Orthogonal Projections to Latent Structures Discriminant Analysis Modeling on in Situ FT-IR Spectral Imaging of Liver Tissue for Identifying Sources of Variability

ARTICLE *in* ANALYTICAL CHEMISTRY · SEPTEMBER 2008

Impact Factor: 5.64 · DOI: 10.1021/ac8005318 · Source: PubMed

5 AUTHORS, INCLUDING:

Hans Stenlund
Umeå University
25 PUBLICATIONS **327** CITATIONS
SEE PROFILE

András Gorzsás
Umeå University
30 PUBLICATIONS **505** CITATIONS
SEE PROFILE

Per Persson
Lund University
144 PUBLICATIONS **4,177** CITATIONS
SEE PROFILE

Johan Trygg
Umeå University
125 PUBLICATIONS **7,329** CITATIONS
SEE PROFILE

Orthogonal Projections to Latent Structures Discriminant Analysis Modeling on in Situ FT-IR Spectral Imaging of Liver Tissue for Identifying Sources of Variability

Hans Stenlund,[†] András Gorzsás,[‡] Per Persson,[†] Björn Sundberg,[‡] and Johan Trygg*,[†]

Computational Life Science Cluster (CLIC), KBC, Umeå University, SE-901 87 Umeå, Sweden, and Department of Forest Genetics and Plant Physiology, Umeå Plant Science Centre, Swedish Agricultural University, SE-901 83 Umeå, Sweden

In this study, the orthogonal projections to latent structures discriminant analysis (OPLS-DA) method was used to assess the in situ chemical composition of two different cell types in mouse liver samples, hepatocytes and erythrocytes. High spatial resolution FT-IR microspectroscopy equipped with a focal plan array (FPA) detector is capable of simultaneously recording over 4000 spectra from 64 × 64 pixels with a maximum spatial resolution of about 5 μm × 5 μm, which allows for the differentiation of individual cells. The main benefit with OPLS-DA lies in the ability to separate predictive variation (between cell type) from variation that is uncorrelated to cell type in order to facilitate understanding of different sources of variation. OPLS-DA was able to differentiate between chemical properties and physical properties (e.g., edge effects). OPLS-DA model interpretation of the chemical features that separated the two cell types clearly highlighted proteins and lipids/bile acids. The modeled variation that was uncorrelated to cell type made up a larger portion of the total variation and displayed strong variability in the amide I region. This could be traced back to a gradient in the high intensity (high-density) areas vs the low intensity areas (close to empty areas) that as a result of normalization had an adverse effect on FT-IR spectral profiles. This highlights that OPLS-DA provides an effective solution to identify different sources of variability, both predictive and uncorrelated, and also facilitates understanding of any sampling, experimental, or preprocessing issues.

The standard procedure to investigate the in situ chemistry, biology, or pathology of medical samples today focuses largely on various staining methods. While these have the advantages of providing high spatial resolution and high specificity, they have certain drawbacks, inherent to their nature. While their specificity is no doubt advantageous, it is also a limitation, as in order to gain additional information, several different stains may need to be used and conclusions drawn from their combination.

With the arrival of focal plan array (FPA) detectors of high spatial resolution, FT-IR microspectroscopy provides an alternative for in situ (i.e., in their original position, without microdissections, chemical extractions with subsequent wet chemistry, etc.) determination of the chemical composition and molecular structures of a wide range of samples. Given that the frequencies of IR radiation are on the scale of micrometers, it can differentiate between individual cells of a certain tissue.[1] On the other hand, a complete FT-IR spectrum is recorded for each pixel of the image, providing a powerful tool for pathology.[2,3] The spatial information is required to encompass inherent heterogeneity in tissue.[4] However, the spatial resolution of the IR image is dependent on the wavelength of the imaging light and as a result the higher wavenumbers have higher spatial resolution compared to the lower wavenumbers.[5] Differences between pixels do not necessarily represent chemical composition or molecular structure differences but can be due to light scattering and differences in spectroscopic path lengths (i.e., additive and multiplicative variation) or due to sample preparation and experimental procedure. For the removal of undesirable systematic variation, spectral preprocessing has traditionally been performed by signal correction methods like Savitzky−Golay smoothing,[6] multiplicative signal correction (MSC),[7] variable selection,[8] and baseline correction.[9] More elaborate methods, for example, the extended multiplicative signal correction (EMSC),[10] estimate effective optical pathlengths, and additive spectral effects. Alternative baseline correction methods and basic filters are minimum-maximum, total sum, and

[*] Corresponding author. E-mail: johan.trygg@chem.umu.se.

[†] Umeå University.

[‡] Swedish Agricultural University.

(1) Krafft, C.; Sergo, V. *Spectrosc. Int. J.* **2006**, *20*, 195–218.

(2) Bambery, K. R.; Schultke, E.; Wood, B. R.; MacDonald, S. T. R.; Ataelmannan, K.; Griebel, R. W.; Juurlink, B. H. J.; McNaughton, D. *Biochim. Biophys. Acta* **2006**, *1758*, 900–907.

(3) Bonnier, F.; Rubin, S.; Venteo, L.; Krishna, C. M.; Pluot, M.; Baehrel, B.; Manfait, M.; Sockalingum, G. D. *Biochim. Biophys. Acta* **2006**, *1758*, 968–973.

(4) Krafft, C.; Shapoval, L.; Sobottka, S. B.; Geiger, K. D.; Schackert, G.; Salzer, R. *Biochim. Biophys. Acta* **2006**, *1758*, 883–891.

(5) Lasch, P.; Naumann, D. *Biochim. Biophys. Acta* **2006**, *1758*, 814–829.

(6) Savitsky, A.; Golay, M. J. E. *Anal. Chem.* **1964**, *36*, 1627–1639.

(7) Geladi, P.; MacDougall, D.; Martens, H. *Appl. Spectrosc.* **1985**, *3*, 491–500.

(8) Baroni, M.; Clementi, S.; Cruciani, G.; Costantino, G.; Riganelli, D.; Oberrauch, E. *J. Chemom.* **1992**, *6*, 347–356.

(9) Barnes, R. J.; Dhanoa, M. S.; Lister, S. J. *Appl. Spectrosc.* **1989**, *43*, 772–777.

(10) Kohler, A.; Bertrand, D.; Martens, H.; Hannesson, K.; Kirschner, C.; Ofstad, R. *Anal. Bioanal. Chem.* **2007**, *389*, 1143–1153.

10.1021/ac8005318 CCC: $40.75 © 2008 American Chemical Society
Published on Web 08/20/2008

quadratic total sum normalizations as well as first and second order differentials.[11] Empty (low intensity) regions, near holes, near fissures, or near margins constitute an additional dimension of complexity and should preferably be removed.[12] However, undesirable systematic variation will still be present especially considering the difference in spatial resolution at different wavenumbers. Recently, the orthogonal projections to latent structures (OPLS) methods (OPLS,[13] OPLS-DA,[14] K-OPLS,[15] and O2PLS[16]) have been put forth as alternative modeling techniques capable of this and thereby facilitate the separation and interpretation of the different types of variation in the data.

