Gregório, I. et al., 2017. Statistical approach for ATR-FTIR screening of semen in sexual evidence. Talanta, 174, pp.853–857.

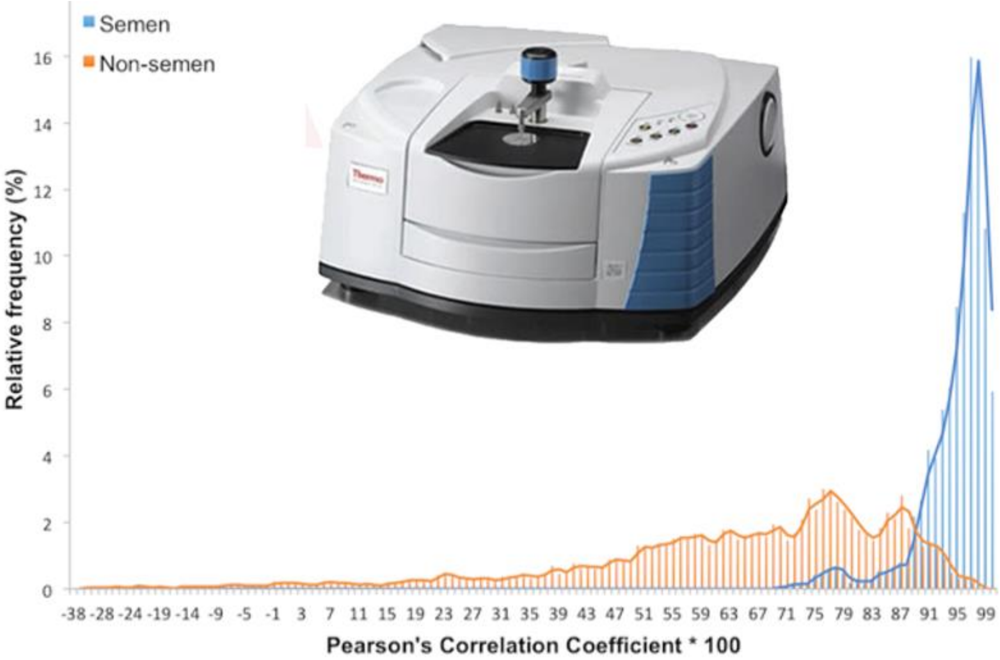
Available at https://doi.org/10.1016/j.talanta.2017.07.016

Talanta 174 (2017) 853-857

Statistical approach for ATR-FTIR screening of semen in sexual evidence

Inês Gregório, Félix Zapata, Mercedes Torre, Carmen García Ruiz*

Department of Analytical Chemistry, Physical Chemistry and Chemical Engineering and University Institute of Research in Police Sciences (IUICP), University of Alcalá, Alcalá de Henares (Madrid), Spain. carmen.gruiz@uah.es; felix.zapata@uah.es;



Cite: I. Gregório, F. Zapata, C. García-Ruiz, Statistical approach for ATR-FTIR screening of semen in sexual evidence, Talanta 174 (2017) 853-857. DOI: 10.1016/j.talanta.2017.07.016

Abstract

Genetic identification has revolutionized the Forensic Sciences, especially in sexual aggression cases. For the successful extraction of the genetic information of a criminal, a crucial step is the prior detection of bodily fluids on evidence. In this article, a method for non-destructive screening of semen samples is reported. Using chemometric tools, bodily fluids can be detected and differentiated without damaging the sample, by correlating the infrared spectra of sexual evidence with previously recorded spectra from undamaged stains of individual bodily fluids. In modern hospitals/laboratories, the proposed method would not require additional equipment/ material nor specialized personnel. Furthermore, the method provides qualitative and reliable results, without requiring human interpretation. Therefore, the proposed method opens a door for a low-cost, fully automated and efficient system for non-destructive screening of semen, which could be easily and massively implemented.

Keywords: Semen; Bodily fluids; Screening; Chemometrics; ATR-FTIR; Pearson's Correlation Coefficient.

1. Introduction

With the application of DNA to forensic sciences described by Gill et al. in 1985 [1], a huge evolution occurred on the resolution of sexual aggression cases. Each sample that entered in contact with semen, during the sexual aggression or on the vaginal discharge during the following days [2], became crucial for the identification of the aggressor through his DNA. One important step of this identification is the detection of semen on supporting material (e.g., swabs, clothes, and hygienic superabsorbent pads, among others) for further DNA profiling.

