Analyst

RSC Publishing

PAPER

FTIR spectroscopy of biofluids revisited: an automated approach to spectral biomarker identification†‡

Cite this: *Analyst*, 2013, **138**, 4092

Julian Ollesch,*[a] Steffen L. Drees,[a] H. Michael Heise,[a] Thomas Behrens,[b] Thomas Brüning[b] and Klaus Gerwert*[a]

The extraction of disease specific information from Fourier transform infrared (FTIR) spectra of human body fluids demands the highest standards of accuracy and reproducibility of measurements because the expected spectral differences between healthy and diseased subjects are very small in relation to a large background absorbance of the whole sample. Here, we demonstrate that with the increased sensitivity of modern FTIR spectrometers, automatisation of sample preparation and modern bioinformatics, it is possible to identify and validate spectral biomarker candidates for distinguishing between urinary bladder cancer (UBC) and inflammation in suspected bladder cancer patients. The current dataset contains spectra of blood serum and plasma samples of 135 patients. All patients underwent cytology and pathological biopsy characterization to distinguish between patients without UBC (46) and confirmed UBC cases (89). A minimally invasive blood test could spare control patients a repeated cystoscopy including a transurethral biopsy, and three-day stationary hospitalisation. Blood serum, EDTA and citrate plasma were collected from each patient and processed following predefined strict standard operating procedures. Highly reproducible dry films were obtained by spotting sub-nanoliter biofluid droplets in defined patterns, which were compared and optimized. Particular attention was paid to the automatisation of sample preparation and spectral preprocessing to exclude errors by manual handling. Spectral biomarker candidates were identified from absorbance spectra and their 1st and 2nd derivative spectra using an advanced Random Forest (RF) approach. It turned out that the 2nd derivative spectra were most useful for classification. Repeat validation on 21% of the dataset not included in predictor training with Linear Discriminant Analysis (LDA) classifiers and Random Forests (RFs) yielded a sensitivity of 93 ± 10% and a specificity of 46 ± 18% for bladder cancer. The low specificity can be most likely attributed to the unbalanced and small number of control samples. Using this approach, spectral biomarker candidates in blood-derived biofluids were identified, which allow us to distinguish between cancer and inflammation, but the observed differences were tiny. Obviously, a much larger sample number has to be investigated to reliably validate such candidates.

Received 17th February 2013
Accepted 8th May 2013

DOI: 10.1039/c3an00337j

www.rsc.org/analyst

Introduction

Fourier transform infrared (FTIR) spectroscopic analysis has been applied for many clinical chemistry applications and proposed for medical diagnosis.[1–3] The infrared absorbance spectrum of a body fluid represents a fingerprint-like integral biochemical status of a patient's sample. The advantage is that no additional markers or labelling are required, and multiplex parameters of the proteome, lipidome, and metabolome are recorded at once. By application of suitable bioinformatics, not only a multi-parameter clinical analysis can be achieved by a single and fast measurement,[4] but also markers for diseases can be extracted from specific spectral band patterns.[5–13]

In the following, we present an automatized FTIR spectroscopic setup of our study on the identification of blood-borne spectral bladder cancer marker candidates. This approach was applied to identify the expected tiny changes in FTIR spectra of blood that may be caused by only a few thousand tumour cells. Previously observed limitations of spectroscopic sample preparation and experimental errors were avoided by using advanced standardized procedures. Particular care was exercised to minimize the user interaction with samples and spectra

[a]*Protein Research Unit Ruhr within Europe (PURE), Ruhr-University Bochum, Department of Biophysics ND04-596, Universitätsstrasse 150, 44780 Bochum, Germany. E-mail: ollesch@bph.rub.de; gerwert@bph.rub.de; Fax: +49 (0)234 32 14849; Tel: +49 (0)234 32 29832*

[b]*Protein Research Unit Ruhr within Europe (PURE), Institute for Prevention and Occupational Medicine of the German Social Accident Insurance, Institute of the Ruhr-Universität Bochum (IPA), Bürkle-de-la-Camp Platz 1, D-44789 Bochum, Germany*

† This paper is part of an *Analyst* themed issue on Optical Diagnosis. The issue includes work which was presented at SPEC 2012 Shedding New Light on Disease, which was held in Chiang Mai, Thailand, November 11–17, 2012.

‡ Electronic supplementary information (ESI) available. See DOI: 10.1039/c3an00337j

by predominant automation. Spectral key procedures such as *e.g.*, 'visual inspection' of data for outliers, or 'manual baseline correction', usually depend on an individual, subjective assessment, which finally influences the prediction outcome. For the ultimate credibility, reproducibility, ease of use and speed, the sample analysis was standardized and automated as far as reasonably achievable.[14]

The specific aim of the study was to distinguish between urinary bladder cancer (UBC) and non-bladder cancer patients based on FTIR spectroscopy of blood samples from suspected bladder cancer patients. UBC is one of the most frequent tumours in the worldwide population, affecting men approximately three times more often than women.[15,16] Risk factors include smoking and occupational exposure to chemical toxins, whereas in developing countries it may also be fostered by infectious diseases.[15–18] Upon early detection and surgery, a high five-year survival rate above 72% is observed.[19] A variety of blood borne marker candidates for UBC has been discussed,[20–28] so that a combination of marker molecules may produce a detectable fingerprint in the mid-infrared spectra of blood serum and plasma.

Cystoscopy by urologists outside of our study substantiated the initial cancer suspicion, and a further cystoscopy for a transurethral resection (TUR) of urinary bladder tissue for a pathological examination was indicated. TUR is an uncomfortable procedure bearing the risks of bleeding, inflammation, thrombosis, embolism, bladder perforation or stricture of the urethra. Usually, three to four days of stationary hospitalization are required. Cystoscopy alone is limited with respect to the detection of particularly early stages of bladder cancer.[29] The patients of the control group mainly suffered from a urinary tract infection, which can be treated with antibiotics. Therefore, a negative non-invasive urine or blood-based test could spare non-cancer patients an unnecessary surgical treatment, the accompanying risks, and avoid hospitalization. Such tests could support the established clinical diagnostics with additional indicators, eventually enabling an earlier onset of therapy.

As the gold standard, UBC cases were confirmed, graded and staged with cytology and histopathology on the collected tissue samples. All control patients were subjected to the appropriate treatment and aftercare outside the frame of this study. Two prostate cancer cases were included because of their clinical relevance; intruding prostate cancer tissue into the bladder is a phenomenon often observed. Blood samples were drawn before TUR. Strict standard operating procedures were defined to ensure unique standards in the sampling process, serum and plasma preparation, transport and sample storage. Patient background, medication and blood status were documented according to the standards of Good Epidemiological Practice.

A specific pattern has to be detected among the immense biological variability of abundant substances as found for blood proteins or metabolites. For an advanced blood-based assessment, the spectra of serum and two plasma preparations were combined into one patient-representative feature vector, along with the respective 1st and 2nd derivative spectra that may reveal subtle band shifts. Differences between serum and plasma are expected due to coagulation *versus* coagulation prevention during preparation. Blood coagulation inhibitors may mask specific band components, justifying two different plasma preparations.

With a limited number of test subjects, dedicated feature selection methods are of particular importance for diagnostic purposes. By selecting classification relevant spectral features, abundant uninformative, uncorrelated variables are removed. Thus, the dimensionality of the classification problem can be reduced, by which the chance of classifier overfitting is lowered as well. In Disease Pattern Recognition (DPR), the feature selection is the crucial process by which the subtle differences between spectra of the disease status are identified.[3]

For the prediction of the patient status based on the selected features, we used a classifying linear discriminant analysis (LDA) as a suitable method requiring only low computer processing power, and a complex random forest (RF) ensemble classifier, which is computationally intensive. Both classifiers were able to predict a patient's disease status with comparable quality.

An RF classifier consists of a collection of decision trees, where each vertex separates the feature space based on a random choice of features. Whenever a feature constitutes the splitting feature at a vertex in a decision tree, one can determine the gain in information either based on entropy or based on the so-called Gini importance, the latter of which is particularly popular for random-forest based feature selection.[30–33] Consequently, we setup one type of RF classifier for the exclusion of most classification irrelevant features, and one ensemble RF classifier for disease status prediction.

The identified spectral patterns were thoroughly validated to assess their diagnostic value. As a compromise to our relatively small dataset of 135 patients, all selection and classification procedures were kept in strict Monte Carlo cross validation (MCCV) schemes with random dataset splits into training and independent test sets to avoid the false detection of dataset-specific randomly correlating features. The final predictors were validated on further MC derived independent test sets. We are aware of the limitations for generalization,[34] but are convinced that the dataset of our current bladder cancer study allows a preliminary evaluation along these lines.

In our study, the technical problems of sampling, sample preparation, spectroscopic measurement, data preprocessing and feature selection were addressed. A conclusive strategy for highly reproducible, automated high throughput FTIR spectroscopy with dedicated equipment and user-independent algorithms was developed, and the results for this special demanding DPR study are reported.

Experimental

The workflow from patient samples to classification is schematically summarized in Fig. 1. Blood was sampled, processed, and sample substrates were robotically prepared. Absorbance spectra were collected and preprocessed by automated procedures.

