Causal inference methods and applications in Bioinformatics: A summary

Thuc Duy Le¹,*

¹School of Information Technology and Mathematical Sciences, University of South Australia, Mawson Lakes, SA 5095, AU

November 15, 2017

Abstract

Causality discovery is the ultimate goal in many research areas, including Bioinformatics. My research focuses on the development of causal inference methods and their applications in Bioinformatics, particularly in gene regulatory networks and cancer subtype discovery. This report summaries my research in the last four years, drawing a road map of what have been done and discussing some possible future works.

1 Causality discovery

Discovering causal relationships is the ultimate goal of many scientific explorations. However, it is not feasible to conduct randomized controlled trials in most cases. Discovering causal relationships in large databases of observational data is therefore very important, but it is also a challenging problem. The pioneering work in this area was rooted in the theory of Bayesian network (BN) learning, which however, is a NP-complete problem. Hence several constraint-based algorithms have been developed to efficiently discover causations in large databases. These methods usually use the idea of BN learning, directly or indirectly, and are focused on causal relationships with single cause variables and still not practical for large datasets. Our aims are to design efficient causal discovery algorithms with the ability to find the causal relationships with multiple causes and/or being interpretable.

1.1 Partial association test

In [1], we propose an approach to mine causal rules in large databases of binary variables. Our method expands the scope of causality discovery to causal
relationships with multiple cause variables, and we utilise partial association
tests to exclude noncausal associations, to ensure the high reliability of discovered causal rules. Furthermore an efficient algorithm is designed for the tests in large databases. We assess the method with a set of real-world diagnostic data. The results show that our method can effectively discover interesting causal rules in large databases.

1.2 Causal association rules

Although significant progress has been made in the field of discovering causal relationships using the Causal Bayesian Network (CBN) theory, the applications of CBNs are greatly limited due to the high computational complexity. In another direction, association rule mining has been shown to be an efficient data mining means for relationship discovery. However, although causal relationships imply associations, the reverse does not always hold. In [11], we study how to use an efficient association mining approach to discover potential causal rules in observational data. We make use of the idea of retrospective cohort studies, a widely used approach in medical and social research, to detect causal association rules. In comparison with the constraint-based methods within the CBN paradigm, the proposed approach is faster and is capable of finding a cause consisting of combined variables. The extension version of [11] with more results is in [12].

1.3 Practical causal exploration methods

In an attempt to present the causal discovery methods that are practical for real world applications, we published a book with Springer [13]. This brief presents four practical methods to effectively explore causal relationships, which are often used for explanation, prediction and decision making in medicine, epidemiology, biology, economics, physics and social sciences. The first two methods apply conditional independence tests for causal discovery. The last two methods employ association rule mining for efficient causal hypothesis generation, and a partial association test and retrospective cohort study for validating the hypotheses. The software and the guidelines of using the four methods are available for download at: http://nugget.unisa.edu.au/Causalbook/

1.4 Combined causes

In recent years, many methods have been developed for detecting causal relationships in observational data. Some of them have the potential to tackle large data sets. However, these methods fail to discover a combined cause, i.e. a multi-factor cause consisting of two or more component variables which individually are not causes. A straightforward approach to uncovering a combined cause is to include both individual and combined variables in the causal discovery using existing methods, but this scheme is computationally infeasible due to the huge number of combined variables. In [16], we propose a novel approach to
address this practical causal discovery problem, i.e. mining combined causes in large data sets. The experiments with both synthetic and real world data sets show that the proposed method can obtain high-quality causal discoveries with a high computational efficiency

