Multi-resolution B-splines data compression improves MIR spectroscopy-based Health diagnostic efficiency

David Martin, Valérie Monbet, Olivier Sire, Maëna Le Corvec, Olivier Loréal

Short Communication

Multi-resolution B-splines data compression improves MIR spectroscopy-based Health diagnostic efficiency.

David Martin[a,b], Valérie Monbet[b*], Olivier Sire[c], Maëna Le Corvec[c,d], Olivier Loréal[a]

a) *NUMECAN, UMR INSERM 1241, CHU Rennes, France*
b) *Univ Rennes, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France.*
c) *IRDL, UMR CNRS 6027, Université Bretagne Sud, Vannes, France*
d) *DIAFIR, Rennes (present address)*

*Corresponding author.
E-mail address : valerie.monbet@univ-rennes1.fr

Keywords : Health diagnostic, MIR spectroscopy, multivariate analysis, B-splines, biomarkers

ABSTRACT

MIR spectroscopy is becoming an increasingly important tool potentially useful for diagnosis purposes especially by studying body fluids. Indeed, diseases induce changes in the composition of fluids modifying the MIR spectra. However, such changes can be difficult to capture if the structure of the data is not considered. Our objective was to improve MIR spectra analysis by using approximation of the spectra by B-splines at different specific resolutions and to combine these spectra representations with a machine learning model to predict hepatic steatosis from serum study. The different resolutions make it possible to identify changes in shape over bands of various widths. The multiresolution model helps to improve the hepatic steatosis prediction compared to conventional approaches where the absorbances are considered as unstructured variables. In addition, B-splines provide both localized and compressed information that can be translated into biochemical terms more easily than with other classical approximation methods (wavelets, Fourier transforms).

1. Introduction

The MIR spectroscopy is becoming an ever-increasing tool to provide health and environmental companion diagnostics due to its ease of use, lack or few needs for sample conditioning and automated protocols [1-3]. A major application in health is dedicated to predictive models intended for rapid, non-invasive diagnostics that can be achieved at the patient's bedside within minutes or through a secured transmission line (e-diagnostic) to benefit UpToDate, evolutive predictive models from some central resource. Such models are built from a calibration spectral data set for a particular disease, such as pre neoplastic states, septic arthritis [4] or NASH [5]. A typical MIR spectrum is made of about 1600 useful absorbance values in the 4 000-800 cm^{-1} frequency range. After different pre-treatments (as for instance multiplicative scatter correction or derivation and vector normalization) [6], MIR spectra feed algorithms for identification of discriminant spectral variables by using Genetic algorithms, Random Forest (RF), and/or Penalized regression.

Raw or derivative spectra are mathematically structured as vectors or matrices where one

thousand of absorbance values are considered as distinct ones with no regard to any information connecting them. Instead, redundancy is emphasized and makes algorithms like principal component analysis (PCA) valuable in drastically reducing the spectrum dimension. Actually, a MIR spectral envelope is made of a linear sum of vibrators, each one featured by a more or less wide spectral distribution. Consequently, in complex samples such as biofluids (sera, synovial fluids or cerebro-spinal fluid), the thousands of distinct biomolecules yield numerous strongly overlapping absorption bands resulting in complex curvatures depending on intensity, width and position band parameters. That is why local tiny changes in the spectral envelope often hold significant information for diagnostic purposes rather than discrete variations at any particular frequency position. To cope with such a partial information, Dynamic Network Biomarkers (DNB) have been proposed to merge and correlate in a single network all the bands or spectral positions that are observed to discriminate a particular health status [7,8].

Here we describe a new multi-resolution application of the well-known B-spline [9, 10] functions to modelize MIR spectra local curvatures in a way that is close to the wavelet approaches [11,12], getting rid of the constraints that bear Fourier Transforms that do not have local features in its regular formulation. The underlying rationale is to no longer consider thousands of discrete *individual* absorbance values to build a predictive model, a spectral fingerprint, but rather a network of patterns that locally describe, or approach, the spectral curvatures at different spectral windows. As quoted previously [13], the *ordering* of the variables has significance. The analogy with pictures captured at different resolutions makes sense to understand how the B-spline approach could bring different but complementary information [14, 10]. The second idea is not to choose a particular resolution but rather to benefit from the information that can be gathered from various (typically three) spectral resolutions. Low resolution, *i.e.* wide spectral windows, will be more sensitive to “large” spectral variations *i.e.* low frequencies, whereas high resolution will be more sensitive to very local intensity changes *i.e* high frequencies. As each B-spline can be defined by a unique coefficient that wears all the information of the local curvature, the pattern, in the considered frequency span, regression algorithms can be fed from these coefficients to identify the relevant frequency domains of the biological signature. Such an approach has been successfully used in Humans and rodent models for Metabolic syndrome and NonAlcoholic Steatohepatitis (NASH) determination.

