Journal of Biomedical Informatics xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Predicting censored survival data based on the interactions between meta-dimensional omics data in breast cancer

Dokyoon Kim, Ruowang Li, Scott M. Dudek, Marylyn D. Ritchie *

Center for Systems Genomics, Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, PA, USA

ARTICLE INFO

Article history:
Received 6 February 2015
Revised 15 May 2015
Accepted 27 May 2015
Available online xxxx

Keywords:
Survival prediction
Data integration
Interaction between multi-omics data
TCGA
Breast cancer

ABSTRACT

Evaluation of survival models to predict cancer patient prognosis is one of the most important areas of emphasis in cancer research. A binary classification approach has difficulty directly predicting survival due to the characteristics of censored observations and the fact that the predictive power depends on the threshold used to set two classes. In contrast, the traditional Cox regression approach has some drawbacks in the sense that it does not allow for the identification of interactions between genomic features, which could have key roles associated with cancer prognosis. In addition, data integration is regarded as one of the important issues in improving the predictive power of survival models since cancer could be caused by multiple alterations through meta-dimensional genomic data including genome, epigenome, transcriptome, and proteome. Here we have proposed a new integrative framework designed to perform these three functions simultaneously: (1) predicting censored survival data; (2) integrating meta-dimensional omics data; (3) identifying interactions within/between meta-dimensional genomic features associated with survival. In order to predict censored survival time, martingale residuals were calculated as a new continuous outcome and a new fitness function used by the grammatical evolution neural network (GENN) based on mean absolute difference of martingale residuals was implemented. To test the utility of the proposed framework, a simulation study was conducted, followed by an analysis of meta-dimensional omics data including copy number, gene expression, DNA methylation, and protein expression data in breast cancer retrieved from The Cancer Genome Atlas (TCGA). On the basis of the results from breast cancer dataset, we were able to identify interactions not only within a single dimension of genomic data but also between meta-dimensional omics data that are associated with survival. Notably, the predictive power of our best meta-dimensional model was 73% which outperformed all of the other models conducted based on a single dimension of genomic data. Breast cancer is an extremely heterogeneous disease and the high levels of genomic diversity within/between breast tumors could affect the risk of therapeutic responses and disease progression. Thus, identifying interactions within/between meta-dimensional omics data associated with survival in breast cancer is expected to deliver direction for improved meta-dimensional prognostic biomarkers and therapeutic targets.

1. Introduction

Translational bioinformatics is one of the most prominent fields that efficiently translate genomic and biomedical data into clinical knowledge for application [3,4,41]. In particular, translational bioinformatics has been playing important roles in cancer research due to the tumor heterogeneity [4]. For example, recent standard-of-care for breast cancer or non-small cell lung cancer includes quantitating panels of gene expression such as Oncotype DX, developed by Genomic Health, or sequencing of genes such as EGFR, respectively, in order to provide therapeutic knowledge for new subtypes of cancer patients [4]. One of the most exciting problems in translational bioinformatics is to predict clinical outcomes using molecular datasets such as somatic mutation, copy number or gene expression data for better diagnostics, prognostics, and further therapeutics [3]. Among problems of predicting clinical outcomes, there is an increasing difficulty in predicting prognosis and therapeutic response prediction [31].

Evaluating survival models is one of the most important attentions in the development of cancer prognostic models, especially based on genomic profiles. One of the common approaches is that patients can be divided into two groups, such as high-risk survival and low-risk survival group, according to a survival-time

* Corresponding author. Tel.: +1 814 863 4467; fax: +1 814 863 6699.
E-mail address: marylyn.ritchie@psu.edu (M.D. Ritchie).

http://dx.doi.org/10.1016/j.jbi.2015.05.019
1532-0464/

threshold, and then a binary classification algorithm can be applied to predict the survival group for each individual patient in a test dataset [24,26,27,52,57]. This approach has an advantage of providing natural performance metrics from two by two contingency tables, along with positive and negative predictive values, to enable unambiguous assessments for survival prediction. However, this approach has a few limitations for predicting survival in cancer. First, it is not easy to take the censored survival information into consideration when building a model. In addition, the performance of binary classification depends on the threshold selected based on patient's survival information, which was used to define the two survival groups [14]. Alternatively, many studies have been using Cox proportional hazards models for cancer prognosis [10]. However, the final model from Cox regression approaches is an additive model. Thus, it is difficult to capture non-linear interactions between genomic features, which might have important roles associated with survival [16]. Even though many studies have shown an association between gene expression data and patient survival using Cox regression approaches [2,15,53], gene expression as a single dimensional genomic data type may not be enough to fully predict survival because cancer could be caused by multiple alterations through meta-dimensional genomic data including genome, epigenome, transcriptome, and proteome [17].

