Computers in Biology and Medicine 100 (2018) 92–99

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Machine learning models to predict the progression from early to late stages of papillary renal cell carcinoma

Noor Pratap Singh[a], Raju S. Bapi[b,c], P.K. Vinod[a,*]

[a] Center for Computational Natural Sciences and Bioinformatics, International Institute of Information Technology (IIIT), Hyderabad, 500032, India
[b] Cognitive Science Lab, International Institute of Information Technology (IIIT), Hyderabad, 500032, India
[c] School of Computer and Information Sciences, University of Hyderabad, 500046, India

ARTICLE INFO

Keywords:
Papillary renal cell carcinoma
Machine learning
Feature selection
Tumour stage prediction
Cell cycle
Chromosome instability

ABSTRACT

Papillary Renal Cell Carcinoma (PRCC) is a heterogeneous disease with variations in disease progression and clinical outcomes. The advent of next generation sequencing techniques (NGS) has generated data from patients that can be analysed to develop a predictive model. In this study, we have adopted a machine learning approach to identify biomarkers and build classifiers to discriminate between early and late stages of PRCC from gene expression profiles. A machine learning pipeline incorporating different feature selection algorithms and classification models is developed to analyse RNA sequencing dataset (RNASeq). Further, to get a reliable feature set, we extracted features from different partitions of the training dataset and aggregated them into feature sets for classification. We evaluated the performance of different algorithms on the basis of 10-fold cross validation and independent test dataset. 10-fold cross validation was also performed on a microarray dataset of PRCC. A random forest based feature selection (varSelRF) yielded minimum number of features (104) and a best performance with area under Precision Recall curve (PR-AUC) of 0.804, MCC (Matthews Correlation Coefficient) of 0.711 and accuracy of 88% with Shrunken Centroid classifier on a test dataset. We identified 80 genes that are consistently altered between stages by different feature selection algorithms. The extracted features are related to cellular components - centromere, kinetochore and spindle, and biological process mitotic cell cycle. These observations reveal potential mechanisms for an increase in chromosome instability in the late stage of PRCC. Our study demonstrates that the gene expression profiles can be used to classify stages of PRCC.

1. Introduction

Mutations or epigenetic modifications are seen as drivers for cancer progression by affecting the pattern of gene expression. The advent of next generation sequencing techniques (NGS) has led to the creation of valuable resources of human cancer which are available from The Cancer Genome Atlas (TCGA) (http://cancergenome.nih.gov) [27]. Understanding the alterations in gene expression can serve as a diagnostic tool to distinguish normal vs cancer tissue, subtypes and stages of cancer [8,58]. Further, it can help to understand the disease mechanism and to identify therapeutic targets. Machine learning methods have been used extensively for cancer prediction and prognosis [18,32]. For this purpose, different feature selection and classification techniques have been applied on multidimensional heterogeneous cancer data to predict the cancer susceptibility, recurrence and survival. Feature selection techniques can be employed in the place of other dimensionality reduction techniques based on projection (e.g. principal component analysis) or compression (e.g. using information theory) to extract features that are subset of original variables allowing for interpretability of feature(s) [12,45].

Recently, TCGA studies reported comprehensive molecular profiles of three major histologically defined types of Renal Cell Carcinoma (RCC): clear cell, chromophobe and papillary, providing an opportunity to further analyse these datasets with an aim to develop predictive tools [9,10,19]. RCC arises from the various parts of the nephron and possesses distinct genetic features and histological characteristics [42,43]. The clear cell RCC is the most common RCC and its dataset has been subjected to different analyses to identify subtypes and to predict the survival and stages of tumour development [4,14,29,55]. Here, we focus on analysing the dataset of Papillary Renal Cell Carcinoma (PRCC), which is second most common histological subtype of RCC accounting for 10%–15% cases [30,43]. PRCC is a heterogeneous disease with two main histologic subtypes. Type 1 tumours consist of papillae and tubular structures covered by small cells with basophilic cytoplasm and small oval nuclei, whereas Type 2 tumours consist of papillae covered by large cells with eosinophilic cytoplasm and large

* Corresponding author.
E-mail address: vinod.pk@iiit.ac.in (P.K. Vinod).

https://doi.org/10.1016/j.compbiomed.2018.06.030
Received 10 March 2018; Received in revised form 12 June 2018; Accepted 25 June 2018

spherical nuclei (with prominent nucleoli) [20]. In some cases, PRCC is indolent and multifocal in nature while in other cases it has aggressive lethal phenotype of solitary tumours. There are still no effective treatments available for PRCC [38].

The TCGA study on PRCC is a significant step forward in understanding the molecular basis of PRCC [10]. This study using 161 samples revealed additional subtypes within type 2 PRCC and identified genes associated with PRCC including MET, NF2, SETD2, TFE3, CDKN2A and Nrf2 pathway genes. There are more samples made available now that can be used for further characterization of PRCC. A multi-genomics study based on renal cell carcinoma found four PRCC subtypes that highly overlapped with earlier subtype designations and histology based classification [13]. Further, supervised analysis using single/multi-genomic data and clinical information of PRCC can be performed to develop a model for predicting the progression from early to late stage of disease. A recent study used a network based approach to find biomarkers associated with pathological stages of tumour development (Stages I, II, III and IV). However, this study used only a subset of available RNA sequencing (RNAseq) dataset (106 samples) [25]. Further characterization of stages of tumour development will aid in early detection and effective treatment.