Thus this note presents a spectral characterization of MIR spectrum variations and how the B-splines fit the corresponding structure of the signal and finally the improvement of the method in establishing a diagnostic is demonstrated from a MIR data set earlier published.

2. Material and Methods

Data set

Data set was previously used and published [15]. Briefly, it described the liver steatosis stage of 68 mice with its corresponding MIR serum. Steatosis has been evaluated histologically with two different scores: expansion of micro-vacuolar steatosis and macro-vacuolar steatosis. Both parameters are returned in percentage of expansion compared to the global liver surface. NAFLD (steatosis) is clinically defined as positive when macro-vacuole of lipids are present in the liver even in small percentages [16]. In this case, we chose a very low threshold (> 1.5%) to detect all mice presenting steatosis. Then, the NAFLD stage is specified by a higher threshold of macro-steatosis (> 5%).

Signal analysis

The signal used for health status fingerprinting is presented (Fig. 1). Usually, the analyzed signal is the MIR second derivative spectrum, smoothed and vector normalized [6]. Fig. 1 presents a typical MIR raw spectrum along with its inverted second derivative (D2). This pre-treatment is useful in emphasizing spectral shoulders that originate in band overlapping. Therefore the second derivative contains all the information about the molecular composition, meaning functional biochemical groups, present in the studied samples.

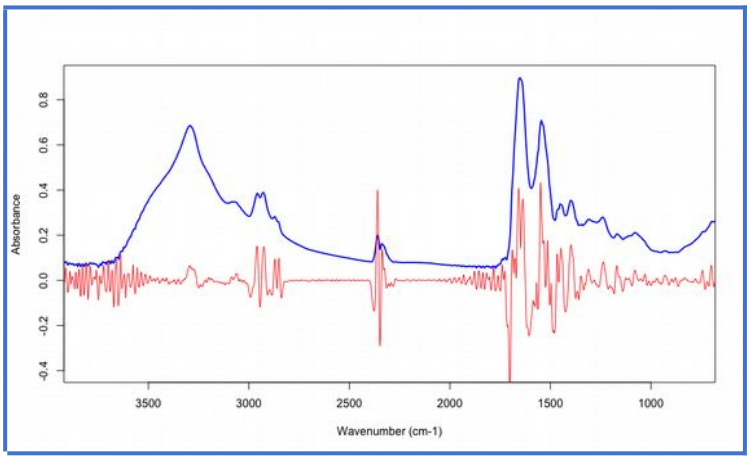

Figure 1. Typical serum MIR spectrum with its second derivative (inverted)

A MIR spectrum results from the linear combination of absorption bands that can be fitted as Voigtian distributions that are themselves a convolution of a Lorentzian band with a gaussian factor to consider the chemical and physical environment of any vibrator [17]. Fig. 2 shows an example of such a spectral decomposition in the 1300-1800 cm^{-1} frequency domain. Such a spectrum displays the importance to focus on band widths and overlaps. Obviously, the small variations within those bands impose tiny changes in the D2 spectral curvatures. The boxplot on the right of Fig. 2 shows the distribution of Full Width at Half Maximum (FWHM) of the 11 recovered bands from spectral curve fitting. The band's width distribution median is 32 cm^{-1} (mean = 31 ± 10 cm^{-1}).

