Computer Methods and Programs in Biomedicine 161 (2018) 45–53

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

Integrating genomic data and pathological images to effectively predict breast cancer clinical outcome

Dongdong Sun[a,1], Ao Li[a,b,1,*], Bo Tang[a], Minghui Wang[a,b]

[a] School of Information Science and Technology, University of Science and Technology of China, 443 Huangshan Road, Hefei 230027, China
[b] Research Centers for Biomedical Engineering, University of Science and Technology of China, 443 Huangshan Road, Hefei 230027, China

ARTICLE INFO

Article history:
Received 5 January 2018
Revised 31 March 2018
Accepted 11 April 2018

Keywords:
Breast cancer survival prediction
Genomic data
Pathological image
Multiple kernel learning

ABSTRACT

Background and objective: Breast cancer is a leading cause of death from cancer for females. The high mortality rate of breast cancer is largely due to the complexity among invasive breast cancer and its significantly varied clinical outcomes. Therefore, improving the accuracy of breast cancer survival prediction has important significance and becomes one of the major research areas. Nowadays many computational models have been proposed for breast cancer survival prediction, however, most of them generate the predictive models by employing only the genomic data information and few of them consider the complementary information from pathological images.

Methods: In our study, we introduce a novel method called GPMKL based on multiple kernel learning (MKL), which efficiently employs heterogeneous information containing genomic data (gene expression, copy number alteration, gene methylation, protein expression) and pathological images. With above heterogeneous features, GPMKL is proposed to execute feature fusion which is embedded in breast cancer classification.

Results: Performance analysis of the GPMKL model indicates that the pathological image information plays a critical part in accurately predicting the survival time of breast cancer patients. Furthermore, the proposed method is compared with other existing breast cancer survival prediction methods, and the results demonstrate that the proposed framework with pathological images performs remarkably better than the existing survival prediction methods.

Conclusions: All results performed in our study suggest that the usefulness and superiority of GPMKL in predicting human breast cancer survival.

1. Introduction

Cancer is a class of diseases in which involves abnormal cell growth and body change [1]. In general, cancer is termed after the body part in which it originated, thus breast cancer is a kind of disease that develops from the breast tissue [2]. Breast cancer in women is still the most common malignancy, with a leading cause of cancer-related deaths throughout worldwide [3]. There are around 3.1 million breast cancer survivors in the United States (U.S.), and the chance of a woman dying from breast cancer is around 1 in 37, or 2.7% according to American Medical News Reports. This highlights the urgent need to design computational methods for a more precise survival prediction of breast cancer and may lead to the development of personalized treatment and management. Accordingly, this would ultimately contribute to reducing overall mortality rate of breast cancer and further improving the quality of life in breast cancer patients.

Towards this goal, during the past few years, many researches have adopted the microarray technology to study gene expression profiles in breast cancer, however only a small fraction shows clear prognostic significance [4,5]. For example, Van't Veer et al. use DNA microarray analysis on primary breast tumors from 117 patients and utilize a supervised classification method to recognize a 70-gene prognostic signature [4]. Further, they test these previously applicable prognostic markers in a series of 295 consecutive breast cancer patients and the results demonstrate the significance of 70-gene prognostic signature [6]. Wang et al. [7] reveal a 76-gene prognostic signature that can accurately predict distant tumor recurrence by clustering the gene expression profiles and correlating them with prognostic values. By using microarray markers, some machine learning classification methods, such as Support Vector Machine (SVM) [8], Bayes classifier [9], Random Forest (RF)

[*] Corresponding author at: School of Information Science and Technology, University of Science and Technology of China, 443 Huangshan Road, Hefei 230027, China.
E-mail addresses: sddchina@mail.ustc.edu.cn (D. Sun), aoli@ustc.edu.cn (A. Li), tb214200@mail.ustc.edu.cn (B. Tang), mhwang@ustc.edu.cn (M. Wang).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.cmpb.2018.04.008
0169-2607/© 2018 Elsevier B.V. All rights reserved.

[10] have also been applied to predict breast cancer survival. For instance, Nguyen et al. [10] propose to diagnose and prognosticate breast cancer based on random forest classifier and feature selection technique, which outperforms previously reported results.

