez D, Miranda-Saavedra D. The repertoires of ubiquitinating and deubiquitinating enzymes in eukaryotic genomes. Mol Biol Evol. 2013;30:1172–87.

Irminger-Finger I. BARD1, a possible biomarker for breast and ovarian cancer. Gynecol Oncol. 2010;117:211–5.

Iyer LM, Burroughs AM, Aravind L. The prokaryotic antecedents of the ubiquitin-signaling system and the early evolution of ubiquitin-like beta-grasp domains. Genome Biol. 2006;7:R60.

Jeram SM, Srikumar T, Pedrioli PG, Raught B. Using mass spectrometry to identify ubiquitin and ubiquitin-like protein conjugation sites. Proteomics. 2009;9:922–34.

Jones J, Wu K, Yang Y, Guerrero C, Nillegoda N, Pan ZQ, Huang L. A targeted proteomic analysis of the ubiquitin-like modifier nedd8 and associated proteins. J Proteome Res. 2008;7:1274–87.

Kaelin WG. Von Hippel-Lindau disease. Annu Rev Pathol. 2007;2:145–73.

Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 2012;40:D109–14.

Kang C, Yi GS. Identification of ubiquitin/ubiquitin-like protein modification from tandem mass spectra with various PTMs. BMC Bioinform. 2011;12 Suppl 14:S8.

Kim W, Bennett EJ, Huttlin EL, Guo A, Li J, Possemato A, Sowa ME, Rad R, Rush J, Comb MJ, et al. Systematic and quantitative assessment of the ubiquitin-modified proteome. Mol Cell. 2011;44:325–40.

Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V, et al. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. Nucleic Acids Res. 2011;39:D1035–41.

Lee KA, Hammerle LP, Andrews PS, Stokes MP, Mustelin T, Silva JC, Black RA, Doedens JR. Ubiquitin ligase substrate identification through quantitative proteomics at both the protein and peptide levels. J Biol Chem. 2011a;286:41530–8.

Lee TY, Chen SA, Hung HY, Ou YY. Incorporating distant sequence features and radial basis function networks to identify ubiquitin conjugation sites. PLoS ONE. 2011b;6:e17331.

Lee WC, Lee M, Jung JW, Kim KP, Kim D. SCUD: Saccharomyces cerevisiae ubiquitination database. BMC Genomics. 2008;9:440.

Lehman NL. The ubiquitin proteasome system in neuropathology. Acta Neuropathol. 2009;118:329–47.

Li H, Xing X, Ding G, Li Q, Wang C, Xie L, Zeng R, Li Y. SysPTM: a systematic resource for proteomic research on post-translational modifications. Mol Cell Proteomics. 2009;8:1839–49.

Lin H, Lin Q, Liu M, Lin Y, Wang X, Chen H, Xia Z, Lu B, Ding F, Wu Q et al. PKA/Smurf1 signaling-mediated stabilization of Nur77 is required for anticancer drug cisplatin-induced apoptosis. Oncogene. 2013.

Linehan WM, Bratslavsky G, Pinto PA, Schmidt LS, Neckers L, Bottaro DP, Srinivasan R. Molecular diagnosis and therapy of kidney cancer. Annu Rev Med. 2010;61:329–43.

Liu Z, Ma Q, Cao J, Gao X, Ren J, Xue Y. GPS-PUP: computational prediction of pupylation sites in prokaryotic proteins. Mol BioSyst. 2011;7:2737–40.

Liu Z, Yuan F, Ren J, Cao J, Zhou Y, Yang Q, Xue Y. GPS-ARM: computational analysis of the APC/C recognition motif by predicting D-boxes and KEN-boxes. PLoS ONE. 2012;7:e34370.

Lu CT, Huang KY, Su MG, Lee TY, Bretana NA, Chang WC, Chen YJ, Huang HD. DbPTM 3.0: an informative resource for investigating substrate site specificity and functional association of protein post-translational modifications. Nucleic Acids Res. 2013;41:D295–305.

Magrane M, Consortium U. UniProt Knowledgebase: a hub of integrated protein data. Database (Oxford). 2011;2011:bar009.

Mandel SA, Fishman-Jacob T, Youdim MB. Modeling sporadic Parkinson's disease by silencing the ubiquitin E3 ligase component, SKP1A. Parkinsonism Relat Disord. 2009;15(Suppl 3):S148–51.

Maor R, Jones A, Nuhse TS, Studholme DJ, Peck SC, Shirasu K. Multidimensional protein identification technology (MudPIT) analysis of ubiquitinated proteins in plants. Mol Cell Proteomics. 2007;6:601–10.

Meierhofer D, Wang X, Huang L, Kaiser P. Quantitative analysis of global ubiquitination in HeLa cells by mass spectrometry. J Proteome Res. 2008;7:4566–76.

Melino G, Gallagher E, Aqeilan RI, Knight R, Peschiaroli A, Rossi M, Scialpi F, Malatesta M, Zocchi L, Browne G, et al. Itch: a HECT-type E3 ligase regulating immunity, skin and cancer. Cell Death Differ. 2008;15:1103–12.

Muller S, Briand JP, Van Regenmortel MH. Presence of antibodies to ubiquitin during the autoimmune response associated with systemic lupus erythematosus. Proc Natl Acad Sci U S A. 1988;85:8176–80.

Nakajima H, Fujiwara H, Furuichi Y, Tanaka K, Shimbara N. A novel small-molecule inhibitor of NF-kappaB signaling. Biochem Biophys Res Commun. 2008;368:1007–13.

Newton K, Matsumoto ML, Wertz IE, Kirkpatrick DS, Lill JR, Tan J, Dugger D, Gordon N, Sidhu SS, Fellouse FA, et al. Ubiquitin chain editing revealed by polyubiquitin linkage-specific antibodies. Cell. 2008;134:668–78.

Ohki Y, Funatsu N, Konishi N, Chiba T. The mechanism of poly-NEDD8 chain formation in vitro. Biochem Biophys Res Commun. 2009;381:443–7.

Orlicky S, Tang X, Neduva V, Elowe N, Brown ED, Sicheri F, Tyers M. An allosteric inhibitor of substrate recognition by the SCF(Cdc4) ubiquitin ligase. Nat Biotechnol. 2010;28:733–7.

Oshikawa K, Matsumoto M, Oyamada K, Nakayama KI. Proteome-wide identification of ubiquitylation sites by conjugation of engineered lysine-less ubiquitin. J Proteome Res. 2012;11:796–807.

Osula O, Swatkoski S, Cotter RJ. Identification of protein SUMOylation sites by mass spectrometry using combined microwave-assisted aspartic acid cleavage and tryptic digestion. J Mass Spectrom. 2012;47:644–54.

Peng J, Schwartz D, Elias JE, Thoreen CC, Cheng D, Marsischky G, Roelofs J, Finley D, Gygi SP. A proteomics approach to understanding protein ubiquitination. Nat Biotechnol. 2003;21:921–6.

Radivojac P, Vacic V, Haynes C, Cocklin RR, Mohan A, Heyen JW, Goebl MG, Iakoucheva LM. Identification, analysis, and prediction of protein ubiquitination sites. Proteins. 2010;78:365–80.

Rape M, Reddy SK, Kirschner MW. The processivity of multiubiquitination by the APC determines the order of substrate degradation. Cell. 2006;124:89–103.

Ren J, Gao X, Jin C, Zhu M, Wang X, Shaw A, Wen L, Yao X, Xue Y. Systematic study of protein sumoylation: Development of a site-specific predictor of SUMOsp 2.0. Proteomics. 2009;9:3409–12.

Rufini A, Fortuni S, Arcuri G, Condo I, Serio D, Incani O, Malisan F, Ventura N, Testi R. Preventing the ubiquitin-proteasome-dependent degradation of frataxin, the protein defective in Friedreich's ataxia. Hum Mol Genet. 2011;20:1253–61.

Sadowski M, Sarcevic B. Mechanisms of mono- and poly-ubiquitination: Ubiquitination specificity depends on compatibility between the E2 catalytic core and amino acid residues proximal to the lysine. Cell Div. 2010;5:19.

Scheffner M, Staub O. HECT E3 s and human disease. BMC Biochem. 2007;8(Suppl 1):S6.

Shi Y, Chan DW, Jung SY, Malovannaya A, Wang Y, Qin J. A data set of human endogenous protein ubiquitination sites. Mol Cell Proteomics. 2011;10:M110 002089.

Shi Y, Xu P, Qin J. Ubiquitinated proteome: ready for global? Mol Cell Proteomics. 2011;10:R110 006882.

Singhal S, Taylor MC, Baker RT. Deubiquitylating enzymes and disease. BMC Biochem. 2008;9(Suppl 1):S3.

Sohns W, van Veen TA, van der Heyden MA. Regulatory roles of the ubiquitin-proteasome system in cardiomyocyte apoptosis. Curr Mol Med. 2010;10:1–13.

Starita LM, Lo RS, Eng JK, von Haller PD, Fields S. Sites of ubiquitin attachment in Saccharomyces cerevisiae. Proteomics. 2012;12:236–40.

Striebel F, Imkamp F, Sutter M, Steiner M, Mamedov A, Weber-Ban E. Bacterial ubiquitin-like modifier Pup is deamidated and conjugated to substrates by distinct but homologous enzymes. Nat Struct Mol Biol. 2009;16:647–51.

Teng S, Luo H, Wang L. Predicting protein sumoylation sites from sequence features. Amino Acids. 2012;43:447–55.

Tung CW. PupDB: a database of pupylated proteins. BMC Bioinformatics. 2012;13:40.

Tung CW, Ho SY. Computational identification of ubiquitylation sites from protein sequences. BMC Bioinform. 2008;9:310.

Udeshi ND, Mani DR, Eisenhaure T, Mertins P, Jaffe JD, Clauser KR, Hacohen N, Carr SA. Methods for quantification of in vivo changes in protein ubiquitination following proteasome and deubiquitinase inhibition. Mol Cell Proteomics. 2012;11:148–59.

Udeshi ND, Svinkina T, Mertins P, Kuhn E, Mani DR, Qiao JW, Carr SA. Refined preparation and use of anti-diglycine remnant (K-epsilon-GG) antibody enables routine quantification of 10,000s of ubiquitination sites in single proteomics experiments. Mol Cell Proteomics. 2013;12:825–31.

Ulrich HD. The fast-growing business of SUMO chains. Mol Cell. 2008;32:301–5.

van der Veen AG, Ploegh HL. Ubiquitin-like proteins. Annu Rev Biochem. 2012;81:323–57.

Wagner SA, Beli P, Weinert BT, Nielsen ML, Cox J, Mann M, Choudhary C. A proteome-wide, quantitative survey of in vivo ubiquitylation sites reveals widespread regulatory roles. Mol Cell Proteomics. 2011;10:M111 013284.

Wagner SA, Beli P, Weinert BT, Scholz C, Kelstrup CD, Young C, Nielsen ML, Olsen JV, Brakebusch C, Choudhary C. Proteomic analyses reveal divergent ubiquitylation site patterns in murine tissues. Mol Cell Proteomics. 2012;11:1578–85.

Wang D, Xu W, McGrath SC, Patterson C, Neckers L, Cotter RJ. Direct identification of ubiquitination sites on ubiquitin-conjugated CHIP using MALDI mass spectrometry. J Proteome Res. 2005;4:1554–60.

Wang J. Cardiac function and disease: emerging role of small ubiquitin-related modifier. Wiley Interdiscip Rev Syst Biol Med. 2011;3:446–57.

Wong CS, Moller A. Siah: a promising anticancer target. Cancer Res. 2013;73:2400–6.

Xu J, He Y, Qiang B, Yuan J, Peng X, Pan XM. A novel method for high accuracy sumoylation site prediction from protein sequences. BMC Bioinform. 2008;9:8.

Xue Y, Ren J, Gao X, Jin C, Wen L, Yao X. GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy. Mol Cell Proteomics. 2008;7:1598–608.

Xue Y, Zhou F, Fu C, Xu Y, Yao X. SUMOsp: a web server for sumoylation site prediction. Nucleic Acids Res. 2006;34:W254–7.