Up to date, forensic approaches for bodily fluid detection and discrimination used by forensic laboratories have not been updated. In fact, the biochemical and immunological techniques commonly used are destructive and specific for one bodily fluid only, being necessary the consecutive application of several tests to determine the bodily fluids involved [3,4]. Regarding semen detection, forensic laboratories perform presumptive tests, as seminal Acid Phosphatase, and confirmatory, such as the observation of spermatozoa by optical microscopy with the Christmas Tree test, or immunological assays of Prostate Specific Antigen (PSA) and Semenogelin Antigen detection [3,5–8]. The principal disadvantage of these techniques is their destructive character that is totally counter-productive, particularly for those samples with low male DNA content, due to low sperm quantity or degraded sperm.

In addition, considering these techniques are specific for only one fluid, a positive detection of semen using these tests does not provide any information about any other bodily fluids mixed with semen. This is a critical setback, since sexual evidence commonly contains mixtures of semen with other bodily fluids (e.g. vaginal fluid), usually in a high ratio non-semen/semen, that often generates a low confidence DNA profile [5,9–12].

Therefore, there is a need for a rapid, confirmatory and non-destructive technique for bodily fluids detection and discrimination. To this aim, few years ago, some studies began to investigate non-destructive vibrational spectroscopic techniques, including Attenuated Total Reflectance-Fourier Transform Infrared Spectroscopy (ATR-FTIR) [3,7,8,13–15]. ATR-FTIR enables a rapid, non-invasive and non-destructive analysis of these samples prior to DNA profiling [16,17]. In fact, this technique is very easy to use, does not need sample preparation, nor any reagent, and it is widely available in forensic laboratories and hospitals. However, according to Gregorio et al. [16], it is not suitable for visually discriminating bodily fluids whose spectra are similar, as occurs with semen and vaginal fluid, since both display similar amide I and amide II vibrational bands from proteins [15]. Thus, the spectral data obtained requires a statistical analysis, focused on acquiring a level of certainty either for further genetic profiling, either for translating the result to the legal court: presence/absence of semen.

The aim of this work was to develop a simple statistical procedure of bodily fluids screening, compatible with both analytical and legal considerations. With this purpose, a statistical approach was performed, comprising a multivariate analysis using Pearson's correlation coefficient followed by Bayesian statistics. Interestingly, it was also possible to aim for a discrimination of spots with higher proportion of semen, so necessary in cases where there is a mixture of the aggressor's and the victim's bodily fluids.

2. Materials and methods

2.1. Spectral data acquisition

Stains of semen, vaginal fluid and urine from healthy volunteers were prepared on white 100% cotton and hygienic superabsorbent pads: feminine sanitary napkins from Ausonia and Evax (Procter & Gamble, Ohio, USA) and Deliplus (SCA Hygiene Products, Tarragona, Spain); panty-liners from Carefree (Johnson & Johnson, New Jersey, USA) and Evax; and diapers from Deliplus and Dodot (Procter & Gamble, Ohio, USA). These volunteers signed an informed consent form and the research was carried out under the Ethical Committee approval. In sum, 6 stains of vaginal fluid from 3 female donors were prepared by placing the samples directly on each supporting material, except on diaper, where vaginal fluid was not analysed since babies do not produce it [16,18]. In addition, 8 stains containing 0.5 mL of urine (1 male and 2 female donors) and 8 stains of 0.5 mL of semen (3 male donors) were analysed. The surface of cotton and the first layer of the superabsorbent pads were measured by ATR-FTIR, using a Thermo Nicolet IS10 with an ATR-FTIR accessory (smart iTR), and the OMNIC software version 9.1.26 (Thermo Fisher Scientific Inc., Massachusetts, USA), according to the methodology described on a previous study [16]. In brief, 8 spots randomly selected within each stain were analysed. Thus, 64 spectra from semen stains, 48 spectra from vaginal fluid stains and 64 spectra from urine stains (placed on a wide variety of supporting materials) were collected in total. Also, unstained cotton and pads were also measured as blank samples (64 spectra)..