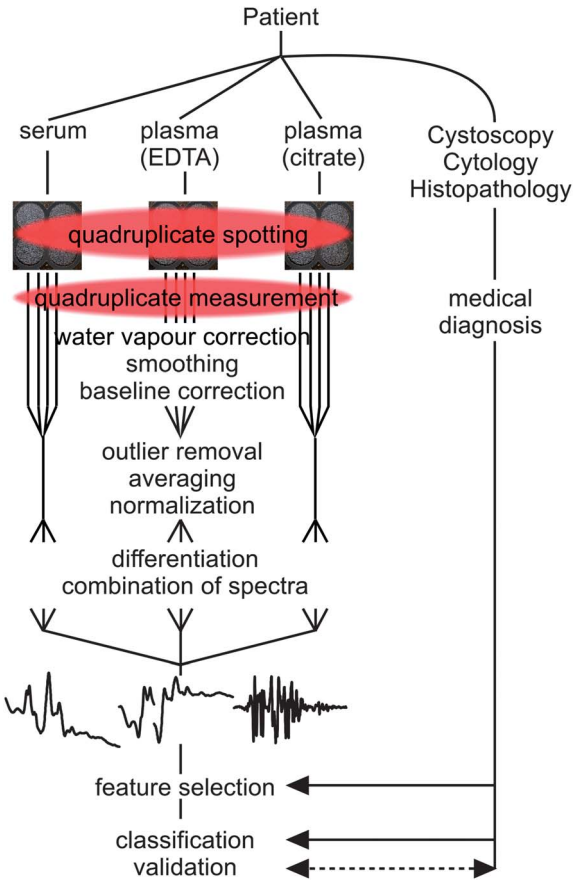

Fig. 1 Scheme that was followed to process each patient's sample. Spectra of three blood preparations of each patient were measured and processed in quadruplicate until outlier removal and averaging. 1st and 2nd derivatives were calculated and concatenated to form a representative complex spectrum of a patient's biochemical blood status. Class assignment obtained by the medical gold standard was used for feature selection and classification.

Blood sample preparation

A total number of 108 men and 27 women participated in the study. The dataset consisted of 89 UBC patients (73.1 ± 11.2 years of age, 72 men, 17 women), of which 30 were recurrent cases (1 urothelial papilloma, 38 G1, 28 G2, 20 G3-G4 (WHO 1973); or according to WHO 2004: 1 urothelial papilloma, 1 papillary urothelial neoplasm of low malignant potential (PUNLMP), 61 low grade, 23 high grade). For nine cancer patients only one of both routine gradings[35–37] was available. Stagings were determined in 79 cases: 42 Ta, 23 T1, 5 T2, 8 T2a, and 1 T4a. The 46 non-bladder cancer controls (72.5 ± 10.8 years of age, 36 men, 10 women) had the following case history: 40 *cystitis cystica*, 1 *cystitis lymphofollicularis*, 1 urethritis, 2 open bladder tumor resections, 2 glandular prostate carcinoma, and one inverse urothelial papilloma.

Blood samples were collected and processed to serum, ethylene diamine tetraacetic acid (EDTA), and sodium citrate stabilized plasma (BD Biosciences, Heidelberg, Germany) at the Marien-Hospital Herne (Herne, Germany) by study nurses following standard operating procedures (SOPs). The SOPs had been developed with the Scientific Epidemiological Study Centre of the IPA (Institute for Prevention and Occupational Medicine of the German Social Accident Insurance, Institute of the Ruhr-Universität Bochum, Germany, member of the research initiative PURE). Processed blood samples were shock frozen within minimal time lapse of less than 30 min for plasma, and 50 min after sampling for serum. Serum and plasma samples were delivered as 400 µl aliquots and stored at −80 °C until experimental use.

For preparation, samples were thawed at 4 °C and 47 µl of the respective liquid were added to 3 µl of filtered KSCN solution (0.5 M), yielding a KSCN spike of 30 mM concentration for FTIR quantification. The samples were then mixed at 1000 rpm for 1 minute and centrifuged for 30 s at 2000 × *g*. A volume of 15 µl of each sample was transferred onto 384 well microtiter plates (Greiner Bio-one GmbH, Frickenhausen, Germany), which were sealed immediately with an adhesive aluminium foil (Greiner) to prevent evaporation. Sealed plates were centrifuged at 2000 × *g* at 10 °C for 2 min to settle the liquid within the wells and to remove air bubbles.

Sample spotting

A compact, benchtop sized robotic spotting system (Instrument2, M2 Automation GmbH, Berlin, Germany) was used to dispense the samples in quadruplicate on 384-well silicon multi-well titer plate (MTP) substrates (Bruker Optics, Ettlingen, Germany). Substrates were cleaned with sodium hypochlorite solution and plasma treatment (Zepto, Diener plasma surface technology, Ebhausen, Germany) before use. Each well covers a circular area of 4 mm diameter, which was almost perfectly covered with the sample. A total volume of 3 µl was loaded at a syringe-pump controlled speed into the dispenser. The single sample film was formed from approximately 50 nl. Subsequently, the sample loaded substrate plates were vacuum-dried for 10 min.

Transmission FTIR spectroscopy

For each measurement, an inner diameter spot of 3 mm was transilluminated at each MTP well position. The measurement was started immediately after vacuum drying in transmission mode at ambient temperature (22 ± 1 °C) on a Vertex 70V vacuum FTIR spectrometer with the HTS-XT MTP reader extension and the Twister robotic plate feeder (Bruker Optics). All parts of the optical path that could not be evacuated were thoroughly dry-air purged (25 l min^{-1}, Parker-Balston, Parker-Hannifin, MA, USA). Interferogram acquisition was double sided – forward/backward with the internal deuterated triglycine sulphate (DTGS) detector; 64 scans were accumulated before Fourier transformation with a spectral resolution of 2 cm^{-1}. Blackman-Harris-3-term apodization, Mertz phase correction and 4× zero filling was applied. An average signal to rms-noise ratio (*S*/*N*) of 20 400 : 1 (root mean square, rms) was determined with the OPUS software for interval data of 2100–1900 cm^{-1} on 20 randomly selected raw spectra of blank silicon MTP wells.

Spectral preprocessing

Water vapour lines from atmospheric absorption were corrected using scaled subtraction.[38] A Fourier transform based low pass

filter with an appropriate Gaussian apodization function removed noise with full width at half height <4 cm^{-1} as reported in earlier studies.[39,40] Outliers were removed based on the Pearson-correlation of the quadruplicate spectra to their median, calculated for the CH-stretching and the fingerprint spectral regions. The two closest spectra were averaged, and the less correlated spectra were discarded. For an adaptive baseline correction, spectra were split up into five overlapping sections (875–2010 cm^{-1}: fingerprint and amide I/II; 1860–2200 cm^{-1}: KSCN; 2100–2700 cm^{-1}: signal free; 2250–3700 cm^{-1}: CH-stretch/amide A; 3650–4000 cm^{-1}: signal free), each of which was then fitted with an individually parameterized adaptive iteratively reweighted penalized least squares (airPLS) baseline.[41–43] Details about the parameterization are given in the ESI.‡ For a concatenated dataset of serum and plasma spectra, the absorbance spectra were min–max–normalized on the KSCN peak between 2150 and 2050 cm^{-1}. The 1st and 2nd derivatives were calculated by Fourier expansion with a Gaussian function for low pass filtering at 6 and 8 cm^{-1} cut off, respectively. The set of 1st derivative spectra was linearly scaled up to a maximum absolute amplitude of 0.5 maintaining interspectral relations. Similarly, the set of 2nd derivative spectra was consistently scaled up to a respective maximum absolute amplitude value of 0.25.

The absorbance spectra of serum, EDTA plasma and citrate plasma were concatenated with the respective 1st and 2nd derivative spectra into one extended feature vector for each patient. The final dataset comprised the intervals of 3200–2800 cm^{-1} and 1750–875 cm^{-1}, covering spectral fingerprint features of all relevant biomolecules as found in blood above noise-level in absorbance at each 11 493 wavenumber variables (Fig. S2–S4‡).

Attenuated total reflection setup

Attenuated total reflection (ATR) absorbance spectra for comparison with transmission spectra (see Fig. 2) were recorded on a Bruker IFS66 spectrometer equipped with a DuraSamplIR II diamond μATR accessory (Smith Detection, Edgewood, MD, USA). The absorbance of 5 μl serum was recorded against a distilled water background. A sample of 5 μl of 1 : 20 diluted serum was dried under a slow N_2 flow and repeatedly recorded *versus* the blank ATR element until no further spectral changes were observed. A total of 512 bidirectional double sided interferograms were co-added at a resolution of 4 cm^{-1}. Again, Blackman-Harris-3-term apodization, Mertz phase correction and 4× zero filling were applied.

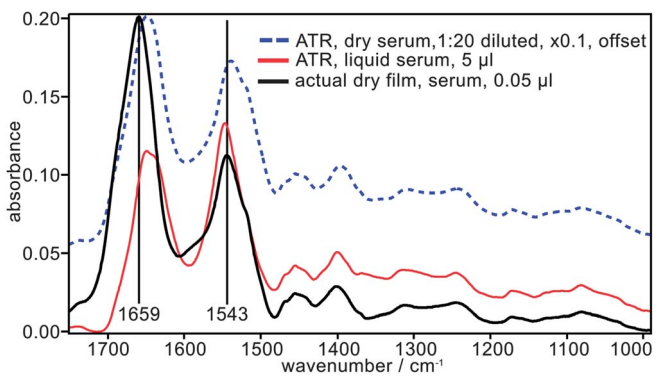

Fig. 2 ATR-spectra of dry (dashed blue) and liquid serum (red), respectively, and a dry-film spectrum measured in transmission (black) are shown for comparison with obvious band shifts and different band shapes. The serum ATR spectra are affected by dispersion effects, whereas the solution spectrum suffers from water absorbance overcompensation as measured *versus* a water background.

Atomic force microscopy

The surface characteristics of a robotically deposited sample spot on the Bruker silicon substrate (approximately 200 μm in diameter) were determined by atomic force microscopy (AFM, Witec alpha 300 AR, Ulm, Germany). An area of 65 × 65 $μm^2$ of a single spot, *i.e.* approximately a quarter, was scanned with a lateral resolution of 0.15 μm, whereas 80 × 200 $μm^2$ of the contact area of overlapping spots was measured in tapping mode with a lateral resolution of 0.625 μm. The removal of bad scan lines, offset correction and the preparation of graphics were done using the Gwyddion software (Version 2.30).

