Analytical Methods

PAPER

View Article Online
View Journal | View Issue

Cite this: *Anal. Methods*, 2019, **11**, 1868

Informatics analysis of capillary electropherograms of autologously doped and undoped blood†

Shiladitya Chatterjee,[a] Sean C. Chapman,[a] George H. Major,[a] Denis L. Eggett,[a] Barry M. Lunt,[a] Christopher R. Harrison[b] and Matthew R. Linford *[a]

An 'Autologous Blood Transfusion' (ABT) is the reinjection of blood previously taken from an athlete to increase its oxygen transport capabilities. Despite the World Anti-Doping Agency's ban on such practices, ABT abuse continues. Autologous blood doping (ABD) is challenging to detect because of the similarities between an individual's doped and undoped blood. Recently, Harrison *et al.* reported that high-speed capillary electrophoresis may identify ABD. In their work, first order derivatives of the electropherograms were used to identify doping. However, this method suffered from false negatives due to the subjective nature of the analysis. Here, we provide an informatics analysis of the data from this study, contrasting the results of traditional statistical methods and less traditional mathematical techniques. First, three well-known multivariate statistical tools: cluster analysis, principal component analysis (PCA), and partial least squares (PLS) are applied to develop calibrations and/or group electropherograms of undoped (0%) and doped (5% and 10%) blood samples. (These doping levels were chosen due to the low physiological effect of doping below 5%, with 10% corresponding to the approximate 'gain' derived from the transfusion of a single unit of blood into an adult.) Different preprocessing and variable selection methods were considered. Due to variation in the electropherograms and the limited sample size, these methods were inadequate. We next considered four less commonly used mathematical/informatics tools: pattern recognition entropy (PRE), the Euclidean distance between vectors, a peak fitting/integration method, and the second moment (SM). Each of these techniques showed some ability to differentiate between the 0, 5, and 10% doped samples. We then evaluated the prediction capabilities of inverse least squares (ILS) models based on these summary statistics. An ILS calibration based on PRE, the Euclidean distance, and peak fitting/integration proved more successful than the PLS model at predicting levels of blood doping from the corresponding electropherograms; the ILS model distinguished between doped (5% and 10%) and undoped (0%) blood. This methodology may be applicable to other challenging informatics problems like determining risk factors for genetically linked diseases, robust pattern finding in peak-like data such as ChIP-seq, or other genomic sequencing for understanding the 3D genome.

Received 24th January 2019
Accepted 22nd February 2019

DOI: 10.1039/c9ay00192a

rsc.li/methods

1. Introduction

Unethical methods for increasing oxygen delivery to skeletal muscle have been in existence for the last four decades despite a ban on such activity by the International Olympic Committee in the mid-1980s.[4] Indeed, according to a World Anti-Doping Agency report,[5] introduction of any quantity of autologous, homologous, or heterologous blood or red blood cells (RBCs) into the circulatory system constitutes doping. Of these doping methods, the detection of autologous blood transfusions (ABTs), *i.e.*, autologous blood doping (ABD), is the most challenging.[6] In an ABT, transfused RBCs are taken from the athlete and stored for reinfusion at a later date. Currently, ABD cannot be directly detected by regular anti-doping tests. Most anti-doping agencies rely on indirect methods, the most common of which consists of maintaining an athlete's 'biological passport'.[7] ABD alters the characteristic biomarkers associated with erythropoiesis (red blood cell production). Thus, the observation of the hematological module and the monitoring of specific biomarkers allows for the detection of ABD. However, biological passport based fingerprinting of every athlete's hematological profile is expensive and time consuming.

Recently, Harrison *et al.* introduced a fast (*ca.* 3 min), direct capillary electrophoresis (CE) based method to detect ABD.[1] This approach relies on a decrease in the zeta potential of stored RBCs, which impacts their mobility. The aging of the blood results in significant rheological changes in the RBCs, particularly a decrease in surface area and volume.[8,9] Harrison's work

[a]*Brigham Young University, Provo, Utah 84602, USA. E-mail: mrlinford@chem.byu.edu*

[b]*San Diego State University, San Diego, CA 92182-1030, USA*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c9ay00192a

demonstrated the ability of CE to respond to changes in RBC distributions, *i.e.*, ABD resulted in changes to the RBC peak envelope, indicating the presence of aged RBCs. Fig. 1 shows the raw data from their study, which included undoped (0%) and simulated (5% and 10%) doped samples from three individuals/subjects: A, B, and C. Each electropherogram consists of a sharp peak at earlier time (*ca.* 1.5 min) followed by a shoulder at longer times (*ca.* 1.8–2.7 min), where the length and height of the shoulder tend to increase with increasing doping levels (see Fig. 1d). Harrison *et al.* presented a first derivative of the data as a mathematical tool for quantifying this difference. Doping was identified by the presence of positive slopes. However, this approach was subjective, where a lack of a clear figure of merit for this approach resulted in false negatives.

The electropherograms in the Harrison study exhibited a substantial amount of variability and complexity, while still showing features that were consistent with doping.[1] Forexample, the initial sharp peak in the electropherograms of the samples from subjects A and C elute close together and with similar standard deviations: 1.47 ± 0.02 min and 1.48 ± 0.02 min, respectively (see Fig. 1a and c). However, this initial sharp peak varies in both shape and position in the subject B samples: 1.49 ± 0.09 min (see Fig. 1b). Overall, the electropherograms of the subject C samples are narrower than those from subjects A and B. The shoulders following the initial peaks in subject B have lower absolute intensities than the shoulders on samples A and C. The raw data suggest that it will be challenging to develop a universal informatics model that is simultaneously applicable to all three subjects and able to differentiate between 0, 5, and 10% doped samples.

In this work, we applied three traditional informatics methods to differentiate between 0, 5, and 10% doping in three subjects. Doping levels of 5% and 10% were chosen because below 5% doping, there are no appreciable physiological effects that increases an athlete's performance.[10] A single unit of ABD blood transfusion into an athlete (assuming 4–5 L of blood for an adult) results in 10% doping.[2] The informatics methods employed in this work included cluster analysis, principal component analysis (PCA), and partial least squares (PLS), which struggled to identify doping due to the limited size of the data set and the large natural variation in the electropherograms that was noted above. For example, cluster analysis achieved separation of the undoped samples from the doped samples at a level of three clusters, but gave meaningless results at a level of two clusters. PCA scores did not show clear clustering of any of the samples, and the PLS calibration showed large error bars after a leave-one-out cross validation of the data. Accordingly, we considered four less traditional methods: pattern recognition entropy (PRE), the Euclidean distance, a peak fitting method (Peak Fit-Integration), and the second moment (SM) to differentiate the electropherograms, all of which showed some success. Combinations of 2, 3 and 4 of the summary statistics generated from these analyses were used in an inverse least squares (ILS) analysis. The resulting ILS calibrations showed solid promise in differentiating between doped and undoped samples and to some extent between different levels of doping. Thus, this approach appears to be able to identify ABD in athletes.