2.2. Data treatment and multivariate analysis

Spectra without any previous treatment were imported as a matrix to The Unscrambler X 10.1 software (CAMO Software, Oslo, Norway), to perform a Principal Components analysis (PCA) focussed on studying the interference of pads' spectra on bodily fluids' spectra so as to ascertain the most discriminatory wavelength range. For this, several PCA considering different spectral ranges were performed using blank samples, semen stains, urine stains and vaginal fluid stains (data not shown). The blanks' spectra were totally differentiated from the bodily fluids' spectra and this difference was best verified within the range 1690–1500 cm−1, which is the region with less supporting material spectral interference as demonstrated in previous studies [16] and in which the major bands of each fluid due to proteins are included. Afterwards, the following pre-processing procedure was optimized. The spectra baseline was corrected by selecting: Baseline Offset, i.e. the value of the lowest point in the spectrum was subtracted from all the variables being set as 0, and Linear Baseline Correction, which transformed the sloped baseline into a horizontal baseline [19]. Normalization by range was performed, and, finally, the data was smoothed by Savitzky-Golay method, with a polynomial order of 2 and 11 smoothing points in a symmetric kernel, which helps to reduce the spectral noise [19].

To perform the unsupervised multivariate PCA, the following parameters were chosen: 5 PCs presented, Nonlinear Iterative Partial Least Squares (NIPALS) algorithm, Leverage correction, a fast approximation to cross validation to estimate the prediction error, and 1/SDev standardization, which gives all variables the same variance, i.e. the same chance for each variable to influence the model. This way all spectra influence the model [17,19].

Then, the treated data were imported as an Excel matrix to Matlab R2016b (Mathworks, Massachussets, USA) to automatically and iteratively calculate Pearson's Correlation Coefficients (designated as R) through the "corrcoef" function [20], so a numerical value of the similarity between spectra could be obtained. These R coefficients, which are related to the covariance of the data compared, were calculated according to Eq. (1) [20]:

$$R(i,j) = \frac{C(i,j)}{\sqrt{C(i,i)C(j,j)}}$$

Within Matlab code, this function was running in iterative cycles to compare each sample with all other samples, one at a time. It should be noted that the comparison of each spectrum with itself was carefully avoided since it would have provided a 100% match, causing double-dipping. The results were then given in a data matrix showing the Pearson's correlation coefficient for each pair of spectra. Finally, the frequency of each Pearson's correlation coefficient was calculated and presented in a histogram.

2.3. Evaluation of results through a Bayesian approach

At last, some Bayesian statistical parameters were calculated to evaluate the data analysis process and try to define possible thresholds. For that, the following percentages were calculated: False negative results (FN), which is the percentage of samples wrongly identified as non-semen; False positives (FP), which is the percentage of samples wrongly identified as semen; True positives (TP), which is the percentage of samples correctly identified as semen; and True negatives (TN), which is the percentage of samples correctly identified as non-semen. With these parameters calculated, it was possible to calculate the Likelihood-ratio (LR) by dividing TP by FP. Also, based on these parameters, a ROC Curve was plotted, which shows the trade-off between sensitivity and specificity.

Finally, the following validation parameters were calculated: Accuracy (percentage of correctly identified samples), Sensitivity (True Positive rate), Specificity (True Negative rate) and Precision or Positive Predictive Value (ratio between the correctly identified semen samples and the total number of semen samples). The formulae [21] used to calculate these parameters are presented below:

$Accuracy = (TP + TN)/(TP + TN + FP + FN) \times 100$

$Sensitivity = TP/(TP + FN) \times 100$

$Specificity = TN/(TN + FP) \times 100$

$Precision = TP/(TP + FP) \times 100$

2.4. Proof-of-concept

For proof-of-concept, the 7 different pads and cotton were impregnated with bodily fluids mixtures, accordingly to the method described in Gregório et al. [16]. On sanitary napkin (Ausonia, Evax and Deliplus) and panty-liner (Carefree and Evax) the three fluids were added at 1:1:1 ratio (semen: urine: vaginal fluid). On diapers (Deliplus and Dodot), semen and urine were added at 1:1 ratio (semen: urine).

The samples were measured by ATR-FTIR and 8 spectra from each pad were analysed according to the method described in the present article. A total of 64 spectra were measured from mixture stains and were used as test set. Each mixture spectra was correlated to all semen spectra, one at a time. As described above, the results were given in a data matrix showing the Pearson's correlation coefficient for each pair of spectra, and the frequency of each Pearson's correlation coefficient was calculated. These frequencies were presented in a histogram.