1.5 Parallel-PC

Discovering causal relationships from observational data is a crucial problem and has applications in many research areas. PC algorithm is the state-of-the-art method in the constraint based causal discovery approach. However, the PC algorithm is worst-case exponential to the number of nodes (variables), and thus it is inefficient when applying to high dimensional data, e.g. gene expression datasets where the causal relationships between thousands of nodes (genes) are explored. On another note, the advancement of computer hardware in the last decade has resulted in the widespread availability of multi-core personal computers. There is a significant motivation for designing a parallelised PC algorithm that is suitable for personal computers and does not require end users’ parallel computing knowledge beyond their competency in using the PC algorithm. In [5], we propose a fast and memory efficient PC algorithm using the parallel computing technique. We apply our method on a range of synthetic and real-world high dimensional datasets. Experimental results on a dataset from DREAM 5 challenge show that the PC algorithm could not produce any results after running more than 24 hours; meanwhile, our parallel-PC algorithm with 4-core CPU computer managed to finish within around 12.5 hours, and less than 6 hours with a 8-coreCPU computer.

In [4], we present an R package, ParallelPC, that includes the parallelised versions of 6 causal exploration algorithms, PC, FCI, RFCI, PC-simple, IDA and Joint-IDA. The parallelised algorithms help speed up the procedure of experimenting big datasets and reduce the memory used when running the algorithms. The package is not only suitable for super-computers or clusters, but also convenient for researchers using personal computers with multi core CPUs. Our experiment results on real world datasets show that using the parallelised algorithms it is now practical to explore causal relationships in high dimensional datasets with thousands of variables in a single multicore computer. ParallelPC is available in CRAN repository at https://cran.rproject.org/web/packages/ParallelPC/index.html.

1.6 Causal decision tree

Uncovering causal relationships in data is a major objective of data analytics. Causal relationships are normally discovered with designed experiments, eg randomised controlled trials, which, however are expensive or infeasible to be conducted in many cases. Causal relationships can also be found using some well designed observational studies, but they require domain experts’ knowledge and the process is normally time consuming. Hence there is a need for scalable
and automated methods for causal relationship exploration in data. Classification methods are fast and they could be practical substitutes for finding causal signals in data. However, classification methods are not designed for causal discovery and a classification method may find false causal signals and miss the true ones. In [14], we develop a causal decision tree where nodes have causal interpretations. Our method follows a well established causal inference framework and makes use of a classic statistical test. The method is practical for finding causal signals in large data sets.

2 microRNAs, transcription factors, messenger RNAs regulatory relationships

MicroRNAs (miRNAs) are important gene regulators. They control a wide range of biological processes and are involved in several types of cancers. Thus, exploring miRNA functions is important for diagnostics and therapeutics. To date, there are few feasible experimental techniques for discovering miRNA regulatory mechanisms. Alternatively, predictions of miRNA-mRNA regulatory relationships by computational methods have increasingly achieved promising results. Computational approaches are proving their ability as effective tools in reducing the number of biological experiments that must be conducted and to assist with the design of the experiments.

In [7], we categorize and review different computational approaches to identify miRNA activities and functions, including the co-regulation of miRNAs and transcription factors. Our main focuses are on the recent approaches that use multiple data types for exploring miRNA functions. We discuss the remaining challenges in the evaluation and selection of models based on the results from a case study. Finally, we analyse the remaining challenges of each computational approach and suggest some future research directions.

In the following sub-sections we summarise and present different approaches of exploring miRNA functions.

2.1 Inferring miRNA-mRNA relationships using causal inference methods

miRNAs are known to play an essential role in the post-transcriptional gene regulation in plants and animals. Currently, several computational approaches have been developed with a shared aim to elucidate miRNA-mRNA regulatory relationships. Although these existing computational methods discover the statistical relationships, such as correlations and associations between miRNAs and mRNAs at data level, such statistical relationships are not necessarily the real causal regulatory relationships that would ultimately provide useful insights into the causes of gene regulations. The standard method for determining causal relationships is randomized controlled perturbation experiments.
In practice, however, such experiments are expensive and time consuming. Our motivation in [6] is to discover the miRNA-mRNA causal regulatory relationships from observational data. The experimental results indicate that the causal discovery method effectively discovers miRNA regulatory relationships in data. Although computational predictions may not completely replace intervention experiments, the accurate and reliable discoveries in data are cost effective for the design of miRNA experiments and the understanding of miRNA-mRNA regulatory relationships.