Many clinical data and meta-dimensional omics data have been generated from large-scale initiatives such as the International Cancer Genome Consortium (ICGC) or The Cancer Genome Atlas (TCGA). The explosion of these unprecedented dataset has provided many opportunities to examine the complex genetic architecture of several cancers and improve the diagnosis, treatment, and ultimately prevention of cancer [21,35,45–47]. Despite these efforts, it is crucial to develop a novel data integration method to better predict cancer clinical outcome, further exploring a global view on the interactions within/between meta-dimensional genomic data [23,24,27,28,39,44,56].

Previously, we proposed many methodological frameworks that predict clinical outcomes by integrating multi-omics data [23,24,27,28]. However, these binary classification approaches have difficulties to directly predict survival data due to the problems of setting threshold and the characteristics of censored observations. In the present study, we propose a novel framework designed to perform three functions simultaneously: (1) predicting censored survival data; (2) integrating meta-dimensional omics data; (3) identifying interactions within/between meta-dimensional genomic features associated with survival outcome. In order to demonstrate the utility of the proposed framework, we applied the framework on a simulation dataset followed by the breast cancer data from TCGA. Breast cancer is an extremely heterogeneous disease [22]. High degree of diversity within/between breast tumors could affect the risk of therapeutic responses and disease progression [36]. In addition, most breast cancer studies based on molecular data have mainly focused on one- or two-dimensions of genomic data, mostly copy number alteration or gene expression profiles [12,42,43]. Thus, identifying interactions within/between meta-dimensional omics data associated with survival outcome in breast cancer is expected to deliver direction for improved meta-dimensional prognostic biomarkers and therapeutic targets.

2. Materials and methods

2.1. Data

Normalized and preprocessed multi-omics datasets in breast cancer were downloaded from TCGA data matrix (http://tcga-data.nci.nih.gov/tcga/) and cBio Cancer Genomics Portal (http://www.cbioportal.org/public-portal/) (Table 1). Four different genomic data types were used for this study to represent each dimension of genomic data; CNA as genome dimension, methylation as epigenome dimension, gene expression as transcriptome dimension, and protein data as proteome dimension. Each genomic dataset was retrieved as a gene-based feature in order to better interpret the results. CNA data was obtained from the cBio Portal in order to retrieve the significantly altered copy number regions across a set of cancer patients using the GISTIC method [7]. For CNA data, 473 genes with log2 copy number value were extracted from 62 significant altered regions. DNA methylation data was also retrieved as a gene-level feature from the TCGA data matrix by choosing the least correlated with gene expression when genes were mapped with multiple methylation probes, from 485,577 methylation probes to 19,943 genes. The beta-value of human methylation 450 BeadChip was used for the elements of methylation data. Gene expression data from RNA-seq consisted of 20,502 unique gene symbols with upper quartile normalized RSEM count estimates [30]. Protein or phosphoprotein levels measured by the reverse phase protein array (RPPA) were retrieved from the cBio Portal [50]. Protein data contains 131 proteins after removing 11 proteins due to the missing data. Patients that have overlap among four types of omics data with available survival and age information, 476 patients, were used for this study.

Table 1
TCGA breast cancer data types used for meta-dimensional analysis.

Data type	Platform	# Features
CNA	Affymetrix SNP 6	473 genes
Methylation	Infinium humanmethylation450 BeadChip	19,943 genes
Gene expression	Illumina GA RNA-seq	20,502 genes
Protein expression	Reverse phase protein array (RPPA)	131 proteins

2.2. Analysis Tool for Heritable and Environmental Network Associations (ATHENA)

ATHENA was developed to uncover the meta-dimensional models that examine the genetic etiology of complex diseases such as cancer. Thus, ATHENA provides three key functions: (1) performing feature selection from categorical or continuous independent variables; (2) modeling single variable and/or interaction effects to predict categorical or continuous clinical outcomes; (3) annotating the candidate models for the interpretation in translational bioinformatics [19,24,51]. ATHENA contains several subcomponents: preprocessing, modeling, and an evolutionary-algorithm based machine learning technique at its core (Fig. 1). The current implementation of ATHENA contains two different evolutionary-algorithm modeling methods, which are Grammatical Evolution Neural Networks (GENN) and Grammatical Evolution Symbolic Regression (GESR). We have extended ATHENA to perform integrative analysis using meta-dimensional omics data to identify models that underlie the multi-layered architecture of cancer. A schematic overview of the ATHENA was shown in Fig. 1. ATHENA can simultaneously analyze meta-dimensional genomic data such as CNA, methylation, gene expression, and protein expression data to build the meta-dimensional models of complex disease. For the further analysis, we used GENN as the modeling component.

2.3. Grammatical Evolution Neural Networks (GENN)

Even though many computational methods such as multifactor dimensionality reduction (MDR) have been proposed to discover interactions between genomic features [9,38], many of them