Computational Biology and Chemistry 59 (2015) 15–31

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research Article

Network regularised Cox regression and multiplex network models to predict disease comorbidities and survival of cancer

Haoming Xu[a,*,1], Mohammad Ali Moni[a,b,1], Pietro Liò[a]

[a] Computer Laboratory, University of Cambridge, Cambridge CB3 0FD, UK
[b] Bone Biology, Garvan Institute of Medical Research, Australia

ARTICLE INFO

Article history:
Received 13 June 2015
Received in revised form 21 August 2015
Accepted 25 August 2015
Available online 19 October 2015

Keywords:
Comorbidity
Network regularised Cox regression
Multiplex network
Survival prediction
Cancer

ABSTRACT

In cancer genomics, gene expression levels provide important molecular signatures for all types of cancer, and this could be very useful for predicting the survival of cancer patients. However, the main challenge of gene expression data analysis is high dimensionality, and microarray is characterised by few number of samples with large number of genes. To overcome this problem, a variety of penalised Cox proportional hazard models have been proposed. We introduce a novel network regularised Cox proportional hazard model and a novel multiplex network model to measure the disease comorbidities and to predict survival of the cancer patient. Our methods are applied to analyse seven microarray cancer gene expression datasets: breast cancer, ovarian cancer, lung cancer, liver cancer, renal cancer and osteosarcoma. Firstly, we applied a principal component analysis to reduce the dimensionality of original gene expression data. Secondly, we applied a network regularised Cox regression model on the reduced gene expression datasets. By using normalised mutual information method and multiplex network model, we predict the comorbidities for the liver cancer based on the integration of diverse set of omics and clinical data, and we find the diseasome associations (disease–gene association) among different cancers based on the identified common significant genes. Finally, we evaluated the precision of the approach with respect to the accuracy of survival prediction using ROC curves. We report that colon cancer, liver cancer and renal cancer share the CXCL5 gene, and breast cancer, ovarian cancer and renal cancer share the CCND2 gene. Our methods are useful to predict survival of the patient and disease comorbidities more accurately and helpful for improvement of the care of patients with comorbidity. Software in Matlab and R is available on our GitHub page: https://github.com/ssnhcom/NetworkRegularisedCox.git.

1. Introduction

Comorbidity refers to the presence of one or more coexisting diseases that appear simultaneously with another disease, or interdependently with each other (Capobianco and Liò, 2013). Sometimes, a comorbidity is considered to be a secondary diagnosis, having been detected at the same time, or one after another, treatment for the principal diagnosis. It is associated with an elevated burden of symptoms, more complex clinical management, decreased length and quality of life and increased health-care costs (Valderas et al., 2009). Comorbidity has a significant predictive value on overall survival (Lagro et al., 2010). For an instance, comorbidity has a main effect on survival in cancer, particularly for cancer of the cervix (Ferrandina et al., 2012). However, there is also inverse comorbidity, which is characterised by a lower-than-expected probability of certain diseases occurring in individuals diagnosed with other health conditions, related to cancer (Tabarés-Seisdedos and Rubenstein, 2013). Most of the diseases associated with inverse cancer comorbidity are related to the Central Nervous System (CNS) or neuropsychiatric disorders (Tabarés-Seisdedos et al., 2011). There are some epidemiological evidences that patients with CNS disorders have a lower than expected probability of developing some types of cancer, such as Alzheimer's disease, Parkinsons disease and Schizophrenia (Ibáñez et al., 2014).

Data concerning the expression of various genes have been widely used for the prediction of the risk of disease. For appropriate disease-specific diagnostic, prognostic, and therapeutic approaches, disease–gene association studies provide valuable information (Tiffin et al., 2009). Understanding relationships between diseases and genes at the molecular level could help us to gain a better understanding of pathogenesis, and it leads to better prevent, treatment, and diagnosis (Du et al., 2009). Diseases are

* Corresponding author.
E-mail addresses: hx228@cam.ac.uk (H. Xu), m.moni@garvan.org.au (M.A. Moni), pl219@cam.ac.uk (P. Liò).
[1] Joint first author.

http://dx.doi.org/10.1016/j.compbiolchem.2015.08.010
1476-9271/

more likely to be comorbid if they share associated gene expression profiles (Park et al., 2009). These associations can be due to direct or indirect causal relationships and the shared risk factors among diseases (Moni and Liò, 2014; Liò et al., 2012). For instance, people with HIV-1 appear to have a markedly higher rate of end-stage renal disease (ESRD) than the healthy people (Kumar et al., 2005). It is because some of the risk factors associated with HIV-1 acquisition are the same as those that lead to kidney disease. Patients with chronic kidney disease increase risk of cardiovascular mortality (de Jager et al., 2014). Thus HIV-1 infections are associated with cardiovascular mortality.

