Journal of Microbiological Methods 98 (2014) 26–30

Contents lists available at ScienceDirect

Journal of Microbiological Methods

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

Feature selection strategies for identification of *Staphylococcus aureus* recovered in blood cultures using FT-IR spectroscopy successive projections algorithm for variable selection: A case study

Aline de Sousa Marques [a], Maria Celeste Nunes de Melo [b], Thiago André Cidral [b], Kássio Michell Gomes de Lima [a,*]

[a] Universidade Federal do Rio Grande do Norte, Instituto de Química, Programa de Pós-Graduação em Química, Grupo de Pesquisa em Química Biológica e Quimiometria, CEP 59072-970 Natal, RN, Brazil
[b] Departamento de Microbiologia e Parasitologia, UFRN, Natal 59072-970, Brazil

ARTICLE INFO

Article history:
Received 17 November 2013
Received in revised form 19 December 2013
Accepted 21 December 2013
Available online 31 December 2013

Keywords:
Blood cultures
ATR-FTIR spectroscopy
S. aureus
LDA
GA–LDA
SPA–LDA

ABSTRACT

Staphylococcus aureus is one of the leading causes of bacteremia, with high levels of accompanying morbidity and mortality. Current gold standard for the detection of *S. aureus* is very time-consuming, typically taking 24 h or longer. We set out to determine whether Fourier-transform infrared spectroscopy (FT-IR) combined with variable selection techniques, such as, genetic algorithm–linear discriminant analysis (GA–LDA) and successive projection algorithm–linear discriminant analysis (SPA–LDA) could be applied to detect this pathogen of bloodstream infection in samples based on the unique spectral "fingerprints" of their biochemical composition. Thirty real blood samples from healthy volunteers were contaminated with five different concentrations (10^7 until 10^3 CFU/mL) of microorganism and it analyzed by IR spectroscopy. The resulting GA–LDA model successfully classified all test samples with respect to their concentration in contaminated blood using only 18 wavenumbers. Discriminant functions revealed that GA–LDA clearly segregated different microorganism concentrations and the variable selected confirmed the chemical entities associated with the microorganism. The current study indicates that IR spectroscopy with feature selection techniques have the potential to provide one rapid approach for whole-organism fingerprint diagnostic microbial directly in blood culture.

1. Introduction

Bloodstream infections (BSIs) represent an important cause of human morbidity and mortality accounting for 30–40% of all cases of severe sepsis and septic shock (Doern et al., 1994). Diagnostic assays for identification of microorganisms and antimicrobial resistance determinants directly from positive blood culture broth are reported. (Pence et al., 2013). Prompt detection of microorganisms circulating in the bloodstream of patients is imperative as it allows clinicians to make decisions on possible therapeutic interventions (Dellinger et al., 2008; Weinstein and Doern, 2011). Automated blood culture systems are the most sensitive approach for detection of the bacteremia causative agent. There are some automated blood culture systems commercially available, such as BACTEC FX (BD, Franklin Lakes, NJ, USA) and VersaTREK (ThermoFisher Scientific, Waltham, MA, USA). However, this procedure typically takes 24 h (e.g., for *Staphylococcus aureus* up to 5 days for *Candida* species) to generate the results (Pence et al., 2013). Moreover none of the currently available molecular methods is sufficiently rapid and accurate (Pence et al., 2013; Waterer and Wunderink, 2001). Because time is essential in preventing the evolution of BSI to severe sepsis or septic shock, faster detection methods are needed.

In the past 10 years, molecular techniques have been explored as tools for the identification of microbial species and subspecies (Wenning and Scherer, 2013). In particular, attenuated total reflection Fourier-transform infrared spectroscopy (ATR-FTIR) can be utilized to determine the microorganism metabolic fingerprint (DNA, RNA, proteins, cell-wall components), emerging as an interesting alternative for a rapid and cost-effective identification of microorganisms (Riding et al., 2012). ATR-FTIR is also characterized by a minimum of sample handling. It requires no extractions and is non-destructive. Moreover, amplifications, labeling, or staining steps of any kind are needless. The metabolic fingerprint generated by ATR-FTIR spectroscopy reflects the balance of some factors such as compositional and quantitative differences of biochemical compounds in microbial cells (Martin et al., 2010).