Xue Y, Zhou F, Zhu M, Ahmed K, Chen G, Yao X. GPS: a comprehensive www server for phosphorylation sites prediction. Nucleic Acids Res. 2005;33:W184–7.

Yunus AA, Lima CD. Lysine activation and functional analysis of E2-mediated conjugation in the SUMO pathway. Nat Struct Mol Biol. 2006;13:491–9.

Zhao C, Collins MN, Hsiang TY, Krug RM. Interferon-induced ISG15 pathway: an ongoing virus-host battle. Trends Microbiol. 2013;21:181–6.

Zhong Q, Gao W, Du F, Wang X. Mule/ARF-BP1, a BH3-only E3 ubiquitin ligase, catalyzes the polyubiquitination of Mcl-1 and regulates apoptosis. Cell. 2005;121:1085–95.

Chapter 10
Identification of Biomarkers for Pharmacological Activity

Guang Hu, Yuqian Li and Bairong Shen

Abstract Biomarkers are kinds of biological signatures of particular physiological state, which also can be validated and qualified as indicators of clinical endpoints, surrogate endpoints, and particular indicator that respond to drug therapy. The rapid development of high-throughput technologies has facilitated the identification of new biomarkers at different systems levels. In this chapter, we will focus on the recent advance in identifying biomarkers based on technologies of genomics, proteomics, and metabolomics, as well as their applications in drug response, and thus achieving personalized medicine. In addition, some well-known examples of pharmacological biomarkers especially for cancers are collected and provided. The last part of this chapter will discuss the key biomarkers-related resources, including web-based databases and bioinformatics tools.

Keywords Biomarkers · Pharmacogenomics · Drug response

10.1 Introduction

Biomarkers are kinds of biological signatures of particular physiological state, which also can be validated and qualified as indicators of clinical endpoints, surrogate endpoints, and particular indicator that respond to pharmacological activity

G. Hu (✉) · B. Shen
Center for Systems Biology, Soochow University, PO. Box 206, No. 1 Shizi Street,
Suzhou 215006, Jiangsu, China
e-mail: huguang@suda.edu.cn

B. Shen
e-mail: bairong.shen@suda.edu.cn

Y. Li
School of Electronic Engineering, University of Electronic Science and Technology
of China, No. 2006, Xiyuan Ave, West Hi-Tech Zone, 611731 Chengdu, China
e-mail: yuqianli@uestc.edu.cn

B. Shen (ed.), *Bioinformatics for Diagnosis, Prognosis and Treatment of Complex Diseases*, Translational Bioinformatics 4, DOI: 10.1007/978-94-007-7975-4_10,
© Springer Science+Business Media Dordrecht 2013

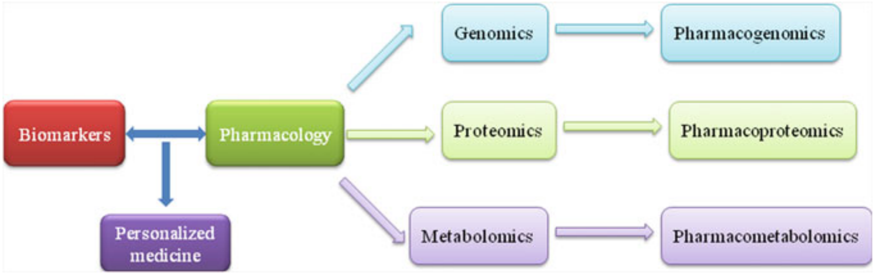


Fig. 10.1 The scheme for applying different "omic" technologies to identify pharmacological biomarkers, achieving the goal of personalized medicine

(Frank and Hargreaves 2003). In particular, the application of biomarkers in drug response is currently occurring at an increasing rate (Matsui 2013). Drug response biomarkers have been used for outcome prediction and assessment in a variety of diseases, especially for cancers (Ludwig and Weinstein 2005). The discovery of novel biomarkers in drug response holds great promise for the future of personalized medicine (De Koning and Keirn 2009; Kelloff and Sigman 2012; Jain 2004), which is a requirement to achieve "the right drug into the right patient".

To facilitate the biomarker discovery, recent biotechnologies such as genomics, proteomics, and metabolomics have grown up (Jain 2010). The fields of using these technologies in biomarker discovery and drug development are termed pharmacogenomics (Mendrick 2008), pharmacoproteomics (Sinha et al. 2007), and pharmacometabolomics (Stewart and Bolt 2011), respectively. With the development of these "omic" technologies, they provide predictive tools to identify biomarkers for pharmacological activity. Thus, the discussion of recent developments for pharmacological biomarkers and their applications in drug development and medical practice in this chapter is particular compelling (Ong et al. 2012). Figure 10.1 shows that how to apply different "omic" technologies to identify pharmacological biomarkers, and then reach the goal of personalized medicine.

The genomic technologies including genome-wide association studies, gene expression analysis, and RNA expression analysis, as well as their applications in indentify biomarkers are introduced in Sect. 10.2. Meanwhile, other two "omic" technologies and their applications including proteomics and metabolomics are discussed in Sect. 10.3 and Sect. 10.4. In Sect. 10.5, we list some examples of pharmacological biomarkers and their value in drug response. Finally in Sect. 10.6 some bioinformatics resources for biomarkers are provided.

10.2 Pharmacogenomic Biomarkers

Pharmacogenetics (Klotz 2007) is the study of how human genetic variation (DNA and RNA) associated with drug response. Pharmacogenetics studies include pharmacokinetics (PK) and pharmacodynamics (PD), which describes drug absorption, distribution, metabolism, and elimination at metabolite levels and the pharmacological effects of a drug on the target biologic pathway, respectively. Pharmacogenomics (Karczewski et al. 2012), a portmanteau of pharmacology and genomics, is the more recent field of applying genome-wide technologies to analysis the makeup affects of genetic to individual's response to drugs. Both pharmacogenetics and pharmacogenomics have the potential for identifying biomarkers in drug response, thus promising personalized medicine (Brandi et al. 2012; De Koning and Keirns 2009; Roses 2000; Ross et al. 2005; Sim and Ingelman-Sundberg 2011). It should be noted that the concept of pharmacogenetics has been included in pharmacogenomics, with a shift from the focus on individual candidate genes to genome-wide association studies. Nowadays, the terms of pharmacogenomics and pharmacogenetics are tending to be used interchangeably. In the following, we will discuss technological advances and applications of these two fields in indentifying biomarkers together.

On the other hand, the fields of next-generation sequencing are in an era of rapid development, reflecting continuous technological advancements in the discovery of novel biomarkers for drug response. The rapid development of genomic techniques brings an unprecedented impact on the pharmaceutical industry, providing powerful tools for mining pharmacogenomic and pharmacogenetics biomarkers. This section will focus on the discovery and developments of genomic methods used in indentify biomarkers in drug response, including genome-wide association studies (GWAS), expression analysis, and next-generation sequencing (NGS). The scheme of using methods into biomarker discovery is shown in Fig. 10.2.

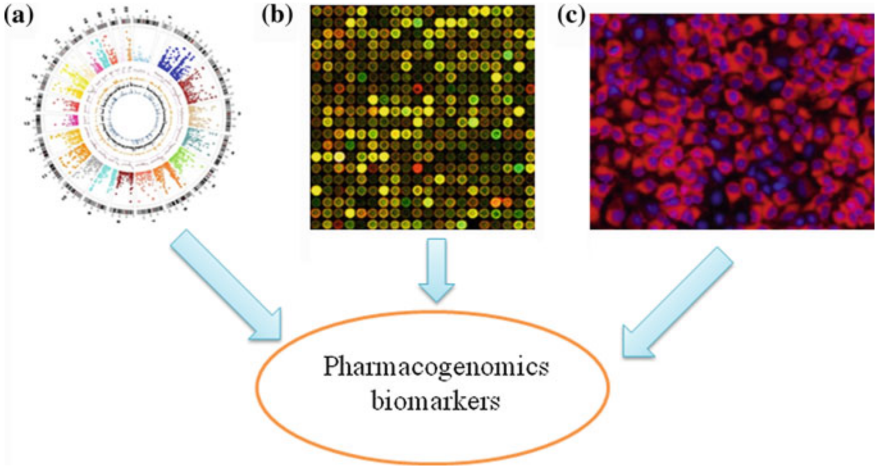


Fig. 10.2 The three pharmacogenomic methods of **a** GWAS, **b** gene expression analysis, and **c** RANi screen for the identification of biomarkers

10.2.1 Genome-Wide Association Study

Genome-wide association study (GWAS) is a family of standard whole genome studies, providing correlations between a continuous trait and genetic information of control sets. GWAS was first introduced in 2005 to investigate patients' age-related macular degeneration. So far, it has been increasingly applied to pharmacogenomics, according to its statistical power and lack of presumptions. In pharmacogenomics research, GWAS aims to search for significant associations between single-nucleotide polymorphisms (SNPs) as biomarkers and traits as drug response. In addition, combined with high-throughput expression analysis and bioinformatics, GWAS offers a more powerful tool in this area.

In a GWAS, hundreds of thousands or millions of genetic variants are probed on a SNP array. The technology of SNP microarray marks each region of the human genome to determine the genotype at each locus, and thus a particular SNP associated with different drug response due to the hidden interaction with an alternate variant of another gene. As contrast, DNA and oligonucleotide arrays are two other methods used in gene expression analysis, which regard for their automated, genome-wide, high-throughput, analysis of a large number of genes, and short oligonucleotide sequence. For the further data analysis, hierarchical clustering, self-organizing maps, multidimensional scaling, and pathway associations are four types of predominant methods.

Li et al. (2009) performed a GWAS in cell-based model system to test a large number of biomarkers that might response to two drugs, i.e. Gemcitabine (dFdC) and AraC. GWAS were also performed on human lymphoblastoid cell lines (LCLs) by Niu et al. (2010). Five biomarkers, C13orf34, MAD2L1, PLK4, TPD52, and DEPDC1B were identified response to radiation therapy in LCLs. Aslibekyan et al. (2012) first applied the GWAS to investigate fenofibrate effects on systemic inflammation. In this work, they have identified several plausible biomarkers for systemic inflammation both before and after fenofibrate treatment. For instance, the rs6517147 locus near the immunologically relevant IFNAR2 gene is associated with IL6 pattern. More recently, Cui et al. (2013) performed a GWAS to Rheumatoid Arthritis (RA) and indentified that CD84 as a biomarker for response to etanercept treatment in RA. In addition, GWAS were also used to indentify biomarker for response to chemotherapeutic agents and radiation therapy.

10.2.2 Gene Expression Analysis

The identification of pharmacogenetic biomarkers using high-throughput gene expression analysis is essential in drug responses that offer new therapeutic approaches of personalized medicine. As mentioned in the above section, DNA microarray analysis has been used to monitor changes in gene expression in response to drug treatments. Other methods include large-scale DNA sequence analysis (or NGS) and comparative genomic hybridization (CGH) technologies have also allowed the genome-scale identification of gene copy number (Majewski and Bernards 2011). DNA sequence analysis provides an alternative method to quantify gene expression, by using high-sequencing technology, which provides a more flexible platform than DNA microarray. In addition, these approaches combined with gene copy number analysis have illuminated core cancer pathways successfully.

Parissenti et al. (2007) have given a comprehensive review on using gene expression for the identification of biomarkers with response to chemotherapy drugs. In this review, a number of anticancer agents have been discussed as genetic biomarkers, including the anthracyclines, alkylating agents, nucleoside analogs taxanes, topoisomerase I and II inhibitors, and vinca alkaloids. Mamdani et al. (2011) have used peripheral gene expression patterns to investigate response to antidepressant treatment in major depressive disorder (MDD). Interferon regulatory factor 7 (IRF7) was found to be an interesting biomarker that associate with antidepressant response. In particular, DNA sequence analyses provide a lot of novel insights into the discovery new cancer biomarkers for drug response for cancers (Jones et al. 2010). Encoding AT-rich interactive domain 1A (ARID1A gene) has been proved to be a useful biomarker in clear-cell ovarian cancers and endometrioid carcinomas (Wiegand et al. 2010).