2. Experimental

2.1 Sample preparation and data collection

Blood samples analyzed in this study were procured from three professional male cyclists and one less active control male

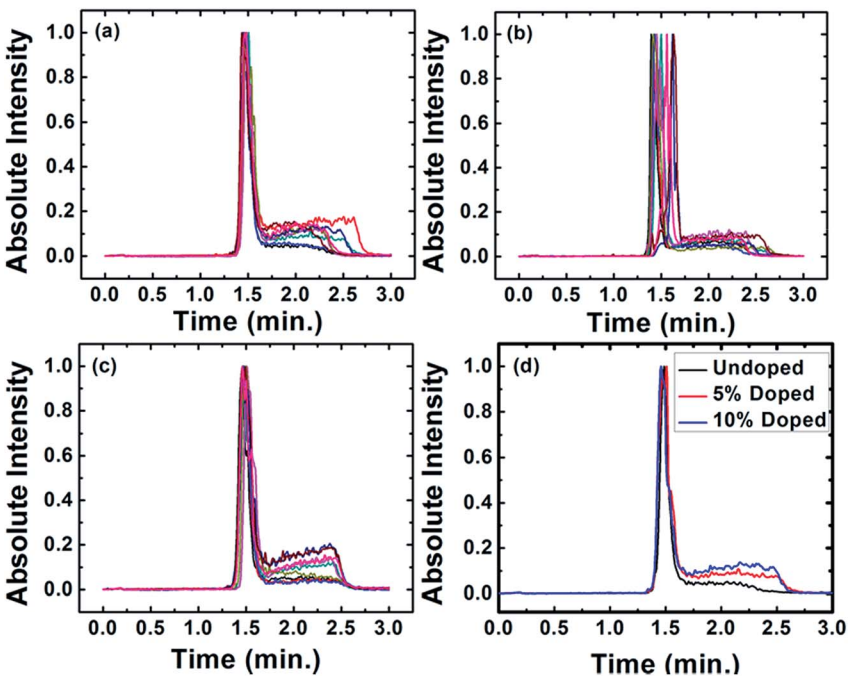

Fig. 1 Capillary electropherograms of undoped (0%) and doped (5 and 10%) blood samples of subjects (a) A, (b) B, and (c) C. Three replicates at each doping level are shown in each panel. (d) Three electropherograms from subject A at 0, 5, and 10% doping levels.

subject according to proper ethical practices. Each of the subjects was provided with a written informed consent document, which included details on procedures and biological data handling. Samples (200 μL) were collected by a fingertip lancing process and stored for 41–42 days at 4 °C before being infused into freshly-drawn blood samples to replicate an autologous blood transfusion. Though the storage of whole blood in the citrate-phosphate-dextrose (CPD) buffer used in this study is potentially not the method of choice used by athletes, it is a reasonable approach for studying autologous blood doping. Indeed, this option was selected because it would not trigger any of the controls set by the Athlete's Biological Passport (ABP) testing regimen, the most common tool used to detect blood doping. Other storage and transfusion methods, such as cryopreservation of RBCs, could trigger the ABP alarm, as the influx of RBCs without compensating for an increase in total blood volume would push an athlete above the hematocrit limit (% RBCs in total blood volume). Thus, while the doping approach taken here may not have been perfect, it was adequate to simulate what could likely take place. This protocol and study had been approved and funded by the World Anti-Doping Agency. The RBCs contained in the transfused samples were then separated and prepared for the CE separation. The RBCs were isolated *via* centrifugation and vortex mixing with phosphate-buffered saline (PBS) and 2.5% glutaraldehyde solutions in PBS (gPBS), after which they were given adequate time to stabilize. The RBCs were further isolated and resuspended in a 45% w/v NaBr solution for the CE separation. A P/ACE™ MDQ capillary electrophoresis system from Beckman Coulter, Inc. (Fullerton, CA) was employed to carry out the subsequent CE analysis using fused capillaries of 365 μm outer diameter and varying internal diameter. Data were acquired every 0.25 s and monitored at 415 nm to identify the RBCs. All separations were performed at a controlled temperature of 25 °C for both the sample compartment and the capillary. Further experimental details associated with the sample preparation and data collection were previously reported in the original paper published by Harrison *et al.*[1]

2.2 Computations and data analysis

Computer programs used to perform the calculations of pattern recognition entropy (PRE) and the Euclidean distance (d_{Eu}) were written in the Matlab computing environment (Version R2015b, Release No. 8.6.0.267246, The Mathworks Inc., 1 Apple Hill Drive, Natick, MA, USA). CasaXPS (Version 2.3.19PR1.0) was used for the peak fitting/area calculations. The computer used for this work was an Intel® Core™ i7-4770 CPU@3.40 GHz with 16.0 GB of RAM on a 64-bit Windows 7 Enterprise Edition operating system. Capillary electropherograms were organized row-wise to construct a data matrix. PCA and cluster analysis were performed using the PLS Toolbox, version 7.9.3 from Eigenvector Research, Inc., Wenatchee, WA, USA in the MATLAB programming environment. Cluster analysis was performed on the preprocessed data (preprocessing described below) using Ward's minimum variance method.

3. Theory

The following is a brief description of the informatics methods used in this study.

3.1 Cluster analysis[11]

Cluster analysis relies on the assumption that related spectra/data vectors will be closer in an n-dimensional space, *i.e.*, similar samples will cluster.[12–14] It is primarily an exploratory analysis method. Spectra/data vectors are aggregated according to the similarity of their features/variables, *i.e.*, a cluster will define group memberships at different levels of aggregation. In particular, the Euclidean distance can be used in a cluster analysis to determine similarity between spectra. Our calculation of Euclidean distances between undoped and subsequently doped blood samples is similar in concept to cluster analysis and a distance analysis we previously reported.[15]

3.2 Principal component analysis (PCA)[16–18]

In PCA, a set of data, *e.g.*, spectra, is expressed in a different coordinate system, which is defined by the eigenvectors, a.k.a., principal components or factors, of the data matrix. The eigenvalues of these eigenvectors provide a quantitative measure of the amount of variance captured by each principal component. PCA can be viewed as plotting spectra as single points in a hyperspace and then rotating the original coordinate system of the data in a way that captures the largest amount of variance possible in the spectra (data points) along new axes as they are sequentially determined. The projections of the data points on the new axes (principal components) are the scores, and the loadings are the contributions of the original axes (variables) to the new axes.

3.3 Partial least squares (PLS)[11,17]

The fundamental aim of PLS is to find factors (latent variables) that can capture the maximum variation present in a data matrix, **X**, for predicting some attribute of the samples, **c**.[19] In this work, **c** represents the doping concentration matrix and **X** represents the electropherograms arranged in a row-wise fashion. A 'leave-one-out' cross-validation was used to test the PLS calibration developed herein.