Random forests for feature selection and classification

RF classifiers were used as single forests and as an ensemble classifier of 1001 RFs. Following the theoretical considerations and practical observations on settings[33] with few data points in a very high dimensional space, the RF classifiers were parameterized as follows: the number of trees per forest was three times the number of features with a maximum of 5000, if more than 1666 features were present. The number of features randomly selected for the split of each tree node was a third of the number of features rounded up to the nearest integer. All other tunable parameters were set by unaltered default of the routines available for download (http://code.google.com/p/randomforest-matlab/, January 30, 2013).

Each RF was trained and validated on an individual MC derived data subset. Our MC algorithm arranged the same number of patients per class for the validation dataset. For feature selection, the total dataset of 135 patients was split into 12 randomly selected sets of 125 (84 UBC, 41 controls). On these, 192 further training-validation pairs of 69/15 UBC and 26/15 control patients were generated. 192 RFs were trained and validated on these, the average classification error rate was registered, and the Gini importance values of the features used in the 192 RFs were accumulated. After removal of the 20% least important features based on their Gini importance, 192 further RFs were evaluated on 192 new MC based cross-validation (MCCV) datasets, until only 4 features remained. The set of features producing the lowest misclassification rate was selected as the optimum set. Then, eleven additional cycles were calculated using the other randomly selected groups of 125 patients.

For calculations with reduced dimensionality using the concatenated absorbance spectra of all three biofluids only, 10 total cycles were calculated on randomly selected subsets of 125 patients with the identical algorithm. The same procedure was carried out separately for first and second derivative spectra with 11 and 13 total cycles, respectively.

The selection frequency of individual features was used as a superior selection criterion. The average classification accuracy of 50 LDA classifiers in a leave-28-out MCCV on the total 135 patients, based on the feature selection frequency under the scheme of a stepwise decreased threshold, was determined. The feature set resulting in the highest average classifier accuracy was selected for additional RF evaluation (see Tables 2 and 3).

The ensemble classifier for disease state prediction consisted of 1001 RFs, of which each was validated in an individual MCCV (107 training, 14 test samples) to perform with an error rate of less than 50%. The majority vote of the 1001 RFs was used as ensemble classifier prediction.

Bioinformatics environment

Random forest calculations were performed within the Matlab environment, Version 2012a with the R-project based[44] Matlab port (as found on http://code.google.com/p/randomforest-matlab/, January 30, 2013) on a High-Performance Computing Server Supermicro SYS-5086B with 8× Intel® Xeon® Westmere EX (E7-8837, 2.66 GHz, 8-Core), 512 GB RAM. Linear discriminant analysis (LDA) was performed with the internal Matlab function ('*classify*') with a quadratic discriminant function. The *a priori* class membership probability was empirically calculated for taking into account the different number of control and UBC patients. Final predictor training was performed on an office PC equipped with Intel Core2Quad CPU Q9650@3.0 GHz, 8 GB RAM (Dell Optiplex 780) running Matlab 2012a.

Results and discussion

The task to collect infrared spectra from liquid samples appears simple at first sight. However, the optimization with regard to analyte detection limit, spectral reproducibility, and speed is a demanding challenge.

First, the most suitable measurement technique has to be selected. There are basically two options for the acquisition of absorbance spectra: attenuated total reflection (ATR) and transmission, which have been frequently described in the literature. Both are suitable for liquid and dry samples with particular limitations and advantages. Here, we face two further alternatives, *i.e.* to sample spectra of the liquid as-is, or to develop advanced preparations, for which the available sample volumes may also be of decisive importance.

Transmission *versus* ATR, liquid or dried?

A simple, very reproducible method is to collect spectra from a drop of body fluid on a micro-ATR device as demonstrated in previous studies.[4,13] Some drawbacks are inherent to this methodology using liquid samples, as the water compensation is imperfect, the detection limit is relatively high because only an approximately micrometer depth of the sample volume is detected, and spectra suffer from dispersion effects. Sample volumes for dry-film preparations were even reduced to sub-microliter volumes using microdispensing devices,[45,46] but the suggested technology was not appropriate for high throughput application.[45] A further limitation was still sample inhomogeneity with crater-shaped dried serum, although this could be partly reduced by the limited ATR probing depth. A different approach was reported recently, when a dried film prepared from 100 μl of plasma or serum was scanned at different sample locations with a microscopic ATR tip,[13] which has also found its use in the low-spatial-resolution analysis of tissue.[47] This approach avoided the possible thickness inhomogeneities of the dried samples with the comparably small (0.0625 mm^2) detecting area of the μATR crystal.

Another solution is to analyse the body fluid with a flow-through cuvette of a fixed path length.[48] This setup suffers again from imperfect water compensation and a high detection limit, due to the chosen path lengths of 5–30 μm, with the consequence that some spectral intervals cannot be analysed due to strong water absorption. An advanced liquid handling system is also required for high throughput application,[49] and cleaning the system between measurements, particularly from adsorbed proteins, requires dedicated and elaborate solutions.[50]

In contrast, transmission spectra of vacuum-dried serum films offer an excellent signal to noise ratio with neither non-compensated water artefacts nor nonlinear absorbance effects (Fig. 2).

High throughput capability

To employ transmission measurements with dried samples, flow cells are to be replaced with a movable solid infrared-transparent sample substrate. Automated sample changers have been developed by several companies (Bruker, PIKE, Specac) and were used in previous research.[5] With those, the sample preparation and the time-consuming cleaning procedure of the substrate are decoupled from the actual spectroscopic measurement. Those setups are usually free of water-associated interference, because the prepared samples can simply be vacuum dried and stored water-free until measurement, which is optimally done in a dry gas-purged or evacuated compartment. The achievable detection limit is low, because the relative large sample volumes condense to highly concentrated, thin films upon water evaporation. After water removal, a relatively large area of the film can be analysed. Spectral artefacts due to infrared transmission are reduced to a minimum and can generally be explained with established models.[51–53] The exchangeable sample substrate is ideal for high throughput solutions, because the modern standard MTP formats can be maintained using the appropriate substrate shape. Due to the mentioned separation of preparative steps from the spectroscopic measurement, the parallel processing of sample substrates can be automated and optimized individually for each biofluid.

Reproducible homogeneous dry films of biofluids

The deposition of a defined volume of a liquid sample onto an infrared transparent substrate appears to be an easy task. It is only necessary to spread the same amount of liquid reproducibly onto one position of the substrate, and then to position the sample reproducibly into the spectrometer beam.

Unfortunately, these three demands are hardly met by manual pipetting devices. All positioning has to happen within

the tolerance of the prepared flat dry films. Pipetting precision in the lower nanoliter range is crucial, because body fluids are highly concentrated solutions of proteins, salt, lipids, and carbohydrates to name only the most abundant compounds. The total protein as found in blood plasma with an average concentration of approximately 70 mg ml^{-1} produces the major absorbance. A mass difference of 1 ng would originate from a 14 pl difference of dispensed volume.

By drying a droplet of such a biofluid, structures termed "coffee rings" are formed,[54] which exhibit an inhomogeneous distribution within the substance film. With insufficient positioning accuracy, recording a spectrum particularly of the droplet border area was shown to produce spectral artefacts, due to a varying film thickness and locally wedge-like geometry of the crater edges.[55,56] Thus, the approach to reproducibly spread an oversized (compared to the detection beam radius) droplet on the MTP substrate aided by robotics is an appropriate compromise to avoid its usual coffee-ring topography.[14] However, subtle variations of positioning and the sample drying process led to a detectable spectral variance in our manually spread samples (see Fig. 3).

One different approach to deposit a minute amount of solution is a nebulizer, which still produced wedge-shaped sample film edges.[57] In two further approaches small droplets were used for dry-film preparation, but the capability to print patterns of sub-nanoliter volumes was not evaluated. In combination with high performance liquid chromatography (HPLC), spots of the analyte were applied to an infrared compatible substrate and further analysed using FTIR microscopy.[58–60] Also, nanoliter volumes were evaluated for the quantitative determination of glucose.[45,46] Particularly, dried down spots of the lowest volumes used in those previous projects appeared to be the most homogeneous, as far as could be judged by the shown topography. Arranging small sample volumes for dry-film patterns on the wells of a MTP substrate was proposed and patented,[61] but not yet evaluated for its use in FTIR spectroscopy.

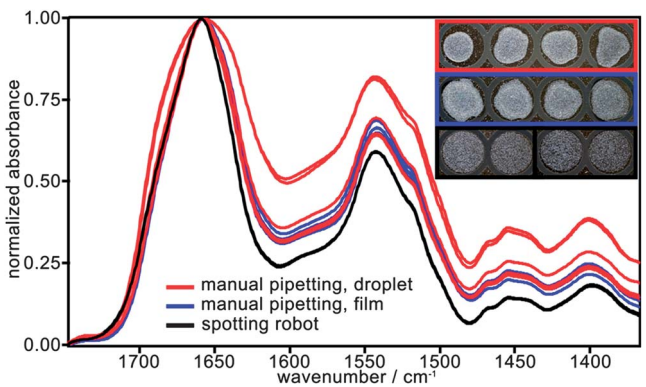

Fig. 3 Manual placing of a 1 µl serum drop (red line and red-framed inset) and 0.5 µl of serum manually spread out on the Si substrate (blue line and blue-framed inset), resulting in highly diverse spectra (min–max–normalized). The highest reproducibility for the quadruplicate samples was achieved with optimized robotic sample dispensing (four superimposed black lines, see also Fig. 4E). The inner well diameter was 4 mm (see the inset).

Here, we report on the application of a robotic dispensing system with a piezo-driven capillary dispenser head for distributing biofluid samples on the MTP silicon substrate. An assortment of different spotting patterns was analysed for the spectral reproducibility as manifested by the average absorbance difference, relative standard deviation, and required spotting time (see Fig. 4 and Table 1). The so far optimum pattern is described by a merged concentric, circular arrangement resulting in a film with holes (Fig. 4E). The pattern consisted of 217 drops of approximately 200 pl each, which were deposited onto each single well of the silicon substrate.

