EBioMedicine xxx (2018) xxx–xxx

Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.ebiomedicine.com

Research Paper

RankProd Combined with Genetic Algorithm Optimized Artificial Neural Network Establishes a Diagnostic and Prognostic Prediction Model that Revealed *C1QTNF3* as a Biomarker for Prostate Cancer

Qi Hou [a,b,1], Zhi-Tong Bing [c,d,2], Cheng Hu [e], Mao-Yin Li [e], Ke-Hu Yang [c,d], Zu Mo [f], Xiang-Wei Xie [f], Ji-Lin Liao [f], Yan Lu [b], Shigeo Horie [b], Ming-Wu Lou [a,*]

[a] Post-Doctoral Research Center, Longgang Central Hospital, Shenzhen Clinical Medical Institute, Guangzhou University of Chinese Medicine, Shenzhen 518116, China
[b] Department of Urology, Juntendo University Graduate School of Medicine, Tokyo 1138421, Japan
[c] Evidence Based Medicine Center, School of Basic Medical Science, Lanzhou University, Lanzhou 730000, China
[d] Key Laboratory of Evidence Based Medicine and Knowledge Translation of Gansu Province, Lanzhou 730000, China
[e] Department of Urology, Third Affiliated Hospital of Sun Yat-sen University, Guangzhou 510630, China
[f] Department of Urology, Longgang Central Hospital, Shenzhen Clinical Medical Institute, Guangzhou University of Chinese Medicine, Shenzhen 518116, China

ARTICLE INFO

Article history:
Received 14 March 2018
Received in revised form 8 May 2018
Accepted 8 May 2018
Available online xxxx

Keywords:
RankProd
Artificial neural network
Genetic algorithm
Prostate cancer
Biomarker

ABSTRACT

Prostate cancer (PCa) is the most commonly diagnosed cancer in males in the Western world. Although prostate-specific antigen (PSA) has been widely used as a biomarker for PCa diagnosis, its results can be controversial. Therefore, new biomarkers are needed to enhance the clinical management of PCa. From publicly available microarray data, differentially expressed genes (DEGs) were identified by meta-analysis with RankProd. Genetic algorithm optimized artificial neural network (GA-ANN) was introduced to establish a diagnostic prediction model and to filter candidate genes. The diagnostic and prognostic capability of the prediction model and candidate genes were investigated in both GEO and TCGA datasets. Candidate genes were further validated by qPCR, Western Blot and Tissue microarray. By RankProd meta-analyses, 2306 significantly up- and 1311 down-regulated probes were found in 133 cases and 30 controls microarray data. The overall accuracy rate of the PCa diagnostic prediction model, consisting of a 15-gene signature, reached up to 100% in both the training and test dataset. The prediction model also showed good results for the diagnosis (AUC = 0.953) and prognosis (AUC of 5 years overall survival time = 0.808) of PCa in the TCGA database. The expression levels of three genes, *FABP5*, *C1QTNF3* and *LPHN3*, were validated by qPCR. *C1QTNF3* high expression was further validated in PCa tissue by Western Blot and Tissue microarray. In the GEO datasets, *C1QTNF3* was a good predictor for the diagnosis of PCa (GSE6956: AUC = 0.791; GSE8218: AUC = 0.868; GSE26910: AUC = 0.972). In the TCGA database, *C1QTNF3* was significantly associated with PCa patient recurrence free survival ($P < .001$, AUC = 0.57). In this study, we have developed a diagnostic and prognostic prediction model for PCa. *C1QTNF3* was revealed as a promising biomarker for PCa. This approach can be applied to other high-throughput data from different platforms for the discovery of oncogenes or biomarkers in different kinds of diseases.