3.4 Pattern recognition entropy (PRE)[18]

PRE is a recent application of Shannon's Information Theory[20–22] that serves as a summary statistic and shape recognition tool for differentiating between spectra. Shannon's entropy (H) of a data stream is defined as:

$$H(x_i) = -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i) \quad (1)$$

where the $p(x_i)$ are the probabilities associated with each data point x_i. H is a measure of the uncertainty in the system and serves as a quantification of the total information present in a data stream. PRE is a modification of Shannon's entropy where 'pseudo-probabilities' in the electropherograms are

obtained by normalizing the data with the 1-Norm. Spectra with more features have higher PRE values (many data points with higher $p(x_i)$ values), and *vice versa*. PRE has been recently shown to be helpful in analyzing X-ray photoelectron spectroscopy (XPS) and time-of-flight secondary ion mass spectrometry (Tof-SIMS) depth profiles.[18] The 'reordered spectrum' is a visual, intuitive tool for better understanding the relationship between normalized spectra and their corresponding PRE values.[23] PRE has been used to select mass chromatograms to prepare high quality total ion current chromatograms in liquid chromatography-mass spectrometry.[24,25] Because the CE spectra from doped and undoped blood differ in shape, PRE can be employed to differentiate and identify the samples. As illustrated in Fig. 1d, the electropherograms of undoped blood tend to be narrower/more 'spike-like', *i.e.*, they should have lower PRE values, with the absence of a wide shoulder arising from an absence of aged RBCs, while the electropherograms of the doped samples tend to be wider/contain more evenly matched values, *i.e.*, they should have higher PRE values.

3.5 Euclidean distance (d_{Eu})

The Euclidean distance (d_{Eu}) of two vectors[15,26] in an n-dimensional space is the length of the line segment connecting them. For two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$, d_{Eu} is defined as:

$$d_{Eu}(u, v) = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2 + (u_3 - v_3)^2 \ldots (u_n - v_n)^2} \quad (2)$$

for example, d_{Eu} for **u** (1, 2, 3, 4) and **v** (2, 3, 4, 5) is,

$$d_{Eu}(\boldsymbol{u}, \boldsymbol{v}) = \sqrt{(1-2)^2 + (2-3)^2 + (3-4)^2 + (4-5)^2} = 2 \quad (3)$$

an electropherogram, which is a set of intensity values at distinct time points, can be considered a vector in an n-dimensional space, where n is the total number of time points at which intensity values are recorded. Accordingly, if two electropherograms are similar, their d_{Eu} value will be closer to zero. On the other hand, d_{Eu} values of less similar electropherograms (or spectra) are expected to be larger.

3.6 Peak fit, integration (PFI)

As the degree of autologous blood doping (ABD) increases, the shoulder to the right of the main signal in the electropherograms generally becomes longer and higher. A commercial peak fitting software package (CasaXPS – see details above) was used to calculate the areas of the entire signals (main sharp peaks and shoulders) and the areas of just the sharp peaks. The difference between these areas was a measure of the degree of ABD. The background chosen for this purpose was the Shirley background with a five-point average, where an 'n' point average in this background defines a '$2n + 1$' window on each side of the region described by the background to establish its starting and ending points. The Shirley background has been widely used in XPS peakfitting.[27] A higher window width for the background is preferred when the data contains a higher noise level. Fig. 2 shows a representative Shirley background under the sharp feature of an ABD electropherogram.

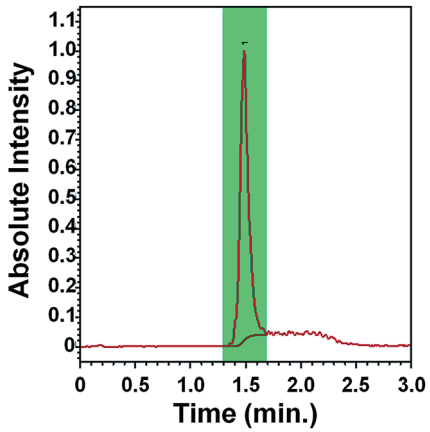

Fig. 2 Analysis of the main, sharp peak (labeled '1' and highlighted in green) centered at *ca.* 1.5 min in an electropherogram of a 5% doped sample from subject A using a Shirley background.

3.7 Second moment (SM)

The second moment (SM), of the electropherograms was calculated using the following formula:

$$\mathrm{SM} = \sum_{1}^{n} y_i t_i^2 \quad (4)$$

where, y_i are the intensity values from the electropherograms and the t_i are the corresponding time points. Here, the square of the time values enhances the intensity values at increased times. It is relatively easy to show that the second moment is not shift invariant. To do so, we consider the second moment of a 'spectrum' composed of two data points: (t_i, y_i) and (t_{i+1}, y_{i+1}):

$$y_i t_i^2 + y_{i+1} t_{i+1}^2 \quad (5)$$

here, it is assumed that the spacing between the times is Δt, such that

$$t_{i+1} = t_i + \Delta t \quad (6)$$

so that we can write eqn (5) as

$$y_i t_i^2 + y_{i+1}(t_i + \Delta t)^2 \quad (7)$$

which is equivalent to

$$y_i t_i^2 + y_{i+1} t_i^2 + y_{i+1} 2 t_i \Delta t + y_{i+1} \Delta t^2 \quad (8)$$

now, it is imagined that this spectrum is shifted by n time increments, *i.e.*, by $n\Delta t$, which converts eqn (7) into:

$$y_i(t_i + n\Delta t)^2 + y_{i+1}(t_i + (n+1)\Delta t)^2 \quad (9)$$

expanding and simplifying this equation gives:

$$\begin{aligned} &y_i t_i^2 + y_{i+1}(t_i + \Delta t)^2 + y_i[t_i 2n\Delta t + n^2 \Delta t^2] \\ &+ y_{i+1}[t_i 2n\Delta t + 2n\Delta t^2 + n^2 \Delta t^2] \end{aligned} \quad (10)$$

if the second moment enjoyed shift invariance, eqn (7) and (10) would be the same. However, it is clear that if n is an integer

greater than 0, $\Delta t > 0$, $y_i \neq 0$, and $x_i \neq 0$, the third and fourth terms in eqn (10) are not zero. Accordingly, we calculated the second moment of our data set starting from the first data point in the series, and also starting just before the sharp peaks that contain useful information.

3.8 Inverse least squares (ILS)

The governing and most simple equation for classical least squares (CLS) is $\mathbf{A} = \mathbf{KC}$, where **A**, **K**, and **C** are matrices containing absorbance spectra, pure component spectra, and concentrations, respectively. As written here, **K** organizes the pure component spectra column-wise. CLS models spectra as linear combinations of pure component spectra. Inverse least squares (ILS) is based on a similar equation: $\mathbf{C} = \mathbf{PA}$. That is, ILS directly relates measured spectra to concentrations through a matrix **P**. To develop an ILS calibration, *i.e.*, to solve for **P** when **C** and **A** are known, one must first right-multiply both sides of $\mathbf{C} = \mathbf{PA}$ by $\mathbf{A}^T$. The resulting matrix ($\mathbf{AA}^T$) will only have an inverse, *i.e.*, not be rank deficient, if it has at least as many columns as it does rows. That is, ILS requires that there be at least as many samples as there are data points in **A**. Many spectra, *e.g.*, electropherograms, contain hundreds or thousands of values, and it is not generally feasible to work with hundreds or thousands of specimens (spectra). Hence, a variable reduction technique is often necessary for ILS to function. In this work, we reduced the electropherograms to four numbers: the PRE, d_{Eu}, PFI, and SM values, to develop an ILS model for predicting doping levels.