In this study, the OPLS-DA method was used to assess the in situ chemical composition of mouse liver cells of two different cell types, hepatocytes and erythrocytes. Liver samples are heterogeneous with different density distributions, which create edge effects, normalization, and baseline correction issues. Hence, we considered the samples to be ideal for the demonstration of the power of OPLS-DA and illustrating its versatility and usability for prediction modeling in biological and medical research and for detecting different sources of variation to gain further insight of the studied system.

NOTATIONS

The following notation has been used throughout. Vectors are denoted by bold, lowercase letters and are assumed to be column vectors unless indicated by a transposition, for example, $\mathbf{p}^T$. Matrices are denoted by bold uppercase letters, for instance, $\mathbf{X}$, and matrix inverses are denoted by $\mathbf{X}^{-1}$.

EXPERIMENTAL SECTION

Sample Preparation. Liver from a 10 week old BALB/c female mouse was fixed in 4% paraformaldehyde in phosphate-buffered saline (PBS) overnight at 4 °C and then equilibrated in 30% sucrose in PBS. The tissue was embedded in Tissue Tec (Sakura) and frozen to −80 °C. Consecutive 10 μM thick cryo-sections were collected on SuperFrost Plus glass slides (Menzel-Gläser) and on rectangular IR-transparent BaF_2 crystals (International Crystal Laboratories). The sections on SuperFrost Plus glass were stained with Mayer's hematoxylin (Dako Corporation) for reference. Thus, sections used in the FT-IR analysis were not identical to the stained ones, but the same area of the liver sample was used in both cases (compare parts a and b of Figure 1, as well as parts e and f of Figure 1).

FT-IR Microspectroscopic Imaging. Spectra were recorded on a Bruker Equinox 55 spectrometer equipped with a microscopy accessory and a 64 × 64 focal plan array (FPA) detector (Hyperion 3000), providing a maximum spatial resolution of approximately 5 μm at about 4000 cm^{-1}. Visual photographs of the samples were taken by the FT-IR microscope for spectral overlay, using a Sony Exwave HAD color digital video camera mounted on the top of the microscope. The sample tray was boxed, and the chamber was continuously purged with dry air. Cryotome sections (10 μm thick) mounted on polished rectangular BaF_2 windows were stored in a desiccator for at least 48 h for drying. Spectra were recorded in transmission mode over the range of 850−3850 cm^{-1} with a spectral resolution of 4 cm^{-1}. For each image, 64 interferograms were coadded to obtain high signal-to-noise ratios. Background spectra were recorded for each sample at a nearby empty spot on the BaF_2 crystal, prior to sample measurement, with the same number of scans. The large region in the upper right corner (i.e., first quartile) of the mouse liver A sample was a blood vessel, and the dark area within the vessel was red blood cells. Similarly, the large round region in the center of the mouse liver B sample was also a blood vessel with red blood cells around and within it (dark areas). With the use of staining (Mayer's hematoxylin) and visual photograph, all other nonempty areas were interpreted as liver tissue of the same cell type (hepatocytes; parts b and f of Figure 1). Similarly, hepatocyte and erythrocyte areas could be found and identified in the visual images of the mouse liver C-L samples (Figure 6).

Data Preprocessing and Normalization of FT-IR Imaging Data. Each hyperspectral image was unfolded to a two-dimensional data matrix (**X**) where each row constitute an FT-IR spectrum at a specific (*x*,*y*)-coordinate in the microspectroscopic image. Each spectrum was then processed by minimum value correction, baseline and offset correction, and total sum normalization as specified in the Supporting Information. A representative training set was constructed based on mouse liver A sample by selecting 80 pixels from identified red blood cell areas (class 1, erythrocytes) and 80 pixels from identified liver tissue areas (class 2, hepatocytes), marked as red and blue boxes, respectively, in Figure 1d. A total of 1776 pixels in the mouse liver A data and 1368 pixels in the mouse liver B data were identified and labeled as empty areas (orange boxes). The remaining 2160 pixels in the mouse liver A data and 2728 pixels in the mouse liver B data were left unassigned and made up the prediction set in our analysis. Similarly, empty areas were excluded from the analysis of mouse liver C-L samples. The remaining (nonempty) pixels were left unassigned.

METHOD

Principal component analysis (PCA)[17,18] is the workhorse of chemometrics. PCA is often used to get an overview of a data table **X**, detect clusters, and identify anomalies and outliers. It is a projection method that captures all systematic variation in a data table **X** [NxK] into a few new "latent" variables, called scores **T**. The loading matrix **P** describes the influence of each of the original variables in the construction of the scores **T**.

$$\text{Model of data:}\quad \mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \tag{1}$$

OPLS[13] is an extension to the supervised PLS[19] regression method featuring an integrated OSC[20]-filter. It uses information in the **Y**

[11] http://www.brukeroptikcs.com/opus.

[12] Krafft, C.; Kirsch, M.; Beleites, C.; Schackert, G.; Salzer, R. *Anal. Bioanal. Chem.* **2007**, *389*, 1133–1142.

[13] Trygg, J.; Wold, S. *J. Chemom.* **2002**, *16*, 119–128.

[14] Bylesjo, M.; Rantalainen, M.; Cloarec, O.; Nicholson, J. K.; Holmes, E.; Trygg, J. *J. Chemom.* **2006**, *20*, 341–351.

[15] Rantalainen, M.; Bylesjö, M.; Cloarec, O.; Nicholson, J. K.; Holmes, E.; Trygg, J. *J. Chemom.* **2007**, *21*, 376–385.

[16] Trygg, J.; Wold, S. *J. Chemom.* **2003**, *17*, 53–64.

[17] Wold, S.; Esbensen, K.; Geladi, P. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52.

[18] Jackson, J. E. *J. Educ. Behav. Stat.* **1995**, *20*, 105–107.

[19] Naes, T.; Isaksson, T.; Fearn, T.; Davies, T. *Multivariate Calibration and Classification*; NIR Publications: Chichester, U.K., 2004.

[20] Wold, S.; Antti, H.; Lindgren, F.; Öhman, J. *Chemom. Intell. Lab. Syst.* **1998**, *44*, 175–185.