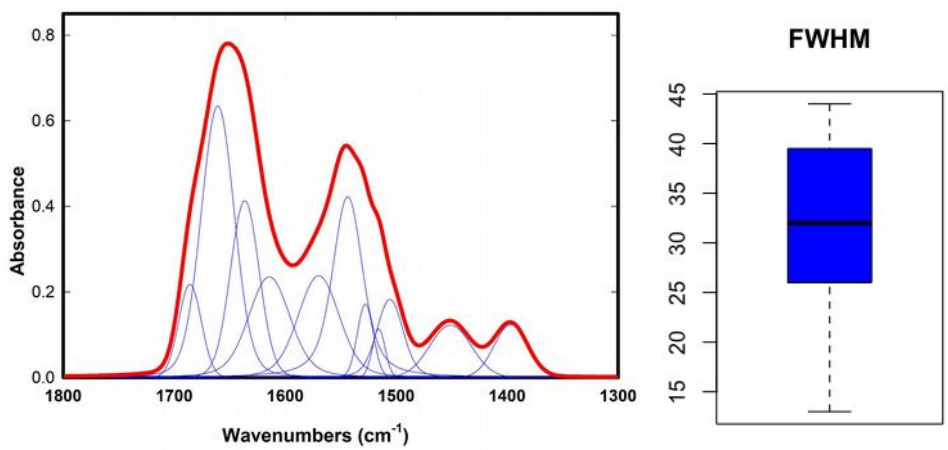

Figure2. MIR Spectrum (red) decomposition as a linear sum of voigtians (blue).

Model set-up

MIR spectra were pre-processed in the 3500-800 cm^{-1} frequency domain. Second derivatives were calculated, smoothed using a 13-point Savitzky–Golay algorithm, and normalized by

vector normalization over the whole spectral range. A principal component analysis was performed on the set of MIR spectra. The 1st plan individual plot shows that there was no outlier. Results are provided in supplement material (Figure S1). Random Forest (RF) models are fitted for hepatic steatosis prediction. Cross-validation was used in this study. More precisely, the model fitting was repeated 30 times. At each iteration, 85% of the data set was used to learn the model. Then, the remaining 15% was used to validate the model. The algorithm performance was evaluated by the Area Under Receiver Operating Characteristic (AUROC) from the 30 validation sets. Boxplots of AUROC are plotted to highlight the variability of the model validation on an independent set. Significant variables were selected as the most important variables returned by the RF (computed as the mean decrease in Gini index) and used as the biochemical signature of each mouse serum.

When spectra are used in prediction models, a dimension reduction usually helps to improve the performances [9]. One simple and effective solution to reduce the dimension of the spectral data set while preserving the important variation scales is to replace each curve by B-spline approximations. The approximation is a piecewise polynomial representation of the spectra. In practice, the original frequency range is split into $p$sub-intervals, defined by $p+1$ values, $t_0 \leq \ldots \leq t_p$, called knots. In this paper, we consider equal length sub-intervals. A spline of order *d* is a function f such that:

- f is a polynomial of order *d* in each subinterval $[t_k, t_{k+1}]$
- f is continuous and has continuous derivatives up to order *d*-2.
- Each B-spline is a spline with localized support: it is positive only on at most *d* consecutive intervals.
- The B-spline basis can be used to define *n=p-1+d* new variables which are characterized by the position of the knots.
- In the sequel, we propose to combine B-splines of order 3 associated with different sets of knots *i.e.* different values of *p* in order to capture different variation scales (from low, mid to high frequencies) that we refer to as B-spline resolutions.

3. Results and discussion

To highlight matching the B-splines patterns with D2 local curvatures, the signal was approached with conventional frequency tools used for waves analysis or more generally time series. In the present B-spline approach, three resolution levels were empirically chosen as they yielded the best discrimination results in the considered application (see below). These frequency frames encompass 7, 5, and 3 points that range, for a 2 cm^{-1} discretization sampling, 14, 10, and 6 cm^{-1} spans. To better estimate these ranges, it is useful to recall that absorbance MIR bands usually exhibit half-height widths around 30 cm^{-1} (see boxplots of Fig. 1). To relate these B-splines (7, 5, 3) with the D2 structure, the partial auto-correlation functions (R core *pacf* function) of D2 spectra and their reconstructions by using each B-spline are displayed for the ten first auto-correlation factors (Fig. 3).

