Made available through a Creative Commons CC-BY-NC-ND License

analytical chemistry

ACS AUTHORCHOICE

pubs.acs.org/ac Article

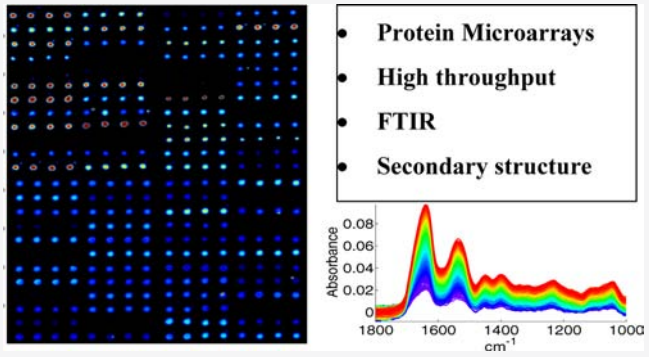
FTIR Imaging of Protein Microarrays for High Throughput Secondary Structure Determination

Joëlle De Meutter and Erik Goormaghtigh*

Cite This: *Anal. Chem.* 2021, 93, 3733−3741 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The paper introduces a new method designed for high-throughput protein structure determination. It is based on spotting proteins as microarrays at a density of ca. 2000−4000 samples per cm^2 and recording Fourier transform infrared (FTIR) spectra by FTIR imaging. It also introduces a new protein library, called cSP92, which contains 92 well-characterized proteins. It has been designed to cover as well as possible the structural space, both in terms of secondary structures and higher level structures. Ascending stepwise linear regression (ASLR), partial least square (PLS) regression, and support vector machine (SVM) have been used to correlate spectral characteristics to secondary structure features. ASLR generally provides better results than PLS and SVM. The observation that secondary structure prediction is as good for protein microarray spectra as for the reference attenuated total reflection spectra recorded on the same samples validates the high throughput microarray approach. Repeated double cross-validation shows that the approach is suitable for the high accuracy determination of the protein secondary structure with root mean square standard error in the cross-validation of 4.9 ± 1.1% for α-helix, 4.6 ± 0.8% for β-sheet, and 6.3 ± 2.2% for the "other" structures when using ASLR.



Downloaded via 109.128.129.192 on March 9, 2021 at 11:36:45 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

Both academic research and applied research need to assess the protein structure in many different experimental conditions. Processes are always more complex and correct folding of proteins must be demonstrated in a wide variety of situations. Enzyme or therapeutic protein efficiency depend indeed on the correct protein structure. For instance, quality check of therapeutic antibodies or biosimilar must be thoroughly assessed in the course of development as well as on production lines. Among the methods available for such an assessment, Fourier transform infrared (FTIR) spectroscopy is most useful as it can be applied in different environments.[1] The secondary structure is usually obtained after a relation has been established between absorbance values in the amide I−amide II bands and secondary structure features. This relationship depends on the state of the sample (dried, solution in H_2O, or in D_2O) and the recording method [transmission, attenuated total reflection (ATR), diffuse reflectance]. It also critically depends on the quality and diversity of the protein library used for calibration.[2]

In the present paper, we introduce a new method designed for the high-throughput protein structure determination. It is based on spotting proteins as microarrays at a density of ca 2000−4000 samples per cm^2 and recording FTIR spectra by FTIR imaging. Even though microscopy has been applied in some particular cases to analyze small protein particles,[3] so far protein microarrays combined with FTIR imaging have not been used for protein secondary structure determination. In a previous paper, we presented a new protein library called cSP92, which contains 92 proteins.[2] It has been designed to cover as well as possible the structural space, both in terms of secondary structures and higher level structures as defined by CATH.[4] High-resolution structures available in the PDB[5] have been carefully analyzed for their quality and for the secondary structure content by the DSSP algorithm.[6] Smaller protein sets have been designed previously for establishing a relationship between FTIR spectra and secondary structures[7−9] and a set of 50 proteins was described in 2003[10] to create a predictive model using ATR−FTIR.[11−13] More recently, a larger protein set counting 84 proteins was used by Wilcox et al. for aqueous protein solution measurements.[14] In the latter research, prediction errors were found to be ca. 12% for α-helix, 7% for β-sheet, and 8% for the other structures.

Different methods and algorithms have been used to relate spectral data to the secondary structure content. Fourier self-deconvolution and curve fitting procedures have been originally preferred[15,16] because they can account for slight

Received: August 30, 2020
Accepted: January 19, 2021
Published: February 12, 2021