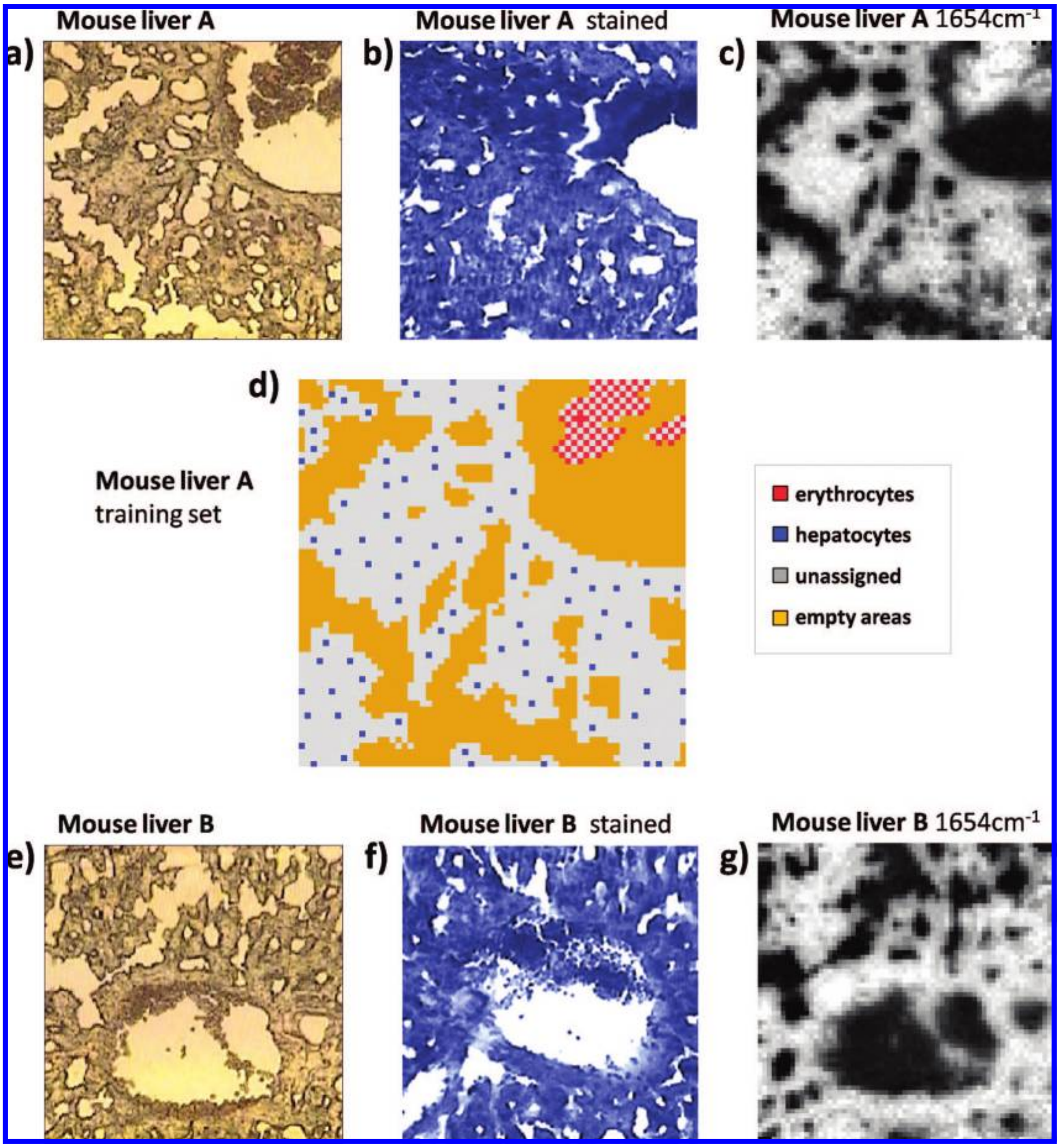

Figure 1. Data analysis. (a, e) Unstained visual images of the mouse liver samples A and B, respectively, captured by the camera of the FT-IR microspectroscope. (b, f) The corresponding stained sections (Mayer's hematoxylin) of the respective samples. (c, g) FT-IR images (intensities at 1654 cm^{-1}) for mouse liver samples A and B, respectively. The pixels were colored from black to white with increasing absorbance values at 1654 cm^{-1}. (d) Subset selection of 80 pixels from each class for training set on the mouse liver A data. Class 1, red, erythrocytes; class 2, blue, hepatocytes.

matrix to decompose the **X** matrix into two blocks of systematic variation, one correlated to **Y** (predictive) and the other orthogonal to **Y** (Y-orthogonal), respectively. OPLS can, analogously to PLS-DA, be used for discrimination (OPLS-DA). OPLS-DA uses information in the categorical response matrix **Y** to decompose the **X** matrix into three distinct parts as described in eq 2, where $\mathbf{T}_p$ denotes the predictive score matrix for **X**, $\mathbf{P}_p$ denotes the predictive loading matrix for **X**, $\mathbf{T}_o$ denotes the corresponding Y-orthogonal score matrix, $\mathbf{P}_o$ denotes the loading matrix of Y-orthogonal components, and **E** denotes the residual matrix of **X**. Further details of the OPLS algorithm are described by Trygg.[13,16]

$$\mathbf{X} = \mathbf{T}_p\mathbf{P}_p^{\mathrm{T}} + \mathbf{T}_o\mathbf{P}_o^{\mathrm{T}} + \mathbf{E} \quad (2)$$

The main benefit of OPLS-DA compared to PLS-DA thus lies in the ability of OPLS-DA to separate predictive from nonpredictive (Y-orthogonal) variation. The Y-orthogonal variation on FT-IR imaging data make up the within-class variance and can be, e.g., light scattering, differences in spectroscopic path lengths, sample preparation, and experimental problems, etc. OPLS-DA effectively separates the discriminatory direction in $\mathbf{t}_{1p}$ from the Y-orthogonal direction $\mathbf{t}_{1o}$ making the corresponding loading vectors $\mathbf{p}_{1p}$ and $\mathbf{p}_{1o}$ straightforward to interpret. The following statistics for the regression models have been calculated (**X** and **Y** column centered).

Explained variance of **X**, of training set:

$$\mathrm{R}^2\mathrm{X} = 1 - SS(\hat{X} - X)/SS(X) \quad (3)$$

The cross validated variance of **Y**:

$$Q^2Y = 1 - SS(\hat{Y}_{\text{pred}} - Y)/SS(Y) \qquad (4)$$

Multivariate analysis was performed with the software SIMCA-P+, version 11.5 (Umetrics AB, Umeå, Sweden). All columns in the data set were mean centered but not scaled. For all score plots, the Hotelling 95% confidence ellipse was drawn.