However, most of gene expression methods are just account for single biomarkers or groups of related biomarkers, but not considerable many-body biologic interactions. More recently, Masica et al. (2013) proposed a new approach, called Multivariate Organization of Combinatorial Alterations (MOCA), and tried to address the shortcoming caused by the response, which is strictly dependent on many simultaneous genetic alterations. This method use Boolean set operations coupled with optimization to combine a large number of genomic alterations into biomarkers of drug response. The test of this algorithm in some pharmacogenomically characterized cancer cell lines is successfully. In particular, it can detect drug response of multigene features, which show higher correlation than the response for single-gene features.

10.2.3 Next-Generation Sequencing

Next-generation sequencing technologies have been extended to differential expression of genes, especially RNA-Seq experiments. RNA expressions provide more sensitive platforms to detect differences between gene expressions and thus open new doors into the field of biomarker discovery. Burczynski and Dorner (2006) have extensively discussed the application of mRNA expression changes in circulating blood cells. Mendrick (2011) also summarized the recent advances of blood-borne biomarkers, including mRNA and microRNA (miRNA), relating disease and drug response. Additionally, some potential mRNA pharmacogenomic biomarkers are suggested that need further verification.

On the other hand, miRNAs are a kind of noncoding RNA (ncRNA) molecules that have emerged as fundamental, post-transcriptional regulators of cognate target gene expression. Till now, several molecular diagnostic technologies including polymerase chain reaction (PCR), liquid chromatography separations, and biochips provide useful tools to miRNA expression profiling. Therefore, the gene expression based on miRNAs can also be used to analyze the correlation with different expression data, especially protein expression data. miRNAs have been reported that they may be served as biomarkers both for cancer and other diseases (Wittmann and Jack 2010). Single-nucleotide polymorphisms based on RNA, termed structural RNA SNPs (srSNPs), have also been found to have important roles in drug metabolism, toxicity, and response. Sadee et al. (2011) discussed some srSNPs relevant to drug response briefly, including CYP3A5, CYP3A4, TPH2, DRD2, and OPRM1.

RNAi screens (Kuiken and Beijersbergen 2010; Majewski and Bernards 2011), as a method of performing genome-scale loss-of-function genetic screens, could help to find causal relationships between genes and phenotypes especially in mammalian cells. RNAi screens have some potential benefits for biomarker discovery: (1) requires less time and less samples to build genes–phenotypes relationship, (2) provide mechanistic insights into the drug operates pathways, (3) combine synthetic lethal RNAi into therapy targets, (4) ensure right translational strategy. Due to these advantages, RNAi screens have not only been used to indentify biomarkers of drug responsiveness but also have been used to find which biomarkers' ruppression that improve the anticancer ability of the particular drug.

10.3 Pharmacoproteomic Biomarkers

Proteomics is the study of protein complement of the genome of an organism, including protein–protein and protein–nucleic acid interactions. Proteomics now have become a key technology to study the effects of drug treatment and metabolism. Pharmacoproteomics is just a field that understands the role of proteomics in drug discovery and development, which is now becoming an increasingly important area for the discovery of biomarkers on the protein level.

Recent advances in proteomics technologies have been applied to the field of Pharmacoproteomics. Witzmann and Grant (2003) have concluded the various proteomics technologies by dividing into two areas. For instances, the combination of Mass Spectrometry (MS) with Two-Dimensional Gel Electrophoresis (2DE) and liquid chromatography separations, and protein microarrays for expression proteomics. In functional proteomics, the main tools and applications are investigating protein–protein interactions (PPIs) for protein complexes and Yeast 2-Hybrid. Among them, 2DE and protein microarrays are two widely used methods to indentify novel biomarkers. SELDI–ToF–MS, as a novel mass spectrometry based technologies, is a promising approach for the identification of novel biomarkers. Seibert et al. (2005) have discussed the usefulness of SELDI-ToF-MS in identification of serum markers for ovarian and gastric cancer. In addition, proteome profiling has been extensively used for identification and monitoring of specific biomarkers for chemical toxicity (Merrick and Bruno 2004). In combination with other methods including functional imaging, biosensors and computational biology, proteomics can predict drug response, resistance and toxicity much better (Ross et al. 2005).

Human saliva proteomics has proven to be a novel approach for validation and discovery of target, efficacy and toxicity of candidate drugs assessment, disease subgroups identification, and response predication of individual patients. Hu et al. (2007) have given a brief overview of the application of human saliva proteome analysis in biomarker detection in human cancers. Based on shotgun proteomics and 2-DE/MS approaches, human epidermal growth factor receptor-2 (her-2) has been found to be a biomarker that response to trastuzumab for breast cancer. Serum proteomics have also found several potential protein biomarkers for prediction of response to biologics in rheumatoid arthritis treated with infliximab (Ortea et al. 2012). Kondo et al. (2013) performed proteomic studies on gastrointestinal stromal tumor and identified that potassium channel tetramerization domain containing 12 (KCTD 12) could be a novel prognostic biomarker. In addition, the combination of pharmacogenomics and iTRAQ-coupled LC-MS/MS based pharmacoproteomics has been used to analyze plasma protein profiles of patients (Saminathan et al. 2010). Research on epoxide reductase complex subunit 1 (VKORC1) gene suggests suggest transthyretin precursor as a potential biomarker that response to warfarin anticoagulant therapy.

10.4 Pharmacometabolomic Biomarkers

As a latest technology, metabolomics complements proteomics and transcriptomics, providing a comprehensive understanding of cellular functions. Metabolomics usually studies drug metabolism at the global level. The technologies in the field of metabolomics used include nuclear magnetic resonance (NMR) spectroscopy, direct infusion mass spectrometry, and/or infrared spectroscopy, and other complied tools. Metabolomics technologies have two main advantages. One is it can analyze bodily fluids, isolated cells, and biopsy material, the other is it can monitor biological samples, as well as analyses multiple pathways and metabolite arrays simultaneously.

Metabolomics has important potential implications for pharmacologic science, which leading to a new area of pharmacometabolomics (Kaddurah-Daouk et al. 2008). The aim of pharmacometabolomics refers to the prediction of the drug response of a particular individual, and then permit continued treatment with personalized medicine, which depends on the variations in their drug metabolism and ability to respond to treatment. Thus, pharmacometabolomics is thought to provide information in conjunction with pharmacogenomics and pharmacoproteomics. In comparison with pharmacogenetic, pharmacometabolomics focuses on the identification of metabolic pathways at the genetic level. Particularly, integrating GWAS data into metabilomic analysis may provide additional information for discovering biomarkers (Robinette et al. 2012).

In clinical practice, it is routine to evaluate certain biomarkers, such as bilirubin, serum alanine aminotransferase (ALT) and aspartate aminotransferase (AST) to diagnose liver disease. Mamas et al. (2011) discussed the application of metabolomics biomarkers in various diseases, including metabolic and cardiovascular disease, as well as cancer. Therefore, pharmacometabolomics is a complementary tool for drug target identification and validation. In a serum metabolomic analysis, Chen et al. (2008) identified stearoyl-CoA desaturase 1 (SCD1) and its related lipid species, which may be served as potential targets for treatment of inflammatory diseases. Suhre et al. (2011) genetically determined metabolite traits were reported with strong association for various diseases. The identification of associated metabolic traits may generate many new hypotheses for biomedical and pharmaceutical research. Among these genes, SLC16A9 (MCT9) was demonstrated as a carnitine efflux transporter responsible for carnitine efflux from absorptive epithelia into the blood. Wei (2011) conducted a target-based metabolomics study to characterize metabolic response of Huh 7.5 cells to genomic perturbation of HIF-1. The results identify a new therapeutic target by confirming HIF-1's regulatory role in tumor energy metabolism. Saliva metabolomic analysis has proven to be a novel approach in the search for metabolite biomarkers for noninvasive detection of human diseases (Zhang et al. 2012a).

10.5 Examples of Biomarkers for Pharmacological Activity

10.5.1 Cancer Biomarkers

Many cancers belong to a type of heterogeneous disease. Cancer biomarkers play an important role in understanding the molecular mechanism of cancer, with the help of the recent technologies advances in pharmacogenomic, pharmacoproteomics, pharmacometabomics, and the combination of them. A large number of biomarkers for drug response in cancer have been studied during the past years. Biomarkers specific for diseases, especially various cancers will be described in this section, as shown in Table 10.1. It should be noted that we just list some examples, not an inclusive list.

Colorectal cancer (CRC) Colorectal cancer is a kind of cancer from uncontrolled cell growth in the colon, in which most of them caused by underlying genetic disorders happened with the lifestyle and increasing age. CRC is the third most common cancer for both gender, following breast, lung, and prostate cancers. Pullarkat et al. (2001) used RT-PCR to analysis the response of thymidylate synthase gene polymorphism to 5-Fluorouracil (5-FU) chemotherapy in CRC, suggesting it could be a potential mRNA biomarker. Cetuximab or panitumumab

Table 10.1 Cancer biomarkers for drug response

Cancers	Drugs	Biomarkers	References
Colorectal cancer	5-FU	Thymidylate synthase	Pullarkat et al. (2001)
Colorectal cancer	Cetuximab	EGFR	De Roock et al. (2008), Lièvre et al. (2008), Tabernero et al. (2011)
Colorectal cancer	Irinotecan	UGT1A1*28 1*28	Hoskins et al. (2007)
Colorectal cancer	Bevacizumab	VEGFR	Strimpakos et al. (2009)
Breast cancer	Trastuzumab	HER-2/neu	Vogel et al. (2002)
Breast cancer	Tamoxifen	HER-2	Paik et al. (2004)
Breast cancer	PF-03084014	HEY2, HEY3, HEY4	Zhang et al. (2012b)
Breast cancer	Tamoxifen	CYP2D6	Punglia et al. (2008)
Breast cancer	Tamoxifen	CYP2C19	Schroth et al. (2007)
Lung cancer	Erlotinib	EGFR	Shepherd et al. (2005)
Lung cancer	Gefitinib	EGFR	Mok et al. (2009)
Lung cancer	Cetuximab	EGFR	Pirker et al. (2009)
Prostate cancer	Tissue-Associated Antigens	IgG	Smith et al. (2011)
Prostate cancer	Dasatinib and Sunitinib	Cav-1	Tahir et al. (2012)
Pancreatic cancer	BSI-201	BRCA2	Fogelman et al. (2011)
Bladder cancer	Platinum-based adjuvant chemotherapy	ERCC1	Bellmunt et al. (2007)
Bladder cancer	Adjuvant chemotherapy	MDR1	Hoffmann et al. (2010)
Thyroid cancer	Motesanib	PlGF and VEGF	Bass et al. (2010)
Renal cell carcinoma	Sunitinib	VEGF, VEGF-2, VEGF-3	De Primo et al. (2007)
Melanoma	Vemurafenib	BRAF V600E	Long et al. (2011)

is efficient in anti-EGFR receptor therapy of CRC. Several genomics analyses based on DNA extraction and RNA expression profiling showed that KRAS mutation is a candidate marker associated with resistance to cetuximab treatment in CRC (De Roock et al. 2008; Lièvre et al. 2008; Tabernero et al. 2011). Overall, the standard agents licensed for use in CRC include conventional cytotoxics, such as fluoropyrimidines, and targeted agents, such as, cetuximab, panitumumab and bevacizumab (Strimpagos et al. 2009). Inhibitors against the EGFR and VEGFR proteins have been demonstrated to be the most common biomarkers for predicting drug response.

Breast cancer Breast cancer is very common in women, and trastuzumab has become an efficient therapy treat for breast cancer. Human epidermal growth factor receptor (HER) 2/neu gene amplification has been proved to be a marker that responds to both trastuzumab and tamoxifen treatments in breast cancer (Paik et al. 2004; Vogel et al. 2002). In breast cancer, variants of the encoding cytochrome P450 2D6 (CYP2D6) gene and CYP2C19 polymorphism could also plays roles in predicting tamoxifen therapy (Schroth et al. 2007). In particular, Punglia et al. (2008) created a Markov model to determine the choice of optimal adjuvant endocrine therapy of CYP2D6 in breast tumor. Despite these achievements, other cytochrome P450 enzymes may be also related to the clinical outcome of tamoxifen-treated breast cancer patients, which needs more investigations. More recently, a small molecule γ-secretase inhibitor PF-03084014 has been applied in breast cancer clinical investigation (Zhang et al. 2012b). Some notch pathway target genes including HEY2, HES4, and HES3 also response to PF-03084014 treatment of breast cancer.