Using the same method but in another direction, we propose a novel method [24] to infer condition-specific miRNA activity by considering (i) the difference between the regulatory behavior that an miRNA has in the condition of interest and its behavior in the other conditions; (ii) the causal semantics of miRNA-mRNA relationships. The method is applied to the epithelial-mesenchymal transition (EMT) and multi-class cancer (MCC) datasets. The validation by the results of transfection experiments shows that our approach is effective in discovering significant miRNA-mRNA interactions. Functional and pathway analysis and literature validation indicate that the identified active miRNAs are closely associated with the specific biological processes, diseases and pathways. More detailed analysis of the activity of the active miRNAs implies that some active miRNAs show different regulation types in different conditions, but some have the same regulation types and their activity only differs in different conditions in the strengths of regulation.

We also investigate the way to utilise domain knowledge into causal inference methods of identifying miRNA targets [26]. We present an integrative framework, CIDER (Causal miRNA target Discovery with Expression profile and Regulatory knowledge), to predict miRNA targets. CIDER is able to utilise a variety of gene regulation knowledge, including transcriptional and post-transcriptional knowledge, and to exploit gene expression data for the discovery of miRNA-mRNA regulatory relationships. The benefits of our framework is demonstrated by both simulation study and the analysis of the epithelial-to-mesenchymal transition (EMT) and the breast cancer (BRCA) datasets. Our results reveal that even a limited amount of either Transcription Factor-miRNA or miRNA-mRNA regulatory knowledge improves the performance of miRNA target prediction, and the combination of the two types of knowledge enhances the improvement further. Another useful property of the framework is that its performance increases monotonically with the increase of regulatory knowledge.

### 2.2 Inferring miRNA-TF-mRNA regulatory relationships

Transcription factors (TFs) and miRNAs are primary metazoan gene regulators. Regulatory mechanisms of the two main regulators are of great interest to biologists and may provide insights into the causes of diseases. However, the interplay between miRNAs and TFs in a regulatory network still remains unearthed. Currently, it is very difficult to study the regulatory mechanisms that
involve both miRNAs and TFs in a biological lab. Even at data level, a network involving miRNAs, TFs and genes will be too complicated to achieve. Previous research has been mostly directed at inferring either miRNA or TF regulatory networks from data. However, networks involving a single type of regulator may not fully reveal the complex gene regulatory mechanisms, for instance, the way in which a TF indirectly regulates a gene via a miRNA.

In [3], we propose a framework to learn from heterogeneous data the three-component regulatory networks, with the presence of miRNAs, TFs, and mRNAs. This method firstly utilises Bayesian network structure learning to construct a regulatory network from multiple sources of data: gene expression profiles of miRNAs, TFs and mRNAs, target information based on sequence data, and sample categories. Then, in order to produce more meaningful results for further biological experimentation and research, the method searches the learnt network to identify the interplay between miRNAs and TFs and applies a network motif finding algorithm to further infer the network.

We apply the proposed framework to the data sets of epithelial-to-mesenchymal transition (EMT). The results elucidate the complex gene regulatory mechanism for EMT which involves both TFs and miRNAs. Several discovered interactions and molecular functions have been confirmed by literature. In addition, many other discovered interactions and bio-markers are of high statistical significance and thus can be good candidates for validation by experiments. Moreover, the results generated by our method are compact, involving a small number of interactions which have been proved highly relevant to EMT. This framework has the potential for application to other heterogeneous datasets to reveal the complex gene regulatory relationships.

In another direction, in [23], we present a causal discovery based framework (called DirectTarget) to infer direct miRNA-TF-mRNA causal regulatory relationships in heterogeneous data, including expression profiles of miRNAs and mRNAs, and miRNA target information. DirectTarget is applied to the Epithelial to Mesenchymal Transition (EMT) datasets. The validation by experimentally confirmed target databases suggests that the proposed method can effectively identify direct miRNA-mRNA regulatory relationships. To explore the upstream regulators of miRNA regulation, we further identify the causal feedforward patterns (CFFPs) of TF-miRNA-mRNA to provide insights into the miRNA regulation in EMT. DirectTarget has the potential to be applied to other datasets to elucidate the direct miRNA-mRNA causal regulatory relationships and to explore the regulatory patterns.