1. Introduction

Prostate cancer (PCa) is the most commonly diagnosed cancer in males and one of the leading causes of cancer mortality in the Western world. An estimated 164,690 Americans will be diagnosed with prostate cancer and 29,430 will die of the disease in the United States in 2018 [30]. In recent years the number of diagnosed prostate cancer patients also increased rapidly in developing countries such as China [41]. Here it is one of the ten most common cancers diagnosed in men, with an estimated 60,300 new cases and 26,600 deaths in 2015 [8]. The 5-year relative survival rate of localized prostate cancer patients approaches 100% but sharply decreases to 28% for patients diagnosed at an advanced stage [22]. Therefore, early detection and precise diagnosis for prostate cancer needs to advance further.

At the moment, prostate specific antigen (PSA) testing is widely used for PCa diagnosis at an early organ-confined stage. However, it occasionally leads to unnecessary biopsies due to its poor specificity [1,21,31]. Therefore, other new biomarkers with high accuracy and specificity are needed to improve diagnosis and prognosis of PCa. Cima et al. employed a proteome method and identified a five-protein

* Corresponding author.
E-mail address: mingwulou@sina.com (M.-W. Lou).
[1] Equal contributors
[2] Equal contributors

https://doi.org/10.1016/j.ebiom.2018.05.010

signature (GALNTL4, FN, AZGP1, GBA and ECM1) in PCa which could be used to improve screening efficacy [10]. Ankerst et al. proposed that detection of PCA3, a PCa-specific gene, combined with PSA testing could improve diagnostic accuracy [2]. In addition, hypermethylation of some critical genes or microRNAs like GSTP1, PITX2, GABRE~miR-452~miR-224, and a methylated site (cg05163709) in Chromosome Y have been proposed as promising biomarkers of PCa [32,39]. Furthermore, a number of commercially available products show potential, but there is still some way to go before there is enough data to convincingly demonstrate the added value of these methods.

High-throughput microarray chip and second-generation sequencing technologies are powerful tools for discovering and studying novel biomarkers for PCa. However, analyses based on high throughput data may encounter the "curse of dimensionality" [23]. This refers to the phenomena that the amount of dependent variables increases greatly while the amount of samples is relative small, resulting in an increase of statistical errors. Fortunately, increasing the sample size and using some machine learning algorithm can effectively improve the problems caused by this "curse" [3,23].

The geometric mean algorithm can integrate ranked lists from various datasets produced by a wide variety of platforms, such as Affymetrix oligonucleotide arrays, two-color cDNA arrays and other custom-made arrays [11,15]. To increase the sample size, RankProd, a non-parametric statistical method which can combine datasets from different origins to increase the power of identification, has been used to datamine various cancers. Suraj Peri et al. integrated various data from kidney tissue microarrays for differential expression of genes (DEG) analysis and identified NF-κB and interferon signatures as clinical features of clear cell renal cell carcinoma (ccRCC) [25]. Many other studies also employed RankProd to extend their sample size [9,19,33].

In recent years, some machine learning algorithms, such as support vector machines (SVM), principal component analysis (PCA), least absolute shrinkage and selection operator (LASSO) and artificial neural network (ANN) also help with solving the curse of dimensionality. When comparing these models and methods with ANN, the latter shows many advantages when handling high dimensional data by dealing with non-linear relationships in data [5]. ANN is based on a collection of connected units called artificial neurons (analogous to axons in a biological brain). Neurons and synapses may also have a weight that varies as learning proceeds, which can increase or decrease the strength of the signal that it sends downstream. This model is suitable for prediction even when the experimental data are not subjected to Gaussian distribution for it is established simply by construction of multiple layers of artificial neurons and utilizes its network connections to deliver and process the required input information. Due to its advantages in processing defective or non-linear data, ANN is currently widely used in the diagnosis of cancer, survival analysis and estimation of intensive care [24,27,29]. Genetic algorithm (GA) is a generally evolutionary algorithm that has already been considered appropriate to solve the general optimization problems [14]. GA is widely used in the selection of the variables resulting in the best fit for the ANN models [7,37]. Although many studies reported the use of ANN in classification gene expression microarray, integrating GA with ANN to establish prediction models and filter candidate genes for cancer samples has few reports.