ACS Publications

© 2021 The Authors. Published by American Chemical Society

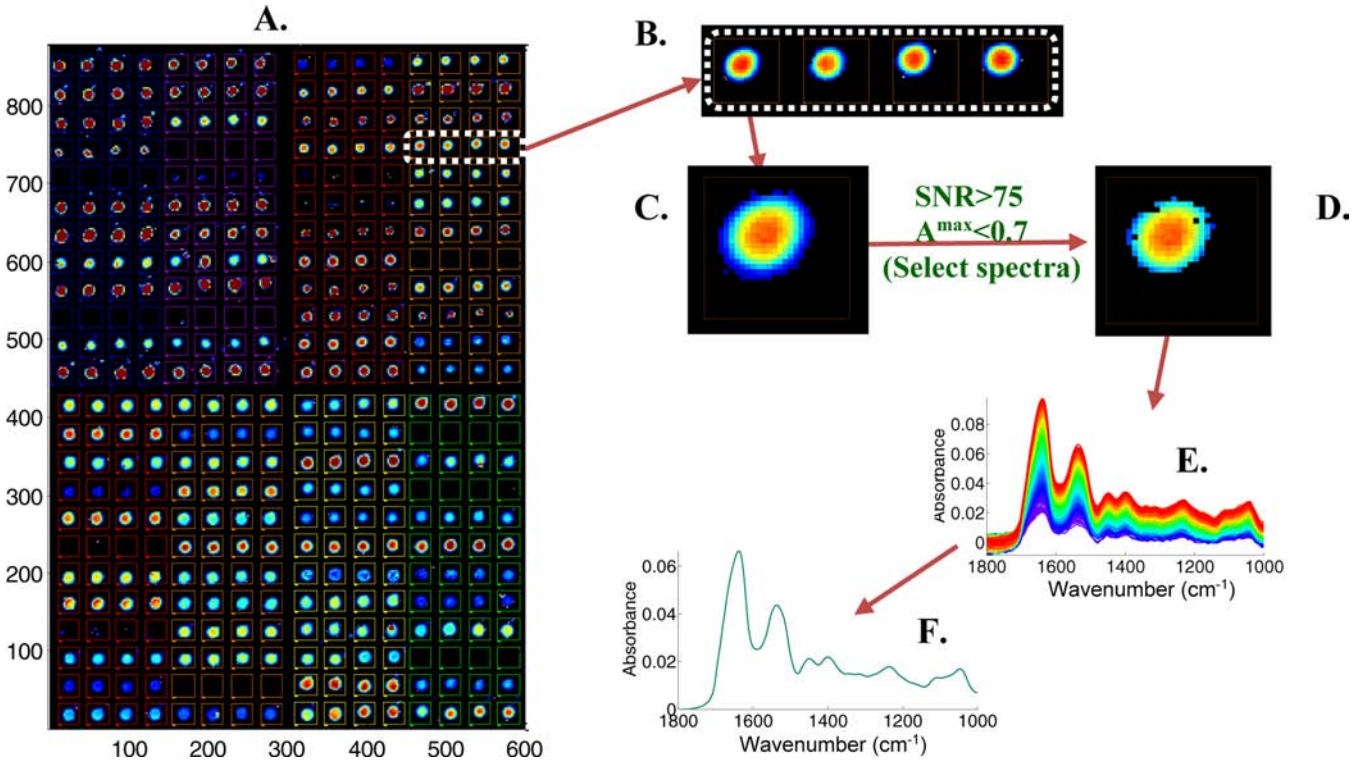


Figure 1. Schematic representation of the FTIR analysis of protein microarrays. (A) Image represents the absorbance at 1654 cm^{-1} of 384 samples, including blanks and controls. Eight grids identified by different colors have been placed on the image. Each grid has 12 rows corresponding to 12 different proteins and 4 columns corresponding to 4 replicates of each protein. The image contains 600 × 880 pixels, i.e. 528,000 full FTIR spectra. Each spot contains ca. 1 ng protein. The whole image covers 3.3 × 4.6 mm^2. Each square covers an area of 154 × 154 μm^2. A SNR = 15 filter has been applied and pixels corresponding to spectra with SNR < 15 have been colored in black. (B) Enlargement corresponding to quadruplicates of a protein sample and (C) subtraction of the background for each square. In this example, all 555 spectra of the noisy black area were averaged and this average was subtracted from all spectra present in the square. (D) More stringent filter SNR > 75 was then defined to collect 229 high-quality spectra for the protein as shown in (E). Spectra were averaged in (F). One average was finally computed for the quadruplicates.

variability in the spectral contribution of the α-helix and β-sheet structures. However, curve fitting by Gaussian/Lorentzian bands is often subject to overfitting and obtaining stable solutions depends on the number of constrains and essentially on the skill of the operator. Simple linear regressions, such as ascending stepwise linear regression (ASLR), are very efficient. The ASLR method introduces, in an ascending stepwise manner, one absorbance at a time in the model. It remains attractive because of its simplicity and the possibility to generate a linear equation, where only 2−5 absorbance values must be introduced to obtain the desired secondary structure content.[12,13] Multivariate methods have been widely used as well. Colinearity is best handled by partial least square (PLS), including interval partial least square (iPLS).[17] PLS is more efficient than principal component regression (PCR); in that way, it is expected to provide the best prediction with less components because the first components are already optimized for a particular structure prediction.[1] Nonlinear methods have been investigated as well, including neural networks[18,19] and genetic algorithm.[20] However, a small data set is usually an issue. Support vector machine (SVM) has been widely used in FTIR analysis of cancer tissues.[21,22] It can easily deal with the limited number of samples and large number of data. Finally, different definitions of protein secondary structures have been proposed in the course of years, and DSSP appears to be best related to FTIR spectral features.[23]

The paper aims at demonstrating that protein microarrays combined with FTIR imaging can be used for high throughput protein secondary structure prediction. The results demonstrate that this combination provides the prediction of the protein secondary structure content as accurate as a reference recording method used to record FTIR spectra such as ATR−FTIR applied on the same protein set. This observation fully validates the microarray approach. The choice of the method, ASLR, PLS, or SVM is not a very critical factor though, in general, ASLR outperforms the others.

EXPERIMENTAL SECTION

All 92 proteins of cSP92 data set are commercially available and were obtained as described in a previous paper.[2] The proteins sorted for increasing α-helix content with their PDB code are listed in the Supporting Information. Protein samples were solubilized at a final concentration of 10−20 mg/mL in 4 mM *N*-(2-hydroxyethyl)piperazine-*N*′-ethanesulfonic acid, 85 mM NaCl buffer solution. Details on their preparation and secondary structure content evaluation are reported in the Supporting Information.

Microarray Printing. In a first step, 10 μL of ethylene glycol were deposited in each well of a 384 well plate before addition of 10 μL protein sample. Prior to printing, the addition of 50% ethylene glycol (filtered through 0.2 μm Millipore filters) is indeed mandatory to obtain adequate viscosity for good printing conditions. The sample solution was then homogenized by pipetting and centrifuged at 5000*g* for 3

https://dx.doi.org/10.1021/acs.analchem.0c03677
Anal. Chem. 2021, 93, 3733−3741

min to eliminate bubbles. Samples were then loaded into the printing head from the 384-well plate using a 12 sample low volume loading device (ArrayJet's Jet Spyder). Microarrays were printed with an Arrayjet Marathon noncontact inkjet Microarrayer (ArrayJet, Roslin, UK) on 40 × 26 × 2 mm^3 BaF_2 slides (Neyco, France). BaF_2 slides were cleaned with Milli-Q H_2O and dried under a nitrogen flow. Drops of ca. 100 pL protein solution were deposited to form regular arrays. Temperature (18−20 °C), relative humidity (50−56%), and HEPA air filtration control were provided by a Jetmosphere cabinet containing the microarrayer to maintain a constant printing environment. Spot-to-spot distances in the X and Y directions were 200 or 220 μm. Spot diameter was about 80 μm, but this may vary according to the nature and concentration of the protein. Ethylene glycol was then evaporated under vacuum before recording FTIR spectra.

FTIR Imaging. Spectra were recorded as the average of 64 scans, between 3650 and 900 cm^{-1} at a nominal resolution of 8 cm^{-1}. FTIR data were collected using an Agilent mid-IR imager equipped with a liquid nitrogen cooled 128 × 128 mercury cadmium telluride focal plane array (FPA) detector and a 15× objective (NA = 0.62). Every element of the FPA acts as an independent and discrete detector from which a full spectrum is obtained. The corresponding pixel covers an area of 5.5 × 5.5 μm^2. Blackman−Harris 4-term apodization and zero filling of 2 were applied. Data were finally encoded every 2 cm^{-1} by linear interpolation. Data were collected in the transmission mode from sample regions of 700 × 700 μm^2 to form one FTIR image (unit image) containing 16,384 spectra. To cover larger sample areas, an automatic tiling combined several FTIR unit images in order to obtain one large mosaic FTIR image. The background image (128 scans per pixel) was acquired in the absence of the sample on a clean surface of the BaF_2 slide.

FTIR Image Processing. Automated spectrum extraction: a grid of squares, each fully including one spot, was placed on the image. The mean spectrum of all the spectra present in each square was then computed automatically. To improve the quality of the spectra, two processes were applied. First, even though a background image was recorded before imaging the sample and subtracted automatically, any variation in environmental conditions (water vapor content, CO_2, temperature, and alignment) results in the unwanted signal in the sample absorbance spectrum. A unique advantage of microarray infrared images is that the empty spaces found between protein spots can be used as a perfect background.[24] Practically, the empty spaces between spots can be most conveniently identified by their low signal-to-noise ratio (SNR). The signal is defined here as amide I maximum absorbance above a baseline drawn between 1720 and 1480 cm^{-1}, and noise is defined as the root mean square in the 2000−1900 cm^{-1} spectral range. The mean spectrum of the spectra identified as belonging to empty spaces represents the local noise, as described in Figure 1. Second, great attention was paid to use only high quality spectra but to avoid signal distortion because of potential signal saturation. Another filter was therefore applied to eliminate from the analysis pixels where the absorbance was larger than 0.7. After the selection of the spectra with SNR > 75 and maximum absorbance <0.7 and gathering the spectra of the quadruplicates, an overall mean spectrum was computed for each protein. Spectra were baseline corrected by the subtraction of straight lines interpolated between the spectral points at 1720 and 1480 cm^{-1}. Scaling was obtained by dividing the spectra by 10^{-4}× their area between 1720 and 1480 cm^{-1}.

