Analytica Chimica Acta xxx (2017) 1–8

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Optimal preprocessing of serum and urine metabolomic data fusion for staging prostate cancer through design of experiment

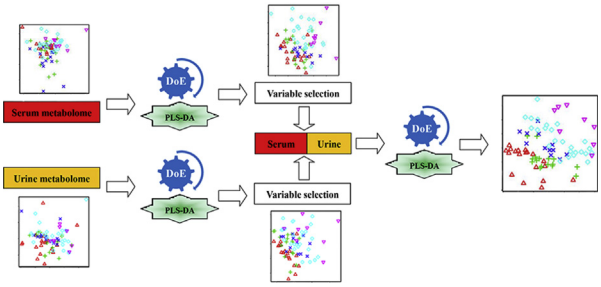
Hong Zheng [a], Aimin Cai [a], Qi Zhou [a], Pengtao Xu [a], Liangcai Zhao [a], Chen Li [a], Baijun Dong [b], Hongchang Gao [a, *]

[a] Institute of Metabonomics & Medical NMR, School of Pharmaceutical Science, Wenzhou Medical University, Wenzhou 325035, China
[b] Department of Urology, Renji Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai 200127, China

HIGHLIGHTS

- NMR metabolomic analysis of body fluids can be used for staging prostate cancer.
- Data preprocessing is an essential step for metabolomic analysis.
- Data fusion improves information recovery for cancer classification.
- Design of experiment achieves optimal preprocessing of metabolomic data fusion.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 9 March 2017
Received in revised form
17 July 2017
Accepted 8 September 2017
Available online xxx

Keywords:
Cancer
Data fusion
Precision medicine
Preprocessing
Metabolomics

ABSTRACT

Accurate classification of cancer stages will achieve precision treatment for cancer. Metabolomics presents biological phenotypes at the metabolite level and holds a great potential for cancer classification. Since metabolomic data can be obtained from different samples or analytical techniques, data fusion has been applied to improve classification accuracy. Data preprocessing is an essential step during metabolomic data analysis. Therefore, we developed an innovative optimization method to select a proper data preprocessing strategy for metabolomic data fusion using a design of experiment approach for improving the classification of prostate cancer (PCa) stages. In this study, urine and serum samples were collected from participants at five phases of PCa and analyzed using a ^{1}H NMR-based metabolomic approach. Partial least squares-discriminant analysis (PLS-DA) was used as a classification model and its performance was assessed by goodness of fit (R^2) and predictive ability (Q^2). Results show that data preprocessing significantly affect classification performance and depends on data properties. Using the fused metabolomic data from urine and serum, PLS-DA model with the optimal data preprocessing ($R^2 = 0.729$, $Q^2 = 0.504$, $P < 0.0001$) can effectively improve model performance and achieve a better classification result for PCa stages as compared with that without data preprocessing ($R^2 = 0.139$, $Q^2 = 0.006$, $P = 0.450$). Therefore, we propose that metabolomic data fusion integrated with an optimal data preprocessing strategy can significantly improve the classification of cancer stages for precision treatment.

1. Introduction

Prostate cancer (PCa) is the second most common cancer in men and the fifth leading cause of cancer death in the world [1]. PCa is

* Corresponding author.
E-mail address: gaohc27@wmu.edu.cn (H. Gao).

http://dx.doi.org/10.1016/j.aca.2017.09.019
0003-2670/

more frequently diagnosed in developed countries, but now its incidence and mortality rates are also rising in developing countries due to diet and lifestyle changes [2]. Currently, PCa diagnosis is mainly based on biopsy Gleason grade [3] and serum prostate-specific antigen (PSA) level [4]. However, these two methods are always inconsistent for individual patients, for example, Walsh reported that as many as 15% of PCa patients had a 'normal' serum PSA level [5]. The misdiagnosis of PCa will result in the over-treatment or under-treatment of patients and thereby cause an increased mortality rate. Thus, improving the classification accuracy of PCa stages is of great importance for PCa diagnosis and treatment.

Precision medicine attempts to treat individual patients differently according to genetic, biomarker or phenotypic characteristic and its near-term purpose will focus on cancers [6]. Omics-based methods are able to provide biological phenotypes at omics levels (gene, protein or metabolite) and have shown a great potential in precision oncology [7]. For example, genomic profiling has been successfully used to classify PCa stages and further improves therapeutic stratification [8,9]. Urinary genomic biomarkers can improve PCa detection, although this method still need to be further developed for clinical practice [10]. In addition, Petricoin et al. [11] reported that serum proteomics may be a promising tool to decide whether prostate biopsy is needed for a man with an increased PSA level. Serum proteomics has also been used for the early detection of PCa [12]. The potential of urine proteomics was shown in the diagnosis and management of PCa [13]. Metabolomics aims to analyze all low-molecular weight metabolites in a biological organism, reflecting events downstream of gene expression [14]. Relative to genomics and proteomics, therefore, metabolomics is closer to the actual phenotype. Moreover, metabolomics possesses its advantages, such as simple sample preparation, rapid detection and relatively low cost. Zhang et al. [15] revealed that the diagnostic ability of PCa using urinary metabolomic biomarkers was close to the PSA test. Furthermore, a panel of 40 metabolic features in serum can detect PCa with an accuracy of 93.0%, which is higher than the PSA test [16]. However, omics-based methods still need to be further developed and validated in order to be useful tools in clinical practice.