The application of IR spectrometry for *S. aureus* microorganism analysis (Amiali et al., 2011; Grunert et al., 2013; Maquelin et al., 2003) has been a target in the past years. Amiali et al. (2011) determined an FTIR spectral region or combination of regions reflecting a specific biochemical feature of a community-associated methicillin-resistant *S. aureus* (CA-MRSA). The authors provided a substitute for descriptive epidemiology in the definition of CA-MRSA strain types. Grunert et al. (2013) studied the potentiality of FTIR spectroscopy for

* Corresponding author. Tel.: +55 84 3215 3828; fax: +55 83 3211 9224.
E-mail address: kassio@ufrnet.br (K.M.G. de Lima).

0167-7012/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.mimet.2013.12.015

differential diagnostic of the most clinically relevant *S. aureus* capsular polysaccharide types. Maquelin et al. (2003) realized a first prospective clinical study in which the causative pathogens (*Staphylococcus aureus*, *Enterococcus faecalis*, *Escherichia coli* and *Pseudomonas aeruginosa*) of blood infections were identified by FTIR spectroscopy.

For the analysis of *S. aureus* bacteria with IR spectroscopy, most of the reports are based on principal component analysis (PCA) for initial data reduction (de Sousa Marques et al., 2013), hierarchical cluster analysis (HCA) for analyzed groups in a set of data on the basis of spectral similarities (Martin et al., 2011), and linear discriminant analysis (LDA) for classify unknown samples into predetermined groups (Cheung et al., 2011). However, when employing full spectrum in the construction of these mathematic models, many variables are redundant and/or non-informative, and their inclusion may affect the performance of the final model. A well-succeeded approach to overcome this drawback is the successive projections algorithm (SPA) (Pontes et al., 2005) in conjunction with LDA and genetic algorithm (GA) (Tapp et al., 2003).

The present paper proposes the determination of an FTIR spectral region, or combination of variables, that reflects a specific biochemical feature of *S. aureus* in blood samples. We employed SPA and GA to select an appropriate subset of wavenumbers for LDA. Other goals were the elucidation of the altered variables using different concentrations of bacteria in the blood and the identification of the altered biochemical-bacteria fingerprint. This novel approach envisions rapid microbial identifications in clinical diagnostic assays.

2. Material and methods

2.1. Bacterial strain

Strain of *S. aureus* ATCC 29213 was cultivated in 2 mL of Brain Heart Infusion (BHI) broth (BHI, Oxoid, Ltd., Basingstoke, Hampshire, England) for 24 h at 35 °C. A microbiological strain suspension was standardized to 0.5 McFarland scale (~10^8 CFU/mL) in sterile saline.

2.2. Sample preparation

For IR measurements blood samples from healthy volunteers were contaminated with *S. aureus* in a microwell plate at five dilutions (1×10^7, 1×10^6, 1×10^5, 1×10^4 and 1×10^3 CFU/mL). The data set consisted of 36 samples that were divided into five for each dilution (30 samples, bacteria group) and six for the control group (uncontaminated blood).