ATR−FTIR. Polished triangular-shaped germanium prisms (4.8 mm × 4.8 mm × 45 mm) were purchased from Neyco (France) and accommodated on the beam condenser from a Golden Gate Micro-ATR from Specac (UK). A top plate with a groove fitting the crystal was used in the replacement of the diamond-bearing plate (WOW Company, Belgium). With one horizontal surface facing upward, in this configuration, a single reflection occurs as for the diamond-fitted plate. Samples (0.5 μL), prepared at ca. 10 mg/mL, as described in the Supporting Information, were deposited on the germanium surface. No ethylene glycol was present for ATR−FTIR. After solvent evaporation under a gentle N_2 stream over an area of ca 4 mm^2, ATR−FTIR spectra (average of 128 scans) were recorded between 4000 and 900 cm^{-1} with a Bruker Equinox 55 spectrophotometer (Bruker, Ettlingen, Germany) equipped with a MTC detector at a resolution of 2 cm^{-1} acquired in the double-sided, forward−backward mode.

Preprocessing of the spectra: subtraction of a reference water vapor spectrum was obtained after scaling on the area between 1738 and 1732 cm^{-1} band. All the spectra were then apodized in the Fourier domain to obtain a resolution of 4 cm^{-1} in the spectral domain. Baseline and scaling were obtained as described above for microarray spectra.

Multivariate Analyses. Multivariate analyses include ASLR,[12,13] PLS regression[25−28] including iPLS regression, which is an extension to PLS developed by Nørgaard et al.,[26,29] and SVM dedicated to solving regression problems.[30] Besides the capability of working with nonlinear problems, the approach can deal with a limited number of samples for high dimension data.[31] The formulation introduced by Suykens et al.[32,33] was used here along with the Matlab toolbox[34] built by the authors. Details on these tools and computation of errors of prediction are reported in the Supporting Information. The determination of the complexity of the models was obtained by the repeated double cross-validation (rdCV) developed in ref 35 as described in details in the Supporting Information under "multivariate analyses". The rdCV method was used to determine the number of wavenumbers for ASLR, number of LVs for PLS and γ and σ^2 values for SVM as well as result reliability using 500 independent test sets as described in the Supporting Information and illustrated by Figures S1−S7.

Image analysis, spectrum processing, and multivariate analyses were all performed with Kinetics, a home-made software running under MatLab (The MathWorks Inc.).

RESULTS

The set of reference proteins used here has been described in a previous paper.[2] FTIR imaging of protein microarrays has been demonstrated in a previous work[24,36] to yield high-quality protein spectra on as little as a few picograms of proteins. However, no attempt was made to predict protein secondary structure. As the noise of the FPA detector is proportional to its area, the very small detectors, used for building FPA, bring little noise, and excellent quality spectra can be acquired. In addition, a local background can be collected from the pixels immediately surrounding the sample spot, which ideally accounts for water vapor contribution as spectra are recorded simultaneously and a few μm away from the sample. The overall process is described in Figure 1.

As for each protein of cSP92, a high-resolution structure is available, the secondary structure contents obtained by

https://dx.doi.org/10.1021/acs.analchem.0c03677
Anal. Chem. 2021, 93, 3733−3741

applying the DSSP algorithm were reported previously.[2] In the following text, we restrict the meaning of α-helix to H and of β-sheet to E as defined by DSSP.[2,6] A category called "others" was created to contain all other structures such as random (essentially) but also the other structures defined by DSSP. The reason is that none of the minor structures could be predicted with sufficient accuracy. The E structure was further split in $E^{\parallel}$ and $E^{anti\parallel}$ standing for parallel and antiparallel β-sheet, respectively. As previously proposed by Kalnin et al.,[37] the H structure was split into "ordered" and "disordered" helices. Accordingly, disordered helix content was obtained by considering the two amino acid residues at the end of each α-helix as a "disordered" helix, while the core of the helix was assigned to "ordered" helix. Figure 2 presents the 92 spectra of

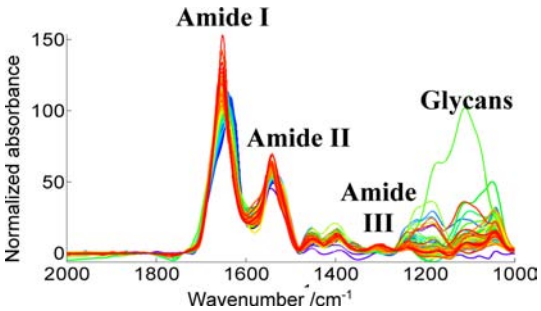


Figure 2. FTIR microarray spectra of the 92 proteins included in the protein library cSP92. Proteins have been sorted from low (blue) to high (red) α-helix content.

cSP92 sorted from low (blue) to high (red) helix content. The large variance observed between 1250 and 1000 cm^{-1} is related to the presence of glycosylation on some proteins. These bands are interesting in their own right as they allow the quantification of the glycosylation content,[38] yet, most information on the secondary structure is contained in the amide I–amide II region of the spectrum located between 1700 and 1500 cm^{-1}. As reviewed elsewhere,[39] amide I is essentially because of ν(C=O) but also includes a significant contribution of ν(C–N) vibration. Amide II is a mix between δ(N–H), ν(C–N), and ν(C–C). Amide III is even more complex and is rarely used for structure determination, in part because it absorbs in a spectral region where interferences of buffer molecule contributions are usually quite significant. Amide A near 3300 cm^{-1} is derived from amide ν(N–H) and amide B from a Fermi resonance between the first overtone of amide II and ν(N–H).

The evaluation of the correlation at each wavenumber between the secondary structure content and absorbance of the 92 spectra indicates a high correlation coefficient in the amide I region (r = 0.87 at 1658 cm^{-1} for α-helix, r = 0.89 at 1636 cm^{-1} for β-sheet) and is almost as good in the amide II range (see Figure S8). Amide III at 1300 cm^{-1} displays correlation coefficients of 0.78 for α-helix and −0.66 for β-sheet. Amide A, at 3316 cm^{-1} for α-helix and 3233 cm^{-1} for β-sheet displays a rather poor correlation (r < 0.4), while amide B at 3085 cm^{-1} has a r = −0.67 for α-helix and 0.65 for β-sheet (Figure S8).

PLS REGRESSION

PLS is a regression method where the latent variables (LVs) are built in the direction of maximum covariance between the matrix of the infrared spectra composed by the predictor variables and the matrix of related fractions of structural elements composed by the predicted or dependent variables.