3.9 Preprocessing

Preprocessing plays an important role in many chemometrics analyses. For example, mean centering consists of taking the average of the values of the electropherograms at a given time and then subtracting that average from each individual value at that time. In other words, the average electropherogram is subtracted from each electropherogram in the data set and the center of the data point cluster (individual electropherograms) is moved to the origin. This is advantageous because otherwise the first principal component (PC 1) in PCA points towards the center of the cloud of data points, *i.e.*, it represents the average spectrum, where this direction may or may not correlate with any chemical trend in the data and PC 1 may have to be discarded. However, with mean centering, the spectral regions (points in time in the electropherograms here) that correspond to greater excursions (spreads) in the data are more heavily weighted in the analysis. Autoscaling overcomes this problem. Autoscaling consists of mean centering the data and then dividing by the corresponding standard deviations, putting the regions of the electropherograms/spectra on equal footing in the analysis. Autoscaling is generally inappropriate for data sets that contain both noisy and signal-containing regions because it gives them equal importance in the analysis. This approach is appropriate for our range-selected data (see below) because the data do not contain regions of significant noise.

4. Results and discussion

The purpose of our work is to find statistical/mathematical tools that differentiate between the electropherograms from doped and undoped blood in the ABD data set in Fig. 1. Believing it would be important to start with well-accepted tools before considering or introducing others, we first applied three well-known chemometrics methods to the data set: cluster analysis, PCA, and PLS. These traditional methods were inadequate because of the large natural variation in the electropherograms and the limited number of samples (spectra), which made variance analysis difficult. Accordingly, we pursued other possible approaches/algorithms. These included pattern recognition entropy (PRE), which we have recently used multiple times,[18,23,24,28] the Euclidean distance, peak fit-integration (PFI), and the second moment (SM). These results were then combined to develop inverse least squares (ILS) calibrations.

4.1 Traditional analyses: cluster analysis, PCA, and PLS

Three different preprocessing methods were applied to the data in the cluster analysis. In the first, a process referred to as 'range selection', the data were selected over the range in which they appear to contain meaningful signal(s) (from about 1.3–2.7 min, see Fig. 1). Range selection is a form of scaling in which the data are multiplied by a weighting factor of 0 or 1. The range selected data points were then normalized with the 1-Norm, where this operation consists of division of each data point in an electropherogram by the sum of the data points in that electropherogram, or in the case under current study, each point in the range-selected electropherogram was divided by the sum of the data points in the range-selected electropherogram. Finally, the data were autoscaled. Replicate runs of each sample introduce correlation into the analysis. However, given the limited sample size and the significant natural variation between the runs, smoothing over this variation by averaging the runs would result in the loss of information. Fig. 3 shows the dendrogram produced from the cluster analysis of the preprocessed data. The results are mixed. Of the three main clusters in Fig. 3, which are delineated by the black, vertical, dashed line, the bottom cluster contains seven mostly 'Clean' (0%) electropherograms with only two that are not (one 5% and one 10% sample). The middle cluster consists of an even mix of 5% and 10% samples (six of each) and one 'Clean' sample, and the top cluster also contains an even mix of 5% and 10% samples (two of each) plus one 'Clean' sample. These results suggest that cluster analysis can fairly reasonably separate doped (5% and 10%) from undoped (0% 'Clean') samples, but that it cannot distinguish between the two levels of doping considered in this study. However, there is an inconsistency in the clustering here that is revealed in the two-cluster model in the dendrogram (see the light blue vertical dashed line). That is, *a priori*, one would expect that of the three clusters suggested by the dendrogram, the two that should be most similar, and that should cluster, would contain mostly 5% and 10% samples, *i.e.*, the upper two clusters in the three-cluster model. However, this is not the

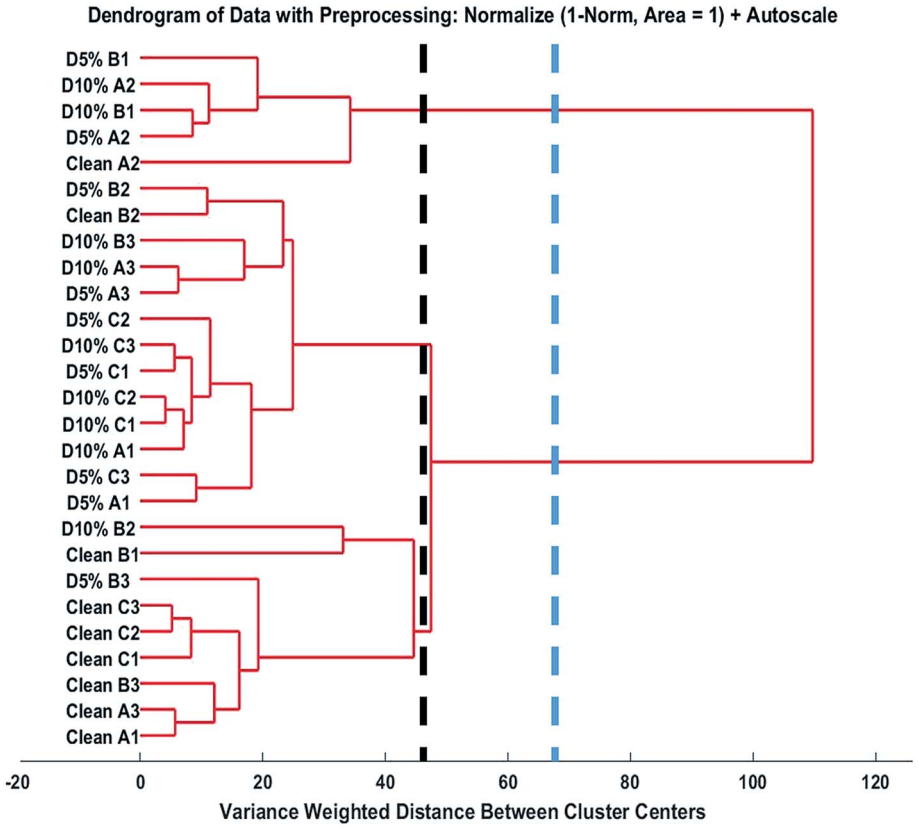

Fig. 3 Dendrogram from a cluster analysis of the 0, 5, and 10% electropherograms under consideration in this study. The data were pre-processed using range selection followed by normalization (1-Norm) and autoscaling. The dashed, vertical, light blue line indicates a two-cluster model, and the dashed, vertical, black line indicates a three-cluster model. 'Clean', 'D5%', and 'D10%' represent the 0, 5, and 10% samples, and 'A', 'B', and 'C' represent the three subjects. Replica runs are represented by the number following the 'A', 'B', or 'C'.

case. The cluster with the larger number of 5% and 10% samples combines with the bottom 'Clean' cluster in the two-cluster model. While this may suggest that the five samples in the top cluster are outliers, it is probably inappropriate to eliminate 5 of our 27 samples in this way.