Lung cancer Lung cancer is the leading cause of cancer death both for men and women. As monoclonal antibodies targeting EGFR, Erlotinib, gefitinib, and cetuximab have been studied extensively in the treatment of Lung cancer. EGFR mutation has been proved to be a predicator of the efficacy of erlotinib (Shepherd et al. 2005), gefitinib (Mok et al. 2009), and cetuximab in lung cancers (Pirker et al. 2009). For the crizotinib treatment in lung cancer, two genes of echinoderm microtubule-associated protein-like 4 (EML4) and the anaplastic lymphoma kinase (ALK) are two promising candidate biomarkers (Soda et al. 2007).

Prostate cancer Prostate cancer is a leading cause of cancer-related death of men through the whole-world. Prostate-specific antigen is the most useful biomarker for detecting prostate cancer. Smith et al. (2011) indentified that IgG responses to a panel of tissue-associated antigens in prostate cancer. Tahir et al. (2012) showed that a serum maker Serum caveolin-1 (Cav-1) could be a biomarker of response to both dasatinib and sunitinib treatment in Prostate cancer.

Pancreatic cancer Fogelman et al. (2011) investigated the drug response of pancreatic cancer treated with niparib (BSI-201), showing that germline BRCA2 mutation should be a premising biomarker.

Bladder cancer Bellmunt et al. (2007) suggested that excision repair cross-complementing 1 (ERCC1) is a biomarker for platinum-based adjuvant chemotherapy in bladder cancer. This biomarker has been confirmed by Hoffmann et al. (2010), as well as multidrug resistance gene 1 (MDR1) is also proposed as a related biomarker.

Thyroid cancer Bass et al. (2010) have shown that serum placental growth factor (PlGF) and vascular endothelial growth factor (VEGF) are two predicting biomarkers in thyroid cancer with treatment of motesanib, which is always accompanied with antiangiogenic therapies.

Renal cell carcinoma (RCC) De Primo et al. (2007) suggested that VEGF, soluble VEGFR-2, and a novel soluble variant of VEGFR-3 could be potential biomarkers that response to sunitinib in RCC.

Melanoma The last cancer shown is melanoma, which is the leading cause of death from skin disease. Long et al. (2011) conducted a survival analysis of how BRAF mutation status correlated with clinicopathologic features and outcome in melanoma. BRAF V600E was shown to have response to vemurafenib treatment in melanoma.

10.5.2 Biomarkers for Other Diseases

Epilepsy is one of the most prevalent chronic neurologic syndromes, which affects an estimated 50 million people worldwide. The meta-analysis (Grover and Kukreti 2013) of the published studies of reported genetic variants from ABCC2 showed that it is a useful biomarker on drug response in patients with epilepsy (PWE). Other pharmacogenetic biomarkers for PWE include transporters ABCC1, and ABCC5. In autoimmune diseases, ATR-107 is an antibody that targets the IL-21 receptor. IL-21 induced phosphorylation of STAT3 (pSTAT3) can be used as a biomarker to evaluate the target engagement of ATR-107 in human whole blood (Zhu et al. 2013). In chronic lymphocytic leukemia (CLL), Saddler et al. (2007) found that the gene p53 is the important biomarker of response to murine double minute 2 inhibitors in CLL. This work analysis genome-wide change of copy number from single-nucleotide polymorphism (SNP) arrays to identify p53 status.

10.6 Bioinformatics for Biomarkers

The open-source data repositories and powerful bioinformatics tools have been grown and developed for supporting new biomarkers discovery research. This section will discuss key information resources: FDA labels and PharmGKB, and a web-based tool: OmniBiomarker, as well as their application in identifying biomarkers for pharmacological activity.

10.6.1 FDA Labels

US Food and Drug Administration (FDA), the biggest genomic biomarker treasure was launched in 2006, which focuses on developing biomarkers for use in regulatory decision making, as well as biomarker discovery (Wagner et al. 2007). FDA classifies current biomarkers involved in drug response into three types: ‘known valid biomarkers’, ‘probable valid biomarkers’ and ‘exploratory or research biomarkers’ (Glauser 2011; Goodsaid and Frueh 2006). Therefore, it requires several prerequisites and evidentiary standards to validate and qualify biomarkers before practice. FDA provides a tool to evaluate qualification data for biomarkers for efficient drug development. The qualification pilot process for biomarkers at FDA was described by Goodsaid and Frueh (2007).

In addition, FDA also identifies drug labels and determines their drug response. FDA-approved drug labels provide a comprehensive list of these markers and links to pharmacogenomic data, which still need genetic testing thereby for reaching the therapeutic decision. In the last decade, there is a significant increase of labels containing such biomarkers and drug information. A tabulated overview of valid genomic biomarkers in the context of approved drug labels are summarized in the FDA website (http://www.fda.gov/default.htm). Here, we just list some examples of FDA-approved drugs for EGFR with pharmacogenomic information in their labels (Table 10.2).

10.6.2 PharmGKB

Pharmacogenomics Knowledge Base (PharmGKB) is a comprehensive database for pharmacogenomic biomarkers (Klein et al. 2001; Hewett et al. 2002; Thorn et al. 2010). PharmGKB can collect pharmacogenomic data from a variety of sources and provides knowledge about the impact of genetic variation on drug response for researchers. The content of the PharmGKB database have variant annotations, drug-centered pathway, very important pharmacogene summaries, clinical annotations, pharmacogenomics-based drug-dosing guidelines, and drug labels with pharmacogenomic information. PharmGKB currently contains over 25,000 genes under study, over 100 pathways and large ontologyies of pharmacogenetics concepts.

PharmGKB allows the application in systematic pharmacogenomic analysis of biomarkers. The pharmacogenomic biomarkers can be accessed from the PharmGKB website (http://www.pharmgkb.org/). Firstly, the genomic variants and their automated annotation, aggregation, and integration can help users to find new drug-gene interactions. Then, very important pharmacogenes provide a concise summary of which gene is very important in differential drug response. Finally, bioinformatics tools, such as text mining, will be used to extract PharmGKB data for the clinical use.

10.6.3 OmniBioMarker

OmniBioMarker (Phan et al. 2009a, b), a knowledge-driven biomarker identification and data combination, is a very famous web-based bioinformatics tool for

Table 10.2 List of EGFR in drug labels (adopted form FDA website)

Drug	Therapeutic area	Biomarkers	Label sections
Cetuximab (1)	Oncology	EGFR	Indications and usage, warnings and precautions, description, clinical pharmacology, clinical studies
Erlotinib	Oncology	EGFR	Clinical pharmacology
Gefitinib	Oncology	EGFR	Clinical pharmacology
Panitumumab (1)	Oncology	EGFR	Indications and usage, warnings and precautions, clinical pharmacology, clinical studies

biomarker discovery, optimization and clinical validation, by high-throughput analyzing various microarray data. Now, omniBioMarker can be accessed via http://omnibiomarker.bme.gatech.edu/. OmniBioMarker provides both the available samples and the selection algorithm for the discovery of new biomarkers. For the application of this software tool in biomarker identification and clinical validation it includes several steps: (1) collecting high-throughput "omics" datasets form microarray gene expression, (2) using previous biological knowledge to guide the feature and algorithm section, and using score that represent maximal biological relevance to rank previously validated genes, (3) applying biotechnology such as real-time polymerase chain reactions (RT-PCR) to validate candidate biomarkers. In particular, omniBiomarker uses the Cancer Gene Index (CGI) as the biomarker knowledge base, to select suitable algorithm for cancers. Phan et al. (2012) have discovered a list of novel biomarkers for renal, prostate, liver, and pancreatic cancer (see Table 1 in Phan et al. 2012).

10.7 Conclusion

Biomarkers are playing an increasingly important role, not only in marking particular physiological state, but also in the prediction of drug response. New high-throughput "omic" technologies including pharmacogenomic, pharmacoproteomic, and pharmacometabolomic are advancing the field of biomarker discovery in drug response. As such, the flood of pharmacological biomarkers is opening the new door into personalized medicine. However, there still a lot of open questions in these new discoveries. Although methods for biomarker discovery are developing rapidly, the combination of these methods is still needed. Using bioinformatics tools for qualification and valuation of biomarkers sets another important challenge. To this end, we are currently using the method of meta-analysis to evaluate the potential function of biomarkers (Yuan et al. 2013). We are looking forward to an exciting future in this area.

Acknowledgments This work was supported by the National Nature Science Foundation of China (21203131), the Natural Science Foundation of the Jiangsu Higher Education Institutions (12KJB180014).

References

Aslibekyan S, Kabagambe EK, Irvin MR, Straka RJ, Borecki IB, Tiwari HK, Tsai MY, Hopkins PN, Shen J, Lai CQ, et al. A genome-wide association study of inflammatory biomarker changes in response to fenofibrate treatment in the genetics of lipid lowering drug and diet network (GOLDN). Pharmacogenet Genom. 2012;22:191–7.

Bass MB, Sherman SI, Schlumberger MJ, Davis MT, Kivman L, Khoo HM, Notari KH, Peach M, Hei YJ, Patterson SD. Biomarkers as predictors of response to treatment with motesanib in patients with progressive advanced thyroid cancer. J Clin Endocrinol Metab. 2010;11: 5018–5020.

Bellmunt J, Paz-Ares L, Cuello M, et al. Gene expression of ERCC1 as a novel prognostic marker in advanced bladder cancer patients receiving cisplatin-based chemotherapy. Ann Oncol. 2007;18:522–8.

Brandi et al. Challenges faced in the integration of pharmacogenetics/genomics into drug development. J Pharmacogenom Pharmacoproteomics. 2012;3:2.

Burczynski ME, Dorner AJ. Transcriptional profiling of peripheral blood cells in clinical pharmacogenomic studies. Pharmacogenomics. 2006;7:187–202.

Chen C, Shah YM, Morimura K, Krausz KW, Miyazaki M, et al. Metabolomics reveals that hepatic stearoyl-CoA desaturase 1 downregulation exacerbates inflammation and acute colitis. Cell Metab. 2008;7:135–47.

Cui J, Stahl EA, Saevarsdottir S, Miceli C, Diogo D, et al. Genome-wide association study and gene expression analysis identifies CD84 as a predictor of response to etanercept therapy in rheumatoid arthritis. PLoS Genet. 2013;9:e1003394.

De Koning P, Keirns J. Clinical pharmacology, biomarkers and personalized medicine: education please. Biomark Med. 2009;3:685–700.

De Primo SE, Bello CL, Smeraglia J, et al. Circulating protein biomarkers of pharmacodynamic activity of sunitinib in patients with metastatic renal cell carcinoma: modulation of VEGF and VEGF-related proteins. J Trans Med. 2007;5:32.

De Roock W, Piessevaux H, De Schutter J, et al. KRAS wild-type state predicts survival and is associated to early radiological response in metastatic colorectal cancer treated with cetuximab. Ann Oncol. 2008;19:508–15.

Fogelman DR, Wolff RA, Kopetz S, Javle M, Bradley C, Mok I, et al. Evidence for the efficacy of Iniparib, a PARP-1 inhibitor, in BRCA2-associated pancreatic cancer. Anticancer Res. 2011;31:1417–20.

Frank R, Hargreaves R. Clinical biomarkers in drug discovery and development. Nat Rev Drug Discov. 2003;2:566–80.

Glauser TA. Biomarkers for antiepileptic drug response. Biomark Med. 2011;5:635–41.

Goodsaid F, Frueh F. Process map proposal for the validation of genomic biomarkers. Pharmacogenomics. 2006;7:773.

Goodsaid F, Frueh F. Biomarker qualification pilot process at the US food and drug administration. AAPS J. 2007;9:E105–8.

Grover S, Kukreti R. A systematic review and meta-analysis of the role of ABCC2 variants on drug response in patients with epilepsy. Epilepsia. 2013;54:936–45.