Due to the extremely low volumes spotted, the formation of a coffee ring[54] as still observed with nanoliter volumes[45,46] was avoided. An atomic force microscopic (AFM) scan confirmed that at such a low volume, an indented cylinder shape is formed instead of an irregular ring structure (Fig. 5A and B). The height difference of the outer and inner spot region is less than 1 µm with most of the sample thickness homogeneously levelled at 2 µm. In merged droplets, the inner droplet surface continues into the centre of the next spot (Fig. 5C and D). Thus, an even but occasionally holey film of approximately 2.5–3 µm thickness was formed. Spectra of these films yielded the highest reproducibility (Fig. 3), so that the residual holes turned out to be irrelevant for the spectral measurement.

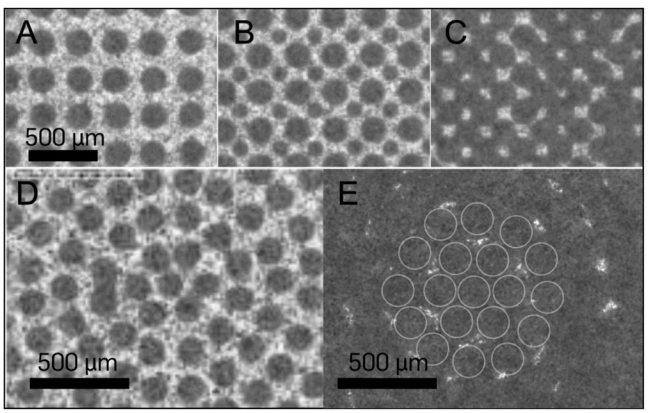

Fig. 4 Using a piezo-electronic sample dispenser, spotting patterns of sub-nanoliter serum sample droplets were tested for spectral reproducibility: rectangular (A–C) and circular (D and E) spotting arrangements. The latter ones covered the MTP-well best and were considerably faster to spot (see also Table 1). Individual spot positions of the blended pattern (E) are indicated for the innermost three circles.

Table 1 Impact of the tested spotting pattern on spectral reproducibility (see Fig. 4) of identical samples: concentric, blended spots in a circular pattern have been found to be the best compromise concerning reproducibility and preparation time. All values were averaged over the intervals of 3200–2800 cm^{-1} and 1750–850 cm^{-1}, particularly, the wavenumber-wise relative standard deviation

Pattern	A	B	C	D	E
Avg. ΔA^a	0.08	0.11	0.25	0.13	0.10
Avg. std. dev./$\Delta A \times 10^3$	1.85	1.56	0.83	0.98	0.96
Spotting time/min 4 wells	4	9	7	4	4

[a] Absorbance min–max–normalized on 2150–2050 cm^{-1}.

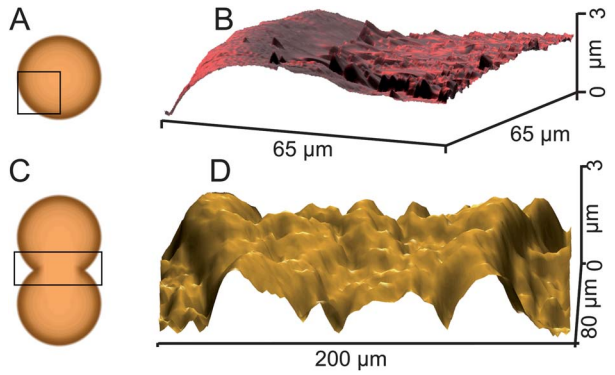

Fig. 5 An approximately 200 pl drop of serum dried with only a rudimentary crater shape ((A), schematic position; (B) AFM scan). Merging droplets (Fig. 2E) formed a continuous film at the maintained thickness of the inner droplets ((C), schematic position; (D), AFM scan).

Highly reproducible sample quadruplicates were repeatedly prepared and measured with optimized spectrometer settings. To eliminate the spectral substrate contribution and localization effects, single channel background spectra for the individual positions were recorded with identical settings of the freshly cleaned MTP substrate. Calculation of the absorbance spectra and further spectral processing was automated, and pre-parameterized algorithms were used to achieve maximum process reproducibility.

Preprocessing spectra

The only spectral distortion observed on the highly reproducible spectra is a minor baseline inconsistency, which is most likely attributable to multi-beam internal interference within the sample film due to back reflected radiation from the substrate, caused by the cone-like beam geometry of a 2× magnifying mirror within the HTS-XT accessory, and the large difference of refractive indices of the sample and the substrate. The spectral effects were described,[51–53] but due to the overlay of these effects, fitting an analytical function would have created another source of error from an underdetermined problem. Instead, the effects can be sufficiently removed by a signal frequency sensitive baseline correction algorithm after splitting the spectral frequencies into regions matching the respective information density.[41–43] Soft baselines can be adapted to a wide range of shapes, but in extreme cases these will affect also broad bands. This is undesirable, *e.g.* for the amide I/II region, where spectral information would be lost. Generally desirable are stiff baselines, which maintain spectral feature fidelity preserving even broad absorption bands. However, those cannot match a broad, slightly non-linear function simultaneously (see Fig. 6A, segment (iii)). Adapting the stiffness of baselines to the particular spectral section yielded the best results (Fig. 6B and C; for parameters and the algorithm see ESI‡).

Discussion of feature selection

Without feature selection, the misclassification error rate of an RF predictor reached an ambiguous value of 50%, a classifying

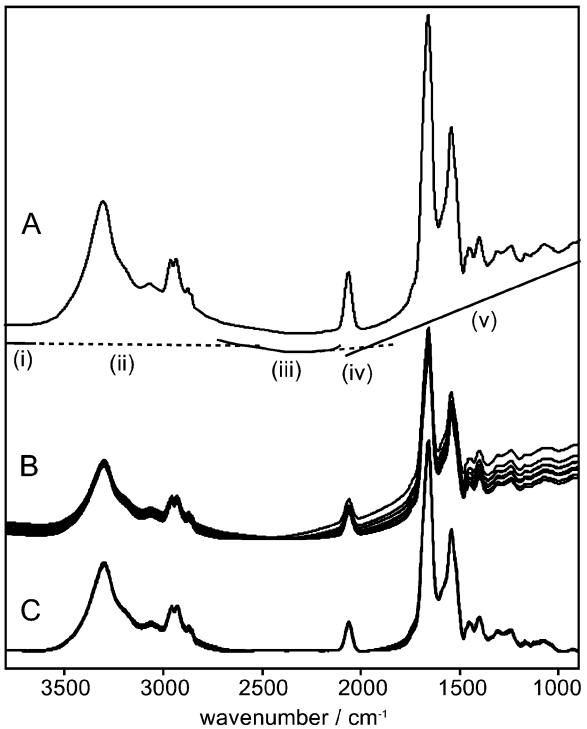

Fig. 6 Absorbance spectra were subdivided into regions containing (i) zero spectral information, (ii) amide A and C–H stretching vibration bands, (iii) zero spectral information, (iv) the KSCN-marker band, and (v) the fingerprint region (A). The efficiency of the algorithm to remove Fresnel-scattering and thin film interference artefacts is demonstrated on 10 spectral replicates of a test serum (B), and the respective corrected spectra (C).

Linear Discriminant Analysis (LDA) of the scores of the first two principal components of the dataset achieved 32% with an obviously insufficient class separation (Fig. S1‡). To improve classification performance, a feature selection was introduced, reducing the amount of redundant and diagnosis-uncorrelated data. Therewith, a more robust classification was expected based on the dimensionality-stratified dataset. Concomitantly, wavenumber variables were identified which contain information most important for classification. Finally, these represent the disease associated spectral biomarker candidates.

A major challenge in both feature selection and classification is constituted by the combination of a small number of subjects (here: $N = 135$) in contrast to a large number of features ($n =$ 11 493, due to the concatenation of wavenumber–intensity pairs of three absorbance spectra and their respective 1st and 2nd derivative spectra). Univariate approaches for feature selection commonly perform particularly poor in this situation, in particular due to the relatively small number of subjects.[62] Thus, it is inevitable to perform a multivariate feature selection, which needs to deal with the exponential growth of the number of m features that can be drawn out of n variables. For obtaining the best possible multivariate selection of features, we combined two recent approaches from the machine learning literature. First, we followed recent studies[32,33] and utilized random forest classifiers[63] for feature selection. Second, we integrated this approach into a "feature-shaving"[30,32] or "wrapper"[11] approach,

which – in contrast to immediately selecting a small number of most significant features – eliminated a fraction of the most insignificant features iteratively, until only a small number of highly significant features remained. The iterative feature-shaving approach was originally termed "gene shaving"[30] due to its origin in gene selection on microarray chips. It was reported to be superior to other methods used on high dimensional genetic micro-array data,[64] on spectral imaging data,[65] and biofluid spectra.[31,32]

For further reduction of the probability of overfitting, a repeated feature selection was proposed.[66] Consequently, we collected the results of 12 consecutive feature selection cycles of the iterative algorithm on Monte Carlo (MC) generated data subsets. The occurrence, *i.e.* the number of feature selection cycles that identified a particular wavenumber variable, was registered and used for a stepwise top-down identification of the optimal feature subset for classification.

The feature selection result for the spectral region depicting the fingerprint regions of absorbance, 1st and 2nd derivative spectra of citrate stabilized plasma is shown in Fig. 7. Five features were repeatedly identified in the 2nd derivative spectrum of citrate plasma (Fig. 7C), whereas none was found in its absorbance or 1st derivative spectra.

Performance results of LDA (Table 2) and RF classifiers (Table 3) indicate a selection of discriminating features of different predictive qualities from the reduced datasets of absorbance, 1st and 2nd derivative alone. It is remarkable that the only feature of the 1st derivative set identified in 11/11 cycles was also chosen in the selection scheme on the total dataset, as were also all four features found significant in the calculations of the isolated 2nd derivative spectra set.