The second preprocessing approach taken for our cluster analysis was to repeat the range selection as was done previously and then apply the 1-Norm to the data. These results are shown in ESI Fig. 1.† Two main clusters were observed. The top cluster contained 12 samples: 2, 6, and 4 of the 0, 5, and 10% samples, respectively, while the bottom cluster contained 15 samples: 7, 3, and 5 of the 0, 5, and 10% samples, respectively. It is difficult to see any distinct separation of the samples in this analysis. Finally, in a third attempt at cluster analysis, the data were range selected and autoscaled. This approach again produced two distinct clusters (see ESI Fig. 2†). The top cluster contained 13 samples: 0, 6, and 7 of the 0, 5, and 10% samples, respectively, while the bottom cluster contained 14 samples: 9, 3, and 2 of the 0, 5, and 10% samples, respectively. That is, with this preprocessing approach, the top cluster only contained doped samples (nearly equal amounts of the 5% and 10% samples), while the bottom cluster contained almost twice as many undoped samples as it did doped samples. While this preprocessing approach is arguably the best of the three methods considered herein, its ability to separate the samples into classes is still arguably weak.

PCA is one of the most commonly used multivariate analysis tools. It is an unsupervised pattern recognition technique, meaning that it requires no prior knowledge of the classes to which objects may belong. PCA has been applied to many different data types from many different types of samples. For example, in our laboratory it has be used to analyze data obtained from the analysis/characterization of alkyl monolayers on silicon,[29] coal samples,[30] mouse livers,[31] nanodiamonds,[32] and chemically treated display glass surfaces.[33] One of the key limitations of PCA is the large sample size required for analysis of variance and determination of correlation structure. Nevertheless, there are numerous reports containing examples of the successful application of PCA to relatively small data sets.[34] We performed PCA of our range-selected, normalized, and autoscaled data. To determine the number of PCs to keep, we examined the root mean square error of cross-validation (RMSECV) and root mean square error of calibration (RMSEC) figures of merit against the number of principal components (PCs) (see ESI Fig. 3†). The RMSECV here was based on a leave-one-out cross validation. As expected, the RMSEC value decreased monotonically as the number of PCs increased, *i.e.*, an increased number of PCs successively captured more of the variance in the data. The RMSECV value decreased by only

a small amount from 1 to 9 components (from 7.513 to 7.251) with only a limited increase in the variance captured. Thus, a one-PC model would appear to be appropriate. However, the resulting scores plot from this one-PC model did not show any reasonable groupings of the samples that corresponded to their degrees of doping (see ESI Fig. 4†). A nine-PC model was then considered. It also failed to show any reasonable groupings of the data on any of the nine PCs (see ESI Fig. 5–13†). Hotelling T^2 *vs.* Q residuals plots were then generated for the one- and nine-PC models (see ESI Fig. 14 and 15†). These plots revealed the distribution of the data both within (Hotelling T^2) and outside of the models (Q residuals). In both cases, most of the data points lie within 95% confidence limits. However, in the one-PC model three data points fell far outside these limits, whereas in the nine-PC model, two data points fell slightly outside the limits. Accordingly, a final attempt was made to analyze the data by PCA in which the three outliers in the one-PC model were removed and the model was recreated. Based on the eigenvalues associated with each PC, a three-PC model appeared appropriate for the remaining data. Unfortunately, none of the PCs in this model showed any reasonable groupings of the data in their scores plots. In summary, multiple attempts with PCA failed to reveal or find any of the expected trends in the electropherograms.

Despite the overall lack of success with PCA, we attempted to use PLS to create a calibration of our data set. For the x-block in this attempt, *i.e.*, the electropherograms, the data were pre-processed by range selection, the 1-Norm, and autoscaling. For the y-block, *i.e.*, the degree of doping, the values were mean centered. Employing leave-one-out cross validation results for this modeling (see ESI Fig. 16†), one-, seven-, or eight-components seemed appropriate (these models showed the lowest RMSEC values). Accordingly, we examined the predictions of the one- and seven-component models. (The eight-component model was not considered because its RMSECV value was essentially identical to that from the seven-component model.) The predictions of the seven-component model were of low quality (see Fig. 4). This model appeared to be able to differentiate between the 0% and 10% doping, but the 5% samples showed strong overlap with both the 0% and 10% samples. The predictions from the one-component model were of a lower quality and were, therefore, useless. We conclude that PLS is fairly unsuccessful in creating the desired calibration between the doping levels and the corresponding electropherograms.

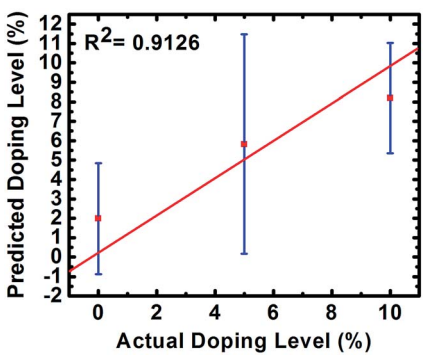

Fig. 4 Seven-component, PLS predictions of doping levels from replicate runs for undoped and doped (5 and 10%) blood samples. Here, a separate seven-component PLS model was created for each data set with one of its samples left out, and that sample was then predicted by the corresponding model.

4.2 Analysis by less traditional tools: PRE, the Euclidean distance, peak fit-integration and the second moment

Because of the inability of the traditional multivariate approaches (cluster analysis, PCA, and PLS) to model the doping levels of the blood samples, we turned to less traditional mathematical/statistical analyses. These were pattern recognition entropy (PRE), the Euclidean distance, peak fit-integration, and the second moment.

First, PRE was performed on the electropherograms under consideration in this study. Fig. 5a shows the average PRE values with standard deviations of the three replicate electropherograms from each subject at each level of doping (see ESI† Fig. 17 for the corresponding raw data). The PRE value, which is a summary statistic, is reflective of the shape of the electropherogram, where the presence of additional peaks/shoulders in the electropherogram, which takes place for the 5% and 10% doped samples, results in higher PRE values. As a result, the PRE values gradually increase with doping – PRE is rather effective at responding to the doping levels of all the samples considered in this study.

The reordered spectrum is a visual, intuitive tool for understanding PRE analysis;[23] the absolute magnitude of PRE values are abstract and a graphical way of understanding it can be helpful. A reordered spectrum sorts the values of a spectrum from high to low. For example, three reordered spectra (electropherograms) of undoped, 5% doped, and 10% doped blood from subject A are shown in Fig. 5b. The reordered electropherogram corresponding to the undoped sample has the sharpest peak, which is consistent with its lower PRE value, while the reordered electropherogram from the 5% and 10% doped samples have higher numbers of data points with larger values, which is consistent with their higher PRE values.

Second, Fig. 5c shows the Euclidean distances (d_{Eu}) between the electropherograms of the clean and 5% or 10% doped samples, *i.e.*, the distance between the 0% and 5% and also the 0% and 10% samples was calculated for each replicate run. These distances between the electropherograms are expected to progressively increase with increasing doping levels. This is another way of saying that the vectors corresponding to the doped and undoped electropherograms are expected to be different, and also increase as the degree of doping increases. It is clear from the results in Fig. 5c that d_{Eu} always shows a difference between the undoped and doped samples. Furthermore, while, on average, the d_{Eu} values for the 10% doped samples are greater than those for the 5% samples, there is enough overlap between these results that it would be difficult to differentiate between these two states with this method.