2.3 Ensemble and other computational methods and software tools for identifying miRNA-mRNA relationships

miRNAs are short regulatory RNAs that are involved in several diseases, including cancers. Identifying miRNA functions is very important in understanding dis-
ease mechanisms and determining the efficacy of drugs. An increasing number of computational methods have been developed to explore miRNA functions by inferring the miRNA-mRNA regulatory relationships from data. Each of the methods is developed based on some assumptions and constraints, for instance, assuming linear relationships between variables. For such reasons, computational methods are often subject to the problem of inconsistent performance across different data sets. On the other hand, ensemble methods integrate the results from individual methods and have been proved to outperform each of their individual component methods in theory.

In [8], we investigate the performance of some ensemble methods over the commonly used miRNA target prediction methods. We apply eight different popular miRNA target prediction methods to three cancer datasets, and compare their performance with the ensemble methods which integrate the results from each combination of the individual methods. The validation results using experimentally confirmed databases show that the results of the ensemble methods complement those obtained by the individual methods and the ensemble methods perform better than the individual methods across different datasets. The ensemble method, Pearson+IDA+Lasso, which combines methods in different approaches, including a correlation method, a causal inference method, and a regression method, is the best performed ensemble method in this study. Further analysis of the results of this ensemble method shows that the ensemble method can obtain more targets which could not be found by any of the single methods, and the discovered targets are more statistically significant and functionally enriched.

Several computational methods of predicting miRNA targets have been proposed using expression data with or without sequence based miRNA target information. A typical procedure for applying and evaluating such a method is i) collecting matched miRNA and mRNA expression profiles in a specific condition, e.g. a cancer dataset from The Cancer Genome Atlas (TCGA), ii) applying the new computational method to the selected dataset, iii) validating the predictions against knowledge from literature and third-party databases, and comparing the performance of the method with some existing methods. This procedure is time consuming given the time elapsed when collecting and processing data, repeating the work from existing methods, searching for knowledge from literature and third-party databases to validate the results, and comparing the results from different methods. The time consuming procedure prevents researchers from quickly testing new computational models, analysing new datasets, and selecting suitable methods for assisting with the experiment design. In [10], we present an R package, miRLAB, for automating the procedure of inferring and validating miRNA-mRNA regulatory relationships. The package provides a complete set of pipelines for testing new methods and analysing new datasets. miRLAB includes a pipeline to obtain matched miRNA and mRNA expression datasets directly from TCGA, 12 benchmark computational methods for inferring miRNA-mRNA regulatory relationships, the functions for validating
the predictions using experimentally validated miRNA target data and miRNA perturbation data, and the tools for comparing the results from different computational methods

2.4 Identifying miRNA modules

Evidence suggests that miRNAs and mRNAs interact collectively in gene regulatory networks. The collective relationships between groups of miRNAs and groups of mRNAs may be more readily interpreted than those between individual miRNAs and mRNAs, and thus are useful for gaining insight into gene regulation and cell functions. Several computational approaches have been developed to discover miRNA-mRNA regulatory modules (MMRM$s$) with a common aim to elucidate miRNA-mRNA regulatory relationships. However, most existing methods do not consider the collective relationships between a group of miRNAs and the group of targeted mRNAs in the process of discovering MMRMs. Our aim is to develop a framework to discover MMRMs and reveal miRNA-mRNA regulatory relationships from the heterogeneous expression data based on the collective relationships.