Hewett M, Oliver DE, Rubin DL, Easton KL, Stuart JM, Altman RB, Klein TE. PharmGKB: the pharmacogenetics knowledge base. Nucleic Acids Res. 2002;30:163–5.

Hoffmann AC, Wild P, Leicht C, et al. MDR1 and ERCC1 expression predict outcome of patients with locally advanced bladder cancer receiving adjuvant chemotherapy. Neoplasia. 2010;12:628–36.

Hoskins JM, Goldberg RM, Qu P, Ibrahim JG, McLeod HL. UGT1A1*28 genotype and irinotecan-induced neutropenia: dose matters. J Natl Cancer Inst. 2007;99:1290–1295.

Hu S, Yen Y, Ann D, Wong DT. Implications of salivary proteomics in drug discovery and development: a focus on cancer drug discovery. Drug Discov Today. 2007;12:911–6.

Jain KK. Role of pharmacoproteomics in the development of personalized medicine. Pharmacogenomics. 2004;5:331–6.

Jain KK. The handbook of biomarkers. New York: Springer Science; 2010.

Jones S, et al. Frequent mutations of chromatin remodeling gene ARID1A in ovarian clear cell carcinoma. Science. 2010;330:228–31.

Kaddurah-Daouk R, Kristal B, Weinshilboum R. Metabolomics: a global biochemical approach to drug response and disease. Annu Rev Pharmacol Toxicol. 2008;48:653–83.

Karczewski KJ, Daneshjou R, Altman RB. Chapter 7: Pharmacogenomics. PLoS Comput Biol. 2012;8:e1002817.

Kelloff GJ, Sigman CC. Cancer biomarkers: selecting the right drug for the right patient. Nat Rev Drug Discov. 2012;11:201–14.

Klein TE, Chang JT, Cho MK, Easton KL, Fergerson R, et al. Integrating genotype and phenotype information: an overview of the PharmGKB project. pharmacogenetics research network and knowledge base. Pharmacogenomics. 2001;J1:167–70.

Klotz U. The role of pharmacogenetics in the metabolism of antiepileptic drugs: pharmacokinetic and therapeutic implications. Clin Pharmacokinet. 2007;46:271–9.

Kondo T, Suehara Y, Kikuta K, Kubota D, Tajima T, Mukaihara K, Ichikawa H, Kawai A. Proteomic approach toward personalized sarcoma treatment: lessons from prognostic biomarker discovery in gastrointestinal stromal tumor. Proteomics Clin Appl. 2013;7:70–8.

Kuiken HJ, Beijersbergen RL. Exploration of synthetic lethal interactions as cancer drug targets. Future Oncol. 2010;6:1789–802.

Li L, Fridley BL, Kalari K, Jenkins G, Batzler A, et al. Gemcitabine and arabinosylcytosin pharmacogenomics: genome-Wide association and drug response biomarkers. PLoS ONE 2009;4: e7765.

Lièvre A, Bachet JB, Boige V, et al. KRAS mutations as an independent prognostic factor in patients with advanced colorectal cancer treated with cetuximab. J Clin Oncol. 2008;26:374–9.

Long GV, Menzies AM, Nagrial AM, et al. Prognostic and clinicopathologic associations of oncogenic BRAF in metastatic melanoma. J Clin Oncol. 2011;29:1239–46.

Ludwig JA, Weinstein JN. Biomarkers in cancer staging, prognosis and treatment selection. Nat Rev Cancer. 2005;5:845–56.

Majewski IJ, Bernards R. Taming the dragon: genomic biomarkers to individualize the treatment of cancer. Nat Med. 2011;17:304–12.

Mamas M, Dunn WB, Neyses L, Goodacre R. The role of metabolites and metabolomics in clinically applicable biomarkers of disease. Arch Toxicol. 2011;85:5–17.

Mamdani F, Berlim MT, Beaulieu M-M, Labbe A, Merette C, Turecki G. Gene expression biomarkers of response to citalopram treatment in major depressive disorder. Transl Psychiatry. 2011;1:e13.

Masica DL, Karchin R. Collections of simultaneously altered genes as biomarkers of cancer cell drug response. Cancer Res. 2013;73:1699–708.

Matsui S. Genomic biomarkers for personalized medicine: development and validation in clinical studies. Comput Math Method M. 2013;2013):865980.

Mendrick DL. Genomic and genetic biomarkers of toxicity. Toxicology. 2008;245:175–81.

Mendrick DL. Transcriptional profiling to identify biomarkers of disease and drug response. Pharmacogenomics. 2011;12:235–49.

Merrick BA, Bruno ME. Genomic and proteomic profiling for biomarkers and signature profiles of toxicity. Curr Opin Mol Ther. 2004;6:600–7.

Mok T, Wu Y, Thongprasert S, et al. Gefitinib or carboplatin–paclitaxel in pulmonary adenocarcinoma. N Engl J Med. 2009;361:947–57.

Niu N, Qin Y, Fridley BL, Hou J, Kalari KR, Zhu M, Wu TY, Jenkins GD, Batzler A, Wang L. Radiation pharmacogenomics: a genome-wide association approach to identify radiation response biomarkers using human lymphoblastoid cell lines. Genome Res. 2010;20:1482–92.

Ong FS, Das K, Wang J, et al. Personalized medicine and pharmacogenetic biomarkers: progress in molecular oncology testing. Expert Rev Mol Diagn. 2012;12:593–602.

Ortea I, Roschititzki B, Ovalles JG, et al. Discovery of serum proteomic biomarkers for prediction of response to infliximab (a monoclonal anti-TNF antibody) treatment in rheumatoid arthritis: an exploratory analysis. J Proteomics. 2012;77:372–82.

Paik S, Shak S, Tang G, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med. 2004;351:2817–26.

Parissenti AM, Hembruff SL, Villeneuve DJ, Veitch Z, Guo B, Eng J. Invited review: gene expression profiles as biomarkers for the prediction of chemotherapy drug response in human tumour cells. Anticancer Drugs. 2007;18:499–523.

Phan JH, Moffitt RA, Stokes TH, Liu J, Young AN, Nie S, Wang MD. Convergence of biomarkers, bioinformatics and nanotechnology for individualized cancer treatment. Trends Biotechnol. 2009a;27:350–8.

Phan JH, Yin-Goen Q, Young AN, Wang MD. Improving the efficiency of biomarker identification using biological knowledge. Pac Symp Biocomput. 2009b;14:427–38.

Phan JH, Young AN, Wang MD. omniBiomarker: a web-based application for knowledge-driven biomarker identification. IEEE Trans Biomed Eng. 2012;PP (99).

Pirker R, Pereira JR, Szczesna A, et al. Cetuximab plus chemotherapy in patients with advanced non-small-cell lung cancer (FLEX): an open-label randomised Phase III trial. Lancet. 2009;373:1525–31.

Pullarkat ST, Stoehlmacher J, Ghaderi V, et al. Thymidylate synthase gene polymorphism determines response and toxicity of 5-FU chemotherapy. Pharmacogenomics J. 2001;1:65–70.

Punglia R, Winer E, Weeks J. Pharmacogenomic variation of CYP2D6 and the choice of optimal adjuvant endocrine therapy for postmenopausal breast cancer: a modeling analysis. J Natl Cancer Inst. 2008;100:642–8.

Robinette SL, Holmes E, Nicholson JK, Dumas ME. Genetic determinants of metabolism in health and disease: from biochemical genetics to genome-wide associations. Genome Med. 2012;4:30.

Roses AD. Pharmacogenetics and the practice of medicine. Nature. 2000;405:857–65.

Ross JS, Symmans WF, Pusztai L, Hortobagyi GN. Pharmacogenomics and clinical biomarkers in drug discovery and development. Am J Clin Pathol. 2005;124:S29–41.

Saddler C, Ouillette P, Kujawski L, et al. Comprehensive biomarker and genomic analysis identifies P53 status as the major determinant of response to MDM2 inhibitors in chronic lymphocytic leukemia. Blood. 2007;111:1584–93.

Sadee W, Wang D, Papp AC, Pinsonneault JK, Smith RM, Moyer RA, Johnson AD. Pharmacogenomics of the RNA world: structural RNA polymorphisms in drug therapy. Clin Pharmacol Ther. 2011;89:355–65.

Saminathan R, et al. VKORC1 pharmacogenetics and pharmacoproteomics in patients on warfarin anticoagulant therapy: transthyretin precursor as a potential biomarker. PLoS One. 2010;5:e15064.

Schroth W, et al. Breast cancer treatment outcome with adjuvant tamoxifen relative to patient CYP2D6 and CYP2C19 genotypes. J Clin Oncol. 2007;25:5187–93.

Seibert V, Ebert MP, Buschmann T. Advances in clinical cancer proteomics: SELDI-TOF-mass spectrometry and biomarker discovery. Brief Funct Genomic Proteomic. 2005;4:16–26.

Shepherd F, Rodrigues Pereira J, Ciuleanu T, et al. Erlotinib in previously treated non-small-cell lung cancer. N Engl J Med. 2005;353:123–132.

Sim SC, Ingelman-Sundberg M. Pharmacogenomic biomarkers: new tools in current and future drug therapy. Trends Pharmacol Sci. 2011;32(2):72–81.

Sinha A, Singh C, Parmar D, et al. Proteomics in clinical interventions: achievements and limitations in biomarker development. Life Sci. 2007;80:1345–54.

Smith HA, Maricque BB, Eberhardt J, Petersen B, Gulley JL, Schlom J, McNeel DG. IgG responses to tissue-associated antigens as biomarkers of immunological treatment efficacy. J Biomed Biotech. 2011;2011:454861.

Soda M, et al. Identification of the transforming EML4–ALK fusion gene in non-small-cell lung cancer. Nature. 2007;448:561–6.

Stewart JD, Bolt HM. Metabolomics: biomarkers of disease and drug toxicity Arch. Toxicol. 2011;85:3–4.

Strimpagos AS, Syrigos KN, Saif MW. Pharmacogenetics and biomarkers in colorectal cancer. Pharmacogenomics J. 2009;9:147–60.

Suhre K, Shin SY, Petersen AK, Mohney RP, Meredith D, et al. Human metabolic individuality in biomedical and pharmaceutical research. Nature. 2011;477:54–60.

Tabernero J, Cervantes A, Rivera F, et al. Pharmacogenomic and pharmacoproteomic studies of cetuximab in metastatic colorectal cancer: biomarker analysis of a phase I dose-escalation study. J Clin Oncol. 2011;28:1181–9.

Tahir SA, Kurosaka S, Tanimoto R, Goltsov AA, Park S, Thompson TC. Serum caveolin-1, a biomarker of drug response and therapeutic target in prostate cancer models. Cancer Biol Ther. 2012;14:117–26.

Thorn CF, et al. Pharmacogenomics and bioinformatics: pharmGKB. Pharmacogenomics. 2010;11:501–5.

Vogel CL, Cobleigh MA, Tripathy D, et al. Efficacy and safety of trastuzumab as a single agent in first-line treatment of HER2-overexpressing metastatic breast cancer. J Clin Oncol. 2002;20:719–26.

Wagner JA, Williams SA, Webster CJ. Biomarkers and surrogate end points for fit-for-purpose development and regulatory evaluation of new drugs. Clin Pharmacol Ther. 2007;8:104–7.

Wei R. Metabolomics and its practical value in pharmaceutical industry. Curr Drug Metab. 2011;12:345–58.

Wiegand KC, et al. ARID1A mutations in endometriosis-associated ovarian carcinomas. N Engl J Med. 2010;363:1532–43.

Wittmann J, Jack HM. Serum microRNAs as powerful cancer biomarkers. Biochim Biophys Acta. 2010;1806:200–7.

Witzmann FA, Grant RA. Pharmacoproteomics in drug development. Pharmacogenomics J. 2003;3:69–76.

Yuan H, Huang J, Lv B, Yan W, Hu G, Wang J, Shen B. Diagnosis value of the serum amyloid a test in neonatal sepsis: a meta-analysis. BioMed Res Int. 2013;520294.

Zhang A, Sun H, Wang X. Saliva metabolomics opens door to biomarker discovery, disease diagnosis, and treatment. Appl Biochem Biotechnol. 2012a;168:1718–27.