Cancer is a group of complex diseases that is caused by abnormalities of biomarker genes. In cancer genomics, gene expression levels provide important molecular signatures for all types of cancer and that can be very useful in predicting the comorbidity and survival of cancer patients. Evidence of an increased comorbidity between CNS (Central Nervous System) disorders and certain cancers has existed for many years. For instance, Down's syndrome (DS) is strongly associated to increase the co-occurrence of cancer, specifically acute leukaemia, testicular cancer and some gastrointestinal cancers (Catalá-López et al., 2014). Mining of high-throughput gene expression data in order to identify biomarker associated with patient survival is an ongoing challenge in complex disease prognostic studies to achieve more accurate prognosis. However, one of the main challenges of microarray gene expression data analysis is the high dimensionality, due to the overwhelming number of measures of gene expression levels compared to the small number of cancer samples (Xu et al., 2010). To tackle this problem, variable selection has been applied to select significant subsets of genes in a microarray gene expression dataset. All the approaches have shown some limitations.

Many methods were proposed for survival analysis on high dimensional gene expression data with highly correlated variables (Van Wieringen et al., 2009; Witten and Tibshirani, 2010). To explore the relations between gene expression data and cancer survival with both censored samples and uncensored samples, Cox's proportional hazards model (Cox et al., 1972) is commonly used. It is the most popular survival model used to describe the relationship between the patient's survival time and predictor variables (Cox, 1992). However, when we have high-dimensional data (e.g. in a microarray study) where the number of predictors (genes) far exceeds the number of subjects (patients), the Cox model cannot be fitted directly unless the high-dimensionality is properly handled. Due to the high-dimensionality of microarray gene expression, a variety of regularisations with different penalties have been proposed including $L1$ penalty in lasso (least absolute shrinkage and selection operator) regression (Tibshirani et al., 1997; Gui and Li, 2005), adaptive lasso (Zhang and Lu, 2007; Zou, 2008) for gene selection and parameter estimation in high-dimensional microarray data. All of these methods can select important variables by shrinking some regression coefficients to equal exactly zero. These penalties can be imposed to individual variables to automatically remove unimportant ones. The lasso shrinks some of the coefficients to zero, and the amount of shrinkage is determined by the tuning parameter, often determined by cross validation. The model determined by this cross validation contains many false positives whose coefficients are actually zero. Hastie and Tibshirani (2004) and Hoshida et al. (2008) introduced an alternative method using $L2$ (Euclidean norm) penalty in ridge regression. Although the $L1$ and the $L2$ penalties have been designed as a statistical technique to solve the high-dimensional data problem it has some drawbacks. The primary one being that these procedures ignore important prior gene structure information regarding modular relations among gene expressions. A better approach is to identify the significant genes that are functionally related. Since it takes into account biological information it leads greater reliability. Groups of genes are co-expressed in different conditions through biological pathway or protein–protein interaction, and it provides prior information to reduce high dimensionally data based on removing confounding factors and statistical randomness for regression models (Chuang et al., 2007; Li and Li, 2008; Tian et al., 2009). Therefore, incorporating prior biological knowledge by exploiting the network structure in a statistical method is expected to improve its performance. The major advantage of the network-based models is the better generalisation across independent studies because the network information is consistent with the conserved patterns in the gene expression data.

In recent years, researchers have focussed on a particular data type, for example mRNA expression, to find profiles that are associated with particular diseases, prognosis, disease comorbidities and drug response. More recently, as the cost of collecting data using highthroughput technologies has decreased, studies have begun to integrate multiple data types collected from the same patient samples (Chalise et al., 2014). In addition, by analysing different types of data in isolation we may miss important information that results from the coordinated activity of biological components at various levels. Several methods have been proposed in the last few years that have aimed to address the issue of integrating multiple data types into a single analysis (Chalise et al., 2014). As data collection at the genomic, transcriptomic, epigenomic and proteomics levels is becoming easier it will become increasingly important to integrate these data in order to predict disease comorbidities. High-throughput omics-data such as messenger RNA (mRNA) expression, DNA copy number alterations, pathway dysregulation, DNA methylation and clinical information can provide a different view of the patients molecular status at various levels. Combining clinical and molecular data types may potentially improve prediction accuracy of disease comorbidity. However, currently there is a shortage of effective and efficient statistical and bioinformatics methods for true integrative data analysis. So far few methods have been proposed to integrate clinical and molecular data to obtain accurate cancer prognosis. Here, we have presented methods of integrating different types of data by modelling association between diseases in a multiplex network (a multilink between nodes and indicates the set of all links connecting these nodes in the different layers (Bianconi, 2013)). The multiplex network allows us to model disease comorbidities by representing each data type as layers in the multiplex network (Estrada and Gómez-Gardeñes, 2014). Importantly, this allows us to capture the interactions between the various types of data, such as the interdependence of pathway regulation with mRNA expression or mRNA expression with miRNA expression. Moreover comorbidity prediction of complex diseases is critical in the field of medicine. As data collection at the genomic, transcriptomic, epigenomic and proteomics levels is becoming easier it will become increasingly important to integrate these molecular data with the clinical information in order to predict disease comorbidities.

2. Materials and methods

In this article, we present the network-regularised Cox proportional hazard and multiplex network models to measure disease comorbidities based on the diverse set of data and to predict survival in cancer. At first each biological data were pre-processed. In the second stage, we apply PCA method to reduce the dimension of initial sample data. After PCA transform, a gene interaction network was built according to gene co-expression information, and the network regularised Cox regression model is used for selecting the significant genes for each cancer microarray gene expression dataset. We have used these identified genes expression data as input to the survival prediction and multiplex network models.