2.3. ATR-FTIR spectroscopy

ATR-FTIR spectroscopy was performed using a Bruker ALPHA FT-IR spectrometer equipped with an ATR accessory. Spectra (8 cm^{-1} spectral resolution giving 4 cm^{-1} data spacing equivalent to 258 wavenumbers, co added for 32 scans) were converted into absorbance by Bruker OPUS software. The time measurement was of 26 s (32 scans) per spectrum. Absorbance spectra of bacterial samples were obtained against the spectrum of sterile blood used as background. Immediately following collection of each background, approximately 0.1 mL of each sample was applied to the ATR crystal using a transfer pipet, ensuring that no air bubbles were trapped on the crystal surface. After each measurement the ATR plate was washed with ethanol (70% v/v) and dried using tissue paper. Cleanliness of the ATR plate was verified by collecting an absorbance spectrum of the crystal using the most recently collected background as a reference. Before and between spectral acquisitions, samples were stored in the dark at ambient temperature. The ATR-FTIR spectrometer was placed in an air-conditioned room (21 °C), and samples were allowed to equilibrate to this temperature before analysis.

2.4. Chemometric methods: PCA, LDA, SPA–LDA and GA–LDA

A data set with many variables can be simplified by performing data reduction which makes the system more easily interpretable. Principal component analysis (PCA) is a well-known way to reduce the number of variables, in which the spectral matrix X is decomposed as:

$$\mathrm{X} = \mathrm{TP}^{\mathrm{t}} + \mathrm{E} \tag{1}$$

where X is the $I \times J$ data matrix, T is the $I \times A$ matrix of score vectors, the score vectors $\mathrm{t_a}$ are orthogonal (i.e., $\mathrm{T^tT} = \mathrm{diag}(\lambda_\mathrm{a})$ and λ_a are eigenvalues of the matrix $\mathrm{X^tX}$), P is the $J \times A$ matrix of loadings vectors, E is the $I \times J$ residual matrix, I is the number of objects, J is the number of variables, and A is the number of calculated components.

LDA is a supervised linear transformation that projects the variables (wavenumbers, for example) into a variable-reduced space which is optimal for discrimination between treatment classes. An LDA seeks for a projection matrix such that Fisher criterion (i.e. the ratio of the between-variance scatter to the within-class variance) is maximized after the projection. The variables created through LDA (factors) are linear combinations of the wavenumber-absorbance intensity values (Martin et al., 2007). Thus, the use of LDA for identification or classification of spectral data generally requires appropriate variable selection procedures (Pontes et al., 2005; Silva et al., 2013). In the present study, the SPA and GA were adopted for this function. In the SPA–LDA and GA–LDA models, the validation set was used to guide the variable selection, a strategy to avoid overfitting. The optimum number of variables for SPA–LDA and GA–LDA was determinate from the minimum of the cost function G calculated for a given validation data set as:

$$G = \frac{1}{N_V}\sum_{n=1}^{N_V} g_n, \tag{2}$$

where g_n is defined as

$$g_n = \frac{r^2\left(x_n, m_{I(n)}\right)}{\min_{I(m)\neq I(n)} r^2\left(x_n, m_{I(m)}\right)} \tag{3}$$

where $I(n)$ is the index of the true class for the nth validation object x_n.

In GA–LDA model, the mutation and reproduction probabilities were kept constant, 10 and 80%, respectively. The initial population was 120 individuals, with 50 generations. The best solution resulting from the three realizations of the GA was kept.

For this study, LDA scores, loading, and discriminant function (DF) values were derived for the biochemical-bacteria fingerprint region. The first LDA factor (LD1) was used to visualize the alterations of the blood sample in 1-dimensional (D) score plots that represented the main chemical alterations. SPA–LDA and GA–LDA were used to detect the biochemical alterations relative to the corresponding vehicle control (uncontaminated blood).

2.5. Software

The data import, pre-treatment, and the construction of chemometric classification models (LDA, SPA–LDA and GA–LDA) were implemented in the MATLAB version 6.5 (Math-Works, Natick, USA). Different preprocessing methods were used, including the baseline correction, derivative, smoothing Savitzky–Golay methods by using a first and second-order polynomial, and varying the number of window points (3, 5, 7 and 15). For SPA–LDA and GA–LDA models, the samples were divided into training (25), validation (6) and test sets (5) by applying the classic Kennard-Stone (KS) uniform sampling algorithm (Kennard and Stone, 1969) to the IR spectra.