The best validation results on our dataset were obtained with a total of fifteen features that were found to be relevant in more than seven of the twelve selection cycles (Fig. 8), one in the serum 1st derivative, four in the serum 2nd derivative, four in the EDTA plasma 2nd derivative, and six in the citrate plasma 2nd derivative of the absorbance spectrum. It is encouraging that neighbouring features were identified, which indicates the classification importance of an actual spectral band. However, not all neighboured features met the threshold (Fig. 7D and S2–S4‡).

Some selected wavenumber variables exhibit a relatively large ordinate distance when spectral class averages are examined, and some are close (Fig. 8). The latter can be thought of as anchor points for classification, whereas features with larger ordinate distances bear the classification relevance.

Including the total spectral dataset enabled the identification of the discriminative feature in the 1st derivative spectra along with the most important 2nd derivative features in the course of a single calculation, which required less time to calculate and validate than to consider separate calculations on three separate datasets. The validation results of the classification by two different classifiers (see below) on the selected feature sets show that a meaningful selection of relevant wavenumber–intensity pairs for the discrimination of UBC and control patients was achieved. However, whether the applied RF based feature selection method proves best has to be further evaluated on an extended dataset and in comparison with a variety of selection techniques.

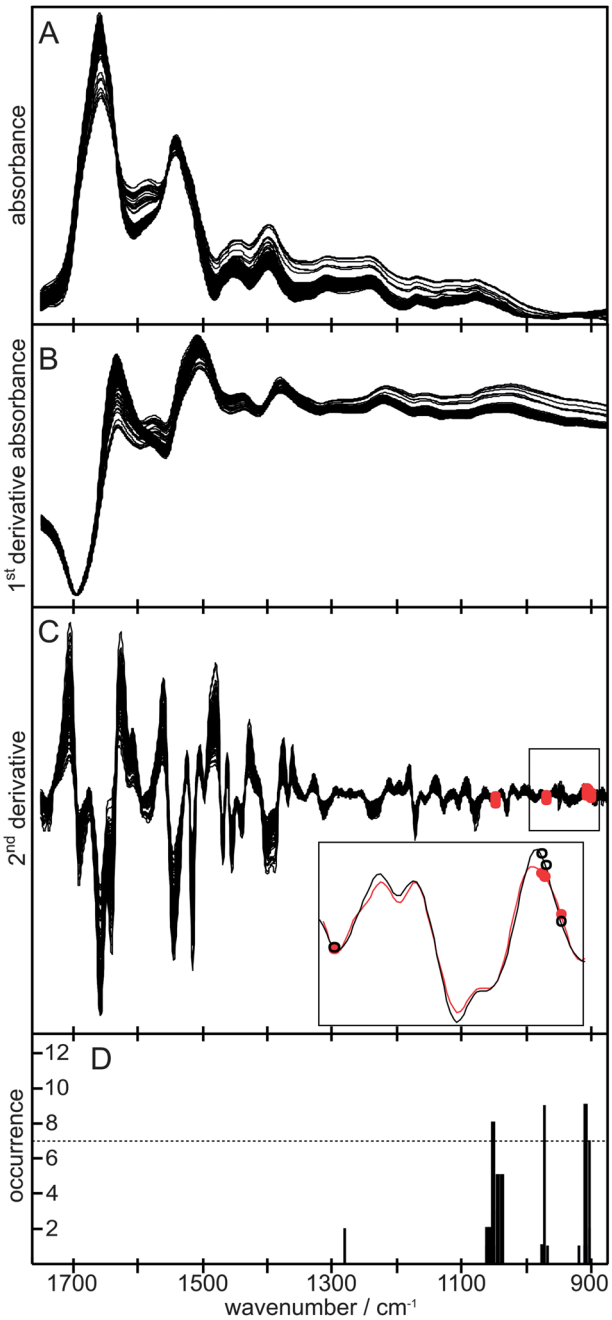

Fig. 7 In the fingerprint region of citrate plasma absorbance (A) and 1st derivative spectra (B), no classification relevant spectral features were identified. Five features were identified in the 2nd derivative spectrum (C) that was repeatedly selected in at least seven of twelve selection cycles (D). Spectral class averages are shown enlarged in the range of 900–1060 cm^{-1} ((C), inset) with the selected features highlighted (red: UBC, black circle: control).

Classification of control samples *versus* bladder cancer

With relevant wavenumber variables identified, we proceeded with the spectral classification of whether the present dataset was sufficient for UBC detection among suspected subjects. As

Table 2 Average performance data of 50 LDA classifiers on optimum feature sets of concatenated absorbance only (abs), 1st derivative only (1st der), and 2nd derivative spectra (2nd der) of a concatenated spectral serum–EDTA plasma–citrate plasma vector in comparison with the performance on features selected from the total dataset. (cyc: threshold of cycles of feature selection, # f: number of features, acc: % accuracy, MER: % average misclassification error rate, sens: % sensitivity, spec: % specificity.)

	Cyc	#f	Acc	MER	Sens	Spec
Abs	≥2/10	25	56 ± 5	14 ± 1	95 ± 6	18 ± 11
1st der	≥3/11	11	55 ± 7	22 ± 4	78 ± 13	33 ± 13
2nd der	≥12/13	4	69 ± 7	18 ± 2	89 ± 9	49 ± 11
Total set	≥7/12	15	66 ± 8	8 ± 2	86 ± 7	45 ± 14

Table 3 Average performance data of 50 ensemble RF classifiers on optimum feature sets of concatenated absorbance only (abs), 1st derivative only (1st der), and 2nd derivative spectra (2nd der) of a concatenated spectral serum–EDTA plasma–citrate plasma vector in comparison with the performance on features selected from the total dataset (see Table 2 for the legend)

	Cyc	#f	Acc	MER	Sens	Spec
Abs	≥2/10	25	58 ± 5	40 ± 2	93 ± 7	23 ± 11
1st der	≥3/11	11	60 ± 6	38 ± 2	93 ± 7	28 ± 11
2nd der	≥12/13	4	64 ± 8	35 ± 2	86 ± 9	42 ± 12
Total set	≥7/12	15	68 ± 7	33 ± 2	93 ± 10	46 ± 18

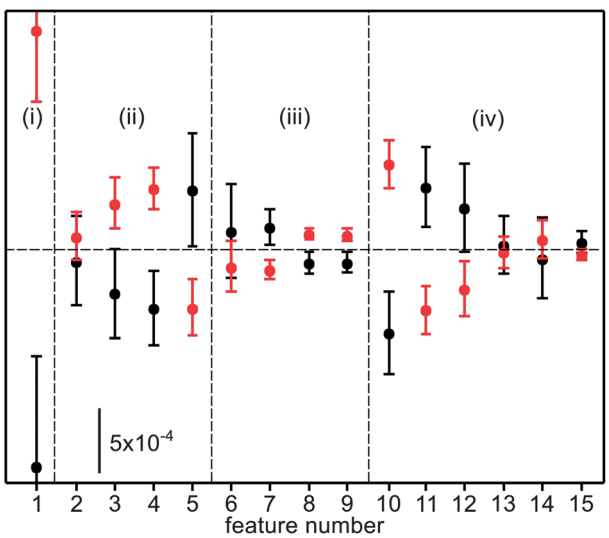

Fig. 8 The fifteen selected feature variables with class-averaged, centred intensities and standard error of mean of the control (black) and UBC (red) groups from serum 1st derivative (i), serum 2nd derivative (ii), EDTA plasma 2nd derivative (iii) and citrate plasma 2nd derivative spectra (iv) are shown, which enabled a disease status prediction by the optimized random forest classifier.

validation is a crucial aspect, 50 training and independent test-sets were randomly chosen from the total dataset with a randomized leave-14-per-class-out procedure to build test sets consisting of 28 subjects (equivalent to 21% of the dataset) which were set aside for validation.

Using each set of split data, 50 LDA and 50 ensemble RF predictors were trained and applied to the respective test sets. The average result of the LDA predictors on the corresponding, independent test sets yielded an accuracy of 66 ± 8% (mean ± standard deviation, Fig. 9), and the random forest predictor achieved a value of 68 ± 7%. The LDA classifiers yielded a sensitivity of 86 ± 7% and a specificity of 45 ± 14%, whereas RF predictors achieved respective values of 93 ± 10% and 46 ± 18%. In summary, a sensitive UBC detection is possible with both classifiers, although both predictors lack specificity in the exclusion of non-UBC controls, which is yet insufficient for clinical use.

Neither LDA nor RF classifiers performed well using absorbance or 1st derivative features only (Tables 2 and 3). Taking the standard deviations into account, the LDA classifiers performed comparably well on four features selected from the 2nd derivative spectra only and on the 15 features selected from the total dataset. The performance based on the 15 features appears worse on average by 3 per cent units, but this is outweighed by a more convincing performance based on 15 features, as shown by the highly significantly lowered, and less than half misclassification error rate (as determined by a paired *t*-test at a 99% confidence level). For illustrative purposes, an exemplary classification result on the scores of the first two principal components of the 15 feature dataset is shown in Fig. 9.

The comparison of 2nd derivative and total dataset features with RF classification yields an unambiguously improved performance on the 15 total dataset derived features with regard to all observed quality parameters (Table 3).