Number of LVs. The number of LVs is usually selected after examining the RMSECV as the number of LVs increases. In cross-validation, the RMSECV starts decreasing when adding more LVs to the model, then increases again because of overfitting (Figure 3 and Figure S1 under "Multivariate

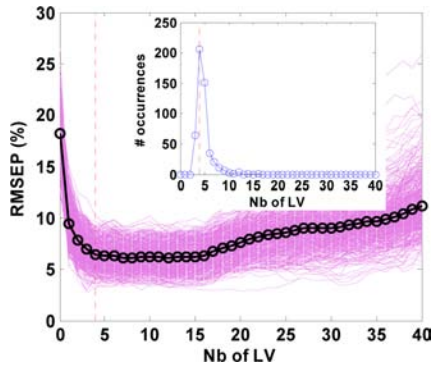


Figure 3. Illustration of the evolution of the RMSEP on the α-helix content prediction as more LVs are added in a PLS model. Here, 500 independent test sets have been evaluated using the rdCV procedure. The inset presents the number of times a given number of LVs has been selected among the 500 tests. The final optimal selected value is indicated by the red dotted line.

analyses" in the Supporting Information). However, it remains difficult to evaluate when the improvement stops to be significant. To address this issue, rdCV was used to evaluate the optimal number of LVs. For 500 independent test sets, the optimal number of LVs was evaluated, keeping into account a parsimony factor as explained in the Supporting Information. The inset in Figure 3 describes the number of times each number of LVs has been selected as optimal over the 500 tests. The most frequent number is retained as the final optimal number, that is, 4 in the case illustrated in Figure 3. The results are presented in Figure S9 for microarray spectra and in Figure S10 for ATR spectra. Typically, 2–5 LVs were found to be optimal.

Spectral Range. Even though PLS is not very sensitive to the spectral region whose variations are not correlated with the dependent variables, such a region may nevertheless degrade the results. Selecting an adequate spectral region is therefore an issue.[26] In a preliminary analysis, we investigated the 1720–1200 cm^{-1} spectral region including amide III, amide I and II (1720–1480 cm^{-1}), amide I only (1720–1600 cm^{-1}), and sub-intervals restricted to amide I and II. Calibration was made in cross-validation. No result outperformed those obtained when amide I and II were used together (data not shown). We therefore decided to restrict the analysis to this spectral region (1720–1480 cm^{-1}). This is in line with an earlier work by Dousseau and Pezolet[40] who already demonstrated that the best region to be considered was 1720–1480 cm^{-1}.

Evaluation of the errors of prediction established by rdCV are reported in Table 1. Once the number of LVs and the spectral range are established, a leave-one-out (LOO) validation sheds light on the accuracy of the approach using the final optimal number of LVs. Figure 4, upper row, reports the predicted versus actual α-helix, β-sheet, and others content obtained on the 1720–1480 cm^{-1} spectral range. It demonstrates that a good prediction of the α-helix and β-

https://dx.doi.org/10.1021/acs.analchem.0c03677
Anal. Chem. 2021, 93, 3733–3741

Table 1. Summary of the Performances of the Different Methods Used to the Predict Protein Secondary Structure from the 1720−1480 cm^{-1} Spectral Range of Protein Microarray and ATR Spectra

	rdCV-RMSECV[a]		LOO-RMSECV[b]		KS-RMSEP[c]		ζ_{rdCV}[d]		ζ_{LOO}[e]		ζ_{KS}[f]		
structure	μarray	ATR	μarray	ATR	μarray	ATR	μarray	ATR	μarray	ATR	μarray	ATR	method
α-helix	4.9 ± 1.1 [5]	4.7 ± 1.0[6]	5.7 [5]	6.0 [6]	4.8 [5]	5.9 [6]	3.8	3.9	3.2	3.0	4.5	3.6	ASLR
	6.4 ± 1.3 [4]	6.8 ± 1.2 [5]	6.3[4]	6.8 [5]	7.2 [4]	7.5 [5]	2.9	2.7	2.9	2.7	3.0	2.9	PLS
	6.6 ± 1.4	6.9 ± 1.8	6.2	7.2	7.2	7.8	2.8	2.6	2.9	2.5	3.0	2.8	SVM
ordered α-helix	5.0 ± 1.0 [5]	5.2 ± 1.1 [5]	5.9 [5]	6.3 [5]	6.0 [5]	5.3 [5]	3.6	3.5	3.1	2.9	3.4	3.9	ASLR
	6.6 ± 1.1 [4]	7.1 ± 1.2 [5]	6.4 [4]	7.1 [5]	8.0 [4]	9.5 [5]	2.7	2.6	2.8	2.5	2.6	2.2	PLS
	6.7 ± 1.2	7.2 ± 1.8	6.3	7.4	8.1	8.7	2.7	2.5	2.9	2.5	2.6	2.4	SVM
β-sheet	4.6 ± 0.8 [4]	4.5 ± 0.8 [4]	5.4 [4]	5.6 [4]	5.3 [4]	4.2 [4]	3.0	3.0	2.5	2.5	2.9	3.6	ASLR
	5.6 ± 0.9 [3]	5.8 ± 0.9 [3]	5.6 [3]	5.8 [3]	6.2 [3]	6.3 [3]	2.4	2.3	2.4	2.4	2.4	2.4	PLS
	5.5 ± 0.8	5.7 ± 1.0	5.3	5.5	6.1	6.7	2.5	2.4	2.6	2.5	2.5	2.3	SVM
antiparallel β-sheet	6.0 ± 1.0 [3]	5.0 ± 0.9 [5]	7.0 [3]	6.1 [5]	5.5 [3]	5.2 [5]	2.3	2.7	2.0	2.2	2.6	2.8	ASLR
	7.2 ± 1.0 [2]	6.8 ± 1.0 [2]	7.0 [2]	6.9 [2]	7.3 [2]	7.8 [2]	1.9	2.0	2.0	2.0	2.0	1.9	PLS
	6.4 ± 1.1	6.4 ± 1.2	6.2	6.2	6.9	7.1	2.1	2.1	2.2	2.2	2.1	2.0	SVM
others	6.3 ± 2.0 [4]	6.1 ± 1.9 [4]	7.4 [4]	7.4 [4]	8.3 [4]	7.8 [4]	1.6	1.7	1.4	1.4	1.8	1.9	ASLR
	7.7 ± 1.9 [4]	7.7 ± 1.9 [4]	7.9 [5]	7.7 [4]	10.6 [4]	10.2 [4]	1.3	1.3	1.3	1.3	1.4	1.5	PLS
	7.2 ± 1.6	8.0 ± 2.3	6.8	7	9.5	9.1	1.4	1.3	1.5	1.5	1.6	1.7	SVM

[a]rdCV-RMSECV reports the error of prediction (in %) in rdCV. [b]LOO-RMSECV reports the error of cross-validation (in %) obtained in a LOO procedure. [c]KS-RMSEP reports the error of prediction obtained on the Kennard−Stone selection containing 30 spectra. [d]ζ_{rdCV}. [e]ζ_{LOO}. [f]ζ_{KS} reports the ratio between the standard deviation of the structure in the particular test set of the methods and the error of prediction for, respectively, the rdCV, LOO, and Kennard−Stone procedures. Data are provided for spectra recorded on protein microarrays (μarray) and by ATR−FTIR (ATR). The number of wavenumbers or of LVs is indicated in brackets for ASLR and PLS, respectively.