Given the complexity and heterogeneity of breast cancer survival prediction, a more practical strategy, as proposed by Brenton et al. [11], is to use both clinical data and gene prognostic markers that may contain some complementary information. In addition, with the rapid development of new technologies in the area of medicine, a large amount of clinical data for breast cancer have been generated and collected. By combining both clinical data and microarray markers, different computational methods have been developed for the accurate survival prediction of breast cancer [9,12–14]. For example, Gevaert et al. develop Bayesian networks [9] to integrate both clinical and those 70 gene information by three different strategies including full, decision or partial integration, and demonstrate that the use of clinical and microarray data has better or comparable performance than the methods with clinical or microarray data, respectively. Khademi et al. [14] propose an interesting strategy to reduce the dimensionality of microarray data by applying manifold learning and deep belief network and integrate the clinical data by a probabilistic graphical model. The extensive experiments show a promising result compared to traditional classification methods. Except for microarray and clinical information, the reference human protein interactive network has also been explored to predict breast cancer survival. For example, Das et al. [15] design an elastic-net-based approach named ENCAPP by combining protein network with gene expression dataset to accurately predict survival for human breast cancer. However, the limitation of ENCAPP is that the accuracy of the survival prediction is highly dependent on the quality of the gene expression dataset. Thus there is still considerable room for the improvement of prediction performance of breast cancer survival by incorporating more cancer-related information.

Currently, with the advance of technology in medical imaging [16], there is a great opportunity to analyze pathological images and well study tumor morphology [17]. Previous studies show that some computational methods have been introduced to predict cancer clinical outcome based on pathological images by assuming that pathological images may provide complementary information related to tumor characteristics. Wang et al. [18] propose an integrated framework for non-small cell lung cancer computer aided diagnosis and survival analysis by using representative markers from histopathology images. Zhu et al. [19] design a prediction model to integrate pathological image features with gene expression signature for lung cancer survival prediction. By collecting 2186 Hematoxylin and Eosin pathological whole-slide images (WSIs) of non-small cell lung cancer [20], Yu et al. [21] further distill 9879 representative image features and employ common classification methods to distinguish shorter-term and longer-term survivors. Despite the good performance of the above mentioned approaches for lung cancer, there is still a lack of researches with pathological images for breast cancer clinical outcome analysis due to the complexity and heterogeneity of this serious disease. Meanwhile, the rapidly increasing number of features from different data sources and the use of heterogeneous features may bring a big challenge on how to effectively combine them to apply into breast cancer survival prediction.

To address these issues, in this study, we conduct a new, powerful method named GPMKL for survival prediction of breast cancer by integrating genomic data (gene expression, copy number alteration (CNA), gene methylation and protein expression) and features distilled from pathological image. By employing those high-quality features, multiple kernel learning is further introduced to integrate and accurately predict survival time of breast cancer patients. To verify the effectiveness of pathological images, GPMKL is compared with different independent models that only use genomic data and the results indicate that pathological images could contribute to the remarkable prediction performance. Further, we also compare our proposed framework with other popular state-of-the-art survival models. The best performance achieved by GPMKL also demonstrates the feasibility of integration of genomic data and pathological images and the usefulness of GPMKL in breast cancer survival prediction.

Table 1
The properties of our breast cancer dataset.

Total population of patients	578
Cut-off (years)	5
Survival time	
Longer-term survivors	133
Shorter-term survivors	445
Average age at diagnosis	57.80
Median survival	40.46

2. Materials and methods

Fig. 1 shows the framework of the proposed method called GPMKL. The entire procedure consists of three steps: (A) data generation; (B) model optimization and feature selection; (C) prediction and evaluation. In detail, the proposed method first integrates genomic data including gene expression, gene methylation, CNA, protein and pathological images. Second, breast cancer patients are randomly divided into training and testing sets. The training set is used to train our model and tune parameters on ten-fold cross validation experiment. Finally, the trained model is utilized to do classification on the testing set. We describe each part of our framework further in the following sections.

2.1. Data preparation

We download the publicly available dataset of breast cancer samples from The Cancer Genome Atlas (TCGA) portal: https://portal.gdc.cancer.gov/. TCGA is a comprehensive resource including thousands of patients' data, which is consisted of gene expression, CNA, gene methylation, protein expression and pathological images. In this paper, the latest TCGA available data is downloaded to conduct our research (June 2017) and all above mentioned data types are retrieved from original dataset for further analysis. This downloaded dataset consists of five sub-data and each sub-data include different number of patients. For example, the gene expression and pathological images contain 1100 and 1054 patients, respectively. Then we use Venn diagram (Fig. 2) to vividly exhibit the detailed number of patients in different data types and finally obtain 585 valid patients which consist of those mentioned data types. This dataset contains 578 female and 7 male patients. Considering that the gender distribution is very biased in our dataset and some previous studies demonstrate numerous gender-specific differences for breast cancer [22,23], we further remove the 7 male patients. The average age at diagnosis is 57.80 and the average survival time of all these patients is 40.46 months. Similar to previous study, we define the breast cancer survival prediction in our study as a binary classification problem by the threshold of 5 years [14,24]. In detail, patients in our study are divided into two classes by their survival time, namely longer-term and shorter-term survivors. Among 578 patients, 445 patients are regarded as shorter-term survivors and 133 patients are regarded as longer-term survivors. Moreover, the shorter-term survivors are labeled as 0 while longer-term patients are labeled as 1. The detailed information about our dataset is illustrated in Table 1.