We propose Discovering Collective group Relationships (DICORE) [2], an effective computational framework for revealing miRNA-mRNA regulatory relationships. We utilize the notation of collective group relationships to build the computational framework. The method computes the collaboration scores of the miRNAs and mRNAs on the basis of their interactions with mRNAs and miRNAs, respectively. Then it determines the groups of miRNAs and groups of mRNAs separately based on their respective collaboration scores. Next, it calculates the strength of the collective relationship between each pair of miRNA group and mRNA group using canonical correlation analysis, and the group pairs with significant canonical correlations are considered as the MMRMs. We applied this method to three gene expression datasets, and validated the computational discoveries. Analysis of the results demonstrates that a large portion of the regulatory relationships discovered by DICORE is consistent with the experimentally confirmed databases. Furthermore, it is observed that the top mRNAs that are regulated by the miRNAs in the identified MMRMs are highly relevant to the biological conditions of the given datasets. It is also shown that the MMRMs identified by DICORE are more biologically significant and functionally enriched.

2.5 Synergy miRNAs and relationships between miRNAs and PPI

Understanding the synergism of multiple miRNAs in gene regulation can provide important insights into the mechanisms of complex human diseases caused by miRNA regulation. Therefore, it is important to identify miRNA synergism and study miRNA characteristics in miRNA synergistic regulatory networks. A
number of methods have been proposed to identify miRNA synergism. However, most of the methods only use downstream target genes of miRNAs to infer miRNA synergism when miRNAs can also be regulated by upstream TFs at the transcriptional level. Additionally, most methods are based on statistical associations identified from data without considering the causal nature of gene regulation. In [20], we present a causality based framework, called mirSRN (miRNA synergistic regulatory network), to infer miRNA synergism in human molecular systems by considering both downstream miRNA targets and upstream TF regulation. We apply the proposed framework to two real world datasets and discover that almost all the top 10 miRNAs with the largest node degree in the mirSRNs are associated with different human diseases, including cancer, and that the mirSRNs are approximately scale-free and small-world networks. We also find that most miRNAs in the networks are frequently synergistic with other miRNAs, and miRNAs related to the same disease are likely to be synergistic and in a cluster linked to a biological function. Synergistic miRNA pairs show higher co-expression level, and may have potential functional relationships indicating collaboration between the miRNAs. Functional validation of the identified synergistic miRNAs demonstrates that these miRNAs cause different kinds of diseases. These results deepen our understanding of the biological meaning of miRNA synergism.

Recent studies have shown that TFs and miRNAs, while independently regulate their downstream targets, collaborate with each other to regulate gene expression. However, their synergistic roles in protein-protein interactions (PPIs) remain mostly unknown. In [21], we present a novel framework (called CoRePPI) for inferring TF and miRNA co-regulation of PPIs. Particularly, CoRePPI is aimed at discovering the co-regulation specific to a condition of interest, by using heterogeneous data, including miRNA and messenger RNA (mRNA) expression profiles, putative miRNA targets, TF targets and PPIs. CoRePPI firstly finds the network motifs indicating the co-regulation of PPIs by TFs and miRNAs in tumor and normal conditions separately. Then by identifying the differential motifs found in one condition but not in the other, it builds the networks consisting of TFs, miRNAs and their co-regulated PPIs specific to different conditions respectively. To validate CoRePPI, we apply it to the PanCancer dataset which includes the expression profiles of 12 cancer types from TCGA. Through network topology analysis, we found that the tumor and normal CoRePPI networks are scale-free. Furthermore, the results of differential and intersected network analysis between the tumor and normal CoRePPI networks suggest that only a small fraction of the regulatory relationships between TFs and miRNAs are conserved in both conditions but they co-regulate different downstream PPIs in tumor and normal conditions; and in different conditions the majority of the regulatory relationships between TFs and miRNAs are different although they may regulate the same PPIs in their respective conditions. The CoRePPI sub-networks constructed for the three types of cancers (breast cancer, lung cancer and ovarian cancer) are all scale-free, and the intersection of these CoRePPI sub-networks can be utilized as the biomarker CoRePPI sub-
network of the three types of cancers. The PPI enrichment analyses of the
tumor and normal CoRePPI networks suggest that the co-regulating TFs and
miRNAs are significantly associated with the specific biological processes, dis-
eases and pathways. In addition, comparing with the two non-condition-specific
approaches, the tumor CoRePPI network is found to have the most enriched
cancer-related PPIs. Altogether, the results uncover the combined regulatory
patterns of TFs and miRNAs on the PPIs, and may provide new insights for
research in cancer-associated TFs and miRNAs.