RESULTS AND DISCUSSION

Principal Component Analysis Modeling. PCA is used to get an overview of the variation in the data and also to detect possible outliers, grouping, and trends in the data. A two component PCA model was calculated on the selected training set based on the data from mouse liver A ($R^2X = 0.80$, $Q^2Y = 0.79$). Figure 2 (upper left; $\mathbf{t}_1-\mathbf{t}_2$ scores plot) includes not only the training set scores but also the predicted scores of the unassigned and empty region pixels. As observed in the $\mathbf{t}_1-\mathbf{t}_2$ scores scatter plot, there is a degree of separation between the hepatocytes and the erythrocytes but only when combining components 1 and 2, otherwise they strongly overlap in any single component. The reconstructed $\mathbf{t}_1$ score image described general intensity differences between high-intensity (high-density) areas and low-intensity areas (close to empty regions), for the hepatocytes (Figure 2, middle left; $\mathbf{t}_1$ scores image). The high intensity hepatocyte areas (positive score values) were also seen to differ from the main part of the erythrocyte

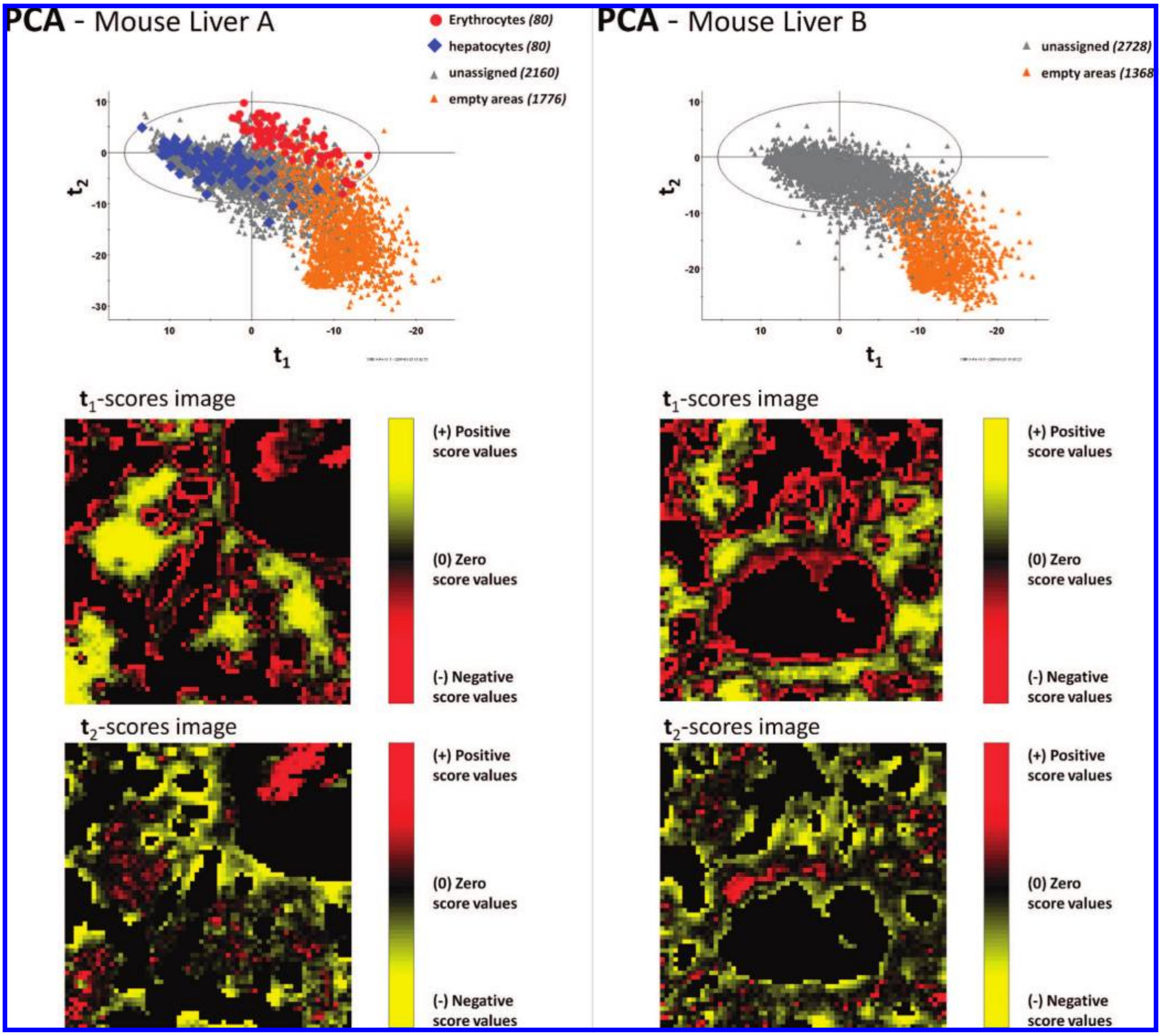

Figure 2. Overview of a two component PCA model on the FT-IR spectra of erythrocytes and hepatocytes from the mouse liver A and B samples. Upper left: $\mathbf{t}_1-\mathbf{t}_2$ scores scatter plot including the predictive scores of the unassigned and empty area FT-IR spectra in the mouse liver A sample. A separation between the two different cell types was observed, but only in the combination of two model components. Middle left: pseudocolored predictive $\mathbf{t}_1$ scores image of mouse liver A sample that display a gradient between high-intensity areas (higher density) and low-intensity areas (near empty areas). Lower left: pseudocolored predictive $\mathbf{t}_2$ scores image of mouse liver A. Upper right: $\mathbf{t}_1-\mathbf{t}_2$ scores of the mouse liver B sample FT-IR spectra that have a similar overall distribution as the mouse liver A FT-IR spectra. Middle right: pseudocolored predictive $\mathbf{t}_1$ scores image for mouse liver B where again the high-intensity vs low-intensity regions are dominating instead of the class separation between the hepatocytes and erythrocytes. Lower right: pseudocolored predictive $\mathbf{t}_2$ scores image for mouse liver B. Empty areas were colored black for the scores images.