Third, Fig. 5d shows the 'Peak Fit-Integration' (PFI) results obtained by measuring the areas of the shoulders in the electropherograms to the right of the main peaks of the doped and

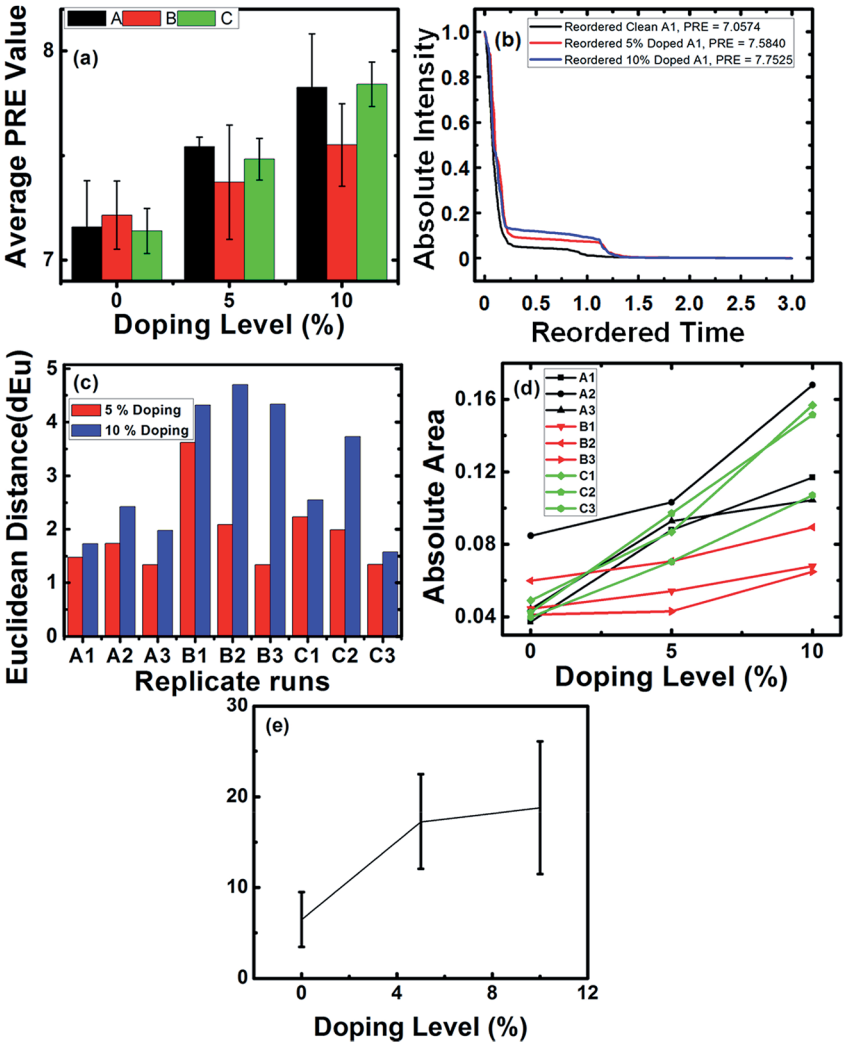

Fig. 5 Results of the less traditional mathematical/informatics methods used to analyze the doping data. (a) (top left) The average PRE values (heights of bars) with standard deviations (error bars) of the electropherograms of subjects A, B and C for 0, 5, and 10% doping levels. (b) (top right) The reordered electropherograms from replicate run 1 of subject A at 0, 5, and 10% doping levels. (c) (middle left) The Euclidean distances between electrophoretic separations of clean and doped (5% and 10%) samples of subjects A, B and C with three replicate runs for each. (d) (middle right) The absolute areas of the broad features (shoulders) to the right of the main peaks from electrophoretic separations of subjects A, B and C for 0, 5, and 10% doping levels. (e) The second moments of the range-selected electropherograms.

undoped samples. Two things are clear here. First, there is some scatter in the results. Second, the area of the shoulder consistently increases with doping level. The average and standard deviation for each set of measurements are 0.05 ± 0.01, 0.07 ± 0.02, and 0.11 ± 0.04 for the undoped, 5% doped, and 10% doped samples, respectively.

The second moment of the electropherograms was calculated in two different ways. In the first case, complete electropherograms were used for SM calculations. This method failed at providing any meaningful difference in the doping levels. In the second, range-selected data were used (Fig. 5e), and the data were preprocessed by normalization (1-Norm), followed by autoscaling. As was the case with some of the other less traditional methods we employed, this approach showed promise in separating the undoped (0%) and doped (5 and 10%) samples, but not in differentiating between different levels of doping.

4.3 Inverse least squares

Inverse least squares (ILS) is an important method for generating calibrations. In general, ILS uses a relatively small number of variables to create calibrations. For example, an ILS regression can be based on principal component regression (PCR), which uses the PCA of a data set to reduce the number of variables in the data set.[17] Here, we chose to construct ILS calibrations using the PRE, d_{Eu}, PFI, and SM summary statistics from the electropherograms. Each of these had demonstrated some ability to differentiate between the samples based on their doping levels. Accordingly, a combination of these metrics would lead to a calibration with greater predictive ability. The coefficient of determination (R^2) was used as the figure of merit. R^2 is defined as the square of the correlation coefficient r (eqn (11)) and is a measure of the percentage variation in one variable as explained by another variable.[35] (We include the formula