3. Results and discussion

After optimizing the spectral pre-processing and, according to PCA results, selecting the spectral range 1690–1500 cm^{-}1 as optimum for discriminating semen, urine and vaginal fluid, the Pearson's Correlation Coefficient (R) was automatically and iteratively calculated. Since semen is the bodily fluid of interest from which subsequently obtain the DNA, this Pearson's Correlation was calculated after defining two groups of samples (semen/non-semen). This way, all spectra were statistically compared with those from semen in such a way that two groups of correlation data were obtained: the intra-variability of semen stains (R among semen spectra) and the inter-variability between non-semen stains (urine or vaginal fluid) with semen stains (R between non-semen spectra and semen spectra) [22].

These Pearson's correlation values were plotted as histogram in terms of percentage (Figure 1). Each group (semen/ non-semen) was displayed in a different colour. Clearly, the two regions, semen and non-semen, overlap between the Pearson's value 70 and 97, both included. Below 70, only non-semen samples were displayed. Over 97 only samples of semen were displayed. From 70–87, the percentage of semen samples is lower than the percentage of non-semen. From 88-97, the percentage of semen samples is higher than the percentage of non-semen, which implies that those unknown spectra compared with the semen reference spectra with an R over 88 are more likely to be semen instead of non-semen. So, the spectrum of each analysed spot of an unknown stain, after being pre-processed might be statistically compared with the semen samples on the database by calculating its Pearson's Correlation Coefficient.

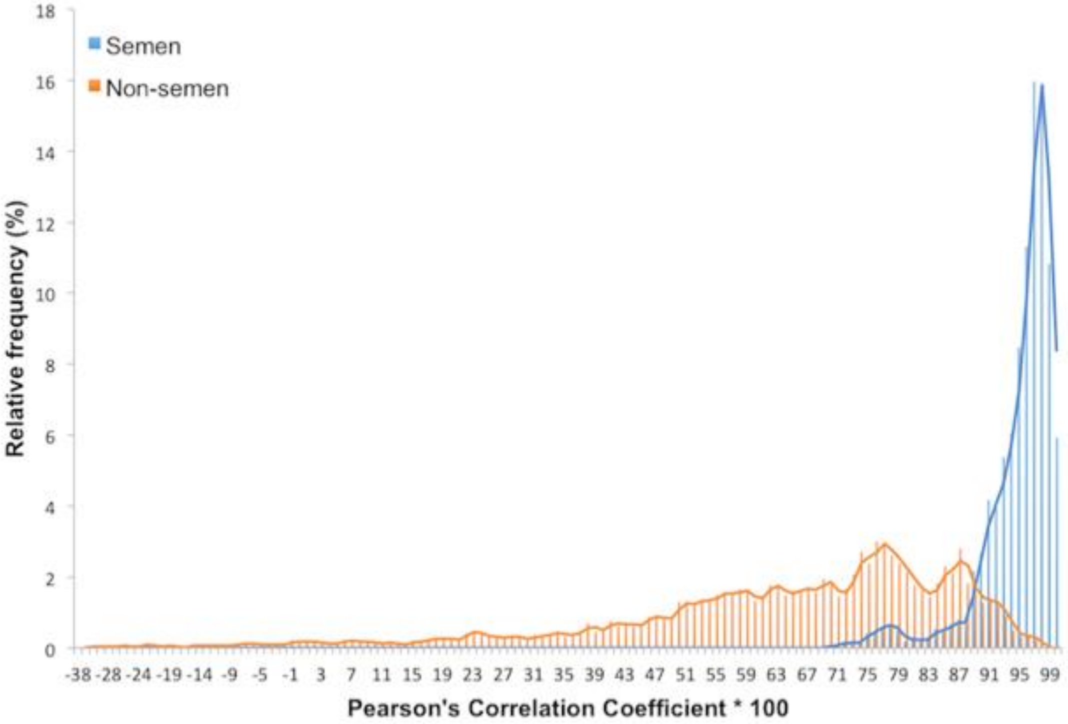


Figure 1. Histogram showing Pearson's Correlation Coefficient frequencies for Semen (Intra-variability) and Non-semen (Inter-variability) samples.