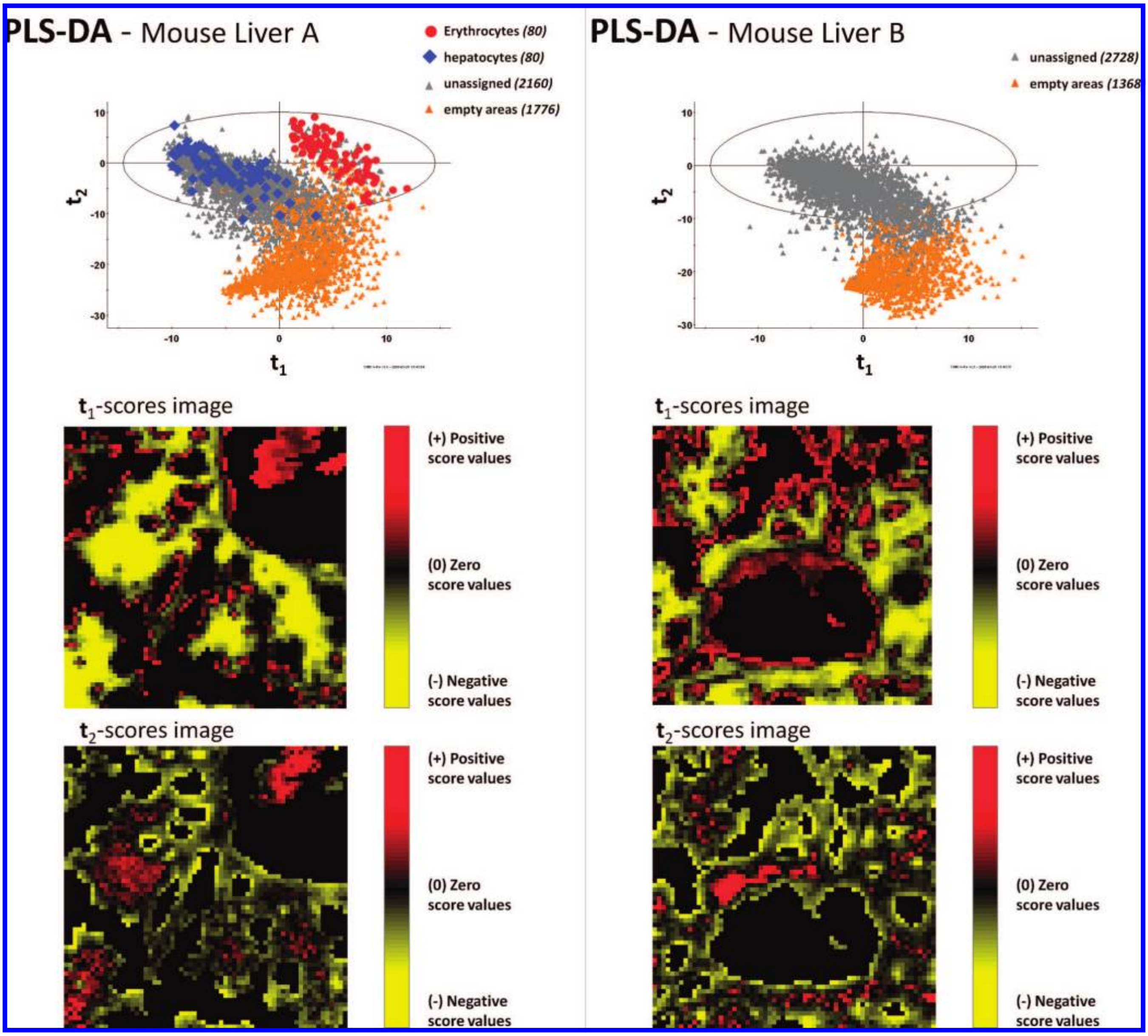

Figure 3. PLS-DA model fitted by four components, for 80 pixels labeled as erythrocytes and 80 pixels labeled as hepatocytes (see Figure 1). Upper left: predictive $\mathbf{t}_1$−$\mathbf{t}_2$ scores for mouse liver A data. Upper right: predictive $\mathbf{t}_1$−$\mathbf{t}_2$ scores for mouse liver B data. Middle left: pseudocolored predictive $\mathbf{t}_1$ scores image for mouse liver A data. Middle right: pseudocolored predictive $\mathbf{t}_1$ scores image for mouse liver B data. Lower left: pseudocolored predictive $\mathbf{t}_2$ scores image for mouse liver A data. Lower right: pseudocolored predictive $\mathbf{t}_2$ scores image for mouse liver B data $\mathbf{X}$.

areas (negative score values). The loadings ($\mathbf{p}_1$) showed a full spectral profile, not only specific bands. This in practice means that it is mainly related to sample density not compositional changes. Had the composition changed, the ratios between bands corresponding to different compounds would also have changed, so the loadings would show certain bands pointing to negative directions, whereas others would point to positive directions. Instead, the entire spectral range (i.e., all bands) points to one direction, meaning that the overall absorbance increased or decreased, owing to thicker or thinner samples at certain pixels. This in practice means that the cell type discriminating information was concealed by sample density effects. The reconstructed $\mathbf{t}_2$ score image showed high similarity to the $\mathbf{t}_1$ score image, except for the erythrocyte areas (Figure 2, lower left; $\mathbf{t}_2$ scores image). Hence, the second component had mainly captured sample density related effects. The effects of light scattering were found to be different at the edges of hepatocyte areas and in areas covered with the small "disks" of deposited erythrocytes, as compared to the dense, homogeneous areas of hepatocytes. In Figure 2 (upper, middle, and lower right) the predicted PCA scores of the mouse liver B data display a similar variability as the mouse liver A data. In summary, the PCA model captured both physical properties, namely, changes in sample thickness/density, as well as compositional differences between the hepatocytes and the erythrocytes. The sample standardization was seen as successful, since the predicted observations showed high resemblance to the prior