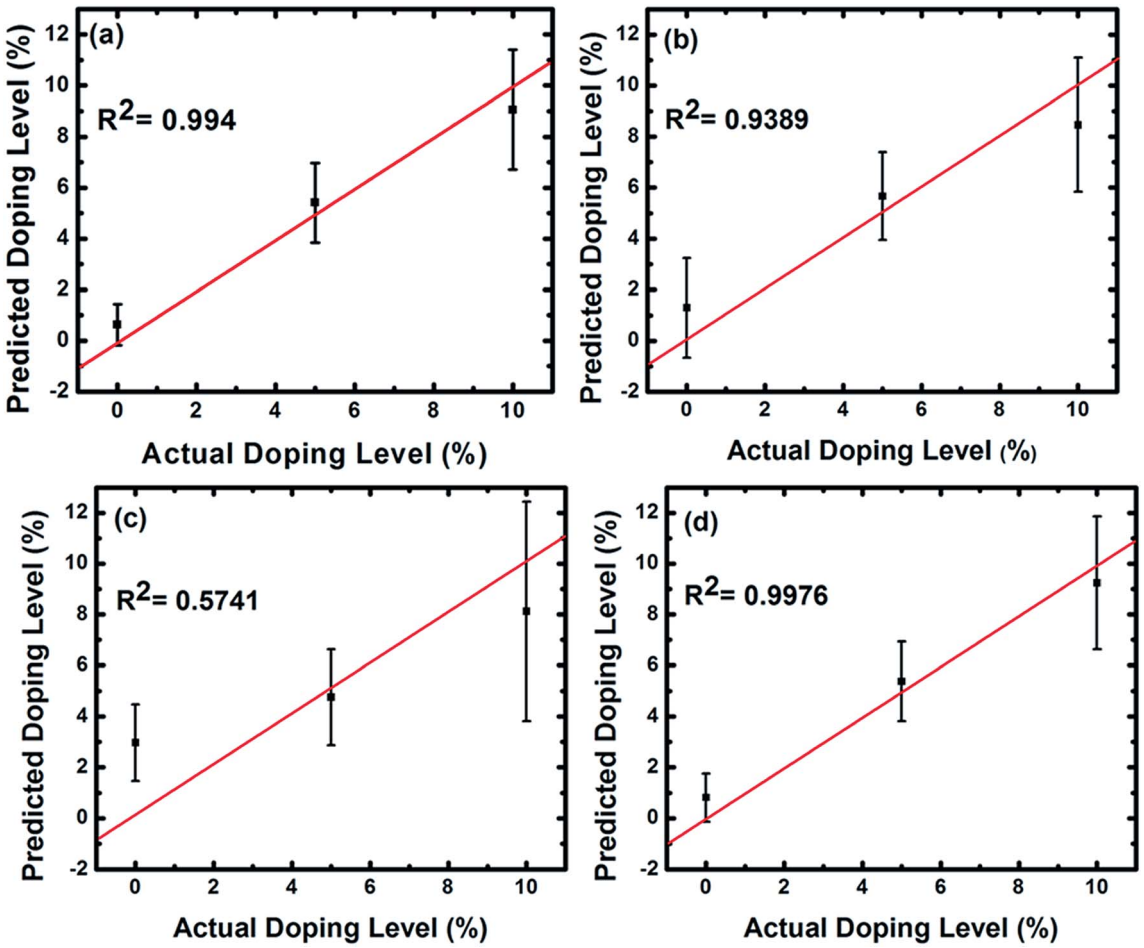

Fig. 6 Leave-one-out predictions from ILS models based on (a) PRE, d_{Eu}, and PFI, (b) PRE, d_{Eu}, and SM, (c) PRE, PFI, and SM, and (d) PRE, d_{Eu}, PFI, and SM summary statistics from replicate runs of A, B and C for undoped (0%) and doped (5% and 10%) blood samples. The R^2 values and data were compared against straight lines $y = x$ (red lines).

for 'R' here because its definition varies in the scientific literature.)

$$R = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{\sqrt{\left[n\sum x^2 - \left(\sum x\right)^2\right]\left[n\sum y^2 - (y)^2\right]}} \quad (11)$$

ILS models were built based on all possible combinations of the four summary statistics. All of the ILS models based on any two of our summary statistics showed very strong overlap in the predictions of all the three doping levels (0, 5 and 10%), *i.e.*, these models were useless. Accordingly, we considered ILS models based on all possible combinations of three summary statistics (see Fig. 6). First, Fig. 6a shows the leave-one-out predictions using the PRE, d_{Eu} and PFI summary statistics. It can easily differentiate between undoped (0%) and doped (5% or 10%) electropherograms, but not between the two levels of doping. This three summary statistic ILS model was the most successful of the four we considered (R^2 value of 0.994). It seems reasonable that it is based on PRE and d_{Eu} because these summary statistics appeared to be the most successful in differentiating between the doping levels (see Fig. 5). Fig. 6b shows the ILS predictions using PRE, d_{Eu} and SM. It was less successful as a model (R^2 value of 0.938), which may be explained by SMs greater struggle to differentiate between the samples (see Fig. 5e). The ILS model based on PRE, PFI and SM in Fig. 6c was quite poor (R^2 value of 0.574), and the ILS model created using d_{Eu}, PFI and SM, which is not shown, was even worse (R^2 value of 0.328). It is evident from these results that PRE and d_{Eu} made the largest contributions to the prediction capabilities of the ILS models. As a final attempt, an ILS model based on all four summary statistics was created (see Fig. 6d). It gave the highest R^2 value (0.997), and like the PRE, d_{Eu}, and PFI-based ILS model, it can clearly differentiate between 0% (undoped) and 5% or 10% levels of doping. Doping below 5% has little physiological effect.[2] Accordingly, the ILS model shows high accuracy in its ability to differentiate between undoped and 'meaningfully' doped blood (see Fig. 6a and d).

5. Conclusions

The detection of autologous blood doping is critical for banning unscrupulous practices used by athletes to gain an unfair

advantage in competition. In this work, we have demonstrated several mathematical techniques that distinguished between doped and undoped blood samples. Capillary electrophoresis was previously used to separate fresh and stored RBCs, serving as a viable alternative to the widely used and more expensive method of monitoring an athlete's biological passport. In Harrison's original work on this topic, a first derivative analysis of slopes was used to detect the presence of doping. However, this method suffered from false negatives, lacking a strong ability to precisely identify doping. In our work, conventional informatics techniques (cluster analysis, PCA, and PLS) had very limited success in distinguishing between electropherograms of samples with different levels of doping. Several preprocessing methods were considered in these analyses. Variance analysis (PCA and PLS) was challenging due to the large natural variation in electropherograms from replicate runs. Four less commonly used summary statistics (PRE, the Euclidean Distance, Peak Fit/Integration, and the Second Moment) were applied to the data. An ILS calibration based on these inputs allowed easy differentiation between undoped and doped samples, and to some degree between the different levels of doping (5 and 10%). We understand that natural variation can exist in the RBCs of athletes due to biological sex, ethnicity, muscle/fat percentage, diet, age, *etc.* In our (Harrison's) broader studies in this area, he only sees minor differences in absolute migration times of RBCs – he has yet to see any significant differences between individuals. Thus, the changes induced by transfused cells appears to be a significant, measurably change to the cell population.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

C. Harrison wishes to offer sincere thanks to the volunteers who participated in this study. Support for the anti-doping study was provided by the World Anti-Doping Agency (reference number 09A23CH). All experiments performed with human blood samples were in compliance with the US Code of Federal Regulations (45 CFR 46; 21 CFR 50) as approved by the San Diego State University Institutional Review Board. Informed consent was obtained from all subjects who participated in the study.

References

1 C. R. Harrison, J. C.-Y. Fang, K. J. Walthall, C. C. Green and V. Porobic, Towards the identification of autologous blood transfusions through capillary electrophoresis, *Anal. Bioanal. Chem.*, 2014, **406**(3), 679–686.

2 M. Nelson, M. Ashenden, M. Langshaw and H. Popp, Detection of homologous blood transfusion by flow cytometry: a deterrent against blood doping, *Haematologica*, 2002, **87**(8), 881–882.

3 D. S. Johnson, A. Mortazavi, R. M. Myers and B. Wold, Genome-wide mapping of *in vivo* protein–DNA interactions, *Science*, 2007, **316**(5830), 1497–1502.

4 J. Mørkeberg, Detection of autologous blood transfusions in athletes: a historical perspective, *Transfus. Med. Rev.*, 2012, **26**(3), 199–208.

5 H. Striegel, D. Rössner, P. Simon and A. Niess, The World Anti-Doping Code 2003-Consequences for Physicians Associated with Elite Athletes, *Int. J. Sports Med.*, 2005, **26**(03), 238–243.

6 O. Salamin, S. De Angelis, J.-D. Tissot, M. Saugy and N. Leuenberger, Autologous blood transfusion in sports: emerging biomarkers, *Transfus. Med. Rev.*, 2016, **30**(3), 109–115.