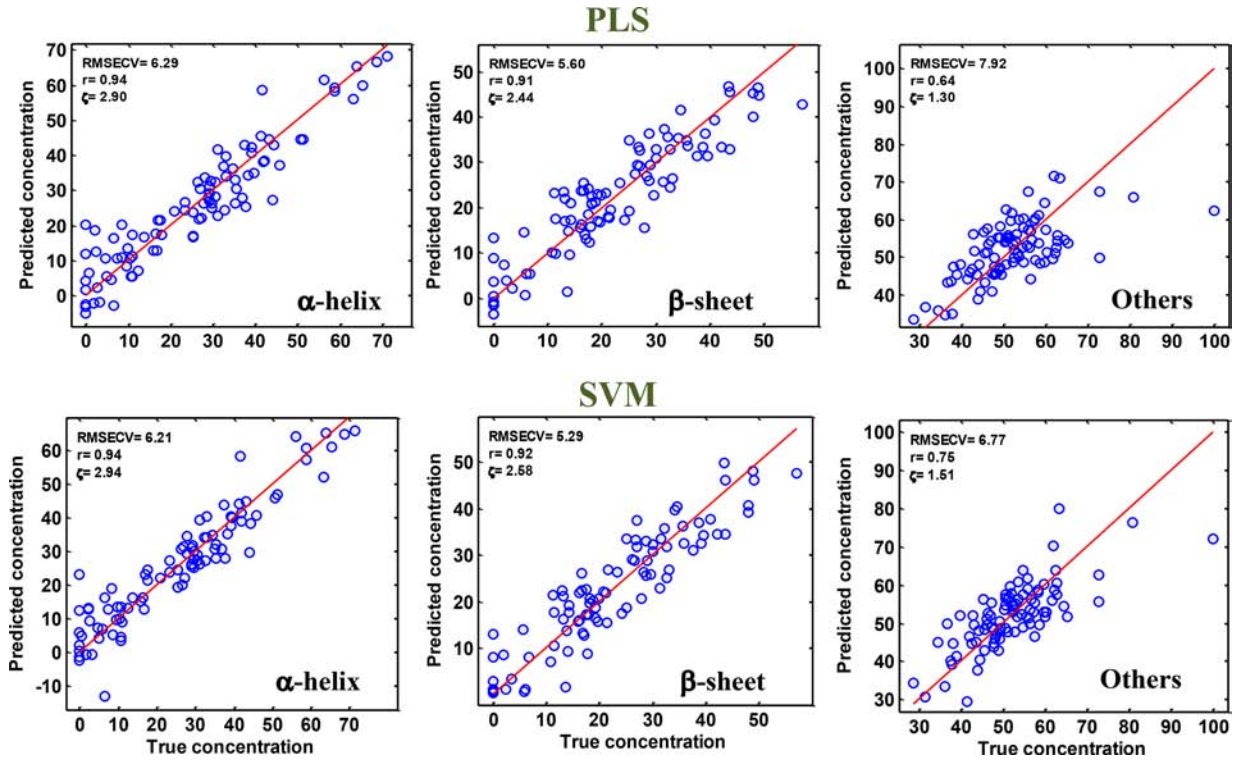


Figure 4. Analysis of protein microarray FTIR spectra: relation between the predicted α-helix (left), β-sheet (middle), and others (right) content and the actual contents obtained by DSSP. The results were obtained in LOO cross-validation. The RMSECV, ζ, and the correlation coefficient r are given in the inset. The upper row refers to results obtained by PLS and the lower row to results obtained by SVM.

sheet structure contents was obtained with a RMSECV of 6.29 and 5.60% for α-helix and β-sheet content, respectively. Predicted versus actual content relation is presented for all structures in Figure S11 for the microarray and ATR spectra. A further validation was obtained upon selecting a 30 spectrum validation set by the Kennard−Stone algorithm,[41] building a model with the remaining 62 spectra. As described in the Supporting Information, the Kennard−Stone algorithm selects samples with a uniform distribution over the predictor space, using the Euclidian distance. It avoids oversampling some structure content ranges that are overrepresented in cSP92. The results of the Kennard−Stone validation are reported in Figure S12. An evaluation of the accuracy of all models is reported in Table 1.

https://dx.doi.org/10.1021/acs.analchem.0c03677
Anal. Chem. 2021, 93, 3733−3741

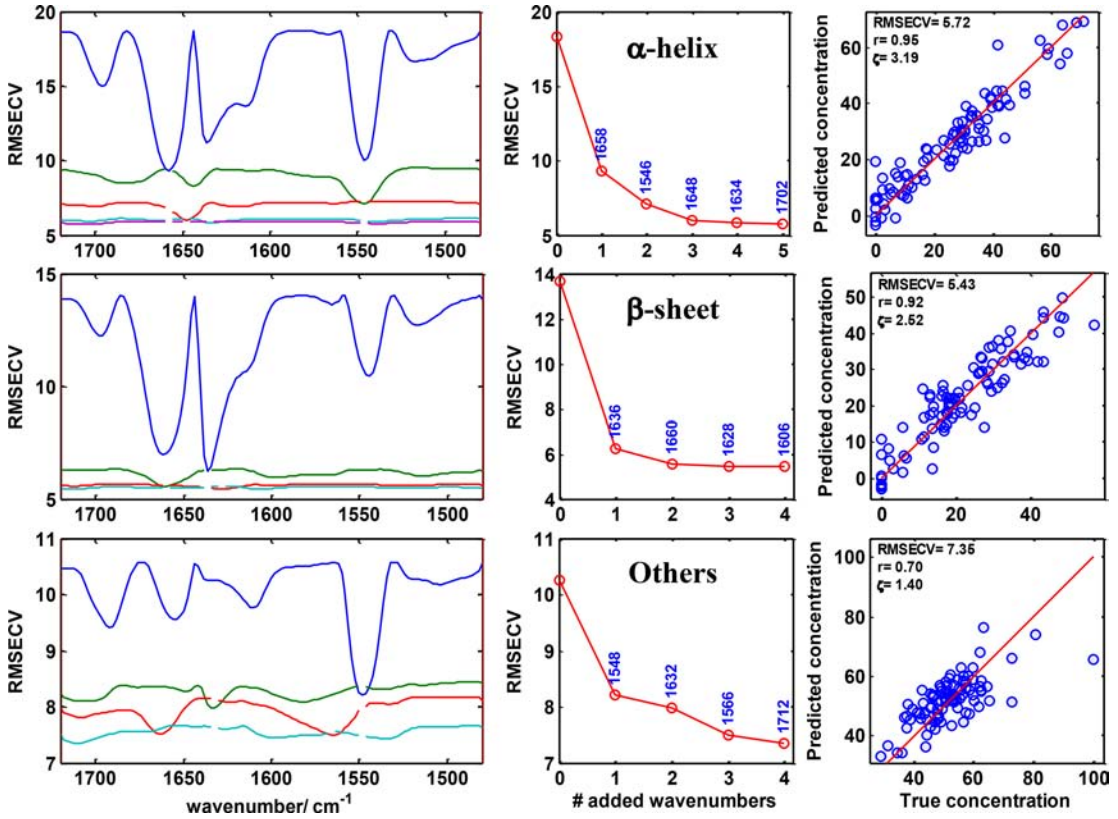


Figure 5. LOO ASLR for protein microarray FTIR spectra using the number of wavenumbers determined by rdCV. Top frame: α-helix, middle frame: β-sheet, bottom frame: "others". The left column presents the profile of the RMSECV (expressed in % of the structure content) when a single wavenumber is used (blue line), when a second is added (green line), a third one (red line), a fourth one (cyan), and a fifth one (magenta) (see note on Figure S20 explaining line crossing). For each curve, the value of the minimum with the corresponding wavenumber is reported in the middle column, the "0" wavenumber is the secondary structure standard deviation. The right column reports the predicted secondary structure content as a function of the actual content for each spectrum for the optimal number of wavenumbers, see middle column and Figure S18. The red line is the diagonal expected for a perfect prediction.