3 ceRNAs and non-coding RNAs: the new play-
ers

Recent findings show that coding genes are not the only targets that miRNAs
interact with. In fact, there is a pool of different RNAs competing with each
other to attract miRNAs for interactions, thus acting as competing endogenous
RNAs (ceRNAs). The ceRNAs indirectly regulate each other via the titra-
tion mechanism, i.e. the increasing concentration of a ceRNA will decrease the
number of miRNAs that are available for interacting with other targets. The
cross-talks between ceRNAs, i.e. their interactions mediated by miRNAs, have
been identified as the drivers in many disease conditions, including cancers. In
recent years, some computational methods have emerged for identifying ceRNA-
ceRNA interactions. However, there remain great challenges and opportunities
for developing computational methods to provide new insights into ceRNA reg-
ulatory mechanisms.

In [9], we review the publically available databases of ceRNA-ceRNA interac-
tions and the computational methods for identifying ceRNA-ceRNA interactions
(also known as miRNA sponge interactions). We also conduct a comparison
study of the methods with a breast cancer dataset. Our aim is to provide a
current snapshot of the advances of the computational methods in identifying
miRNA sponge interactions and to discuss the remaining challenges.

In [19], we propose a novel in silico method, called miRSM (miRNA Sponge
Module) to infer miRNA sponge modules in breast cancer. We apply miRSM to
the breast invasive carcinoma (BRCA) dataset provided by The Cancer Genome
Atlas (TCGA), and make functional validation of the computational results.
We discover that most miRNA sponge interactions are module-conserved across
two modules, and a minority of miRNA sponge interactions are module-specific,
eexisting only in a single module. Through functional annotation and differen-
tial expression analysis, we also find that the modules discovered using miRSM
are functional miRNA sponge modules associated with BRCA. Moreover, the
module-specific miRNA sponge interactions among miRNA sponge modules
may be involved in the progression and development of BRCA. Our experi-
mental results show that miRSM is comparable to the benchmark methods in
recovering experimentally confirmed miRNA sponge interactions, and miRSM outperforms the benchmark methods in identifying interactions that are related to breast cancer.

In [22], we propose a multi-step method called miRSCoPPI to infer miRNA sponge co-regulation of PPIs. We focus on investigating breast cancer (BRCA) related miRNA sponge co-regulation, by integrating heterogeneous data, including miRNA, long non-coding RNA (lncRNA) and messenger RNA (mRNA) expression data, experimentally validated miRNA-target interactions, PPIs and lncRNA-target interactions, and the list of breast cancer genes. We find that the inferred BRCA-related miRSCoPPI network is highly connected and scale free. The top 10% hub genes in the BRCA-related miRSCoPPI network have potential biological implications in breast cancer. By utilizing a graph clustering method, we discover 17 BRCA-related miRSCoPPI modules. Through pathway enrichment analysis of the modules, we find that several modules are significantly enriched in pathways associated with breast cancer. Moreover, 10 modules have good performance in classifying breast tumor and normal samples, and can act as module signatures for prognostication. By using putative computationally predicted miRNA-target interactions, we have consistent results with those obtained using experimentally validated miRNA-target interactions, indicating that miRSCoPPI is robust in inferring miRNA sponge co-regulation of PPIs in human breast cancer.

4 Cancer subtype discovery and prognostic - the personalised medicine framework

Identifying cancer subtypes is an important component of the personalised medicine framework. An increasing number of computational methods have been developed to identify cancer subtypes. However, existing methods rarely use information from gene regulatory networks to facilitate the subtype identification. It is widely accepted that gene regulatory networks play crucial roles in understanding the mechanisms of diseases. Different cancer subtypes are likely caused by different regulatory mechanisms. Therefore, there are great opportunities for developing methods that can utilise network information in identifying cancer subtypes.