In summary, the inclusion of the eleven additional features contributed to an improved result with both classifiers. Evidently, the overall poor specificity can also be attributed to the imbalance of 89 UBC *versus* 46 control samples in the total dataset. With continued recruitment of patients, an extended and more balanced dataset should become available. Both LDA

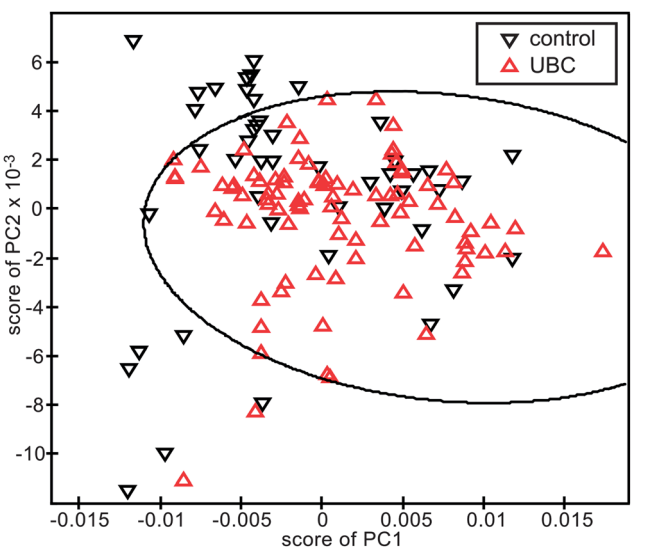

Fig. 9 For illustrative purposes the LDA classifier separating controls *versus* UBC based on the scores of the first and second principal components (PC) of spectra comprised of the selected features is displayed. The misclassification error rate was calculated as 24%, which is worse than the error of 11% from direct LDA classification on the fifteen selected features.

and RF are established classifiers. With both, the spectral separability of blood samples drawn from UBC and control patients in the current state of the study was demonstrated. The more complex RF algorithm performs slightly better on the current data. However, whether these are the optimum performing classifiers has to be evaluated in a future study on an extended dataset.

Conclusion

Blood samples were collected and processed obeying the highest clinical standards. We developed a largely automated procedure for the infrared spectroscopic analysis of body fluids. The highest reproducibility was achieved with minimized user interaction in sample preparation and data processing. Unavoidable caveats such as sample inhomogeneity, pipetting errors, spectral artefacts and classifier overfitting were tackled and suitable solutions were presented. After excluding all possible experimental errors and using state of the art bioinformatics to identify spectral biomarker candidates in body fluids, we ended up with very small spectral differences between the UBC and the control group. Classification relevant wavenumber variables were not found within the absorbance spectra of serum or plasma. Fourteen of fifteen relevant features were identified in the 2^{nd} derivative spectra in calculations comprising all available data. The identified features (see Fig. 8) indicate that only very subtle spectral differences at the detection limit distinguish a UBC patient from the control group. The FTIR spectroscopic classification of biofluids under such a scheme is extremely challenging, but can be performed using the utmost advanced sampling, sample preparation tools, spectrometer hardware with high-throughput accessories and modern bioinformatics software. For further validation of this approach, larger sample numbers are necessary for a final validation of a spectral biomarker detection scheme derived from body fluids. A unification of spectrum collections from groups researching the same diseases and combining them would provide an extensive database within a short time. Nevertheless at the moment, the main barrier to achieve this goal is the lack of standardization of the sample preparation and spectral measurement. The application of the Bruker HTS-XT system has nowadays frequently been reported, but sample preparation was often done manually or with dedicated, lab-customized robots not commonly available.

The automated sample preparation, highly reproducible spectrum recording, and user independent spectrum processing presented here, could be a milestone for building multicenter FTIR-spectral databases for reliable disease diagnosis from body fluids.

Acknowledgements

These studies were made available by support of PURE (Protein research Unit Ruhr within Europe), financed by the state of North Rhine-Westphalia. The participation of all members of the PURE consortium is acknowledged. Particularly for the organization/support of the clinical study with regard to bladder cancer, we are grateful to: Jan Hovanec (ethics vote and privacy policy), Nadine Bonberg, Antje Müller, Isabelle Groß, and Volker Harth (epidemiology and sampling) of the Institute for Prevention and Occupational Medicine of the German Social Accident Insurance, Institute of the Ruhr-Universität Bochum (IPA) (director Thomas Brüning). The authors also thank Thomas Deix and Katharina Braun of the Urological Clinics (director Joachim Noldus), of the Marien-Hospital Herne, for collaboration. The authors are grateful to all patients participating in the reported study. Furthermore, we are grateful to Axel Mosig for assistance with programming and critically reviewing the manuscript, to Laven Mavarani for assistance with AFM experiments, and to Eckhard Nordhoff and Martin Müller of M2 Automation for continued support.

References

1 H.-U. Gremlich and B. Yan, *Infrared and Raman spectroscopy of biological materials*, M. Dekker, New York, 2001.
2 M. Diem, J. M. Chalmers, and P. R. Griffiths, *Vibrational spectroscopy for medical diagnosis*, John Wiley & Sons, Chichester, England; Hoboken, NJ, 2008.
3 P. Lasch and J. Kneipp, *Biomedical vibrational spectroscopy*, Wiley-Interscience, Hoboken, N.J., 2008.
4 G. Hoşafçı, O. Klein, G. Oremek and W. Mäntele, *Anal. Bioanal. Chem.*, 2007, **387**, 1815–1822.
5 P. Lasch, J. Schmitt, M. Beekes, T. Udelhoven, M. Eiden, H. Fabian, W. Petrich and D. Naumann, *Anal. Chem.*, 2003, **75**, 6673–6678.
6 W. Petrich, K. B. Lewandrowski, J. B. Muhlestein, M. E. H. Hammond, J. L. Januzzi, E. L. Lewandrowski, B. Dolenko, J. Früh, W. Köhler and R. Mischler, *Analyst*, 2009, **134**, 1092–1098.
7 T. C. Martin, J. Moecks, A. Belooussov, S. Cawthraw, B. Dolenko, M. Eiden, J. Von Frese, W. Kohler, J. Schmitt, R. Somorjai, T. Udelhoven, S. Verzakov and W. Petrich, *Analyst*, 2004, **129**, 897–901.
8 D. I. Ellis and R. Goodacre, *Analyst*, 2006, **131**, 875–885.
9 M. Beekes, P. Lasch and D. Naumann, *Vet. Microbiol.*, 2007, **123**, 305–319.
10 G. Bellisola and C. Sorio, *Am. J. Cancer Res.*, 2012, **2**, 1–21.
11 J. Trevisan, P. P. Angelov, P. L. Carmichael, A. D. Scott and F. L. Martin, *Analyst*, 2012, **137**, 3202–3215.
12 P. Carmona, M. Molina, M. Calero, F. Bermejo-Pareja, P. Martínez-Martín and A. Toledano, *J. Alzheimers Dis.*, 2013, **34**, 911–920.
13 K. Gajjar, J. Trevisan, G. Owens, P. J. Keating, N. J. Wood, H. F. Stringfellow, P. L. Martin-Hirsch and F. L. Martin, *Analyst*, 2013, DOI: 10.1039/c3an36654e.
14 J. Moecks, G. Kocherscheidt, W. Koehler, and W. H. Petrich, in *Proc. SPIE*, ed. A. Mahadevan-Jansen, M. G. Sowa, G. J. Puppels, Z. Gryczynski, T. Vo-Dinh and J. R. Lakowicz, San Jose, CA, 2004, vol. 5321, pp. 117–123.
15 D. M. Parkin, *Scand. J. Urol. Nephrol., Suppl.*, 2008, **42**(s218), 12–20.
16 M. Ploeg, K. K. H. Aben and L. A. Kiemeney, *World J. Urol.*, 2009, **27**, 289–293.