For omics-based methods, the development of big data analytic approaches would facilitate medical researchers to draw an accurate inference and to make a precise clinical treatment plan for each patient. Data fusion as a very commonly used method in omics data analysis is the process of combining multiple sources of data to achieve a better inference than using a single data source [17]. Yet, selecting an appropriate data preprocessing method is an essential step during omics data analysis and greatly affects the final results [18,19]. Additionally, data preprocessing depends on data properties and no single method can be generally used [20]. Taken together, prior to data fusion, different data preprocessing methods should be considered for different sources of data. We suggest an optimization method to solve this issue. Design of experiment (DoE) is applied to identify significant influencing factors and optimize these factors for reaching the desired outcome with a minimal number of experiments with a statistical certainty [21]. In our previous study, we used a time-saving DoE approach to optimize software parameter setting for improving the processing of LC-MS-based metabolomic data [21]. DoE method has also been applied to select optimum calibration model parameters [22]. Moreover, Gerretzen et al. [23] developed an optimal preprocessing strategy for chemometric data analysis using a DoE approach. In the present study, therefore, we aimed to develop an innovative optimization method to select an appropriate data preprocessing strategy for serum and urine metabolomic data fusion using a DoE approach in order to improve the classification of PCa stages.

2. Materials and methods

2.1. Clinical sample collection

In total, we selected 80 participants from the Renji Hospital of Shanghai Jiao Tong University, including 19 benign prostatic hyperplasia (BPH), 16 early PCa, 12 advanced PCa, 25 metastatic PCa and 8 castration-resistant PCa (CRPC). PCa was diagnosed and graded through the pathological examination and PSA level. Detailed pathological and clinical data for participants are provided in Table S1. Participants were fasted for 12 h and then had blood drawn from the antecubital vein (approximately 5 mL). Serum samples were separated using centrifugation at 1024 g for 10 min at 4 °C. In addition, urine samples were collected from the first morning urine before the breakfast. All samples were frozen immediately after collection and stored in −80 °C until NMR metabolomic analysis. This study was approved by the Ethics Committee of Shanghai Jiao Tong University School of Medicine.

2.2. NMR-based metabolomics analysis

^{1}H NMR spectra were measured at 600.13 MHz on a Bruker AVANCE III 600 MHz NMR spectrometer with a 5-mm TXI probe (Bruker BioSpin, Rheinstetten, Germany). Urine samples were thawed and vortexed, and then 400 μL of the sample was diluted with 100 μL of phosphate buffer (0.2 mM Na_2HPO_4/NaH_2PO_4, pH = 7.4) containing 0.5% sodium trimethlysilyl propionate-d_4 (TSP) in D_2O. The diluted urine sample was mixed by vortex and centrifuged at 10,000 g for 15 min at 4 °C to remove insoluble substance. Then, 500 μL of supernatant was transferred into a 5 mm NMR tube for NMR analysis. NMR spectra were acquired by a standard single-pulse experiment "ZGPR" with pre-saturation of the water resonance at 25 °C. The main acquisition parameters included: data points, 256 K; relaxation delay, 4 s; spectral width, 10,822.5 Hz; acquisition time, 3.03 s per scan. Serum sample was also thawed and vortexed, and then 200 μL of the sample was diluted with 250 μL of phosphate buffer (0.2 mM Na_2HPO_4/NaH_2PO_4, pH = 7.4) and 50 μL of D_2O. The diluted serum sample was mixed by vortex and centrifuged at 10,000 g for 15 min at 4 °C. Afterward, 500 μL of supernatant was transferred into a 5 mm NMR tube for NMR analysis. To minimize the line-broadening effect of macromolecules such as proteins and lipids, the Carr-Purcell-Meiboom-Gill "CPMG" pulse sequence with a fixed receiver-gain value was used to acquire NMR spectra of serum samples at 37 °C. Moreover, the main acquisition parameters were set as follows: data points, 256 K; relaxation delay, 4 s; spectral width, 12,335.5 Hz; acquisition time, 2.66 s per scan.

All NMR spectra were automatically preprocessed by phase and baseline corrections in the Topspin 3.0 software (Bruker BioSpin, Rheinstetten, Germany). The spectra of urine were referenced to TSP peak at 0 ppm, while the spectra of serum were referenced to the anomeric signal of α-glucose at 5.23 ppm. The "*icoshift*" method was applied to align NMR spectra in MATLAB (R2012a, The Mathworks Inc., Natick, MA, USA) [24]. The spectral region from 0.0 to 9.0 ppm excluding the residual water signals from 4.7 to 5.0 ppm for urine and serum samples was subdivided and integrated to binning data with a size of 0.01 ppm for further analysis.

Typical ^{1}H NMR spectra obtained from human serum and urine were illustrates in Fig. 1A and B, respectively. Metabolites were assigned using Chenomx NMR suite 7.7.2 software (Chenomx Inc., Alberta, Canada) as well as our previous data [25,26].

2.3. Optimization of metabolomic data fusion

The procedure of optimal preprocessing for metabolomic data

Please cite this article in press as: H. Zheng, et al., Optimal preprocessing of serum and urine metabolomic data fusion for staging prostate cancer through design of experiment, Analytica Chimica Acta (2017), http://dx.doi.org/10.1016/j.aca.2017.09.019