Zhang CC, Pavlicek A, Zhang Q, et al. Biomarker and pharmacologic evaluation of the gamma-secretase inhibitor PF-03084014 in breast cancer models. Clin Cancer Res. 2012b;18:5008–19.

Zhu M, Pleasic-Williams S, Lin TH, Wunderlich DA, Cheng JB, Masferre JL. pSTAT3: a target biomarker to study the pharmacology of the anti-IL-21R antibody ATR-107 in human whole blood. J Transl Med. 2013;11:65.

Chapter 11
Network Biomarkers for Diagnosis and Prognosis of Human Prostate Cancer

Jiajia Chen and Bairong Shen

Abstract Prostate cancer is one of the most lethal malignancies worldwide, owing to the lack of precise markers for early diagnosis. Researchers are now routinely identifying biomarkers for prostate cancer using whole-genome expression profiling along with proteomic technologies. Although there has been some success in this field, many efforts have been complicated by the fact that individual markers are highly divergent. Prostate cancer is a systems biology disease that results from the accumulated mutations acting in concert. Hence the individual markers would fail to capture the heterogeneity of carcinogenesis. As molecular interaction networks become available for human, network-level biomarker evolves as a promising methodology that can address this challenge. In this chapter we first describe some foundations of network analysis, and then introduce the recent progress in network biomarker discovery for diagnosis and prognosis of human prostate cancer.

Keywords Prostate cancer · Biomarker · Network · Diagnosis · Prognosis

11.1 Introduction

Prostate cancer (PCa) is the most common cancer among males and one of the leading causes of cancer deaths worldwide (Siegel et al. 2012). It is estimated by the cancer statistics (Siegel et al. 2013) that there will be 238,590 new cases and 29,720 deaths from PCa in the United States in the year 2013. The mortality and recurrence rate are projected to continue rising. This has rendered PCa a public health problem which is in need of sensitive diagnostic and prognostic markers.

J. Chen · B. Shen (✉)
Center for Systems Biology, Soochow University, No. 1 Shizi Street, Postbox 206,
215006 Suzhou, Jiangsu, China
e-mail: bairong.shen@suda.edu.cn

B. Shen (ed.), *Bioinformatics for Diagnosis, Prognosis and Treatment of Complex Diseases*, Translational Bioinformatics 4, DOI: 10.1007/978-94-007-7975-4_11,

© Springer Science+Business Media Dordrecht 2013

Biomarkers are unique molecules which could serve as the indicators of disease occurrence and progression. Sensitive biomarkers hold great potential for early diagnosis and in some cases they may represent potential drug targets. Since its introduction two decades ago, prostate specific antigen (PSA) screening has been the mainstay for early detection of prostate cancer (Barry 2001). Nonetheless, screening for PSA remains controversial due to the poor specificity. Elevated serum PSA level may be observed in both malignant tumor and non-malignant prostatic disorders. Moreover PSA-based screening has been criticized for over detection and overtreatment of benign tumors which may otherwise never have been diagnosed without PSA screening (Venderbos and Roobol 2011). The major limitation with PSA highlights the need for more reliable and sensitive biomarkers for diagnosis and staging of PCa.

11.2 Current Prostate Cancer Biomarker Discovered by Genomics and Proteomics Technologies

During the past two decades, there have been intense interests in identifying biomarkers for prostate cancer, as represented by the numbers of published papers found in Pubmed (Fig. 11.1).

Recent progress in high throughput genomic technologies, such as array-based methods and next generation sequencing (NGS), enable us to interrogate the prostate cancer genome with higher throughput and improved accuracy. A mature body of studies has characterized the gene expression profiles in prostate cancer. Many differentially expressed genes have been identified for use in the diagnosis, prognosis, subtype classification, as well as the prediction of therapeutic response of PCa.

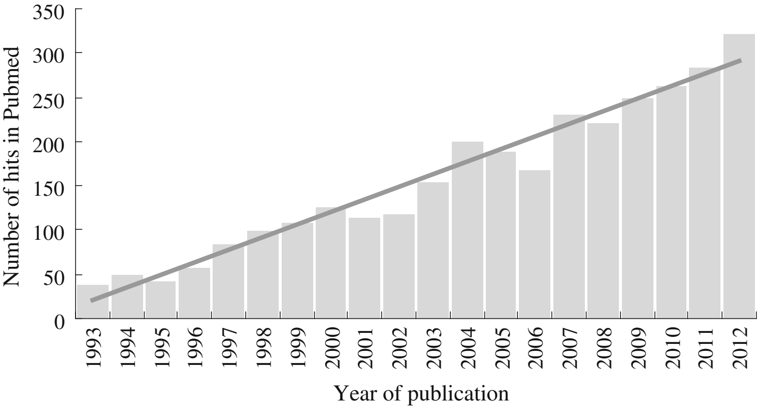


Fig. 11.1 Number of PubMed hits for the query "prostate cancer[ti] and *marker*[tiab]" in the past two decades

With recent advances in proteomic methods, much effort has gone into proteome-scale discovery of novel biomarkers. Current proteomic practice relies on either gel-based [e.g. 2DE, 2D-DIGE (Marouga et al. 2005)] or gel-free separation techniques [e.g. liquid chromatography, cICAT (Hassan et al. 2011), iTRAQ (DeSouza et al. 2005), SILAC (Ong et al. 2002)], followed by mass spectrometry [e.g. MALDI (Kang et al. 2005); SELDI (Merchant and Weinberger 2000) or tandem mass spectrometry] to identify protein biomarkers. These various approaches have been reviewed elsewhere (Domon and Aebersold 2006). Most applications in biomarker discovery aim to determine differential protein expression profiles between malignant and benign samples. Through proteome-wide screening in cancer patients, a list of candidate PCa markers has been identified. These studies used various statistics to quantify the level of differential expression for individual proteins, which are then scored for their discriminative ability between different disease status (e.g. tumors vs. controls, primary vs. metastasis, good vs. poor prognosis, subtype 1 vs. subtype 2). The top-scoring proteins were normally selected as markers.

Some of most prominent diagnostic marker identified to date include alpha-methylacyl-CoA racemase (AMACR) (Luo et al. 2002), prostate cancer gene 3 (PCA3) (Tinzl et al. 2004), early prostate cancer antigen (EPCA)-2 (Leman et al. 2007), hepsin (Luo et al. 2001), kallikrein-related peptidase 2 (KLK2) (Darson et al. 1997) and polycomb group protein enhancer of zeste homolog 2 (EZH2) (Varambally et al. 2002). These molecules were claimed to be more sensitive and specific for PCa detection than PSA, and thus provide a potential complement to PSA for the early diagnosis of PCa.

11.3 Pathway-Level Analysis of Prostate Cancer

While the number of genome-based biomarker analysis is growing exponentially, single gene-based differential expression analysis faces serious challenges owing to limited prediction accuracy, poor reproducibility, and unclear biological relevance. Many of the single gene markers fail to achieve similar performance in validation studies and few molecules will make it to the routine clinical practice. In addition, the marker lists obtained by different research groups do not coincide with each other and share few common candidates (Ein-Dor et al. 2006). A further limitation of these signatures is that they provide poor insight into the molecular mechanisms underlying the carcinogenesis.

These problems are thought to arise as the result of intra-tumor and inter-tumor heterogeneity. In solid tumor diseases such as prostate cancer, it is difficult to separate tumor from normal cells. Therefore, a pure tumor cell population is not easily available. The mixed cell population may dilute the expression profile and make it difficult to get a distinct expression signature. Moreover, the inter-tumor heterogeneity across patients complicates this problem. Inter-tumor heterogeneity refers to the disparity across patients. It’s observed that no single marker is predictive of the phenotype of all patients.

In addition, single gene based approach is thought to be simple and intuitive. These approaches ignore the dependency between genes. It is possible that some of the selected gene markers may be functionally related hence contain redundant information that could reduce the overall prediction power.

There is growing appreciation that cancer is a complex disease. A cancer phenotype is rarely caused by an abnormality in individual genes or proteins, but reflected by functionally related groups of genes or proteins that act in a concerted manner (Chen et al. 2012).

To address the aforementioned limitations, a more effective means of marker identification is needed. Extensive work has been done that extends the level of analysis from an individual gene to groups of functionally related genes, such as pathways. Pathways can be viewed as an ensemble of successive events among a set of genes towards a defined functional outcome. Depending on the scenario, such pathway maps can involve signaling cascades, transcriptional regulation, or metabolic reactions.

Pathway analysis typically correlates seemingly disparate molecular changes together into a common pattern. This is achieved by projecting them onto well-defined biological processes. Known pathways can be readily drawn from pathway databases (listed in Table 11.1).

Predefined biological processes or pathways are then checked for enrichment of differentially expressed genes (DEGs). Statistic approaches using a hypergeometric distribution could be used for enrichment analysis. Enriched pathways including more DEGs than expected by chance are more likely to be the potential candidate markers. Figure 11.2 provides the flowchart of the pathway-based biomarker discovery.

Table 11.1 Prominent pathway databases

Database	Website	Description
Biocarta	http://www.biocarta.com/	The pathway section present gene interactions as dynamic graphical models
DAVID	http://david.abcc.ncifcrf.gov/	Integrates and optimize pathway annotations from BioCarta & KEGG
GO	http://www.geneontology.org/	Provides gene product annotation data and data processing tools
Ingenuity	http://www.ingenuity.com/	High-quality pathway analysis of complex omics data
KEGG	http://www.genome.jp/kegg/	A collection of manually drawn pathway maps representing knowledge on the molecular interaction and reaction networks
MetaCore	http://www.genego.com/	Manually curated databases for pathway analysis and data mining
MetaCyc	http://metacyc.org/	A database of nonredundant, experimentally elucidated metabolic pathways
MSigDB	http://www.broadinstitute.org/gsea/msigdb/index.jsp	A collection of annotated gene sets or pathways from other databases for use with GSEA software

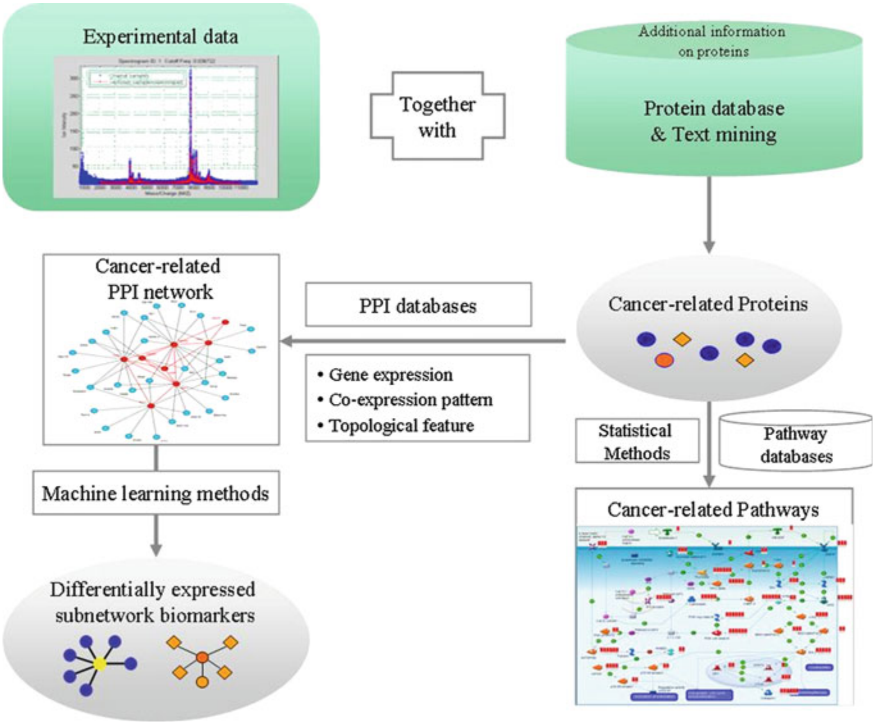


Fig. 11.2 The research pipeline for pathway-based and network-based biomarker discovery

Some studies have provided insight into pathways with relevance to the pathophysiology of prostate cancer. In the seminal work by Rhodes et al. (2002), a statistical model was proposed to meta-analyze independent gene expression datasets. The model was then applied to 4 microarray datasets for prostate cancer and yielded a consistent list of DEGs. Dysregulated genes were subsequently projected to functional annotations and highlighted polyamine and purine biosynthesis as key regulatory pathways with alteration in PCa development. Glinsky et al. (2004) reported that activation of the Wnt signaling pathway along with up-regulation of Wnt5A and down-regulation of KFL6 (COPEB) suggest poor clinical outcome in prostate cancer. Our colleagues, Wang et al. (2011) also sought to identify pathway-level biomarkers by meta-analyzing 10 public prostate cancer microarray expression profiles. Pathways from KEGG and MetaCore were evaluated for enrichment of dysregulated genes. As a result, endothelin-1/EDNRA transactivation of the EGFR pathway was found to be associated with prostate cancer. In a conceptually similar study, Kumar et al. (2011) compared the mutation spectrum of castration-sensitive and castration-resistant PCa cell lines, revealing the role of the Wnt pathway in castration resistance. In a more recent study, Taylor et al. (2010) conducted an integrative genomic profiling of human prostate cancer.

