Canadian Journal of Biotechnology

Poster Presentation   OPEN ACCESS

Category: Cancer Genomics

Identification of Biomarkers for Stage Prediction in Papillary Renal Cell Carcinoma

Noor Pratap Singh, Raju S. Bapi and P.K. Vinod

International Institute of Information Technology, Hyderabad, INDIA

Presenting author: noorpratap.singh@research.iiit.ac.in

Abstract

Papillary Renal Cell Carcinoma (PRCC) is a heterogeneous disease accounting for 10%-15% of renal cell carcinomas. A comprehensive analysis is required to find the genes that are responsible for the stage progression in PRCC. The advent of next generation sequencing techniques (NGS) has produced a lot of high throughput data from patients that can be analyzed to address this problem. The low sample size, noise and high dimensionality of the data though enhances the complexity, requiring the use of sophisticated methods. In our study we propose a machine learning pipeline fulfilling a two-fold objective: 1) To find suitable genes that could serve as potential biomarkers for stage progression in PRCC. 2) To build a classifier using the above biomarkers that can predict the stage of a given patient. The RNA-Seq data of PRCC was taken and divided into training set (80%) and test set (20%). Different groupings of training data were created and on each group different feature selection algorithms. The features (genes) extracted were then combined based on voting. The selected features from each feature selection algorithm were then used to train the classifiers on the training data. The performance of the model on the test data was evaluated using various measures. To further check the quality of the genes a 10 fold cross validation was performed on microarray cohort of PRCC. The selected genes we get are robust with overlap among the features derived from the various feature selection algorithms. The best of the classifiers trained above gave an accuracy > 86 % and area under Receiver Operating Curve (AUC) > 0.8. The 10 fold cross validation on microarray data using the above features yields best accuracy > 85 % and AUC > 0.84 enhancing our confidence on our gene sets. The feature sets which we get could be further investigated for identifying their role in stage progression. Further our pipeline could be used for analyzing other cancer data sets.