More insights in the spectral sub-regions, which bring valid information about secondary structure content, can be obtained by systematically analyzing small spectral intervals. This is conveniently obtained by iPLS as developed by Norgaard et al.[26] iPLS divides the spectral range into a series of smaller intervals. As suggested by Navea et al., sub-interval in the amide I and II spectral regions may significantly outperform the global model.[17,42] However, results reported in Figure S13 indicate that the global model (1720−1480 cm^{-1}) outperforms any single region taken separately except for a very minor gain for the 1662−1642 cm^{-1} interval for the β-sheet in microarray spectra, at the price of using 4 LVs instead of 3 for the global model. It can be safely concluded that using small intervals does not improve significantly the predictions in this spectral range.

LEAST-SQUARES SVMS

While PLS assumes a linear relationship between concentration and the intensity of the absorption bands, SVM is able to deal with nonlinear problems that may occur in spectroscopy. It is used here to evaluate whether a nonlinear approach[43] would be able to improve secondary structure prediction from the FTIR spectra. The model parameters γ and σ^2 were evaluated using the rdCV approach in 500 independent predictions. The 500 best values of γ and σ^2 selected in the rdCV validation were plotted as a histogram and the median value was selected as the final optimal value. The histograms are reported in Figure S14 for microarray spectra and S15 for ATR spectra. The mean RMSE and the error on this value as judged from the independent repetitions appear on Figures S14 and S15 and are summarized in Table 1.

As before, a LOO cross-validation was performed. Figure 4, lower row, reports the actual versus predicted concentration for the 92 proteins. SVM provides results that are in general only slightly better than the ones obtained using PLS (Figure 4). The full results are presented in Figure S16 for microarray and ATR spectra. The results of a Kennard−Stone validation appear in Figure S17. They are summarized in Table 1.

ASCENDING STEPWISE LINEAR REGRESSION

The ascending stepwise regression has been described in detail before.[12,13] It identifies the relevance of every wavenumber in the infrared spectrum to create a simple linear model to predict a secondary structure content (also see Supporting Information). In the cross-validation procedure, a validation set is removed from the full set, a linear regression model is built and the spectra that did not participate in the building of the model are evaluated. Once this has been repeated at every wavenumber, the wavenumber providing the smallest RMSECV is retained (see Figure 5). Once the best wavenumber is identified, it is kept in the model and the second one is added following the same procedure. The full procedure is repeated for each additional wavenumber added. The optimal number of wavenumbers has been determined by

https://dx.doi.org/10.1021/acs.analchem.0c03677
Anal. Chem. 2021, 93, 3733−3741

rdCV. The results of rdCV are presented in Figure S18 for spectra recorded on microarrays and in Figure S19 for spectra recorded by ATR. The full LOO analysis is reported in Figures S20 and S21 for microarray and ATR spectra, respectively. The results of a LOO ASLR using the optimal number of wavenumbers determined by rdCV are reported in Figure 5 for α-helix, β-sheet, and "others" structures. Interestingly, the ASLR approach reveals the spectral profile of the information content for each structure (Figure 5, left column). For the α-helix structure, three major areas are found to contain relevant information around 1658, 1546, and 1648 cm^{-1} (Figure 5). It is interesting to note that the shape of the blue curve for α-helix parallels the one found for the β-sheet structure. The minima found in the blue curve at 1696 and 1634 cm^{-1} are obviously related to the antiparallel β-sheet structure as described in ref 44 but contain almost as much information for predicting the α-helix content as the typical helix band at 1658 cm^{-1}. This is due to the fact that the α-helix and β-sheet contents are strongly correlated in the protein set (negative correlation coefficient 0.83). The central column of Figure 5 reports the RMSECV obtained for the different numbers of wavenumbers in the left column. In the absence of any prediction model, the standard deviation for each secondary structure content is reported at "0 added wavenumber", as shown in Figure 5. The drop observed between the 0 and 1 added wavenumbers is the benefit brought by a single absorbance value. It is apparent that 3 wavenumbers are in general sufficient to extract the most available information relevant to the secondary structure content. In the case of the α-helix, once the first wavenumber (1658 cm^{-1}) is added to the model, the addition of a second wavenumber brings improved prediction with the RMSECV having a minimum at 1546 cm^{-1}, demonstrating that amide II brings independent additional information useful for α-helix content prediction (Figure 5, left column, green curve). The addition of a third wavenumber (1648 cm^{-1}) provides further information. A small improvement was found to be significant for the fourth and fifth wavenumbers by the rdCV procedure. The global relation between the reference concentration and predicted concentration is presented in the right column of Figure 5. The "others" structure is rather poorly quantified, most probably because it represents a complex variety of different structures.

The results for "ordered" helices and antiparallel β-sheet, presented in Figure S20 for protein microarray spectra and S21 for ATR spectra, show that the prediction is not improved with respect to the full α-helix or β-sheet content. While in previous research studies, the ordered alpha helix was much better predicted than the total alpha helix content,[37] this observation is not confirmed when working with a larger protein set.

Beyond the amide I–amide II range, amide III brings very significant contributions near 1300 cm^{-1} (not shown). However, it is redundant with the information present in the 1720–1480 cm^{-1} range. Amide B at 3081 cm^{-1} is also well related to α-helix or β-sheet content but again does not contain information not already present in the amide I–amide II range. As the wavenumbers outside the amide I–amide II range did not bring outstanding contributions, while they are often problematic because of interferences with the buffer contributions, we restricted the discussion to the 1720–1480 cm^{-1} spectral region.

The results of the Kennard–Stone test set are reported in Figure S22. The equations, derived from the Kennard–Stone training set, allowing the prediction of the secondary structure contents are presented in Table 2. Once the appropriate coefficients are determined, it is sufficient to plug a few absorbance values in a linear equation to obtain the secondary structure. Similar equations obtained for the 5 structures can be found in Figures S20 and S21 for LOO models and in Figure S22 for Kennard–Stone validation. While models based on the Kennard–Stone training set are adequately validated (Figure S22), the full models (Figures S20 and S21) may be more robust but do not have independent validation.

Table 2. Equations Allowing the Computation of the α-Helix, β-Sheet, and "Others" Structure Contents Obtained from the Kennard–Stone Selected Training Set[a]

$$\alpha\text{-helix}\ (\%) = -326.6 + 3.37 \times A^{1546} + 2.69 \times A^{1662} + 0.64 \times A^{1624} + 1.89 \times A^{1698} - 1.55 \times A^{1660}$$

$$\beta\text{-sheet}\ (\%) = -37.06 + 0.64 \times A^{1634} - 0.56 \times A^{1666} + 0.37 \times A^{1644} + 0.75 \times A^{1604}$$

$$\text{others}\ (\%) = 230.6 - 3.08 \times A^{1548} - 1.04 \times A^{1636} + 1.15 \times A^{1566} + 0.40 \times A^{1642}$$

[a]These equations can be used as such in the absence of the database, provided that the spectra are baseline corrected and scaled as described in this work.