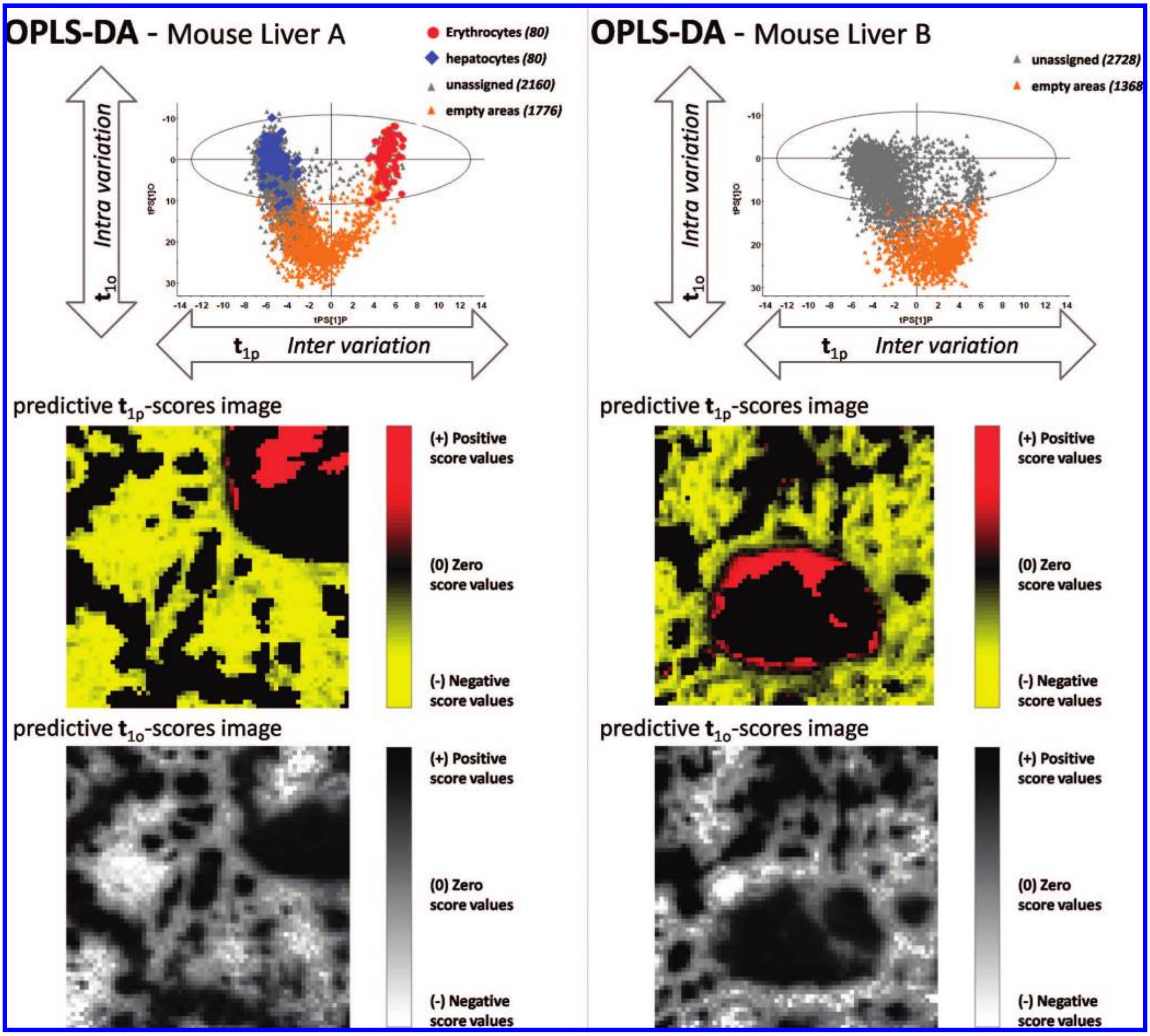

Figure 4. OPLS-DA model fitted by one predictive and three Y-orthogonal components, for 80 pixels labeled as erythrocytes and 80 pixels labeled as hepatocytes (see Figure 1). Upper left: predictive $\mathbf{t}_{1p}-\mathbf{t}_{1o}$ scores for mouse liver A data. Upper right: predictive $\mathbf{t}_{1p}-\mathbf{t}_{1o}$ scores for mouse liver B data. Middle left: pseudocolored predictive $\mathbf{t}_{1p}$ scores image for mouse live A data. Middle right: pseudocolored predictive $\mathbf{t}_{1p}$ scores image for mouse liver B data. Lower left: pseudocolored predictive $\mathbf{t}_{1o}$ scores image for mouse liver A data. Lower right: pseudocolored predictive $\mathbf{t}_{1o}$ scores image for mouse liver B data $\mathbf{X}$. For the $\mathbf{t}_{1p}$ score images, empty areas were colored black.

knowledge (staining) and no large unwanted effects had emerged during the process.

Classification by PLS-DA Modeling. In order to separate the hepatocyte areas from the erythrocyte areas, a four component PLS-DA model was calculated on the selected training set based on the data from mouse liver A. Out of the total variation in **X**, the first and second components constitute 53% ($R^2X[1] = 0.53$) and 26% ($R^2X[2] = 0.26$), respectively. The four component PLS-DA model described 86% of the total variation in **X**, and the model predictive ability was as high as 98%. In Figure 3 (upper left; $\mathbf{t}_1-\mathbf{t}_2$ scores plot), the hepatocytes and the erythrocytes were clearly separated by the two main components, both for the modeled areas and the predicted areas. In Figure 3 (middle left; $\mathbf{t}_1$ scores image), the reconstructed $\mathbf{t}_1$ scores image shows that the main part of the unassigned pixels was correctly assigned by the PLS-DA model. However for the hepatocyte areas, most of the pixels neighboring the empty regions were not properly assigned. The loadings ($\mathbf{p}_1$) showed a full spectral profile, identical to the first loading profile from the PCA modeling. This indicated again that sample density related effects were captured, affecting the interpretations of compositional differences. In Figure 3 (lower left; $\mathbf{t}_2$ scores image), the reconstructed $\mathbf{t}_2$ scores image shows that both the high-density areas of the hepatocytes as well as the high-density areas of the erythrocytes were predicted to positive score values, and the remaining areas were predicted to negative score values. In Figure 3 (upper right, middle right, and lower right), all FT-IR spectra from the mouse liver B sample are predicted using the PLS-DA model and the result display similar variability as for the mouse liver A data. In

