Chemometrics and Intelligent Laboratory Systems 142 (2015) 219–230

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Quality assessment of boar semen by multivariate analysis of flow cytometric data

Hamid Babamoradi[a,*], José Manuel Amigo[a], Frans van den Berg[a], Morten Rønn Petersen[b], Nana Satake[c], Gry Boe-Hansen[c]

[a] Spectroscopy & Chemometrics, Dept. of Food Science, University of Copenhagen, DK-1958 Frederiksberg, Denmark
[b] The Fertility Clinic, University Hospital of Copenhagen, Copenhagen, Denmark
[c] School of Veterinary Science, University of Queensland, Gatton, Queensland 4343, Australia

ARTICLE INFO

Article history:
Received 6 October 2014
Received in revised form 26 January 2015
Accepted 8 February 2015
Available online 14 February 2015

Keywords:
Flow cytometry
k-Means
OPTICS
ASCA
Normalization
Boar semen

ABSTRACT

Flow cytometry (FCM) has become very powerful over the last decades, enabling multi-parametric measurements of up to thousands of cells per second. This generates massive amounts of data on individual cell characteristics that need to be analyzed in an efficient manner from both physiological and chemical points of view. In this study, a methodology of analysis for FCM data was comprehensively studied to assess quality changes in semen extracted from boars. The proposed methodology combines new automated multi-dimensional data normalization, a density-based clustering method for identification of cell populations, and multivariate methods for post-analysis of the identified populations, enabling the exploratory evaluation and prediction/classification of subpopulations within the experimental data set. The performance of the suggested methodology was compared with the performance of an existing automated clustering method.

1. Introduction

Analytical flow cytometry (FCM) has been widely used in a wide range of applications such as cell function [1], cell death [2], DNA and RNA analysis [3], and immunophenotyping [4]. FCM is a fluorescence-based technique that analyzes multiple parameters on a single cell level or within a population of cells, enabling physical and chemical characterization of up to thousands of cells per second. The technique is used extensively, both in research and routine settings, generating a vast amount of data for a single experiment that need to be analyzed efficiently and effectively in order to utilize the systems to their full potential.

FCM data analysis is usually performed in two steps: first, identification of cell populations; and second, post analysis of data extracted from the identified cell populations.

The first step is related to the identification of cell populations (i.e., gating). Automated gating methods have been extensively studied and many have been proposed [5]. Nevertheless, every automated method must be adapted to the different kind of FCM data and, therefore, manual gating is extensively used. Manual gating requires operator decision making to determine cell populations, having disadvantages with non-generalizable solutions (i.e., time-consuming and subjective, often providing erroneous and irreproducible results) [6,7]. On the other hand, semi-automated methods that are based on mathematical and statistical approaches are preferred since manual gating is minimized without losing accuracy.

The second step is directly linked to the analysis of the detected cell populations. For this analysis, one- and two-way analysis of variance (ANOVA) is used [8–11]. ANOVA assumes that the data extracted from the populations (represented by variables) are independent. Nevertheless, in most cases there are only a few sources of variation among populations and each variable or attribute extracted from the populations is simply a different reflection of these few sources of variation [6,12–16]. Therefore, this perspective makes the univariate analysis of the extracted data and thus the interpretation extremely cumbersome. The multivariate philosophy though aims at extracting information on the principal direction of the variations, improving the signal to noise ratio and highlighting the main sources of variance in the collected data [17]. Interpretation of such few main sources of variation is much easier and enables a better understanding of the underlying biology. This is only possible by employing multivariate analysis.

From a biological perspective, semen in its natural state is a heterogeneous cell suspension consisting of different sub-populations of spermatozoa with variable attributes [18]. Hence, sperm function analysis by FCM methodology with commercially available fluorescent probes has been extensively used [19]. Data analysis by objective clustering methods has been previously used to examine other sperm attributes

* Corresponding author. Tel.: +45 35333500; fax: +45 35333245.
E-mail address: hamba@life.ku.dk (H. Babamoradi).

http://dx.doi.org/10.1016/j.chemolab.2015.02.008
0169-7439/

that does not permit FCM assessment. Data obtained through computer assisted semen analysis (CASA) technique assess multiple kinematic measures of sperm motion [20,21], traditionally tested on averages of the whole population. However, multi-parameter CASA data permit cluster analysis to be used to enhance statistical analysis due to the heterogeneity of the sperm motility within an ejaculate [22–24]. Therefore, an objective approach to analyze spermatozoa derived FCM data would help to understand population dynamics within an ejaculate and describe differences in subpopulations within and between individuals. These differences are important in determination of the semen quality among ejaculates. Semen quality is a contributing factor to fertility that can be used in evaluation of boar fertility.

This study examines clustering methods for automated data preprocessing and semi-automated identification of cell populations for automated gating, and multivariate analytical methods for the study of the detected cell populations. For this purpose, an experimentally controlled FCM dataset from boar semen was used in order to assess the seasonal and within-boar variation of the quality of the semen. The aim of the analysis is to study the effect of breed type and seasonal change on the quality of boar semen. The dataset consisted of flow cytometric measurements of semen ejaculates collected from five boars over five periods of time (25 ejaculates), where four attributes were measured for each ejaculate. This dataset was chosen for this study to introduce a comprehensive workflow for processing large and complex data obtained from flow cytometric systems, where different challenges are addressed. The results of clustering were compared between the proposed clustering method and a completely automated method.

