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,

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,

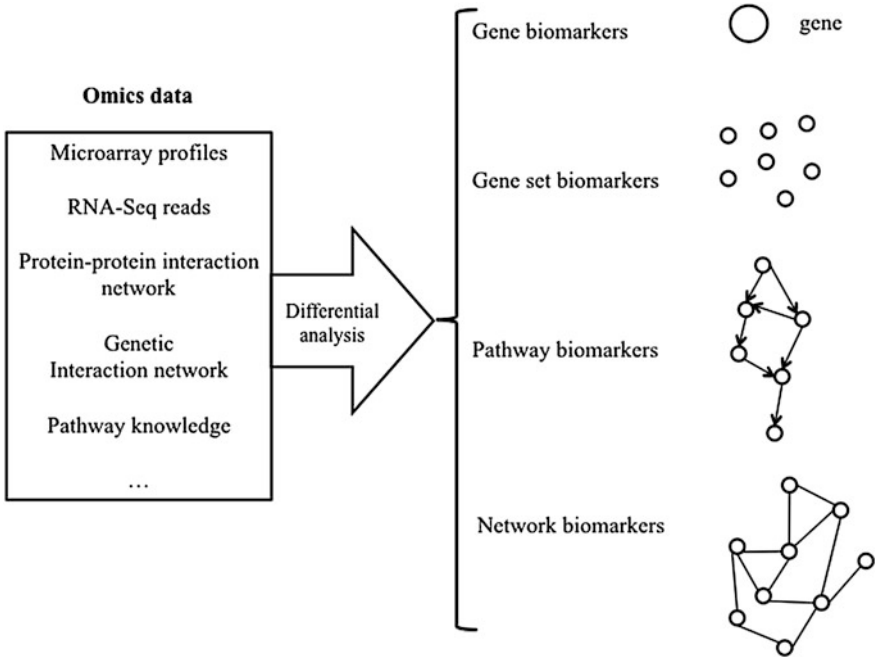

Fig. 2.1 Biomarkers identified based on different data

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 multi-dimensional 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.

http://www.springer.com/978-94-007-7974-7

Bioinformatics for Diagnosis, Prognosis and Treatment of Complex Diseases

Shen, B. (Ed.)

2013, VI, 220 p. 34 illus., 27 illus. in color., Hardcover
ISBN: 978-94-007-7974-7