It should be important to define a threshold value that can guarantee the identification of semen with a certain confidence degree. Ideally, and taking into account the forensic screening, the method should be able to detect every sample that contains semen (i.e. those sample's spectra with R > 70). However, a large number of non-semen stains would be also included constituting false positives. Therefore, an evaluation through Bayesian reasoning was performed. For every Pearson value, the rate of TP, FP, TN and FN values were calculated, as well as the LR. The LR is here inferred comparing the probability of correctly identifying a sample as semen (i.e. TP) against the probability of wrongly identifying a non-semen sample as semen (i.e. FP) [23]. In Figure 2, the rates of these parameters and the LR are represented for each Pearson's value between 70 and 100. For instance, the lower the threshold, the higher the TP, but also the lower the TN, which means they are inversely correlated [22]. For instance, selecting a threshold of 70 of Pearson's correlation would detect the 100% of semen samples, but with a percentage of FP within 50% and 30%. From the Pearson's value 88, the percentage of FP is below 10% and decreasing, however the percentage of TP also decreases from 93% until 48%, when Pearson's value is 97. The LR starts to increase above 10 after the Pearson's value 88, and at 94 is already superior to 50, meaning that with a 94 Pearson's value is 50 times more probable that the unknown sample is semen than non-semen.

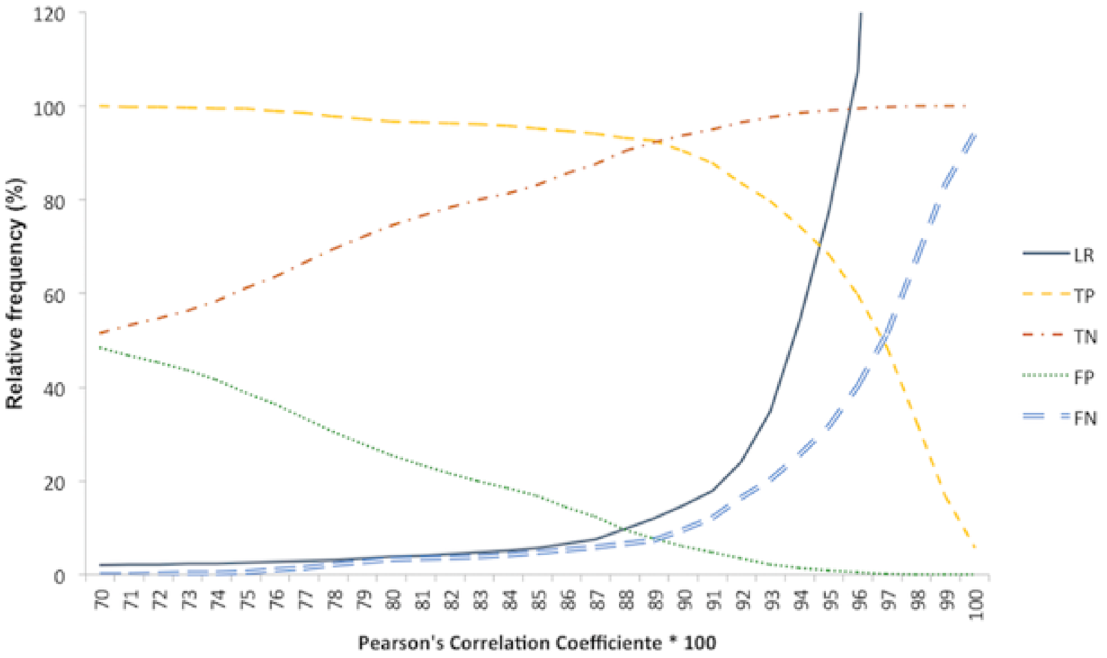


Figure 2. Plot with True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) rates and Likelihood ratio (LR) for each Pearson's Coefficient value.

Through these rates (TP, FP, TN, FN), it was also possible to calculate the qualitative validation parameters of this statistical approach: Accuracy, Sensitivity, Specificity and Precision. To study how the selection of the threshold affects these parameters, the validation parameters were calculated for every Pearson value. As an example, Table 1 summarizes four values in which one or several validation parameters were maximum. These results demonstrate the importance of defining a threshold, finding a balance between these validation parameters. By choosing a low Pearson's correlation value, a high sensitivity is obtained, meaning that the TP rate is high, but all the other validation parameters decrease. If choosing a high value, the specificity increases, as the FP rate decreases, as well as the precision, but the sensitivity of the technique is lost.

Table 1. Pearson's Correlation Coefficient – qualitative validation parameters and their correlation with the confidence grade.

		Pearson's Correlation Coefficient values			
		70	**88**	**96**	**100**
Parameters (%)	**Accuracy**	75.76	91.88	79.61	53.28
	Sensitivity	100.00	93.28	59.77	5.93
	Specificity	51.52	90.47	99.44	100.00
	Precision	67.35	90.73	99.08	100.00

In order to comprehensively visualize these changes along the different thresholds of Pearson value, the ROC Curve of sensitivity in function of the 100-Specificity (Figure 3) might be studied. As previously seen in Figure 2, the true positive rate (sensitivity) is much higher than the false positive rate (100-specificity) for every Pearson value, which makes the ROC curve closer to the upper left corner. The area under the curve suggests that this analysing approach has a good accuracy, meaning it could presumptively differentiate the samples being tested into semen and non-semen.