The authors pinpointed the PI3K, RAS/RAF, and AR pathways as potentially contributing to metastasis of PCa.

Unlike single-gene level markers, pathway level markers are more reproducible and tend to have higher discriminative power. In addition, pathways incorporate biological knowledge and thus provide a strong functional interpretation of the genomics data. Although pathway-based approach holds great promise for biomarker discovery, a remaining hurdle is that currently known pathways only cover a small fraction of human genes. As a result, the important genes that have not yet been assigned to a definitive pathway may be excluded from further analysis. Besides, this approach fails to account the extensive correlation between different pathways. Thus, there is a growing trend to shift from a strictly pathway-centric marker discovery toward an integrative network-based approach.

11.4 Network-Based Biomarker Discovery

Network-based analyses assume that gene products associated with cancer often appear as hot spots in PPI networks. This concept is exploited to investigate the overall behavior of genes connected in a larger human protein–protein interaction (PPI) network. There is growing interest in identifying differentially expressed subnetwork, that is, functionally associated genes with coordinate expression changes as novel markers.

The recent availability of systematic yeast two-hybrid and transcriptional interaction screens (Kim et al. 2005) have increased the coverage and quality of human protein interaction database. This improvement in turn, enables further opportunities for molecular characterization of cancer. Network-based approaches have found application in multiple cancer types, including the breast cancer (Chuang et al. 2007; Taylor et al. 2009; Su et al. 2010), colorectal cancer (Nibbe et al. 2010; Chowdhury and Koyuturk 2009), hepatocellular carcinoma (Zhang et al. 2011) and gastric cancer (Liu et al. 2011).

The network-based method has several major advantages over traditional approaches that study individual genes or loci. First, the networks provide functional insight into the mechanistic bases of cancers. Once a list of candidate markers is inferred, a primary task is to interpret them in the biological context. Network data provide abundant information that could be employed for this purpose. Second, network-based method may lead to the identification of non-discriminative cancer genes. Many disease-causing genes are not differentially expressed *per se*, and are often regarded as non-discriminative or low scoring by conventional analysis. However, they are essential for interconnecting many hot spot proteins, forming an integral network whose overall activity is discriminative. Network analysis is able to uncover these non-discriminative genes which, in turn, offer novel targets for drug development. Third, network markers are reported to be more reproducible across different cancer datasets and can improve the prediction accuracy. In a meta-analysis of breast cancer datasets, Chuang et al. (2007) showed that subnetwork markers were significantly more reproducible between patient cohorts than individual marker genes (12.7 vs. 1.3 %). They also ascertained that subnetwork markers are more accurate in the prediction of metastasis in breast cancer.

In the remainder of this chapter, we will describe some foundations of network analysis, and then address recent developments in the identification of network biomarkers that are discriminative of cancer phenotypes. Finally we highlight 4 case studies on prostate cancer and explain how the emerging network-based approaches can discover robust, specific biomarkers of prostate cancer.

11.5 Research Pipelines for Network Biomarker Identification

A network-based approach aims to find significant subnetworks whose members are coordinately dysregulated in cancer samples. The common tasks could be summarized as follows:

1. Discovery and scoring of protein interaction subnetworks which is discriminative of cancer;
2. Subnetworks are used as features to train a classifier;
3. Cross validation experiments are performed to test the effectiveness of network markers.

The general scheme for the network-based biomarker analysis is also outlined in Fig. 11.2.

11.5.1 Searching for PPI Subnetwork with Discriminate Potential

Network construction is fundamental to network-based marker identification. The generation of PPI networks requires a pooled dataset comprising interactions among proteins. The interaction data could be derived from high-throughput experiments, inferred via homology and co-citation, or culled from prior literatures. In these protein interaction networks, nodes are proteins whereas edges are functional correlations that link the proteins.

Recent years have witnessed an exceptional growth in human-specific protein networks. High-throughput experiments, such as affinity purification mass spectrometry, yeast two-hybrid and protein microarrays, have accelerated the determination of protein–protein interaction data. Most of the PPI data are documented in scientific literatures. To make this information more readily available, some initiations have set out to mine PPI information from literatures and to store the curated interactions in various protein interaction databases. Some prominent protein interaction databases are listed in Table 11.2.

Table 11.2 Protein–protein interaction databases

Name	URL	Features	Reference
BIND (Biomolecular interaction network database)	http://www.bind.ca	Biomolecular interaction and pathways with experimental evidence	Bader et al. (2003)
DIP (Database of interacting proteins)	http://dip.doe-mbi.ucla.edu	Experimentally verified protein interactions	Salwinski et al. (2004)
HPRD (Human protein reference database)	http://www.hprd.org	Manually curated human protein interaction networks	Keshava Prasad et al. (2009)
HPID (Human protein interaction database)	http://www.hpid.org	Integration of the protein interactions in BIND, DIP and HPRD	Han et al. (2004)
IntAct	http://www.ebi.ac.uk/intact	Manually curated molecular interaction data from the literature	Kerrien et al. (2007)
MINT (Molecular interaction database)	http://mint.bio.uniroma2.it/mint	Protein interactions for mammals with experimental evidence, mined from the literature	Licata et al. (2011)
STRING (Search tool for the retrieval of interacting genes/proteins)	http://string.embl.de	Both curated and predicted protein–protein interactions	von Mering et al. (2003)
Reactome	http://www.reactome.org	Curated, peer-reviewed resource of human biological processes	Joshi-Tope et al. (2005)
MIPS (Mammalian protein–protein interaction database)	http://mips.helmholtz-muenchen.de/proj/ppi/	Experimental protein interaction data in mammalian organisms	Pagel et al. (2005)
BioGRID (Biological general repository for interaction datasets)	http://thebiogrid.org	Protein and genetic interactions from major model organisms	Chatr-Aryamontri et al. (2012)
Interpro	http://www.ebi.ac.uk/interpro/	Integration of PROSITE, PRINTS, Pfam for protein families, domains, and functional sites	Apweiler et al. (2001)
Orphid (Online predicted human interaction database)	http://ophid.utoronto.ca	Predicted human protein–protein interactions	Brown and Jurisica (2005)

Network-based discovery process begins by obtaining differentially expressed proteins possibly involved in the phenotype. The resultant proteins are then used as seeds to greedily grow subnetworks from them. The seed proteins could be imported into bioinformatics tools that mine the curated protein–protein interactions database or literatures. In this way, the proteins that are closely connected to the seed proteins are extracted. This provided us with a rough protein interaction network that includes both cancer-responsive proteins and their interactive partners.

11.5.2 Scoring Subnetworks

The next step is to search for subnetworks whose activities were highly discriminative of cancer condition. This requires a proper scoring method to quantify the activity of a candidate subnetwork. Several scoring methods have previously been proposed that rank the activity of a subnetwork in response to given condition. The existing methods define activity from different aspects: expression level, coexpression pattern, or the topological features of the subnetwork.

It is a most popular way to use the gene expression data to infer the subnetwork activity. Several works have superimposed gene expression data onto corresponding proteins in the network. Then the aggregate expression levels of member genes in each network are summarized into activity score. Top-scoring subnetwork regions that show significant changes in expression are viewed as potential active subnetworks.

Other groups try to measure the activity using coherent expression patterns between the network members. These scoring methods are based on the hypothesis that expression correlation implies interaction coherence of the protein. For instance, Ideker et al. (2002) proposed an statistical measure for distinguishing condition-relevant modules by the co-expression of the genes members encoding the network. They sums up the standard normal inverse of a single gene's *P*-value (z-score) adjusted for the size of the subnetwork. Chen and Yuan (2006) developed an network-partitioning algorithm that takes into account functional relationship between the proteins. More recently, Taylor et al. (2009) calculated the variations in interaction coherence between members in a subnetwork under investigated condition.

Recently, a novel scoring scheme, edge-based scoring function, has been proposed (Nibbe et al. 2009). The edge-based scoring function utilizes protein structure information (e.g. proximity and connectivity) of the interactome to search for significant subnetworks implicated in cancer. This approach makes an improvement over the previous methods in that it captures both gene expression data and post-transcriptional activity.

11.5.3 Training Classifiers

Once the activity scores are obtained, they are used as feature values to train a classifier that distinguishes the phenotypes with high accuracy. A variety of statistical tools have been established for pattern recognition and classification. Such as logistic regression, support vector machines (SVM), Bayesian networks, k-nearest neighborhood, decision tree, artificial neural networks (ANN) and clustering.

11.5.4 Performance Evaluation

Finally, resultant classifiers should be evaluated for their discriminative power. To assess the classification performance, classifiers are subjected to cross-validation. In cross-validation, the available patient data are divided into 3 subsets: training sets, test sets and validation sets. Training sets are used to build the classifier whereas the test sets are used to evaluate the predictive accuracy of the classification model. For the validation set, an Area under ROC curve (AUC) is reported to optimize the number of markers used in the classifier. Finally, AUC on the test set is calculated as the final classification performance against a random prediction.

11.6 Network-Based Biomarkers in Cancers

Network-based biomarker discovery is emerging rapidly. We have seen a growing number studies on the exploration of cancer related networks. For example, Chuang et al. (2007) conducted a pioneering study to identify subnetwork biomarkers for metastatic breast cancer. Using proteins with high discriminative power as seeds, they searched for interaction networks from protein interaction databases. According to the authors, subnetwork activity is a function of expression of genes in a given subnetwork. The discriminative power of a subnetwork was quantified in terms of mutual information between the phenotype and subnetwork activity. Subnetworks were used as features to train a classifier based on logistic regression. The author demonstrated that subnetwork markers are more accurate than single gene markers and more reproducible between datasets. Taylor et al. (2009) integrated multiple microarray datasets to identify networks responsible for breast cancer prognosis. The authors proposed to search for coordinate dysregulation between genes in the human interactome. They found that the subnetwork markers displayed favorable performance than previous predictors and suggested altered interaction coherence to be potential indicator of breast cancer outcome. Su et al. (2010) proposed a method to identify discriminative paths containing coexpressed differential genes from PPI networks. The linear paths were then greedily combined to obtain reliable subnetworks that can predict breast cancer metastasis. The concept of coordinate dysregulation is also adopted by Nibbe et al. (2010), who integrated protein and mRNA expression data to identify subnetworks for late stage colorectal cancer. Zhang et al. (2011) applied a network-based approach to the diagnosis of. The authors combined expression profiles with topological features to assess the network activities. The resultant network was reported to enhance the diagnostic ability in hepatocellular carcinoma. Chowdhury and Koyuturk (2009) introduced another network approach. They used binarized gene expression profiles to retrieve confident subnetworks for predicting colorectal cancer metastasis. In a recently published paper, Liu et al. (2011) proposed a novel approach to score dysregulated networks as biomarkers. The method was proved to be useful for the prediction of network activities in gastric cancer.