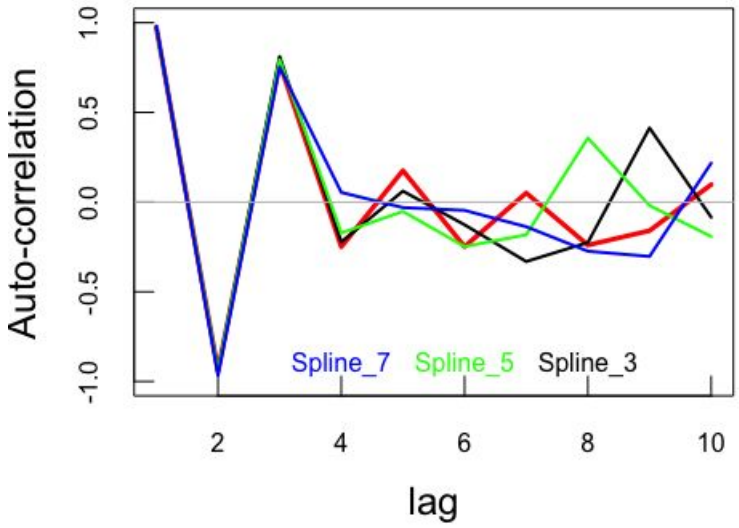

Figure 3. Spectral auto-correlations coefficients of second derivative (red) and splines approximations.

This plot clearly shows that, as expected, coherent information is conveyed by D2 signals and by the B-splines that locally approach the signal. To get a better insight of the D2 intrinsic *frequency* modes that constitute the signal, individual periodograms of the data set were calculated (R core *spectrum* function from **stats** package) and averaged (Fig. 4).

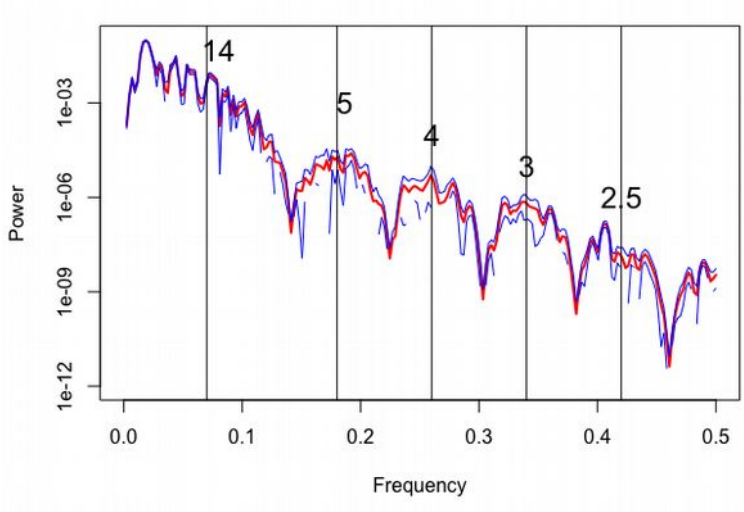

Figure 4. Averaged periodograms of the 68 D2 spectra. *(red: mean; blue: standard deviation); vertical lines and numbers point to the five first frequency modes and their corresponding periods T = 1/f). In the time series analogy, the 2 cm^{-1} signal discretization corresponds to a 0.5 Hz data sampling*.

Fig.4 firstly shows that the main modulation modes of D2 signals are rather stable for a given data set since they present only small oscillations, if any, that allow defining a common *frequency* pattern for all individuals considered in a particular study. In this case, the main signal modulations are observed at periods of 14, 5, 4, 3, and 2.5 cm^{-1} which closely match the optimized spline's frames of 7, 5 et 3 cm^{-1}. The "14" mode is the exact harmonic of "7", the "5" and "3" modes exactly match those which have been selected for their predicting

efficiency, while the “2.5” is half the “5” mode. These agreements between the modulation modes of D2 signals and the B-splines widths give consistency for the B-splines approach for reducing D2 spectra from B-splines coefficients that map any D2 spectrum as a network of patterns that enfold the signal modulations as soon as the B-splines frames are tuned to the signal intrinsic *frequency* modes. Therefore a mere frequency analysis (periodograms) would reveal the *frequency* modes present in the D2 signals set and allow tuning the B-splines frames (here: “7”, “5” and “3”) to further perform reconstructions and regression analyses. Note that *frequency* (in italic) refers to intrinsic D2 modulation modes, not to wavenumbers.

As a test for the efficiency of B-splines coefficients to reveal spectral biomarkers from patterns (the B-splines), a previously published data set was used that were mice sera exhibiting different levels of steatosis in a nutritional context (High fat +/- high carbohydrate, or control diet) [15].