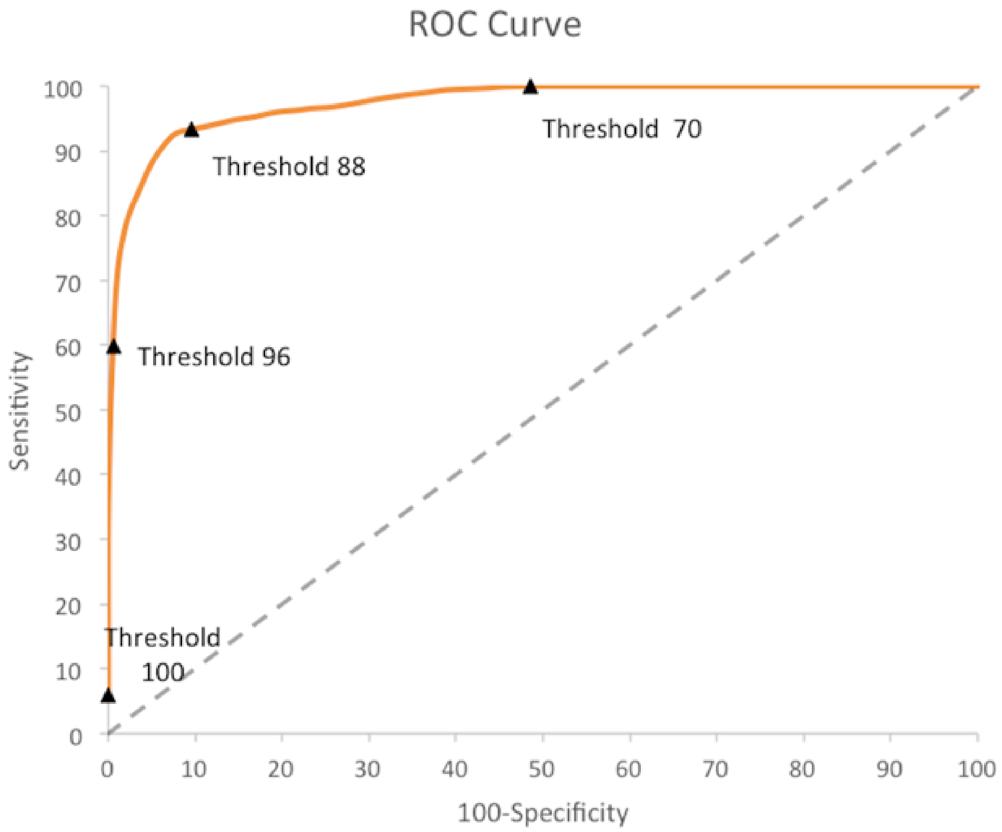


Figure 3. ROC Curve: sensitivity plotted in function of the 100-specificity, including some threshold Pearson values (70, 88, 96 and 100) as an example.

In a further step, the purpose of the analysis must be considered: to answer the question “Does this unknown sample contain semen?”, the lowest threshold, (according to our database a value of 70) should be considered. This way, the analyst would be able to detect every sample containing semen as semen, although this would increase also the number of false positives. In conclusion, a low threshold should be considered for an initial screening of the sample in order to avoid missing any evidence containing semen. However, considering that sexual evidence usually contains mixtures of the victim's and the aggressor's bodily fluids, and that it is common that the female DNA overlaps the male DNA during the DNA profiling, hindering it [9–11], it would be interesting to analyse different spots of the stain with the aim of detecting those with higher concentration of semen. For this second aim, the threshold should be higher (e.g., equal or superior to 94), so it would have a higher precision and confidence that those spots contain semen in a large proportion.

A preliminary proof-of-concept was performed, by analysing the ATR-FTIR spectra of mixtures of bodily fluids on the different supporting materials in different spots. The results demonstrated that it was possible to presumptively discriminate those spots containing larger proportion of semen because they provided higher Pearson Correlation Coefficient, as displayed in the two examples of Figure 4. According to this Figure, the spot “a” provided much higher Pearson Correlation values when compared to semen than spot “b”, i.e. the histogram of spot “a” was included within the range of semen samples whereas the histogram of spot “b” was included within the range of non-semen samples. This finding clearly evidences the major ratio of semen in spot “a”. Thus, spot “a” should be selected to perform the DNA profiling. Interestingly, it was also checked that almost all spots from all the mixtures analysed were positive to semen, considering the Pearson's correlation value of 70 as threshold, (i.e. highest sensitivity).