11.7 Network-Based Biomarkers for Prostate Cancer

Recent studies have also demonstrated the utility of network biomarkers in the molecular diagnosis or prognosis of human prostate cancer. For example, Jin et al. (2009) proposed a biomarker discovery pipeline that integrates expression profiles in both genomic and proteomic levels. Using eight microarray expression datasets and one proteomics dataset, they identified 474 genes and proteins associated with prostate cancer. Then they searched for interactions among these molecules to build a prostate-cancer-related network (PCRN). Based on PCRN, a set of candidate network biomarkers were identified that can reliably distinguish the prostate cancer from the normal conditions.

Guo et al. (2007) suggested an edge-based scoring method for identifying condition-responsive PPI subnetworks. In this work, the authors used interactions (edges) instead of proteins (nodes) to capture relevant protein interaction behaviors. An active score was first computed for each edge based on the gene expression profiles. Then, an overall subnetwork score was obtained by all the edges in the subnetwork. Simulated annealing was employed as the search algorithm. This approach, in contrast to node-based methods, constructed a genuine subnetwork with specific active interactions. In addition, this method evaluated the functional importance of the candidate subnetwork in a systematic manner. The authors then applied this edge-based method to gene expression datasets of prostate cancer and identified potential diagnostic markers from the human PPI network.

Ergun et al. (2007) applied the network biology approach to find mediators in prostate cancer metastasis. This work combined expression data with reverse-engineered gene networks to find associated pathways and networks. The authors proposed an algorithm called mode-of-action by network identification (MNI). The MNI algorithm first used microarray data obtained from multiple samples to train a network model of regulatory interactions between genes. Subsequently, the reverse-engineered network was used to filter the condition-related genes from the differentially expressed genes. The algorithm was applied to both non-recurrent primary prostate cancer and metastatic prostate cancer datasets, and identified AR pathway as a significant mediator in metastatic prostate cancer.

In a more recent study, Ummanni et al. (2011) coupled highly sensitive two-dimensional differential gel electrophoresis (2D-DIGE) and MALDI-TOF–MS/MS to investigate the protein expression patterns in prostate cancer. The differentially expressed proteins were mapped into major pathways involved in PCa using MetaCore^{TM} (GeneGO) and ingenuity pathway analysis (IPA) program. A protein network was built for analyzing highly interconnected shortest pathways. A master global network was created according to published annotations on all differentially expressed proteins. Using expression levels as inputs, functional subnetworks were revealed with altered expression in PCa. The major hubs of the significant subnetworks were further validated by real-time PCR analysis.

11.8 Conclusions

The evolving field of network-based biomarkers is destined to revolutionize the clinical practice of prostate cancer. On the other hand, the studies to date are still preliminary. Advanced computational frameworks are required to handle with the ever-growing network-level information. Finally, the network biomarkers need to be rigorous validated before they are translated into clinical application.

References

Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MD, et al. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res. 2001;29:37–40.

Bader GD, Betel D, Hogue CW. BIND: the biomolecular interaction network database. Nucleic Acids Res. 2003;31:248–50.

Barry MJ. Clinical practice: prostate-specific-antigen testing for early diagnosis of prostate cancer. N Engl J Med. 2001;344:1373–7.

Brown KR, Jurisica I. Online predicted human interaction database. Bioinformatics. 2005;21:2076–82.

Chatr-Aryamontri A, Breitkreutz BJ, Heinicke S, Boucher L, Winter A, Stark C, Nixon J, Ramage L, Kolas N, O'Donnell L, et al. The BioGRID interaction database: 2013 update. Nucleic Acids Res. 2012;41:D816–23.

Chen J, Yuan B. Detecting functional modules in the yeast protein–protein interaction network. Bioinformatics. 2006;22:2283–90.

Chen J, Wang Y, Guo D, Shen B. A systems biology perspective on rational design of peptide vaccine against virus infections. Curr Top Med Chem. 2012;12:1310–9.

Chowdhury SA, Koyuturk M. Identification of coordinately dysregulated subnetworks in complex phenotypes. Pac Symp Biocomput. 2009. p. 133–44.

Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. Mol Syst Biol. 2007;3:140.

Darson MF, Pacelli A, Roche P, Rittenhouse HG, Wolfert RL, Young CY, Klee GG, Tindall DJ, Bostwick DG. Human glandular kallikrein 2 (hK2) expression in prostatic intraepithelial neoplasia and adenocarcinoma: a novel prostate cancer marker. Urology. 1997;49:857–62.

DeSouza L, Diehl G, Rodrigues MJ, Guo J, Romaschin AD, Colgan TJ, Siu KW. Search for cancer markers from endometrial tissues using differentially labeled tags iTRAQ and cICAT with multidimensional liquid chromatography and tandem mass spectrometry. J Proteome Res. 2005;4:377–86.

Domon B, Aebersold R. Mass spectrometry and protein analysis. Science. 2006;312:212–7.

Ein-Dor L, Zuk O, Domany E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proc Natl Acad Sci USA. 2006;103:5923–8.

Ergun A, Lawrence CA, Kohanski MA, Brennan TA, Collins JJ. A network biology approach to prostate cancer. Mol Syst Biol. 2007;3:82.

Glinsky GV, Glinskii AB, Stephenson AJ, Hoffman RM, Gerald WL. Gene expression profiling predicts clinical outcome of prostate cancer. J Clin Invest. 2004;113:913–23.

Guo Z, Wang L, Li Y, Gong X, Yao C, Ma W, Wang D, Zhu J, Zhang M, Yang D, et al. Edge-based scoring and searching method for identifying condition-responsive protein–protein interaction sub-network. Bioinformatics. 2007;23:2121–8.

Han K, Park B, Kim H, Hong J, Park J. HPID: the human protein interaction database. Bioinformatics. 2004;20:2466–70.

Hassan AH, Mahmoud S, El-Hamidy A. Quantitative analysis of total proteins and carbohydrates in the digestive gland-gonad complex (DGG) and hemolymph of the freshwater prosobranch snail Lanistes carinatus. J Egypt Soc Parasitol. 2011;40:303–10.

Ideker T, Ozier O, Schwikowski B, Siegel AF. Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics. 2002;18 Suppl 1:S233–40.

Jin G, Zhou X, Cui K, Zhang XS, Chen L, Wong ST. Cross-platform method for identifying candidate network biomarkers for prostate cancer. IET Syst Biol. 2009;3:505–12.

Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath GR, Wu GR, Matthews L, et al. Reactome: a knowledgebase of biological pathways. Nucleic Acids Res. 2005;33:D428–32.

Kang P, Mechref Y, Klouckova I, Novotny MV. Solid-phase permethylation of glycans for mass spectrometric analysis. Rapid Commun Mass Spectrom. 2005;19:3421–8.

Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R, et al. IntAct–open source resource for molecular interaction data. Nucleic Acids Res. 2007;35:D561–5.

Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, et al. Human protein reference database: 2009 update. Nucleic Acids Res. 2009;37:D767–72.

Kim TH, Barrera LO, Zheng M, Qu C, Singer MA, Richmond TA, Wu Y, Green RD, Ren B. A high-resolution map of active promoters in the human genome. Nature. 2005;436:876–80.

Kumar A, White TA, MacKenzie AP, Clegg N, Lee C, Dumpit RF, Coleman I, Ng SB, Salipante SJ, Rieder MJ, et al. Exome sequencing identifies a spectrum of mutation frequencies in advanced and lethal prostate cancers. Proc Natl Acad Sci USA. 2011;108:17087–92.

Leman ES, Cannon GW, Trock BJ, Sokoll LJ, Chan DW, Mangold L, Partin AW, Getzenberg RH. EPCA-2: a highly specific serum marker for prostate cancer. Urology. 2007;69:714–20.

Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E, Sacco F, Palma A, Nardozza AP, Santonico E, et al. MINT, the molecular interaction database: 2012 update. Nucleic Acids Res. 2011;40:D857–61.

Liu X, Liu ZP, Zhao XM, Chen L. Identifying disease genes and module biomarkers by differential interactions. J Am Med Inform Assoc. 2011;19:241–8.

Luo J, Duggan DJ, Chen Y, Sauvageot J, Ewing CM, Bittner ML, Trent JM, Isaacs WB. Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expression profiling. Cancer Res. 2001;61:4683–8.

Luo J, Zha S, Gage WR, Dunn TA, Hicks JL, Bennett CJ, Ewing CM, Platz EA, Ferdinandusse S, Wanders RJ, et al. Alpha-methylacyl-CoA racemase: a new molecular marker for prostate cancer. Cancer Res. 2002;62:2220–6.

Marouga R, David S, Hawkins E. The development of the DIGE system: 2D fluorescence difference gel analysis technology. Anal Bioanal Chem. 2005;382:669–78.

Merchant M, Weinberger SR. Recent advancements in surface-enhanced laser desorption/ionization-time of flight-mass spectrometry. Electrophoresis. 2000;21:1164–77.

Nibbe RK, Markowitz S, Myeroff L, Ewing R, Chance MR. Discovery and scoring of protein interaction subnetworks discriminative of late stage human colon cancer. Mol Cell Proteomics. 2009;8:827–45.

Nibbe RK, Koyuturk M, Chance MR. An integrative-omics approach to identify functional subnetworks in human colorectal cancer. PLoS Comput Biol. 2010;6:e1000639.

Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol Cell Proteomics. 2002;1:376–86.

Pagel P, Kovac S, Oesterheld M, Brauner B, Dunger-Kaltenbach I, Frishman G, Montrone C, Mark P, Stumpflen V, Mewes HW, et al. The MIPS mammalian protein–protein interaction database. Bioinformatics. 2005;21:832–4.

Rhodes DR, Barrette TR, Rubin MA, Ghosh D, Chinnaiyan AM. Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. Cancer Res. 2002;62:4427–33.

Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The database of interacting proteins: 2004 update. Nucleic Acids Res. 2004;32:D449–51.

Siegel R, Naishadham D, Jemal A. Cancer statistics. CA Cancer J Clin. 2012;62:10–29.

Siegel R, Naishadham D, Jemal A. Cancer statistics. CA Cancer J Clin. 2013;63:11–30.

Su J, Yoon BJ, Dougherty ER. Identification of diagnostic subnetwork markers for cancer in human protein–protein interaction network. BMC Bioinform. 2010;11 Suppl 6:S8.

Taylor IW, Linding R, Warde-Farley D, Liu Y, Pesquita C, Faria D, Bull S, Pawson T, Morris Q, Wrana JL. Dynamic modularity in protein interaction networks predicts breast cancer outcome. Nat Biotechnol. 2009;27:199–204.

Taylor BS, Schultz N, Hieronymus H, Gopalan A, Xiao Y, Carver BS, Arora VK, Kaushik P, Cerami E, Reva B, et al. Integrative genomic profiling of human prostate cancer. Cancer Cell. 2010;18:11–22.

Tinzl M, Marberger M, Horvath S, Chypre C. DD3PCA3 RNA analysis in urine–a new perspective for detecting prostate cancer. Eur Urol. 2004;46:182–6 (discussion 187).

Ummanni R, Mundt F, Pospisil H, Venz S, Scharf C, Barett C, Falth M, Kollermann J, Walther R, Schlomm T, et al. Identification of clinically relevant protein targets in prostate cancer with 2D-DIGE coupled mass spectrometry and systems biology network platform. PLoS ONE. 2011;6:e16833.

Varambally S, Dhanasekaran SM, Zhou M, Barrette TR, Kumar-Sinha C, Sanda MG, Ghosh D, Pienta KJ, Sewalt RG, Otte AP, et al. The polycomb group protein EZH2 is involved in progression of prostate cancer. Nature. 2002;419:624–9.

Venderbos LD, Roobol MJ. PSA-based prostate cancer screening: the role of active surveillance and informed and shared decision making. Asian J Androl. 2011;13:219–24.

von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B. STRING: a database of predicted functional associations between proteins. Nucleic Acids Res. 2003;31:258–61.

Wang Y, Chen J, Li Q, Wang H, Liu G, Jing Q, Shen B. Identifying novel prostate cancer associated pathways based on integrative microarray data analysis. Comput Biol Chem. 2011;35:151–8.

Zhang Y, Wang S, Li D, Zhnag J, Gu D, Zhu Y, He F. A systems biology-based classifier for hepatocellular carcinoma diagnosis. PLoS ONE. 2011;6:e22426.