A summary of the data described so far is presented in Table 1. To evaluate adequately the added value brought by the spectroscopic data, it is necessary to take into account the width of the structure distribution in the protein set. This was provided in Figure 5, at 0 wavenumber added, as the standard deviation of the structure content, that is, the error on the prediction that would be done by always guessing the prediction is the mean value of the content within the protein set. Each RMSE was therefore compared with this standard deviation. This comparison was best obtained by dividing the standard deviation of the distribution of the structures by the RMSE for the predicted values. These ratios are defined as ζ_{rdCV}, ζ_{LOO}, and ζ_{KS}, respectively, for the rdCV test, the LOO cross-validation, and Kennard–Stone test set and are reported in Table 1. It is surprising to observe that, in general, ASLR provides better results than PLS and SVM.

The protein structures analyzed in Table 1 are the only ones for which a reasonable prediction could be obtained. Structures such as "unordered helix" or parallel β-sheet or all the minor structures described by DSSP have ζ values close to 1. This is largely because of a lack of variance within the protein set.

DISCUSSION

Using an inkjet-type printer to create protein microarrays in the presence of 50% ethylene glycol raises the question of the validity of the method. Ethylene glycol is a molecule akin to glycerol, which is well known to be harmless to proteins and can even stabilize proteins. It has the advantage on being volatile. It can therefore be easily eliminated within a few hours under mild vacuum. Friction that may occur in the tubing system of the printer remains another potential concern. For this reason, we repeated all the experiments on the same proteins but recorded the ATR–FTIR spectra. ATR–FTIR is a standard method whose efficiency to predict protein secondary structure has been demonstrated before.[1,13,20,45,46] We therefore recorded the ATR–FTIR spectra of the 92 proteins in the absence of ethylene glycol. The results obtained by ASLR, PLS, and SVM are very similar to the ones presented for the spectra obtained on microarrays (Table 1). This result validates the microarray approach. An alternative approach

https://dx.doi.org/10.1021/acs.analchem.0c03677
Anal. Chem. 2021, 93, 3733–3741

would be to work with aqueous solutions. Working with aqueous solutions has important constraints because of the high absorbance of water in the amide I region of the spectrum,[14,47] with an optimal cell path length of 3−4 μm in H_2O and 40−60 μm in D_2O.[48] Subtracting water contribution whose intensity is an order of magnitude larger than the amide I signal remains an issue. Dealing with slow H/D exchange in D_2O, which strongly affects amide I and amide II is another unsolved problem. In general, working with the solution or dried proteins has been shown to provide similar results unless proteins are dried under vacuum.[49,50] In particular, ATR−FTIR[46,51] has been used extensively, and secondary structures were found to be predicted as well as in transmission FTIR, as directly compared in ref 13 on 45 proteins. The reasons for the validity of dried protein samples have been discussed at length elsewhere, for example, ref 51 and the amount of water left in "dry" proteins has been evaluated to be significant.[52]

Beside regular processing such as baseline subtraction or scaling, it has been suggested that working with second derivatives would "enhance" the "resolution" of the components underlying amide I and amide II bands as reviewed in ref 53 The quotation marks remind that the information content is certainly not enhanced. We repeated all the analyses described in the paper using spectrum second derivatives. No improvement of the secondary structure prediction could be obtained.

Table 1 also reveals that the Kennard−Stone validation results are much poorer than those of rdCV and LOO validations for the "others" structure. This appears to be related to the presence of a few unique proteins. Note 2 in Figure S20 discusses this observation and the reasons to keep these proteins in the study. The Kennard−Stone validation has the merit of using a homogenous spread of the structure contents in the test sets, while some values are over represented in the LOO and rdCV procedures, reflecting the distribution in the protein set.

CONCLUSIONS

In conclusion, a new large protein library has been used to test protein microarray for the protein secondary structure quantification. This approach has been found to yield results of the same quality as the reference ATR−FTIR method, suggesting that this approach can be used for high throughput determination of protein secondary structures. Using this large data set, we also demonstrated that structures such as parallel β-sheet, disordered α-helices, 3_{10}-helix (G), π-helix (I), helix-turn (T), β-bridge (B), and other/loop (L) could not be quantified. The reason is likely to be their low abundance and especially low variance in the data set. Surprisingly, the simple ASLR method usually yields better results than PLS and SVM. This is of interest as use of tunable quantum cascade laser is growing rapidly in FTIR imaging. In that respect, ASLR has the advantage of requiring the measurement of absorbance values at a small number of wavenumbers.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.0c03677.

List of the proteins and a note on their preparation, description of the multivariate analysis techniques used in this work, correlation between FTIR spectra, PLS parameter optimization, LOO validation, Kennard-Stone validation, iPLS profile, SVM parameter optimization, SVM parameter optimization, LLO validation, and ASLR parameter optimization (PDF)

AUTHOR INFORMATION

Corresponding Author

Erik Goormaghtigh − *Center for Structural Biology and Bioinformatics, Laboratory for the Structure and Function of Biological Membranes, Campus Plaine, Université Libre de Bruxelles, B1050 Brussels, Belgium;* orcid.org/0000-0002-2071-2262; Phone: +32 2 650 53 86; Email: egoor@ulb.ac.be

Author

Joëlle De Meutter − *Center for Structural Biology and Bioinformatics, Laboratory for the Structure and Function of Biological Membranes, Campus Plaine, Université Libre de Bruxelles, B1050 Brussels, Belgium*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.0c03677

Author Contributions

The manuscript was written through contributions of all the authors. All the authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the Fonds National de la Recherche Scientifique—FNRS under grant no. 001518F (EOS-convention # 30467715). We thank the Walloon Region (SPW, DGO6, Belgium) for supporting the ROBOTEIN project within the frame of the EQUIP2013 program. E.G. is Research Director with the National Fund for Scientific Research (Belgium).

REFERENCES

(1) Wang, Y.; Boysen, R. I.; Wood, B. R.; Kansiz, M.; McNaughton, D.; Hearn, M. T. W. *Biopolymers* **2008**, *89*, 895−905.
(2) De Meutter, J.; Goormaghtigh, E. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 1864−1876.
(3) Schack, M. M.; Møller, E. H.; Friderichsen, A. V.; Carpenter, J. F.; Rades, T.; Groenning, M. *J. Pharm. Sci.* **2019**, *108*, 1117−1129.
(4) Orengo, C.; Michie, A.; Jones, S.; Jones, D.; Swindells, M.; Thornton, J. *Structure* **1997**, *5*, 1093−1109.
(5) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. *J. Mol. Biol.* **1977**, *112*, 535−542.
(6) Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577−2637.
(7) Prestrelski, S. J.; Byler, D. M.; Liebman, M. N. *Proteins Struct. Funct. Genet.* **1992**, *14*, 440−450.
(8) Pribic, R.; Vanstokkum, I. H. M.; Chapman, D.; Haris, P. I.; Bloemendal, M. *Anal. Biochem.* **1993**, *214*, 366−378.
(9) Hering, J. A.; Innocent, P. R.; Haris, P. I. *Proteomics* **2004**, *4*, 2310−2319.
(10) Oberg, K. A.; Ruysschaert, J.-M.; Goormaghtigh, E. *Protein Sci.* **2003**, *12*, 2015−2031.
(11) Oberg, K. A.; Ruysschaert, J.-M.; Goormaghtigh, E. *Eur. J. Biochem.* **2004**, *271*, 2937−2948.
(12) Goormaghtigh, E.; Ruysschaert, J.-M.; Raussens, V. *Biophys. J.* **2006**, *90*, 2946−2957.