The major objective of our study is to identify biomarkers and build classifiers to discriminate early and late stages of PRCC from gene expression profiles. Supervised machine learning algorithms were employed to analyse RNAseq dataset for both feature extraction and classification. In that process, we evaluated the performance of different algorithms and compared the results. We show that the features extracted from gene expression profiles can be used to efficiently classify the stages of tumour development. A maximum area under Precision Recall Curve (PR-AUC) of 0.81 and MCC of 0.71 were obtained with independent RNASeq test dataset. Shrunken Centroid classifiers performed the best having high PR-AUC and MCC on the test dataset, followed by Random Forest and Naive Bayes. We also validated the models and feature sets by performing a 10-fold cross validation using a microarray dataset of PRCC. The features extracted are related to cellular components centromere, kinetochore and spindle, and biological process mitotic cell cycle.

2. Materials and methods

2.1. Dataset

RNASeq dataset and clinical information of PRCC were downloaded from Genomics Data Commons (GDC) data portal (https://portal.gdc.cancer.gov/). We used level-3 RNAseq data which was obtained using Illumina Hiseq RNAseqV2 platform. The dataset includes samples obtained from 32 normal and 289 patients. There were 31 matched normal and tumour samples. The pathological stage (I, II, III and IV) of each tumour sample was obtained from the clinical information available for each patient. The pathological stage information is only available for 260 samples with the following distributions: Stage I-172, Stage II-22, Stage III-51, and Stage IV-15. It can be seen here that there is a severe imbalance in distribution of samples across the pathological stages, posing potential challenges for machine learning algorithms. The dataset was divided into training (80%) and test (20%) datasets. The training dataset was further divided randomly into four folds/partitions. We formed four groups using the above four folds such that $group_i$ contains samples from each fold excluding $fold_i$. This was done to obtain reliable features with different distributions of samples and to obtain features that possibly account for the heterogeneity of PRCC. The raw count data was normalised using variance stabilizing transformation (VST) [2]. The PRCC microarray dataset with accession no: GSE2748 obtained from Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/) was also used to evaluate the performance of classifiers. This dataset was obtained using Affymetrix HGU133 Plus 2.0 arrays platform and includes 19 and 15 samples in early (excellent survival) and late (poor survival) stage of PRCC, respectively. We used the pre-processed data obtained by robust multichip average (RMA) algorithm, which performs background correction, quartile normalization and summarization of microarray dataset [28].

2.2. Feature selection and characterization

We used four different feature selection algorithms: DESeq2, SAMseq, Shrunken Centroids and varSelRF, to extract features between different tumour stages of PRCC [22,33,35,52,53]. DESeq2 was also used to extract features between matched pair of normal and tumour samples. DESeq2 assumes a negative binomial distribution of reads whereas SAMseq assumes a non-parametric read distribution. We applied different $\log_2$fold change ($\log_2$FC) criteria to extract features for DESeq2 and SAMseq. For DESeq2, we only considered features with Benjamini-Hochberg adjusted p-value < 0.05 while for SAMseq we considered features with q-value < 0.05. Shrunken Centroids finds class-specific genes by computing a class-wise centroid and a t-statistic for each gene for each class, expressing the class-wise centroid in terms of overall centroid and t-statistic. It then shrinks the t-statistic by soft thresholding giving shrunken centroids, leaving only genes with a non-zero t-statistic as the class specific gene [52]. varSelRF is a Random Forest based recursive feature selection algorithm where feature importance is computed first and then features are removed at each iteration. The iteration that yields the least number of genes with an out-of-bag (OOB) error comparable to the iteration yielding the lowest OOB error is chosen [22]. R-package implementations of these algorithms were used for the analysis. All the feature selection algorithms were applied on the four groups of data created from the training dataset. The feature sets were created by aggregating features obtained in at least 1, 2, 3 or all the groups, which are represented as AF1, AF2, AF3 and AF4, respectively. This was done for each of the feature selection algorithms. We performed functional enrichment analysis to obtain Gene Ontology (GO) terms, cellular components and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways associated with feature sets using the Database for Annotation, Visualization and Integrated Discovery (DAVID 6.8) [21]. Benjamini-Hochberg adjusted p-value < 0.05 was used to select the biological processes and components. Further, STRING database was used for extracting the known protein-protein interactions between features [50].

2.3. Classification models and their performance

Different supervised machine learning algorithms: Random Forest, Naive Bayes, SVM, KNN and Shrunken Centroid classifier were used to create models for predicting the tumour stages of PRCC using RNAseq dataset [1,6,16,24,52]. These classifiers were trained on each aggregated feature set: AF1, AF2, AF3 and AF4 (obtained for each feature selection algorithm) and their performances were evaluated by 10-fold cross validation. The metrics such as Accuracy, PR-AUC, MCC, Sensitivity, Specificity and F-value were used to quantify the performance of models [36,46,48]. Since class imbalance exists in our dataset, we have used Matthews Correlation Coefficient (MCC) and Precision Recall AUC (PR-AUC) in our study [15]. MCC considers mutually accuracies and error rates on both classes [5]. Further, precision recall plot is also suggested to be more informative in evaluating binary classifier on imbalanced datasets compared to ROC-AUC [46]. We evaluated the performance of classifiers on an independent test dataset. Further, we also used PRCC microarray dataset (GSE2748) [56] to perform a 10-fold cross validation (training-cum-validation) of different classifiers with the aggregated feature sets extracted from RNAseq data. This also helps to validate the feature sets that can be reliably used for classifying the stages of PRCC. The codes for analysing PRCC samples (feature extraction and classification) are provided in the repository https://github.com/NPSDC/BISP.