A combination of B-spline resolutions (7, 5, and 3 here) increases hepatic steatosis prediction power whatever the chosen threshold. Indeed AUROC of prediction > 1.5% macro-steatosis increases from 0.91 (using D2 as input) to 0.94 (using 3 combined B-spline resolutions as input) (Fig. 5A). For sake of comparison, figures including results for approximations based on discrete wavelets and Fourier transform are provided in Supplementary Material. The mean AUROC is about 0.91 for the wavelets and 0.82 for the Fourier transform. Then prediction > 5% macro-steatosis is improved by combining B-spline resolutions (7, 5, and 3): AUROC is closed to 0.92 using combined B-splines as input compared to an AUROC closed to 0.88 using D2 as input.

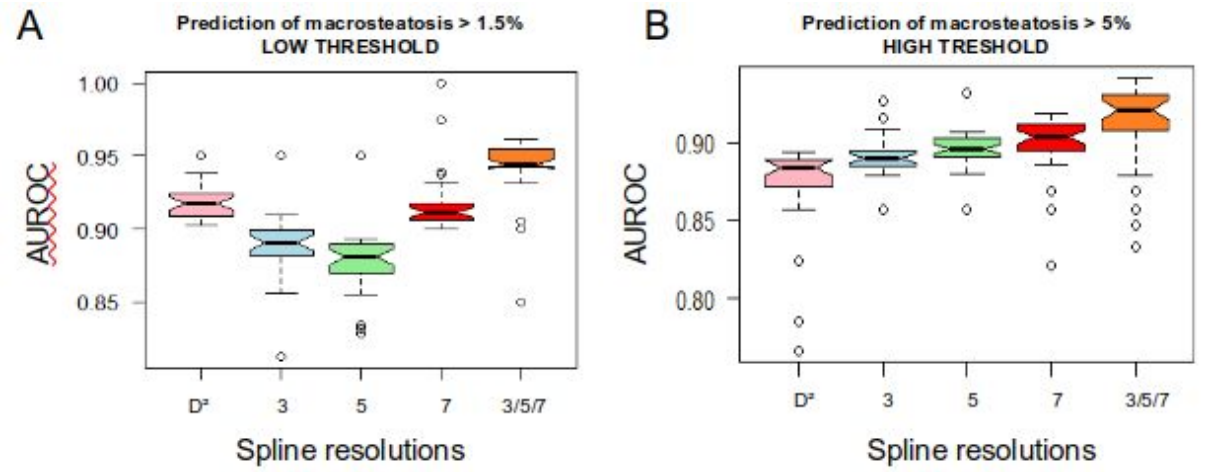

Figure 5. Performance of hepatic steatosis predictions depending on macrosteatosis threshold. *A) Low threshold > 1.5% of macro-steatosis. AUROC of validation sets (n = 30 x 12). B) High threshold > 5% of macro-steatosis. AUROC of validation sets (n = 30 x 12). Combined B-spline resolutions (orange boxplot) are statistically different compared to all other groups. Similar scores obtained from discrete wavelet and Fourier transform approximations are provided In supplementary material (Figure S2).*

Moreover, prediction of the disease by MIR metabolic fingerprint informs about important spectral bands, yielding relevant biological information about the disease. Figure 6 shows the importance of the variables for the different models. Remark that about 15 variables have an importance that is significantly higher than the others. In this context, we have confirmed that combined B-spline resolutions improve clinical understanding of the disease compared to raw data or a single B-spline resolution. Based on a previous study [18], we

might associate bands in the region of ester C=O stretching mode provided by resolutions 3, 5, and 7 to distinct molecules. Indeed, the large bands detected by resolution 7 are associated with a phospholipid, while medium width bands (detected by resolution 5) are more likely related to unsaturated lipids (i.e. Triolein) or LDL (i.e. cholesteryl linoleate) and low width bands (detected by resolution 3) are associated to saturated lipids (i.e. Tripalmitin, cholesteryl palmitate). Therefore, B-spline resolutions provide an advanced comprehension of the disease, in particular of the bands in the region of ester C=O stretching mode.