7 P.-E. Sottas, N. Robinson and M. Saugy, The athlete's biological passport and indirect markers of blood doping, in *Doping in Sports: Biochemical Principles, Effects and Analysis*, Springer, 2010, pp. 305–326.

8 Y. X. Huang, Z. J. Wu, J. Mehrishi, B. T. Huang, X. Y. Chen, X. J. Zheng, W. J. Liu and M. Luo, Human red blood cell aging: correlative changes in surface charge and cell properties, *J. Cell. Mol. Med.*, 2011, **15**(12), 2634–2642.

9 A. Rolfes-Curl, L. L. Ogden, G. M. Omann and D. Aminoff, Flow cytometric analysis of human erythrocytes: II. Possible identification of senescent RBC with fluorescently labelled wheat germ agglutinin, *Exp. Gerontol.*, 1991, **26**(4), 327–345.

10 N. Gledhill, Blood doping and related issues: a brief review, *Med. Sci. Sports Exercise*, 1982, **14**(3), 183–189.

11 M. Otto, *Chemometrics: statistics and computer application in analytical chemistry*, John Wiley & Sons, 2016.

12 T. P. Auf der Heyde, Analyzing chemical data in more than two dimensions: a tutorial on factor and cluster analysis, *J. Chem. Educ.*, 1990, **67**(6), 461.

13 L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*, John Wiley & Sons, 2009, vol. 344.

14 P. Yu, Applications of hierarchical cluster analysis (CLA) and principal component analysis (PCA) in feed structure and feed molecular chemistry research, using synchrotron-based Fourier transform infrared (FTIR) microspectroscopy, *J. Agric. Food Chem.*, 2005, **53**(18), 7115–7127.

15 J. D. Bagley, H. Dennis Tolley and M. R. Linford, Reevaluating the conventional approach for analyzing spectroscopic ellipsometry psi/delta *versus* time data. Additional statistical rigor may often be appropriate, *Surf. Interface Anal.*, 2016, **48**(4), 186–195.

16 S. Wold, K. Esbensen and P. Geladi, Principal component analysis, *Chemom. Intell. Lab. Syst.*, 1987, **2**(1–3), 37–52.

17 R. Kramer, *Chemometric techniques for quantitative analysis*, CRC Press, 1998.

18 S. Chatterjee, B. Singh, A. Diwan, Z. R. Lee, M. H. Engelhard, J. Terry, H. D. Tolley, N. B. Gallagher and M. R. Linford, A perspective on two chemometrics tools: PCA and MCR, and introduction of a new one: pattern recognition entropy (PRE), as applied to XPS and ToF-SIMS depth profiles of

organic and inorganic materials, *Appl. Surf. Sci.*, 2018, **433**, 994–1017.
19 S. Wold, M. Sjöström and L. Eriksson, PLS-regression: a basic tool of chemometrics, *Chemom. Intell. Lab. Syst.*, 2001, **58**(2), 109–130.
20 C. E. Shannon, A mathematical theory of communication, Part I, Part II, *Bell Syst. Tech. J.*, 1948, **27**, 623–656.
21 C. E. Shannon, A mathematical theory of communication, *ACM SIGMOBILE Mobile Computing and Communications Review*, 2001, **5**(1), 3–55.
22 J. Van Der Greef, Method and system for identifying and quantifying chemical components of a mixture,Utility, US20040267459A1, 2004, Int. Cl G06F 19/00, U.S. Cl. 702/30.
23 S. Chatterjee and M. R. Linford, Reordered (Sorted) Spectra. A Tool for Understanding Pattern Recognition Entropy (PRE) and Spectra in General, *Bull. Chem. Soc. Jpn.*, 2018, **91**(5), 824–828.
24 S. Chatterjee, G. H. Major, B. Paull, E. S. Rodriguez, M. Kaykhaii and M. R. Linford, Using pattern recognition entropy to select mass chromatograms to prepare total ion current chromatograms from raw liquid chromatography-mass spectrometry data, *J. Chromatogr. A*, 2018, **1558**, 21–28.
25 S. Chatterjee, S. C. Chapman, B. M. Lunt and M. R. Linford, Using Cross-Correlation with Pattern Recognition Entropy to Obtain Reduced Total Ion Current Chromatograms from Raw Liquid Chromatography-Mass Spectrometry Data, *Bull. Chem. Soc. Jpn.*, 2018, **91**(12), 1775–1780.
26 I. Ragnemalm, The Euclidean distance transform in arbitrary dimensions, *Pattern Recogn. Lett.*, 1993, **14**(11), 883–888.
27 D. A. Shirley, High-resolution X-ray photoemission spectrum of the valence bands of gold, *Phys. Rev. B: Solid State*, 1972, **5**(12), 4709.
28 S. Chatterjee, S. C. Chapman, B. M. Lunt and M. R. Linford, Using Cross-Correlation with Pattern Recognition Entropy to Obtain Reduced Total Ion Current Chromatograms from Raw Liquid Chromatography-Mass Spectrometry Data, *Bull. Chem. Soc. Jpn.*, 2018, **91**(12), 1775–1780.
29 L. Yang, Y.-Y. Lua, M. Tan, O. A. Scherman, R. H. Grubbs, J. N. Harb, R. C. Davis and M. R. Linford, Chemistry of olefin-terminated homogeneous and mixed monolayers on scribed silicon, *Chem. Mater.*, 2007, **19**(7), 1671–1678.
30 L. Pei, G. Jiang, B. J. Tyler, L. L. Baxter and M. R. Linford, Time-of-flight secondary ion mass spectrometry of a range of coal samples: a chemometrics (PCA, cluster, and PLS) analysis, *Energy Fuels*, 2008, **22**(2), 1059–1072.
31 L. Yang, R. Bennett, J. Strum, B. B. Ellsworth, D. Hamilton, M. Tomlinson, R. W. Wolf, M. Housley, B. A. Roberts and J. Welsh, Screening phosphatidylcholine biomarkers in mouse liver extracts from a hypercholesterolemia study using ESI-MS and chemometrics, *Anal. Bioanal. Chem.*, 2009, **393**(2), 643–654.
32 B. Singh, S. J. Smith, D. S. Jensen, H. F. Jones, A. E. Dadson, P. B. Farnsworth, R. Vanfleet, J. K. Farrer and M. R. Linford, Multi-instrument characterization of five nanodiamond samples: a thorough example of nanomaterial characterization, *Anal. Bioanal. Chem.*, 2016, **408**(4), 1107–1124.
33 C. V. Cushman, J. Zakel, B. S. Sturgell, G. I. Major, B. M. Lunt, P. Brüner, T. Grehl, N. J. Smith and M. R. Linford, Time-of-Flight Secondary Ion Mass Spectrometry of Wet and Dry Chemically Treated Display Glass Surfaces, *J. Am. Ceram. Soc.*, 2017, **100**(10), 4770–4784.
34 A. M. Martínez and A. C. Kak, PCA *versus* LDA, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2001, (2), 228–233.
35 J. P. Barrett, The coefficient of determination—some limitations, *Am. Stat.*, 1974, **28**(1), 19–20.

