Chemometrics and Intelligent Laboratory Systems xxx (2017) 1–7

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Combining gene essentiality with feature selection method to explore multi-cancer biomarkers

Ziyan Huang [a], Yongcheng Dong [b], Yan Li [a], Qifan Kuang [a], Daichuan Ma [c], Yizhou Li [a,*], Menglong Li [a,**]

[a] College of Chemistry, Sichuan University, Chengdu, China
[b] College of Life Sciences, Sichuan University, Chengdu, China
[c] Analytical & Testing Center, Sichuan University, Chengdu, China

ARTICLE INFO

Keywords:
Multi-cancer biomarkers
Gene essentiality
VCPA
Feature selection
Protein–protein interaction network

ABSTRACT

Biomarker discovery plays an important role in cancer diagnosis and prognosis assessments. The biomarkers that could be applied among different cancer types are highly useful. Although many traditional feature selection algorithms have shown their power on picking discriminative genes, they are incapable of identifying biologically meaningful biomarkers. Here, on the hypothesis that gene essentiality would be disrupted in cancers, we estimated the gene sets with significant essentiality alteration in six cancer types. We found that different cancer types would share some common gene essentiality alterations. Then, the variable combination population analysis (VCPA) algorithm was applied to identify the potential biomarkers from these common genes, which were used to construct prediction models and exhibited satisfactory classification ability (averaged accuracy: 0.9752) among six cancer types. Interestingly, these biomarkers would tend to cluster as a subnetwork and be characterized by high centrality values in the protein–protein interaction network. They were significantly enriched in the cell cycle and DNA replication pathway which are hallmark signatures of cancers. Several biomarkers have been even verified by the literature searching, reported having roles in chromosome instability and aberrantly expressed between cancer/normal samples. An additional comparison analysis between the VCPA and other six feature selection methods in WEKA suggested biomarkers by VCPA perform superior over those by other methods. These results suggested that our method is promising in identifying the potential multi-cancer biomarkers.

1. Introduction

Cancer is recognized as a generally complex disease. In the past years, most researches focused on the cancer that originating in the same tissue, which was heterogeneous and could be divided into several subtypes [1–3]. Some common characteristics were also found in different cancer types, such as the microsatellite instability in colorectal and endometrial cancers [4,5] and the somatic inactivation of the BRCA1-BRCA2 pathway in both basal-like breast cancer and serous ovarian cancers [6]. Recently, a research from The Cancer Genome Atlas (TCGA) has also indicated that even specific subtypes across different cancer types can share some common features [7]. Thus, for the deeper understanding cancer mechanism, it is desired to design reliable strategies for exploring these commonalities among different cancer types.

Biomarker discovery for cancer based on the molecular factor, such as proteins and genes has become a major strategy in biomedical fields [8]. Traditionally machine learning methods were applied to find cancer biomarkers [9]. Unfortunately, even for the same cancer type few consentaneous gene sets were reported by different researches [10–12]. The subtle alterations on the driver genes might be amplified on those downstream effectors, which might vary from different patients. Though the feature selection methods have shown their power on picking discriminative genes, the biological significance of these selected genes was unclear [13]. Some studies [14,15] focused on differentially expressed genes, while other noted the non-differentially expressed genes could also play a central role by interacting with other genes [16]. To solve the issue, many groups proposed a more promising strategy based on some prior biological knowledge, such as the pathway or the network to reveal the coherent expression patterns [16–19]. However,

* Corresponding author.
** Corresponding author.
E-mail addresses: liyizhou@scu.edu.cn (Y. Li), liml@scu.edu.cn (M. Li).

https://doi.org/10.1016/j.chemolab.2017.11.007
Received 12 May 2017; Received in revised form 16 October 2017; Accepted 2 November 2017
Available online xxxx
0169-7439/

most of them focused on the prognosis and diagnosis characteristics of biomarkers in an individual cancer. Recently, some attentions have been focused on exploring multi-cancer biomarkers across different cancers, due to their potential on the clinical settings and even on the drug target development [20–24]. Kaczkowski et al. [25] identified a set of multi-cancer biomarkers recurrently perturbed in cancer samples based on differential expression profiles. Martinez-Ledesma et al. [26] used a network-based algorithm method to find multi-cancer biomarkers.

The essential genes are critical for cell viability under certain context, which might tend to be highly expressed, take part in the crucial pathway and involve in more protein–protein interactions. These genes may contribute to exploring targets for cancer therapies [27]. Lately, Jiang et al. proposed the NEST (Network Essentiality Scoring Tool) to estimate the gene essentiality in functional genomics experiments [28]. The essentiality scores are closely correlated with their neighboring genes expression levels in the biological network. Furthermore, through a patient survival analysis, they proposed that a gene neighboring over-expressed ones in the cancer samples is more likely the oncogene with associated survival risk. Thus, we speculated that the gene essentiality alterations between cancer/normal samples could provide new clues for discovering potential multi-cancer biomarkers and common biological mechanisms among cancers.