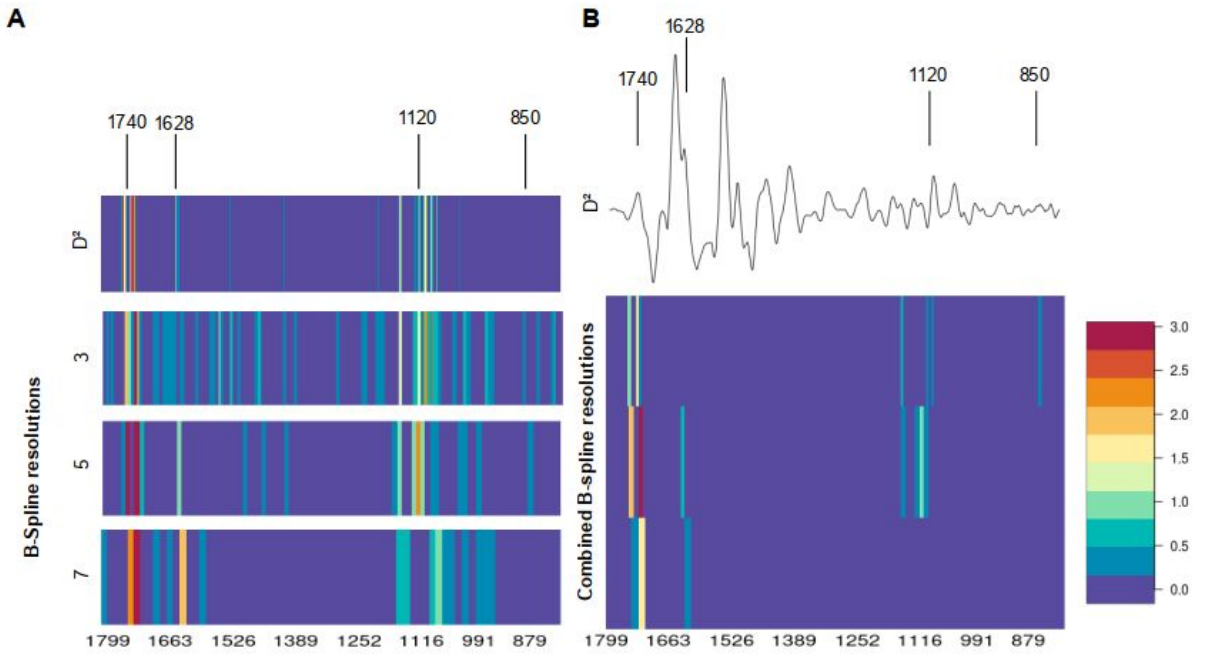

Figure 6. Combined splines resolution and important variables for steatosis low threshold prediction. *A) Discriminative spectral bands selected from varible importance outputs of RF algorithm in 4 different single B-spline resolutions (D2 , 3, 5 or 7). B) Important spectral bands from RF algorithm for combined B-spline resolutions (7, 5, and 3 B-spline resolutions of the same Data set). Gradient color displays the weight of each spectral domain. In Supplementary Material, important variables are provided for the models based on discrete wavelets and Fourier transform (Figure S3).*