In [18], we propose a method, weighted similarity network fusion (WSNF), to utilise the information in the complex miRNA-TF-mRNA regulatory network in identifying cancer subtypes. We firstly build the regulatory network where the nodes represent the features, i.e. the miRNAs, TFs and messenger RNAs (mRNAs) and the edges indicate the interactions between the features. The interactions are retrieved from various interatomic databases. We then use the network information and the expression data of the miRNAs, TFs and mRNAs to calculate the weight of the features, representing the level of im-
portance of the features. The feature weight is then integrated into a network fusion approach to cluster the samples (patients) and thus to identify cancer subtypes. We applied our method to the TCGA breast invasive carcinoma (BRCA) and glioblastoma multiforme (GBM) datasets. The experimental results show that WSNF performs better than the other commonly used computational methods, and the information from miRNA-TF-mRNA regulatory network contributes to the performance improvement. The WSNF method successfully identified five breast cancer subtypes and three GBM subtypes which show significantly different survival patterns. We observed that the expression patterns of the features in some miRNA-TF-mRNA sub-networks vary across different identified subtypes. In addition, pathway enrichment analyses show that the top pathways involving the most differentially expressed genes in each of the identified subtypes are different. The results would provide valuable information for understanding the mechanisms characterising different cancer subtypes and assist the design of treatment therapies. All datasets and the R scripts to reproduce the results are available online at the website: http://nugget.unisa.edu.au/Thuc/cancersubtypes/.

In [17], we introduce CancerSubtypes, an R package for identifying cancer subtypes using multi-omics data, including gene expression, miRNA expression and DNA methylation data. CancerSubtypes integrates four main computational methods which are highly cited for cancer subtype identification and provides a standardized framework for data pre-processing, feature selection, and result follow-up analyses, including results computing, biology validation and visualization. The input and output of each step in the framework are packaged in the same data format, making it convenient to compare different methods. The package is useful for inferring cancer subtypes from an input genomic dataset, comparing the predictions from different well-known methods and testing new subtype discovery methods, as shown with different application scenarios in the Supplementary Material. Availability: The package is implemented in R and available under GPL-2 license from the Bioconductor website(http://bioconductor.org/packages/CancerSubtypes/).

In [25], we propose the Survival Causal Tree (SCT) method. SCT is designed to discover patient subgroups with heterogeneous treatment effects from censored observational data. Results on TCGA breast invasive carcinoma and glioma datasets have shown that for each subtype identified by SCT, the patients treated with radiotherapy exhibit significantly different relapse free survival pattern when compared to patients without the treatment. With the capability to identify cancer subtypes with heterogeneous treatment responses, SCT is useful in helping to choose the most suitable treatment for individual patients.

Relevant to the topic, in [15] we developed a method for cross view dimensionality reduction. Cross-view data are collected from two different views or sources about the same subjects. As the information from these views often consolidate and/or complement each other, cross-view data analysis can gain more
insights for decision making. A main challenge of cross-view data analysis is how to effectively explore the inherently correlated and high-dimensional data. Dimension reduction offers an effective solution for this problem. However, how to choose right models and parameters involved for dimension reduction is still an open problem. In this paper we propose an effective sparse learning algorithm for crossview dimensionality reduction. A distinguished character of our model selection is that it is non-parametric and automatic. Specifically, we represent the correlation of cross-view data using a covariance matrix. Then we decompose the matrix into a sequence of low-rank ones by solving an optimization problem in an alternating least squares manner. More importantly, a new and non-parametric sparsity-inducing function is developed to derive a parsimonious model. Extensive experiments are conducted on real-world data sets to evaluate the effectiveness of the proposed algorithm. The results show that our method is competitive with the state-of-the-art sparse learning algorithms.

References