Here, we attempted to combine the gene essentiality information, as the prior biological knowledge, with the feature selection algorithm VCPA to explore the potential multi-cancer biomarkers. VCPA is a recently proposed method to investigate NIR spectral datasets. It has been proved as a good variable selection strategy when compared with several other well-established variable selection methods [29]. We expected that such combination strategy could produce discriminative genes between cancer/normal samples shared by different cancer types.

In our pipeline, firstly, the gene essentiality score alterations were estimated between cancer and the control samples in the six cancer types (breast, colorectal, gastric, lung, liver and pancreatic), respectively. >37% overlaps were observed among the most altered genes (top 500) for each cancer type. The results indicated that these alterations tended to reflect the common characteristics. These common genes included both common differentially expressed genes among six cancer types and the non-differentially expressed genes closely connected with the differentially expressed genes. Then, the VCPA algorithm was applied on this gene set for picking out the most informative ones. VCPA was performed fifty times against each cancer type and the output gene lists were then combined for the most frequently selected gene. They were used as the features for a 10-fold cross validation PLS [30] model training. The gene set with the best performance for all these six cancer types was regarded as the candidate multi-cancer biomarkers. Finally, a comparison analysis was carried out between VCPA and other six feature selection methods from WEKA. The superior performance of our candidate biomarkers further validated our method. Additionally, these candidate biomarkers could be highly supported by biological pathway analysis, the previously reported multi-cancer biomarkers, the literature searching and the network analysis. These results demonstrated that our method would be promising not only to find the potential multi-cancer biomarkers, but also shed light on understanding the commonly disturbed biological mechanisms among different cancer types.

2. Material and methods

2.1. Data source and preprocessing

The gene expression data of six cancer types were collected from the GEO database (GEO accession numbers in Table 1). The Robust Multiarray Average (RMA) normalization method contained four steps: 1) Background correction. 2) Normalization (across arrays). 3) Probe level intensity calculation. 4) Probe set summarization. We downloaded the expression matrix for each cancer and then converted the gene id to symbol names according to their corresponding experiment platform.

2.2. The combination of gene essentiality score and feature selection method

The workflow was shown in Fig. 1. Firstly, Gene essentiality scores were estimated by the method proposed by Jiang [28] based on STRING protein–protein interaction (PPI) network [31]. By speculating that the gene essentiality score would be alternative under different conditions, we estimated such alteration for each gene by the subtraction of its averaged essential scores over the samples in the cancer/control samples. Then, the 500 most essentiality altered genes for each cancer types were intersected and got 189 common genes.

Then, variable combination population analysis (VCPA) was applied to select the informative genes in each cancer types. Referring to the original paper for the VCPA, we ran the method fifty times for each cancer type and then merged all the output gene lists for further informative gene selection. Here, the ACC value wsa taken instead of the RMSECV (root mean squares error of cross validation) for estimating the performance of a feature set in this classification task. Additionally, the optimized gene set was determined by ACC values from the cross-validation for the gene set in each EDF (exponentially decreasing function) run. For more details about the VCPA algorithm, refer to [29]. The details about the modified pipeline please see Supplementary Fig. 1.

Further, we analyzed the classification ability of the genes of different frequency to classify cancer/control samples. The PLS classification models by using the genes of frequency greater than 65 to 88 were built on each cancer datasets. For each PLS model, both the threshold and the number of latent variables were optimized. The prediction performance was evaluated by using the accuracy (ACC), the area under the ROC curve (AUC), specificity (SPE) and sensitivity (SEN).

To validate the superiority of VCPA feature selection algorithm, the 6 feature importance evaluators were applied on the 189 common genes to select the informative genes in each cancer types, including ChiSquaredAttributeEval, ReliefFAttributeEval, InfoGainAttributeEval, GainRatioAttributeEval, OneRAttributeEval, SymmetricalUncertAttributeEval from Waikato Environment for Knowledge Analysis (WEKA) [32]. We constructed the PLS classification methods by using the same number of the top ranked feature genes to evaluate the classification ability for VCPA feature selection algorithm and the 6 feature importance evaluators from WEKA. The prediction performance was evaluated by using the accuracy (ACC), the area under the ROC curve (AUC), specificity (SPE) and sensitivity (SEN).

Table 1
The six cancer datasets used in our analysis.

Dataset	Cancer type	Abbreviation	Dataset type	Normalized method	Tumor samples	Control samples	Gene number
GSE5847	breast	BRC	microarray	RMA	48	47	13435
GSE21510	colorectal	CRC	microarray	RMA	124	24	21212
GSE27342	gastric	GAC	microarray	RMA	80	80	17492
GSE19804	lung	NSCLC	microarray	RMA	60	60	21212
GSE77314	liver	LIC	RNA-seq	log2	50	50	29095
GSE28735	pancreatic	PDAC	microarray	RMA	45	45	23306