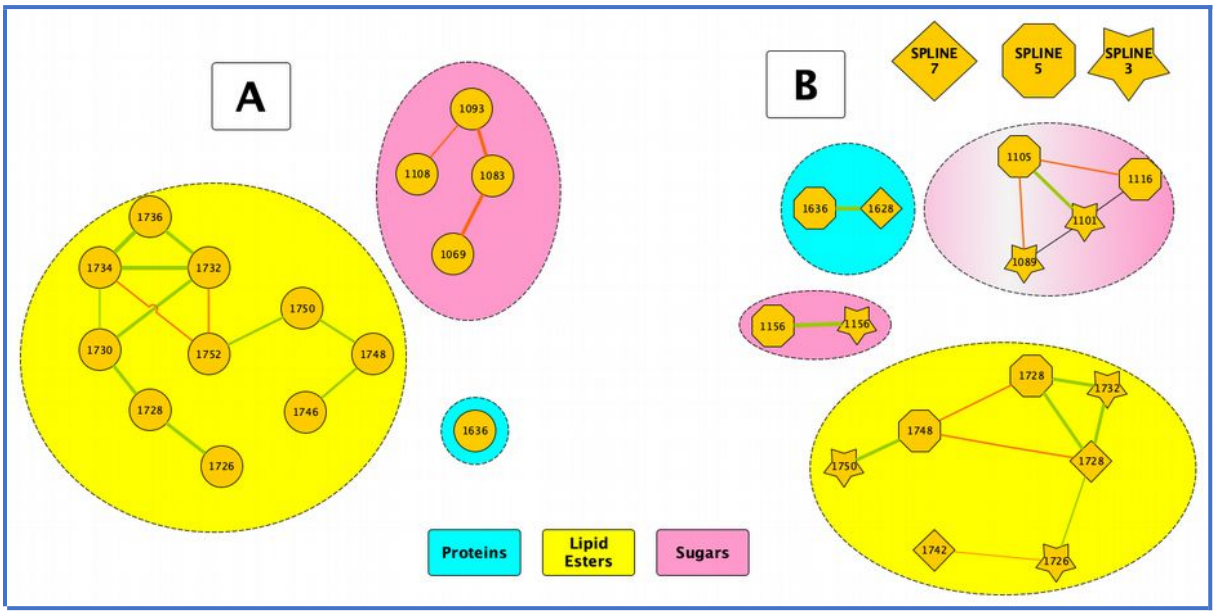

Fig.7 Biomarker's networks as identified from raw spectra (A) and from 3 resolutions B-splines reconstructions (B). *The edges represent the most significant partial correlations between variables. Green edges show positive partial correlations, red ones negative partial correlations; the width of the lines is a function of the partial correlation values.*

Using D2 spectra or 3 B-splines resolutions leads to variable selections where lipids esters, glucids and a protein biomarker (1636 cm-1) show up. Fig. 7 highlights the dependency structure of these spectral biomarkers. More precisely, the graphs, obtained from Graphical Lasso [18], represent the significant partial correlations of the 15 most important variables. Partial correlation measures the degree of association between two variables, with the effect of the other random variables removed.

In the case of D2 spectra (Fig. 7A) it is obvious that the edges mainly represent redundancy of information, due to the proximity of the variables. Indeed, the variables, especially in the lipid esters cluster, are contiguous. In contrast, the network that emerges from B-spline reconstruction (Fig.7B) captures the dependencies between more distant variables, indicating thereby meaningful correlations between different classes of molecules.

4. Conclusion

The novel approach for data compression that we developed allows us to unveil networks of biomarkers from B-splines reconstruction that significantly improve the sensibility and specificity of MIR spectroscopy approach to discriminate clinical situations by studying serum. Moreover, a frequency analysis through a time series analysis of the second derivative signals shows that particular B-splines are preferentially tuned to the inherent D2 periodograms. A few modulations arise in the second derivative that appear to be steady among a particular data set. The developed multi-dimensional resolution matches - *is tuned on*- these *frequency* modes and benefits from the information that is carried on by the signal's continuity, which is not the case of algorithms that only consider spectral variables individually. AUROCs demonstrate that multi-resolution improve the discrimination between cases and controls. Multi-resolution B-splines also improve the understanding of the disease by showing that the main differences between control and cases groups are in saturated and unsaturated ester lipid absorbance bands. This B spline process opens new perspectives in patient management.

Declaration of competing interest

O. Loréal is cofounder of Diafir company and shareholder

Acknowledgments

O. Loréal and O. Sire acknowledge financial support from Institut Carnot *Agrifood*.

Author statement

V. Monbet: multi-resolution conceptualization and software, writing and editing; **D. Martin**: computing, data curation, writing; **M. Le Corvec**: data acquisition; **O. Sire**: conceptualization, writing and editing; **O. Loréal**: Diagnostic conceptualization.

References

[1] S. De Bruyne, Applications of mid-infrared spectroscopy in the clinical laboratory setting, Critical Reviews in Clinical Laboratory Sciences, 55 (2017) 1-20.

[2] L. Bel'skaya, Use of IR Spectroscopy in Cancer Diagnosis. A Review, J. App. Spect., 86 (2019) DOI:10.1007/s10812-019-00800-w

[3] H. Tariel, O. Sire, MIR spectroscopy: the medical diagnosis swiss knife ?, PhotonicsViews, 1 (2021) DOI: 10.1002/phvs.202100022.