17 P. Boffetta, *Scand. J. Urol. Nephrol., Suppl.*, 2008, **42**(s218), 45–54.
18 S. M. Cohen, T. Shirai and G. Steineck, *Scand. J. Urol. Nephrol., Suppl.*, 2000, 105–115.
19 M. Adibi, R. Youssef, S. F. Shariat, Y. Lotan, C. G. Wood, A. I. Sagalowsky, R. Zigeuner, F. Montorsi, C. Bolenz and V. Margulis, *Int. J. Urol.*, 2012, **19**, 1060–1067.
20 J. L. Summers, J. S. Coon, R. M. Ward, W. H. Falor, A. W. Miller 3rd and R. S. Weinstein, *Cancer Res.*, 1983, **43**, 934–939.
21 S. Ramakumar, J. Bhuiyan, J. A. Besse, S. G. Roberts, P. C. Wollan, M. L. Blute and D. J. O'Kane, *J. Urol.*, 1999, **161**, 388–394.
22 I. Osman, *Clin. Cancer Res.*, 2006, **12**, 3374–3380.
23 J. Villanueva, *J. Clin. Invest.*, 2005, **116**, 271–284.
24 L. C. Kompier, A. A. G. van Tilborg and E. C. Zwarthoff, *Urol. Oncol.*, 2010, **28**, 91–96.
25 C. J. Marsit, D. C. Koestler, B. C. Christensen, M. R. Karagas, E. A. Houseman and K. T. Kelsey, *J. Clin. Oncol.*, 2011, **29**, 1133–1139.
26 N. Putluri, A. Shojaie, V. T. Vasu, S. K. Vareed, S. Nalluri, V. Putluri, G. S. Thangjam, K. Panzitt, C. T. Tallman, C. Butler, T. R. Sana, S. M. Fischer, G. Sica, D. J. Brat, H. Shi, G. S. Palapattu, Y. Lotan, A. Z. Weizer, M. K. Terris, S. F. Shariat, G. Michailidis and A. Sreekumar, *Cancer Res.*, 2011, **71**, 7376–7386.
27 A. A. G. van Tilborg, L. C. Kompier, I. Lurkin, R. Poort, S. El Bouazzaoui, K. van der Keur, T. Zuiverloon, L. Dyrskjot, T. F. Orntoft, M. J. Roobol and E. C. Zwarthoff, *PLoS One*, 2012, **7**, e43345.
28 M. R. Karagas, A. S. Andrew, H. H. Nelson, Z. Li, T. Punshon, A. Schned, C. J. Marsit, J. S. Morris, J. H. Moore, A. L. Tyler, D. Gilbert-Diamond, M.-L. Guerinot and K. T. Kelsey, *Hum. Genet.*, 2011, **131**, 453–461.
29 Y. Lotan and C. G. Roehrborn, *Urology*, 2003, **61**, 109–118.
30 T. Hastie, R. Tibshirani, M. B. Eisen, A. Alizadeh, R. Levy, L. Staudt, W. C. Chan, D. Botstein and P. Brown, *Genome Biol.*, 2000, **1**, research0003.1–research0003.21.
31 B. H. Menze, W. Petrich and F. A. Hamprecht, *Anal. Bioanal. Chem.*, 2007, **387**, 1801–1807.
32 B. H. Menze, B. M. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich and F. A. Hamprecht, *BMC Bioinf.*, 2009, **10**, 213.
33 R. Genuer, J. M. Poggi and C. Tuleau-Malot, *Pattern Recogn. Lett.*, 2010, **31**, 2225–2236.
34 Y. Saeys, I. Inza and P. Larranaga, *Bioinformatics*, 2007, **23**, 2507–2517.
35 R. Montironi and A. Lopez-Beltran, *Int. J. Surg. Pathol.*, 2005, **13**, 143–153.
36 W. Otto, S. Denzinger, H.-M. Fritsche, M. Burger, W. F. Wieland, F. Hofstädter, A. Hartmann and S. Bertz, *BJU Int.*, 2011, **107**, 404–408.
37 Z. Chen, W. Ding, K. Xu, J. Tan, C. Sun, Y. Gou, S. Tong, G. Xia, Z. Fang and Q. Ding, *PLoS One*, 2012, **7**, e47199.
38 E. Goormaghtigh and J.-M. Ruysschaert, *Spectrochim. Acta, Part A*, 1994, **50**, 2137–2144.
39 J. Ollesch, E. Künnemann, R. Glockshuber and K. Gerwert, *Appl. Spectrosc.*, 2007, **61**, 1025–1031.
40 K. Elfrink, J. Ollesch, J. Stöhr, D. Willbold, D. Riesner and K. Gerwert, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 10815–10819.
41 Z.-M. Zhang, S. Chen and Y.-Z. Liang, *Analyst*, 2010, **135**, 1138–1146.
42 B. D. Prakash and Y. C. Wei, *Analyst*, 2011, **136**, 3130–3135.
43 K. H. Liland, E.-O. Rukke, E. F. Olsen and T. Isaksson, *Chemom. Intell. Lab. Syst.*, 2011, **109**, 51–56.
44 A. Liaw and M. Wiener, *R News*, 2002, **2**, 18–22.
45 E. Diessel, S. Willmann, P. Kamphaus, R. Kurte, U. Damm and H. M. Heise, *Appl. Spectrosc.*, 2004, **58**, 442–450.
46 E. Diessel, P. Kamphaus, K. Grothe, R. Kurte, U. Damm and H. M. Heise, *Appl. Spectrosc.*, 2005, **59**, 442–451.
47 F. L. Martin, J. G. Kelly, V. Llabjani, P. L. Martin-Hirsch, I. I. Patel, J. Trevisan, N. J. Fullwood and M. J. Walsh, *Nat. Protoc.*, 2010, **5**, 1748–1760.
48 T. Vahlsing, U. Damm, V. Radhakrishna Kondepati, S. Leonhardt, M. D. Brendel, B. R. Wood and H. M. Heise, *J. Biophotonics*, 2010, **3**, 567–578.
49 H. Fabian, P. Lasch and D. Naumann, *J. Biomed. Opt.*, 2005, **10**, 031103.
50 H. M. Heise, R. Marbach, T. Koschinsky and F. A. Gries, *Appl. Spectrosc.*, 1994, **48**, 85–95.
51 R. N. Jones, D. Escolar, J. P. Hawranek, P. Neelakantan and R. P. Young, *J. Mol. Struct.*, 1973, **19**, 21–42.
52 H. M. Heise, *Asian Chem. Lett.*, 2009, **13**, 163–170.
53 M. Miljkovic, B. Bird and M. Diem, *Analyst*, 2012, **137**, 3954–3964.
54 R. D. Deegan, O. Bakajin, T. F. Dupont, G. Huber, S. R. Nagel and T. A. Witten, *Nature*, 1997, **389**, 827–829.
55 T. Hirschfeld, *Appl. Spectrosc.*, 1985, **39**, 426–430.
56 A. Bittner and H. M. Heise, in *AIP Conference Proceedings*, American Institute of Physics, New York, Athens, Georgia, USA, 1998, vol. 430, pp. 278–281.
57 J. L. Jarman, S. I. Seerley, R. A. Todebush and J. A. de Haseth, *Appl. Spectrosc.*, 2003, **57**, 1078–1086.
58 M. Haberkorn, J. Frank, M. Harasek, J. Nilsson, T. Laurell and B. Lendl, *Appl. Spectrosc.*, 2002, **56**, 902–908.
59 I. Surowiec, J. R. Baena, J. Frank, T. Laurell, J. Nilsson, M. Trojanowicz and B. Lendl, *J. Chromatogr., A*, 2005, **1080**, 132–139.
60 S. Armenta and B. Lendl, *Anal. Bioanal. Chem.*, 2010, **397**, 297–308.
61 M. Boese, Bruker BioSpin, GmbH, *US Pat.*, 7 267 838 B2, 2007.
62 H. Peng, F. Long and C. Ding, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005, **27**, 1226–1238.
63 L. Breiman, *Mach. Learn.*, 2001, **45**, 5–32.
64 R. Díaz-Uriarte and S. Alvarez de Andrés, *BMC Bioinf.*, 2006, **7**, 3.
65 A. Kallenbach-Thieltges, F. Großerüschkamp, A. Mosig, M. Diem, A. Tannapfel and K. Gerwert, *J. Biophotonics*, 2013, **6**, 88–100.
66 S. Baek, C. A. Tsai and J. J. Chen, *Briefings Bioinf.*, 2009, **10**, 537–546.

Supplemental Information to:

FTIR spectroscopy of biofluids revisited: An automated approach to spectral biomarker identification

Julian Ollesch, Steffen L. Drees, H. Michael Heise, Thomas Behrens, Thomas Bruening, and Klaus Gerwert

The baseline correction algorithm that was used on each of the five spectral regions (Fig. 6) is available for download at https://code.google.com/p/airpls/ as on April 25th, 2013.[1–3] The parameterization of the airPLS.m function was
[~,baselines(i)]=airPLS([dataset],a(i),b(i),c(i),d(i));
with the parameters
a=[50e11 10e10 10e4 10e9 10e1];
b=[2 2 2 2 10];
c=[0.05 0.05 0.05 0.05 0.5];
d=[0.05 0.05 0.05 0.05 0.5];
The regions were ordered by increasing wavenumber, *e.g.,* a(i) parameterizes region (v) etc. (see Fig. 6).

1. P. H. C. Eilers, *Anal. Chem.*, 2003, **75**, 3631–3636.
2. H. F. M. Boelens, R. J. Dijkstra, P. H. C. Eilers, F. Fitzpatrick, and J. A. Westerhuis, *J. Chromatogr. A*, 2004, **1057**, 21–30.
3. F. Gan, G. Ruan, and J. Mo, *Chemom. Intell. Lab. Syst.*, 2006, **82**, 59–65.

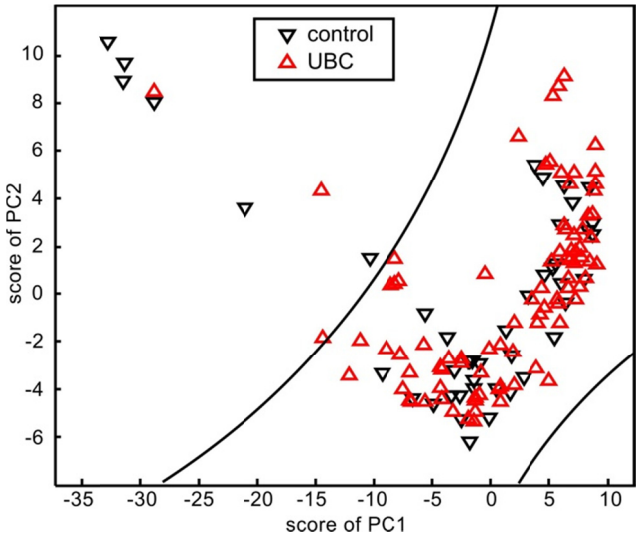

Figure S1: For illustration, the LDA classifier separating controls versus UBC based on the scores of the first and second principal components (PC) of spectra comprised of the full spectral feature set is displayed with a misclassification error rate of 32 %.

In the following, our full spectral dataset from 135 probands as used for feature selection is presented. Apart from the serum dataset, some wavenumber variables were found classification relevant within the absorbance and first derivative spectra, but most did not meet the selection frequency threshold of ≥ 7 of 12 feature selection cycles. For the spectra of EDTA- and citrate-plasma (Fig. S2 and S3), relevant variables were exclusively identified in the second derivative spectra.