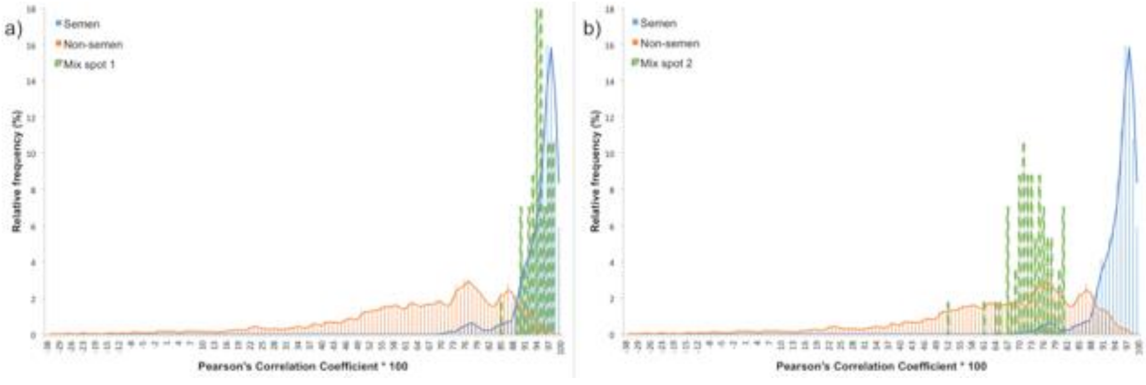


Figure 4. Two examples of the presumptive results of analysing a spot presenting high semen ratio (a) and a spot presenting low semen ratio (b) on a sanitary napkin.

Finally, this analysis approach was applied to the spectral data obtained by ATR-FTIR [16] and by External Reflection FTIR [17] from semen, urine and vaginal fluid stains prepared on 100% cotton cloths in previous studies, with demonstrated efficacy (data not shown). Interestingly, the Pearson correlation was performed against two different spectral databases: the whole database containing stains of bodily fluids placed on all tested materials (sanitary napkins, pantyliners, diapers and also cotton) and a new one containing only stains of bodily fluids on cotton. As it was expected, better results were obtained using the database that only contained the stains on cotton. Therefore, two databases are recommended, one for stains on cotton and another for stains on superabsorbent pads, for a better accuracy. Furthermore, an ideal method would be that one which only considers the stains of bodily fluids over the same material as the unknown stain to perform the Pearson correlation (i.e. a particular spectral database for each type of pad).

4. Conclusions and future trends

The combination of the analytical vibrational spectroscopy and the statistical analysis of spectral data here described shows a high discriminatory power in a non-destructively and presumptively way, so necessary in the forensic field of sexual aggressions to go beyond detection, and to be able to differentiate bodily fluids, specially semen, on mixtures, very usual in sexual aggressions; although, larger spectral databases with a higher number of spectra from stains of semen and other bodily fluids need to be created to refine and escalate the model. As a probabilistic method, the more number of elements, the better and more realistic result.

In addition, one future trend to improve this statistical approach will be the preparation and analysis of mixtures spots with known and different concentrations of bodily fluids. This way, it might be determined the exact ratio of semen which is being detected.

This data analysis approach may be applied in hospitals to perform an initial screening of samples previous to send evidence to the forensic laboratories. It will have economic impact, as the pre-selection of samples will allow freeing resources and manpower within the overloaded forensic laboratories; but also, it will have great social impact, as even samples without sperm cells may be presented as semen with an objective numeric probabilistic confidence to the jury. This low-cost automated approach may be also applied to other fields, such as clinical diagnosis.

Finally, this technique may be easily translated to conflict and poor zones worldwide. For this, it would be ideal the combination of a portable ATR-FTIR with software able to perform the analysis here described in an automated and non-destructive manner.

Declaration of interest

The authors declare they have no competing interests.

Acknowledgements

The authors thank IUICP (University Institute of Research of Forensic Sciences) for the project IUICP-2016-002. Félix Zapata also thanks the Spanish Ministry of Education (FPU14/00790) for his Ph.D. research grant.