[4] J.D. Albert, M. Le Corvec, O. Berthoud, C. David, X. Guennoc, E. Hoppe, S. Jousse-Joulin, B. Le Goff, H. Tariel, O. Sire, A. Jolivet-Gougeon, G. Coiffier, O. Loréal, Ruling out septic arthritis risk in a few minutes using mid-infrared spectroscopy in synovial fluids, Rheumatology, 60 (2021) 1158-1165.

[5] R. Anty, M. Morvan, M. Le Corvec, C. M. Canivet, S. Patouraux, J. Gugenheim, S. Bonnafous, B. Bailly-Maitre, O. Sire, H. Tariel, J. Bernard, T. Piche, O. Loréal, J. Aron-Wisnewsky, K. Clément, A. Tran, A. Iannelli, P. Gual, The mid-infrared spectroscopy: A novel non-invasive diagnostic tool for NASH diagnosis in severe obesity, JHEP Reports, 1 (2019) 361-368.

[6] L. Lovergne, J. Lovergne, P. Bouzy, V. Untereiner, M. Offroy, R. Garnotel, G. Thiéfin, M. J. Baker, G.D. Sockalingum, Investigating pre-analytical requirements for serum and plasma based infrared spectro-diagnostic, J. Biophotonics, 12 (2019) e201900177.

[7] R. Liu, K.i Aihara, L. Chen, Dynamical network biomarkers for identifying critical transitions and their driving networks of biologic processes, Quantitative Biology 1 (2013) 105–114.

[8] M. Le Corvec, C. Jezequel, V. Monbet, N. Fatih, F. Charpentier, H. Tariel, C. Boussard-Plédel, B. Bureau, O. Loréal, O. Sire, E. Bardou-Jacquet, Mid-infrared spectroscopy of serum, a promising non-invasive method to assess prognosis in patients with ascites and cirrhosis, PLoS ONE, 12 (2017) e0185997.

[9] K. Alsberg & O. M. Kvalheim, Compression of *n*th-order data arrays by B-splines. Part 1: theory, J. chemometrics 7 (1993) 61-73.

[10] Alsberg, B. K., E. Nodland, and O. M. Kvalheim. "Compression of Nth-Order Data Arrays By B-Splines. 2. Application to 2nd-Order Ft-Ir Spectra." Journal of Chemometrics 8.2(1994):127-145

[11] B. Walczak, B. van den Bogaert, D. L. Massart, Application of wavelet packet

transform in pattern recognition of near-IR data, Anal. Chem. 68 (1996) 1742-1747.B.

[12] D. Wu, X. Chen, P. Shi, S. Wang, F. Feng, Y. He, Determination of a-linolenic acid and linoleic acid in edible oils using near-infrared spectroscopy improved by wavelet transform and uninformative variable elimination, Analytica Chimica Acta, 634 (2009) 166-171.

[13] F. Rossi, D. François, V. Wertz, M. Meurens, M. Verleysen, Fast selection of spectral variables with B-spline compression, Chemometrics and intelligent laboratory systems, 86 (2007) 208-218

[14] S. Mallat, Multiresolution approximation and wavelets, Trans. Amer. Math. Soc. 315 (1989) 69-88.

[15] [dataset] M. Le Corvec, C. Allain, S. Lardjane, T. Cavey, B. Turlin, A. Fautrel, K. Begriche, V. Monbet, B. Fromenty, P. Leroyer, P. Guggenbuhl, M. Ropert, O. Sire, O. Loréal, Mid-infrared fibre evanescent wave spectroscopy of serum allows fingerprinting of the hepatic metabolic status in mice, Analyst, 141, 6259-6269.

[16] W. Liang, A.L. Menke, A. Driessen, G.H. Koek, J.H. Lindeman, R. Stoop, A.M. van den Hoek, Establishment of a general NAFLD scoring system for rodent models and comparison to human liver pathology. PloS one, 9 (2014) e115922.

[17] Di Rocco, H.O.; Iriarte, D.I.; Pomarico, J. General Expression for the Voigt Function That is of Special Interest for Applied Spectroscopy. Appl. Spectrosc. (2001), 55, 822–826.

[18] M. Nara, M. Okazaki, H. Kagi, Infrared study of human serum very-low-density and low-density lipoproteins. Implication of esterified lipid CO stretching bands for characterizing lipoproteins, Chem. Phys. Lipids, 117 (2002) 1-6.