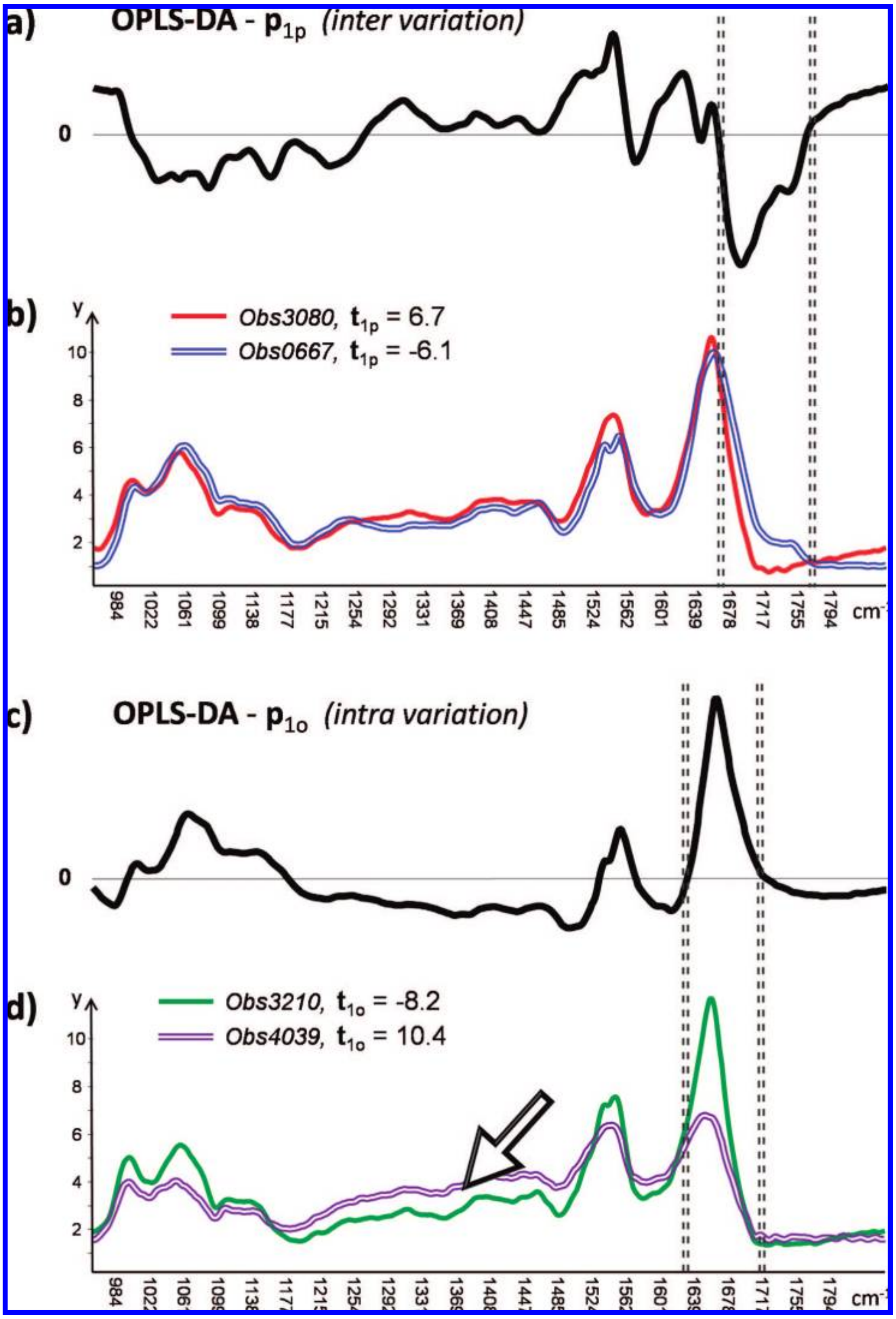

Figure 5. (a) OPLS-DA predictive loading $\mathbf{p}_{1p}$ highlight the 1680−1720 cm^{-1} region to be inversely correlated to the erythrocytes. (b) Two FT-IR spectra that span the variation in the predictive component. Note their similar overall shape (red is an erythrocytes pixel, and blue line is a hepatocytes pixel). (c) OPLS-DA Y-orthogonal loading profile, $\mathbf{p}_{1o}$, highlights a peak in the amide I region, 1630−1720 cm^{-1}. This indicates a strong variability in amide I group concentrations that is not related to cell type, since it represents Y-orthogonal variation. (d) Two FT-IR spectra that span the variation in the Y-orthogonal component. Note the clear amplitude difference in the amide I region that in turn distorted the overall spectrum, e.g., offset in the midregion (1150−1485 cm^{-1}) as a result of the normalization to equal sum.

summary, the outcome of the PLS-DA model showed high resemblances to the outcome of the PCA model, due to the large effects in **X** uncorrelated to class discrimination. The interpretability of the model was clearly compromised, since a multitude of components had to be taken into consideration. However, the predictive ability of the PLS-DA model for strictly classification purposes was successful.

Classification by OPLS-DA Modeling. A four component (1 + 3) OPLS-DA model was calculated, based on cross-validation.[17] The OPLS-DA model had one predictive (class discriminating) and three Y-orthogonal (uncorrelated to class discriminations) components ($R^2Xp = 0.4$, $R^2Xo = 0.50$, and $Q^2Y = 0.98$). Out of the total variation in **X**, the predictive variation makes up 40% and the Y-orthogonal variation constitutes 50%. This means that a

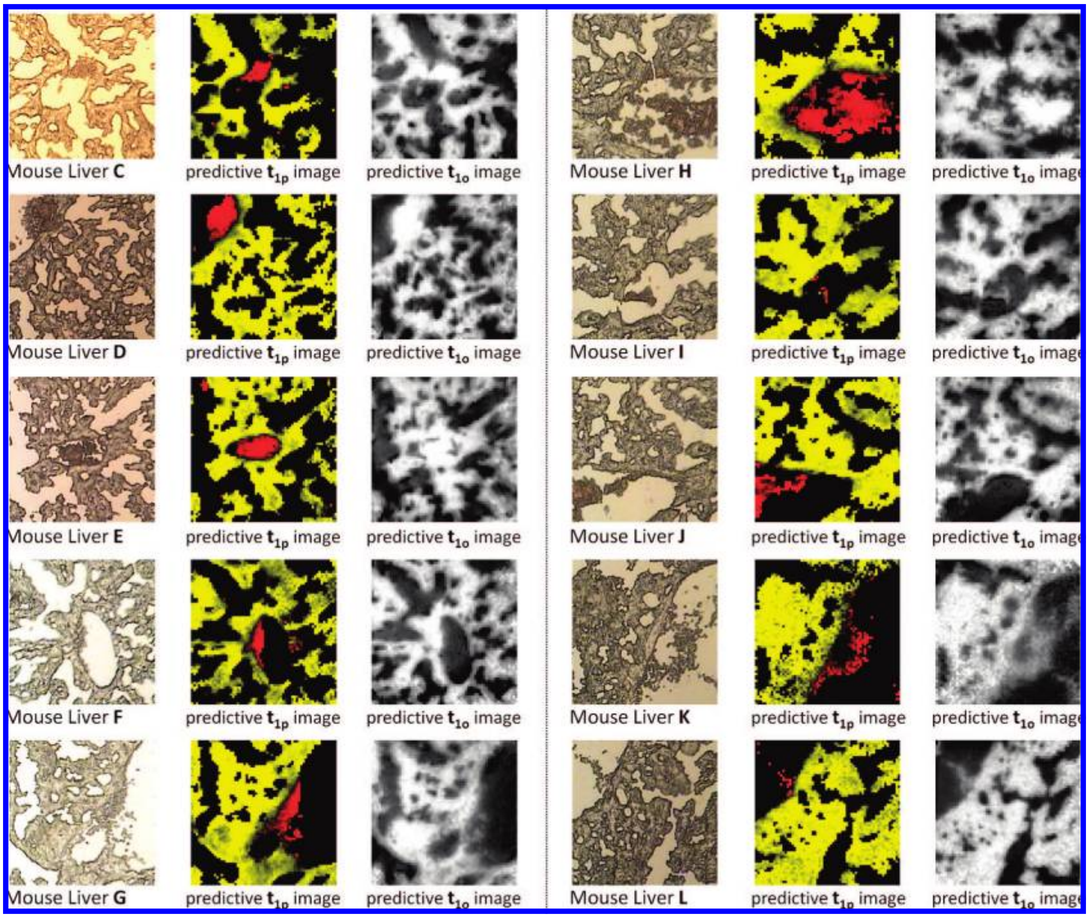

Figure 6. Visual, pseudocolored predictive $\mathbf{t}_{1p}$ scores and pseudocolored (grayscale) predictive $\mathbf{t}_{1o}$ scores images for the samples mouse liver C-L. In the case of the predictive $\mathbf{t}_{1p}$ images, the coloring is the same as in Figure 4, which means that red denotes areas predicted to be erythrocytes, while yellow corresponds to areas predicted to be hepatocytes. Predictions are based on the OPLS-DA model created from the sample mouse liver A (Figure 4). For the predictive $\mathbf{t}_{1p}$ score images, empty areas were colored black.

smaller fraction of the total modeled variation is correlated to the discrimination between hepatocytes and erythrocytes and a larger portion is uncorrelated to class discrimination. In Figure 4 (upper left; $\mathbf{t}_{1p}-\mathbf{t}_{1o}$ scores plot), the predictive component ($\mathbf{t}_{1p}$) clearly separated the hepatocytes from the erythrocytes and the predicted scores. In Figure 4 (middle left; $\mathbf{t}_{1p}$ scores image) the reconstructed predictive $\mathbf{t}_{1p}$ scores image shows that also the unassigned pixels have been correctly assigned by the OPLS-DA model. Interpretation of the chemical features that separated the two cell types can be made by Figure 5a, where the predictive loading vector ($\mathbf{p}_{1p}$) highlighted the wavenumber interval 1680−1720 cm^{-1}, which represents a −C=O stretch originating from proteins, to be inversely correlated to the erythrocytes. Figure 5b shows two typical FT-IR spectra of hepatocytes and erythrocytes. A clear band was seen at 1730 cm^{-1} for the class 2 area (hepatocytes), which was absent for class 1 (erythrocytes). This was again originating from −C=O stretching vibrations,[21−23] most likely of lipids/bile acids and not proteins.[24] In addition, a clear difference was also seen in the amide II stretching vibrations centered around 1550 cm^{-1}, indicating different protein composition for the two different cell types. In Figure 4 (top right and middle right), all FT-IR spectra from the mouse liver B sample are predicted using the OPLS-DA model. Again, the score scatter plot reveal similarity to the mouse liver A sample, and the reconstructed predictive $\mathbf{t}_{1p}$ score image further validates our findings that the OPLS-DA model correctly assigns the cell types.