References

[1] P. Gill, A.J. Jeffreys, D.J. Werrett, Forensic application of DNA 'fingerprints', Nature 318 (1985) 577–579.

[2] R.M. Jobin, M. De Gouffe, The persistence of seminal constituents on panties after laundering. Significance to investigations of sexual assault, Can. Soc. Forensic Sci. J. 36 (2003) 1–10.

[3] F. Zapata, I. Gregório, C. García-Ruiz, Body fluids and spectroscopic techniques in forensics: a perfect match?, J. Forensic Med. 1 (2015).

[4] H. Yang, B. Zhou, H. Deng, M. Prinz, D. Siegel, Body fluid identification by mass spectrometry, Int. J. Leg. Med. 127 (2013) 1065–1077.

[5] J.M. Butler, Advanced Topics in Forensic DNA Typing: Methodology, Elsevier Academic Press, San Diego, 2012.

[6] S.A. Harbison, R.I. Fleming, Forensic body fluid identification: state of the art, Res. Rep. Forensic Med. Sci. 6 (2016).

[7] F. Zapata, M.A. Fernández de la Ossa, C. García-Ruiz, Emerging spectrometric techniques for the forensic analysis of body fluids, Trends Anal. Chem. 64 (2015) 53–63.

[8] K. Virkler, I.K. Lednev, Analysis of body fluids for forensic purposes: from laboratory testing to non-destructive rapid confirmatory identification at a crime scene, Forensic Sci. Int. 188 (2009) 1–17.

[9] R.C. Giles, Improved methods for the elution and extraction of spermatozoa from sexual assault swabs, Forensic Mag. 5 (2008) (14, 16, 19, 20 to 21).

[10] P. Wiegand, M. Schurenkamp, U. Schutte, DNA extraction from mixtures of body fluid using mild preferential lysis, Int. J. Leg. Med. 104 (1992) 359–360.

[11] N. Hu, B. Cong, S. Li, C. Ma, L. Fu, X. Zhang, Current developments in forensic interpretation of mixed DNA samples (Review), Biomed. Rep. 2 (2014) 309–316.

[12] J.M. Butler, Advanced Topics in Forensic DNA Typing: Interpretation, Elsevier Academic Press, San Diego, 2015.

[13] K. Virkler, I.K. Lednev, Raman spectroscopy offers great potential for the non-destructive confirmatory identification of body fluids, Forensic Sci. Int. 181 (2008) e1–e5.

[14] K.M. Elkins, Rapid presumptive "fingerprinting" of body fluids and materials by ATR FT-IR spectroscopy, J. Forensic Sci. 56 (2011) 1580–1587.

[15] C.M. Orphanou, The detection and discrimination of human body fluids using ATR FT-IR spectroscopy, Forensic Sci. Int. 252 (2015) e10–e16.

[16] I. Gregorio, F. Zapata, C. Garcia-Ruiz, Analysis of human bodily fluids on superabsorbent pads by ATR-FTIR, Talanta 162 (2017) 634–640.

[17] F. Zapata, M.A. de la Ossa, C. García-Ruiz, Differentiation of body fluid stains on fabrics by External Reflection Fourier Transform Infrared Spectroscopy (FTIR) and Chemometrics, Appl. Spectrosc. (2016).

[18] S.L. Elvik, Vaginal discharge in the prepubertal girl, J. Pediatr. Health Care: Off. Publ. Natl. Assoc. Pediatr. Nurse Assoc. Pract. 4 (1990) 181–185.

[19] CAMO, The Unscrambler X Help contents, Oslo, Norway, 2009.

[20] Matworks, Correlation Coefficient by Matlab.

[21] E. Szymánska, J. Gerretzen, J. Engel, B. Geurts, L. Blanchet, L.M.C. Buydens, Chemometrics and qualitative analysis have a vibrant relationship, Trends Anal. Chem. 69 (2015) 34–51.

[22] C. Muehlethaler, G. Massonnet, T. Hicks, Evaluation of infrared spectra analyses using a likelihood ratio approach: a practical example of spray paint examination, Sci. Justice: J. Forensic Sci. Soc. 56 (2016) 61–72.

[23] P. Danaher, R.L. White, E.K. Hanson, J. Ballantyne, Facile semi-automated forensic body fluid identification by multiplex solution hybridization of NanoString(R) barcode probes to specific mRNA targets, Forensic Sci. Int.: Genet. 14 (2015) 18–30.