(13) Goormaghtigh, E.; Gasper, R.; Bénard, A.; Goldsztein, A.; Raussens, V. *Biochim. Biophys. Acta Protein Proteonomics* **2009**, *1794*, 1332−1343.
(14) Wilcox, K. E.; Blanch, E. W.; Doig, A. J. *Biochemistry* **2016**, *55*, 3794−3802.
(15) Byler, D. M.; Susi, H. *Biopolymers* **1986**, *25*, 469−487.
(16) Goormaghtigh, E.; Cabiaux, V.; Ruysschaert, J.-M. *Fed. Eur. Biochem. Soc. J.* **1990**, *193*, 409−420.
(17) Navea, S.; Tauler, R.; Juan, A. d. *Anal. Biochem.* **2005**, *336*, 231−242.
(18) Severcan, M.; Haris, P. I.; Severcan, F. *Anal. Biochem.* **2004**, *332*, 238−244.
(19) Hering, J. A.; Innocent, P. R.; Haris, P. I. *Appl. Bioinf.* **2004**, *3*, 9−20.
(20) Smith, B. M.; Oswald, L.; Franzen, S. *Anal. Chem.* **2002**, *74*, 3386−3391.
(21) Baker, M. J.; Trevisan, J.; Bassan, P.; Bhargava, R.; Butler, H. J.; Dorling, K. M.; Fielden, P. R.; Fogarty, S. W.; Fullwood, N. J.; Heys, K. A.; Hughes, C.; Lasch, P.; Martin-Hirsch, P. L.; Obinaju, B.; Sockalingum, G. D.; Sulé-Suso, J.; Strong, R. J.; Walsh, M. J.; Wood, B. R.; Gardner, P.; Martin, F. L. *Nat. Protoc.* **2014**, *9*, 1771−1791.
(22) Bergner, N.; Romeike, B. F. M.; Reichart, R.; Kalff, R.; Krafft, C.; Popp, J. *Analyst* **2013**, *138*, 3983−3990.
(23) De Meutter, J.; Goormaghtigh, E. *Anal. Chem.* **2021**, *93*, 1561−1568.
(24) De Meutter, J.; Vandenameele, J.; Matagne, A.; Goormaghtigh, E. *Analyst* **2017**, *142*, 1371−1380.
(25) Dreissig, I.; Machill, S.; Salzer, R.; Krafft, C. *Spectrochim. Acta, Part A* **2009**, *71*, 2069−2075.
(26) Nørgaard, L.; Saudland, A.; Wagner, J.; Nielsen, J. P.; Munck, L.; Engelsen, S. B. *Appl. Spectrosc.* **2000**, *54*, 413−419.
(27) Geladi, P.; Kowalski, B. R. *Anal. Chim. Acta* **1986**, *185*, 1−17.
(28) Wold, S.; Sjöström, M.; Eriksson, L. *Chemom. Intell. Lab.* **2001**, *58*, 109−130.
(29) Leardi, R.; Nørgaard, L. *J. Chemom.* **2005**, *18*, 486−497.
(30) Ghorbani, M.; Zargar, G.; Jazayeri-Rad, H. *Petroleum* **2016**, *2*, 301−306.
(31) Tange, R.; Rasmussen, M. A.; Taira, E.; Bro, R. *J. Near Infrared Spectrosc.* **2015**, *23*, 75−84.
(32) Suykens, J. A. K.; Vandewalle, J. *Neural Process. Lett.* **1999**, *9*, 293−300.
(33) Suykens, J. A. K.; De Brabanter, J.; Lukas, L.; Vandewalle, J. *Neurocomputing* **2002**, *48*, 85−105.
(34) Pelckmans, K.; Suykens, J.; Van Gestel, T.; De Brabanter, J.; Lukas, L.; Hamers, B.; De Moor, B. LS-SVMlab: A Matlab/C Toolbox for Least Squares Support Vector Machines. *Tutorial. KULeuven-ESAT*, 2002.
(35) Filzmoser, P.; Liebmann, B.; Varmuza, K. *J. Chemom.* **2009**, *23*, 160−171.
(36) De Meutter, J.; Derfoufi, K.-M.; Goormaghtigh, E. *Biomed. Spectrosc. Imag.* **2016**, *5*, 145−154.
(37) Kalnin, N. N.; Baikalov, I. A.; Venyaminov, S. Y. *Biopolymers* **1990**, *30*, 1273−1280.
(38) Derenne, A.; Derfoufi, K.-M.; Cowper, B.; Delporte, C.; Goormaghtigh, E. *Anal. Chim. Acta* **2020**, *1112*, 62−71.
(39) Goormaghtigh, E.; Cabiaux, V.; Ruysschaert, J.-M. *Subcell. Biochem.* **1994**, *23*, 329−362.
(40) Dousseau, F.; Pezolet, M. *Biochemistry* **1990**, *29*, 8771−8779.
(41) Kennard, R. W.; Stone, L. A. *Technometrics* **1969**, *11*, 137−148.
(42) Navea, S.; Tauler, R.; Goormaghtigh, E.; de Juan, A. *Proteins: Struct., Funct., Bioinf.* **2006**, *63*, 527−541.
(43) Cortes, C.; Vapnik, V. *Mach. Learn.* **1995**, *20*, 273−297.
(44) Arrondo, J. L. R.; Muga, A.; Castresana, J.; Goñi, F. M. *Prog. Biophys. Mol. Biol.* **1993**, *59*, 23−56.
(45) Scheirlinckx, F.; Raussens, V.; Ruysschaert, J.-M.; Goormaghtigh, E. *Biochem. J.* **2004**, *382*, 121−129.
(46) Srour, B.; Bruechert, S.; Andrade, S. L. A.; Hellwig, P. *Methods Mol. Biol.* **2017**, *1635*, 195−203.
(47) Rahmelow, K.; Hübner, W. *Appl. Spectrosc.* **1997**, *51*, 160−170.
(48) Venyaminov, S. Y.; Prendergast, F. G. *Anal. Biochem.* **1997**, *248*, 234−245.
(49) Prestrelski, S. J.; Tedeschi, N.; Arakawa, T.; Carpenter, J. F. *Biophys. J.* **1993**, *65*, 661−671.
(50) Roy, I.; Gupta, M. N. *Biotechnol. Appl. Biochem.* **2004**, *39*, 165−177.
(51) Goormaghtigh, E.; Raussens, V.; Ruysschaert, J.-M. *Biochim. Biophys. Acta* **1999**, *1422*, 105−185.
(52) Goormaghtigh, E.; de Jongh, H. H. J.; Ruysschaert, J.-M. *Appl. Spectrosc.* **1996**, *50*, 1519−1527.
(53) Yang, H.; Yang, S.; Kong, J.; Dong, A.; Yu, S. *Nat. Protoc.* **2015**, *10*, 382−396.

https://dx.doi.org/10.1021/acs.analchem.0c03677
Anal. Chem. 2021, 93, 3733−3741