The Y-orthogonal components reveal systematic variation that is uncorrelated to the separation of the two cell types. Here, they represent a larger portion of the modeled variation compared to the class discriminating variation. Understanding the origin of the Y-orthogonal variation can therefore be of great value in assessing the robustness and validity of our results. In Figure 4 (upper left; $\mathbf{t}_{1p}-\mathbf{t}_{1o}$ scores plot), the Y-orthogonal scores ($\mathbf{t}_{1o}$) are plotted on the *y*-axis and represent the largest within-class variation in the model. The reconstructed Y-orthogonal $\mathbf{t}_{1o}$-score image in Figure 4 (bottom left; $\mathbf{t}_{1o}$ scores image) facilitated the understanding of this Y-orthogonal variation. It was clear that the variability in the Y-orthogonal scores represented overall intensity differences between high intensity areas (high-density) and the low intensity areas (empty or close to empty area regions), similar to what the first PCA component showed previously. A similar finding was made in the predictions of the mouse liver B data; see Figure 4 (bottom right; $\mathbf{t}_{1o}$ scores image).

[21] Naumann, D.; Schultz, C. P.; Helm, D. In *Infrared Spectroscopy of Biomolecules*; Mantsch, H. H., Chapman, D., Eds.; Wiley: New York, 1996.

[22] Jackson, M.; Ramjiawan, B.; Hewko, M.; Mantsch, H. H. *Cell. Mol. Biol.* **1998**, *44*, 89–98.

[23] Lasch, P.; Boese, M.; Pacifico, A.; Diem, M. *Vib. Spectrosc.* **2002**, *28*, 147–157.

[24] Goro, C.; Kimiko, M.; Kumiko, A.; Shigeko, S. *Chem. Pharm. Bull.* **1962**, *10*, 1190–1195.

A deeper analysis of our findings was thereafter made to better understand the origin of the Y-orthogonal variation besides being an overall density/intensity problem. In Figure 5c, the $\mathbf{p}_{1o}$ loading profile reveals the influence of the wavenumber variables and it highlights the amide I region (1630−1720 cm^{-1}). This indicates a strong variability in the amide I group that is not related to cell type, since it represents Y-orthogonal variation. To further explore this, we decided to plot the two most diverse FT-IR spectra in the first Y-orthogonal component in Figure 5d. Interestingly, there was a clear amplitude difference in the amide I region, which in turn distorted the overall offset in the midregion (1150−1485 cm^{-1}) as a result of the normalization to equal sum. This created a negative dip artifact in the loading profile. Hence, OPLS-DA highlighted that normalization on FT-IR spectra with large spectral differences (other than offset or overall intensity) had an adverse effect on the outcome and produced spectral differences as those observed in Figure 5c,d.

The FT-IR spectra in the predictive component, see Figure 5b, did not suffer from this problem. However, the problem is not the normalization itself, since that merely passed on the already existing problems with edge effects, imperfect baseline correction, resolution variation issues, etc. that cannot be easily fixed. On the other hand, OPLS-DA provides the opportunity to understand this uncorrelated, Y-orthogonal variation to improve experimental, analytical, and data-analytical procedures.

The robustness and reproducibility of the method was thereafter tested on 10 more images (Figure 6). These were recorded at various spots of different mouse liver sections, approximately 1 year later than the original samples (mouse liver A and B). Predictions were then made using the original model. These new images provided different spectroscopic challenges, too. In some cases (mouse liver E, H, and J), the red blood cells form a dense, continuous deposit, whereas in others (notably mouse liver F, G, K, and L), the erythrocytes are scattered around. This, naturally, affects their spectral behavior, such as overall intensity, light scattering, etc. As can be seen in Figure 6 (predictive $\mathbf{t}_{1p}$ images), all new images were predicted correctly by the OPLS-DA model. This demonstrated that the proposed method was robust and provided reproducible results even over an extended period of time when the different parts of the experimental equipment were replaced. The orthogonal variation (predictive $\mathbf{t}_{1o}$ image) for all predicted images were also identical in nature to what had been observed for the original mouse liver A sample, i.e., corresponding mainly to overall intensity (density) variation.

In conclusion, OPLS-DA modeling could successfully differentiate class 1 and class 2 areas and correctly predict the class membership of unassigned areas. The Y-orthogonal variation in the OPLS-DA model revealed a problem related to overall intensity differences in the sample that could be derived from the sampling and experimental procedure. The preprocessing methods, albeit necessary, were unable to remove these. The OPLS-DA method, on the other hand, was able to highlight this problem and provide better opportunities for the researcher to minimize its effects.

CONCLUSIONS

Challenges attributable to the nature of biological and medical samples, such as sample heterogeneity, edge effects, biodiversity, etc., are commonly tackled by spectral standardization procedures followed by multivariate analysis. In this study, the OPLS-DA classification method was used to assess the in situ chemical composition of mouse liver cells of two different cell types, hepatocytes and erythrocytes, and correctly predicted class membership for unassigned areas of the images. One key advantage with OPLS-DA lies in its ability to separate predictive variation (between cell types) from variation that is uncorrelated to cell type in order to facilitate understanding of different sources of variation. The total adverse effect of sample density, imperfect baseline correction, edge-effects, noise, and biological variation could therefore be separately analyzed. These effects made up the larger part of the total variation, even after spectral preprocessing and standardization was performed. This highlights that OPLS-DA provides an effective solution to not only filter out unwanted variation but also to pinpoint any experimental and data processing errors that cannot effectively be removed in the preprocessing and normalization steps. The FT-IR imaging technique combined with OPLS modeling is powerful, versatile, and efficient, complementing already existing histopathological methods to aid in diagnosis and prognosis.

ACKNOWLEDGMENT

H.S. and A.G. contributed equally to this work. The authors would like to thank Marie Vejgården and Björn Eriksson (Betagenon, Umeå, Sweden) for providing the mouse liver sections. This work was financially supported by grants from FORMAS, The KEMPE Foundation, The Swedish Research Council (VR), and MKS Umetrics.

SUPPORTING INFORMATION AVAILABLE

Additional information as noted in text. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review March 13, 2008. Accepted July 14, 2008.

AC8005318