2. Semen sample collection and data acquisition

Ejaculates of mature Large White boars (n = 5; Premier Pig Genetics, Wacol Pig AB Centre, Australia) were collected using the gloved-hand technique (Hancock 1967), extended in a commercial semen extender (1:1 semen to extender), Androstar (Minitube), and transported to the laboratory and stored at ambient temperature. Ejaculates were collected from the same boars over five dates over a 4 month period (06/12, 27/12, 24/01, 21/02, 14/03). Flow cytometric analysis was carried out using a Beckman Coulter Gallios™ flow cytometer (Beckman Coulter PTYLTD, Gladesville, Australia) on the extended semen samples on arrival to the laboratory for the following fluorescent attributes: Viability (PI/SYBR14), acrosome integrity (PI/FITC-PNA), membrane fluidity (YoPro-1/M540), and reactive oxygen species (ROS) production (PI/H2DCFDA). A prototypical mature ejaculated spermatozoa population would for these attributes have the following functional attributes; viable (SYBR14 positive), intact acrosome (FITC-PNA negative), impervious membrane fluidity and non-apoptotic; (YoPro-1 and M540 negative) and low in ROS production (H2DCFDA negative).

Sperm suspensions were loaded with fluorescent probes for flow cytometric analysis, performed on a Gallios™ flow cytometer (Beckman Coulter Pty Ltd, Gladesville, Australia). 10,000 events were assessed per semen sample, at the rate of 50–200 events per second. Forward (FS) and side (SS) light scatter in linear mode was used. All fluorescent probes utilized within this study were excited using a 15 mW argon ion solid state laser at 488 nm and the fluorescence emissions of stained spermatozoa were analyzed in logarithmic mode using appropriate detectors for the emission spectra of the fluorescent probes. Dual staining for 10 min at 38 °C, were carried out in order to assess the following sperm functions: Viability using 100 nM of SYBR14 (em_{max} 516 nm) and 12 μM of propidium iodide (PI; em_{max} 625 nm); acrosome integrity using 5 μg/ml of FITC-PNA (em_{max} 530 nm) and 12 μM of PI; membrane fluidity using 25 nM of merocyanine 540 (M540, em_{max} 590 nm) and 25 nM of YoPro-1 (em_{max} 509 nm); and ROS production using 1 μM of carboxyfluoresceindiacetate (H_2DCFDA, em_{max} 525 nm) and 12 μM of PI. Fluorescence emission of FITC-PNA, Yo Pro-1, and H_2DCFDA were detected using bandpass filter 525 nm, FL1; emission of PI was detected on bandpass filter 620/30 nm, FL3; and emission of M540 was detected on bandpass filter 575/30 nm and 620/30 nm, FL2 and FL3 respectively. The results were transformed into listmode files (.lmd) using Kaluza Analysis Software (v1.2 Beckman Coulter Australia Pty Ltd, Gladesville, Australia). The MATLAB code used for OPTICS algorithm is available at http://chemometria.us.edu.pl/ while ASCA was applied using Eigenvector's PLS-toolbox for MATLAB. The source code for SWIFT can be downloaded from: http://www.ece.rochester.edu/projects/siplab/Software/SWIFT.html.

The semen samples were collected between December 2010 and March 2011, and approved by the University of Queensland Animal Ethics Committee; SVS/555/09/PORK CRC.

The data contains 100 measurements, 25 measurements per attribute. A measurement is recordings of scatter and fluorescence for 10,000 cells per ejaculate.

3. Identification of cell population

3.1. Automated multi-channel data normalization

FCM datasets may typically be affected from small changes in the fluorescence signal across different measurements, i.e., the positions of the same cell population may shift between measurements [25]. This issue hinders an accurate data analysis since measurements are analyzed simultaneously and the same population in two measurements might be incorrectly assigned to two different cell categories due to instrumental differences. Therefore, the data need to be preprocessed in order to minimize the shifts (technical variations). This is even more crucial when automated methods are utilized to analyze FCM data since they cannot distinguish between biological and technical variations. *Normalization* [25], *Standardization* [26], *quality assessment* and *control* [27,28] are used for referring the procedures for minimizing technical variations in FCM data. *Normalization* will be used here since it is the specific term for minimizing technical variations across measurements.

The most straightforward way of normalization is to adjust the data for each channel (fluorochrome) independently [25]. This strategy does not take into consideration the correlation between channels in minimizing the shifts, which might be acceptable when data of each channel is analyzed also independently. However, most FCM analysis procedures use several channels simultaneously. It is thus reasonable to take this correlation also into account during data normalization, assuring that no artifacts are introduced. Typically, two channels are analyzed simultaneously [29] where the data is visualized through scatter plots to identify cell populations and perform a two-dimensional normalization to minimize the shifts. In this work multi-dimensional normalization is proposed using a combination of *k*-means clustering [30] and Procrustes analysis [31]. The algorithm estimates the cluster centers once for the entire dataset to define target centers. Then, it estimates the cluster centers for each measurement using target centers as an initial estimate for the *k*-means algorithm. It calculates Procrustes analysis components – including a translation, orthogonal rotation, and scaling operation (no reflection) – to match the centers of each measurement to the target centers. For each measurement, the components are applied to the data of all cells to normalize the data towards the target centers. The inputs for the algorithm are the entire dataset and number of clusters (k). The algorithm works as follows:

1. Compute centers of k clusters as target centers using entire measurements (input is measurements concatenated across the row direction — a matrix of size 25 × 10,000 by number of channels)
2. Use the target centers as the initial estimate for *k*-means clustering to compute centers of k clusters for a measurement (input is a matrix of size 10,000 by number of channels)
3. Translate the centers towards the target centers
4. Orthogonally rotate the centers to match the target centers
5. Scale the centers so that their Frobenius norm matches the norm of the target centers