To solve the curse of dimensionality in high-dimensional gene expression data, we were trying to establish a data processing system using these two methods. We firstly integrated data from different independent datasets to expand the sample size by RankProd. Based on that, we employed a GA-ANN model to establish a prediction model and screen for candidate genes of PCa. The combination of RankProd and GA-ANN in this study, allows us to develop a promising processing approach for discovering novel biomarkers and candidate gene patterns for the diagnosis and prognosis prediction of PCa.

2. Materials and Methods

2.1. Data and Sample Collection

Public data was collected from the Gene Expression Omnibus (GEO) dataset and The Cancer Genome Atlas (TCGA) dataset. Only microarray data that met the following criteria were included; (1) Datasets were produced by Genome-wide mRNA expression profiling by microarray, (2) The experimental platform was single-channel; (3) All cases were pathologically diagnosed to be prostate cancer tissues while the controls were identified as para-carcinoma or normal prostate tissues; (4) The minimum number of cases and controls was 3. Finally, available datasets from the following cohort were included. Wallace et al. contained gene expression profiles of primary prostate tumors resected from 33 African-American and 36 European-American patients. It also contained 18 normal prostate tissues from 7 African-American and 11 European-American patients [35]. Wang et al. contained 148 prostate samples [36]. Planche et al. [26] contained 6 prostate cancer and matched normal samples. Taylor et al. contained 218 PCa samples and 149 matched normal samples from patients treated by radical prostatectomy [34]. Ross-Adams et al. contained 99 prostate cancer samples from patients with follow-up data [28]. For the TCGA dataset, we included the TCGA-PRAD which contained 500 PCa patients.

We also analyzed 69 primary prostate cancer patients and paired adjacent normal tissues by a tissue microarray obtained from Shanghai Outdo Biotech, China (Supplementary file 1: Table S1). Another 28 independent PCa and paired adjacent normal tissues were analyzed by qPCR and western blot (Supplementary file 2: Table S2). All fresh tissues were obtained with informed consent from patients hospitalized at the Department of Urology, Longgang Central Hospital and the Department of Urology, Third Affiliated Hospital of Sun Yat-Sen University. All tissue specimens were confirmed by pathology and immediately frozen in liquid nitrogen. All experiments in this study were approved by the ethics committee of Longgang Central Hospital and Third Affiliated Hospital of Sun Yat-Sen University.

2.2. Individual Participant Data Processing

In order to integrate microarray data from different platforms, meta-analysis was carried out by RankProd. The annotation files corresponding to the types of microarrays were downloaded from the official Affymetrix website. To pre-process Affymetrix microarray data, RMAExpress1.0.5 was introduced for background adjustments, normalization was done by Quantile and summarization by Median Polish. The output files were composed of the normalized expression values of every probe. Shared probes were extracted from different platforms using Perl 5.10 and RankProd package installed in R (v3.4.0) was run. Probe signals with percentage of false prediction (pfp) value lower than 0.05 would be considered as DEGs. GO enrichment and KEGG analysis were carried out using clusterProfiler package in R (v3.4.0) [40].

2.3. Development of GA-ANN PCa Prediction Model

After acquiring the DEG list, we constructed the ANN model in MATLAB (MathWorks, Massachusetts, USA) by setting the clinical phenotype of 163 microarray samples as the output variable (normal or cancer patients) and the expression values of the top 500 up- and down-regulated probes as the input variables. A training set was built with 100 randomly selected microarray samples and the other 63 microarray samples were used as a test set. The model was composed of 3 layers with 1000 nodes as the input layer (each representing an expression value of a probe) and 1 node as the output layer (the clinical phenotype). We set the maximum recursive time to 100 and the threshold of mean square error to 0.005. The weight-corrected learning rate was 0.1 and the transfer function from input layer to hidden layer was tansig while purelin was configured as the transfer function from

Please cite this article as: Hou Q, et al, RankProd Combined with Genetic Algorithm Optimized Artificial Neural Network Establishes a Diagnostic and Prognostic Prediction Mo..., EBioMedicine (2018), https://doi.org/10.1016/j.ebiom.2018.05.010