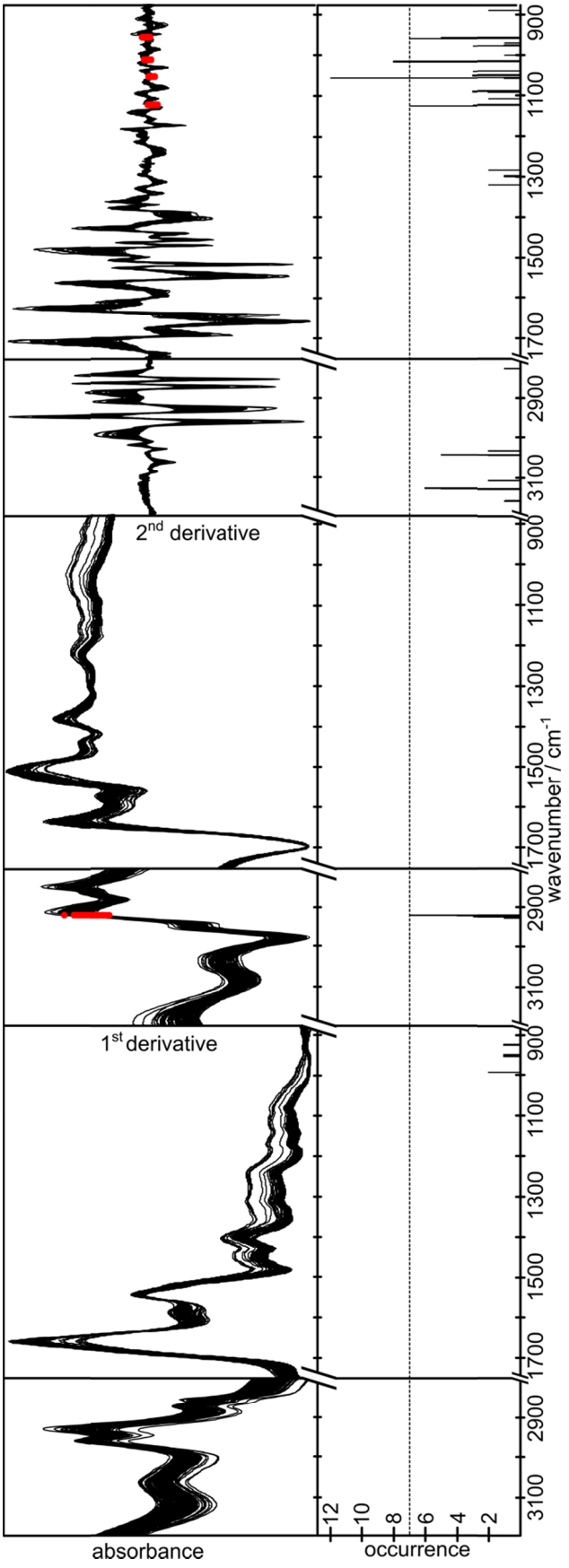

Figure S2: Serum absorbance, 1st derivative, 2nd derivative and selection frequency of features important for classification (labeled in red). Wavenumber segments were scaled individually for optimum display.

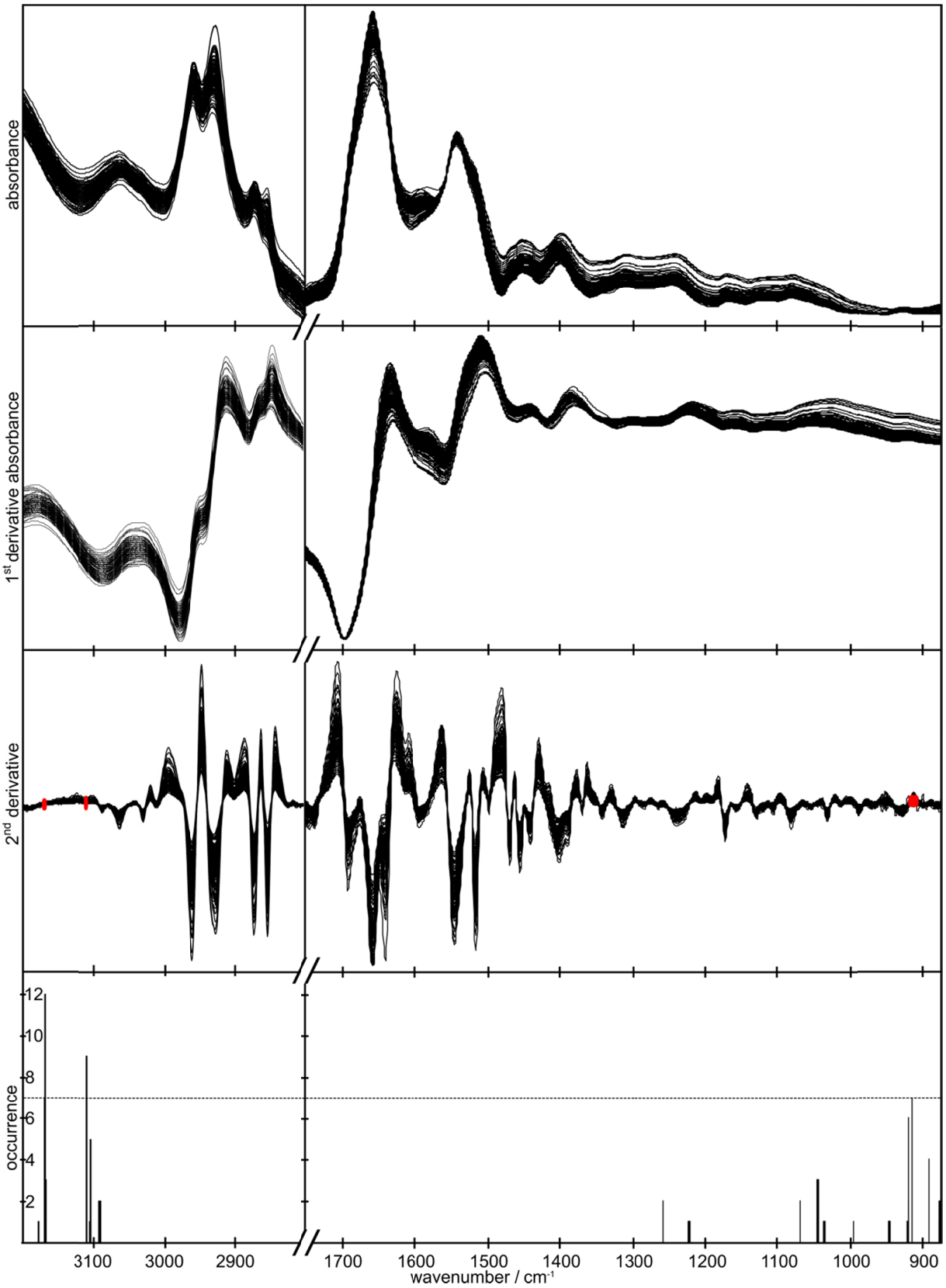

Figure S3: EDTA-plasma absorbance, 1st derivative, 2nd derivative and selection frequency of features important for classification. No features were identified within absorbance or 1st derivative spectra. Wavenumber range segments were scaled individually for optimum display.

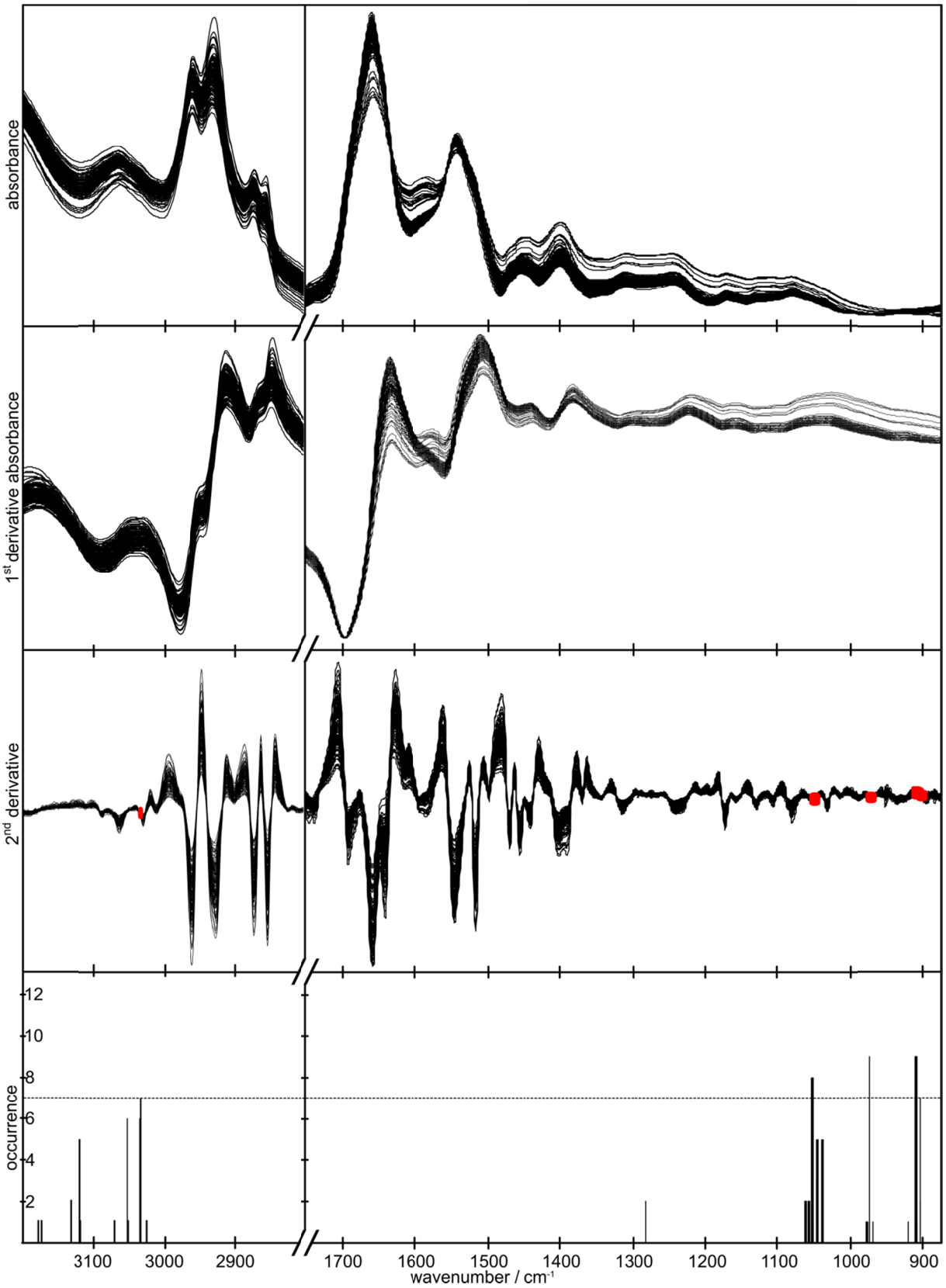

Figure S4: Sodium citrate stabilized plasma absorbance, 1st derivative, 2nd derivative and selection frequency of features important for classification. No features were identified within absorbance or 1st derivative spectra. Wavenumber range segments were scaled individually for optimum display.