Talanta 205 (2019) 120084

Contents lists available at ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta

FTIR-based hierarchical modeling for prediction of average molecular weights of protein hydrolysates

Kenneth Aase Kristoffersen[a,b,*], Kristian Hovde Liland[c], Ulrike Böcker[a], Sileshi Gizachew Wubshet[a], Diana Lindberg[a], Svein Jarle Horn[b], Nils Kristian Afseth[a]

[a] Nofima - Norwegian Institute of Food, Fisheries and Aquaculture Research, P.O. Box 210, N-1431, Ås, Norway
[b] Faculty of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences (NMBU), P.O. Box 5003, N-1432, Ås, Norway
[c] Faculty of Science and Technology, Norwegian University of Life Sciences (NMBU), P.O. Box 5003, N-1432, Ås, Norway

ARTICLE INFO

Keywords:
Enzymatic protein hydrolysis
FTIR
Data analysis
Hierarchical modeling

ABSTRACT

In the presented study, Fourier-transform infrared (FTIR) spectroscopy is used to predict the average molecular weight of protein hydrolysates produced from protein-rich by-products from food industry using commercial enzymes. Enzymatic protein hydrolysis is a well-established method for production of protein-rich formulations, recognized for its potential to valorize food-processing by-products. The monitoring of such processes is still a significant challenge as the existing classical analytical methods are not easily applicable to industrial setups. In this study, we are reporting a generic FTIR-based approach for monitoring the average molecular weights of proteins during enzymatic hydrolysis of by-products from the food industry. A total of 885 hydrolysate samples from enzymatic protein hydrolysis reactions of poultry and fish by-products using different enzymes were studied. FTIR spectra acquired from dry-films of the hydrolysates were used to build partial least squares regression (PLSR) models. The most accurate predictions were obtained using a hierarchical PLSR approach involving supervised classification of the FTIR spectra according to raw material quality and enzyme used in the hydrolysis process, and subsequent local regression models tuned to specific enzyme-raw material combinations. The results clearly underline the potential of using FTIR for monitoring protein sizes during enzymatic protein hydrolysis in industrial settings, while also paving the way for measurements of protein sizes in other applications.

1. Introduction

FTIR spectroscopy has become an established method for protein and peptide structural characterization over the last few decades. This is due to the detailed structural information found in FTIR spectra, where the repeated amino acid building blocks of proteins and peptides give rise to nine distinctive infrared (IR) absorption bands (i.e., the amide bands) [1,2]. The inherent ability of FTIR spectroscopy to monitor the protein backbone can also provide a range of possibilities to study parameters related to protein secondary structures. These parameters include hydration and solvent effects, pH and peptide size [3–8]. Protein size estimations can potentially have practical applications in a range of different fields, one of them being enzymatic protein hydrolysis. This process represents an efficient and suitable method to extract protein from food processing residuals, involving the breakdown of proteins into peptides and free amino acids. The success of a hydrolysis process can be measured by its ability to produce maximum possible yield of a high-quality product within the required specifications and with limited or no batch-to-batch variation. Currently, there is a lack of fast and reliable analytical monitoring tools that can be used to achieve such process control.

In enzymatic protein hydrolysis reactions, a proteolytic enzyme catalyzes the hydrolysis of peptide bonds. The reaction results in the formation of C-terminals (COO^-) and N-terminals (NH_3^+), consequently changing both the primary and secondary structure of the protein or peptide. Several studies on pure model proteins like hemoglobin, β-lactoglobulin, β-casein, and bovine serum albumin have demonstrated that FTIR spectroscopy can be used to monitor proteolytic reactions. [9–13] This was extended to more complex protein-rich matrices when FTIR was employed as a tool to predict degree of hydrolysis (DH%) values for trypsin-catalyzed hydrolysis of whey proteins [14]. Recently, the applicability of monitoring proteolysis using FTIR was further expanded to salmon and poultry-based substrates [15]. Wubshet et al. reported an FTIR-based multivariate approach for monitoring the change in weight average molecular weight (M_w) during enzymatic hydrolysis of chicken by-products [16]. In that study, M_w was

* Corresponding author. Nofima - Norwegian Institute of Food, Fisheries and Aquaculture Research, P.O. Box 210, N-1431, Ås, Norway.
E-mail address: Kenneth.kristoffersen@nofima.no (K.A. Kristoffersen).

https://doi.org/10.1016/j.talanta.2019.06.084
Received 26 January 2019; Received in revised form 19 June 2019; Accepted 21 June 2019
Available online 22 June 2019

calculated from size exclusion chromatography (SEC), establishing it as a reference method for FTIR spectra calibration models and as a measure of the extent of protein hydrolysis.

Several studies have revealed that amide absorptions (i.e., amide I at ~1650 cm^{-1}), NH_3^+ deformation (1516 cm^{-1}), and COO^- stretching (1400 cm^{-1}) are important for prediction of M_w or DH% [14–16]. From these studies, it is also apparent that protein hydrolysates originating from different raw materials and even different enzymes will display different FTIR fingerprints. However, when the aim is to establish a generic prediction model for M_w, such spectral differences may be a challenge. This was illustrated in a recent study by Wubshet et al. [16]. Here, the coefficients of determination between FTIR spectra and M_w were reduced when samples of different raw material origins were combined in a single model, compared to separate modeling according to raw material origin. Different spectral signatures of the raw materials were designated as the main reason for this. Removing these types of spectral variations through mathematical pre-processing is hard and, in most cases, not possible. An alternative approach will therefore be to exploit these spectral differences. This can be achieved through a two-level model, where the spectra are assigned to predefined groups consisting of known enzyme-raw material combinations in the first level using a supervised classification model. On the second level the spectra can be subjected to a local regression model tuned to specific enzyme-raw material combinations. Hierarchical modeling through two-level strategies has previously been exploited in methods such as hierarchical cluster-based partial least squares regression (HC-PLSR) and hierarchically ordered taxonomic classification by partial least squares (Hot PLS) in applications including nonlinear dynamic models and taxonomic classification, respectively [17,18]. Successful applications of FTIR-based two-level partial least squares (PLS) modeling have been demonstrated in the determination of clinical parameters such as urea and glucose as well as complex protein structures [19,20].

In the present study, the objective was to establish and study the relationship between M_w and FTIR spectra in an extensive set of protein hydrolysates. Hence, a total of 885 hydrolysates from enzymatic protein hydrolysis of poultry and fish by-products using five different enzymes were studied. The two-level regression model, tuned to various combinations of raw material and enzymes, was compared to the standard PLSR model in order to demonstrate the power of a two-level regression approach. To the best of our knowledge, this is the first time a link between M_w of proteins and FTIR spectra has been studied and established for an extensive set of samples.

2. Materials and methods

2.1. Materials

Protease from *Bacillus licheniformis* (Alcalase, 2.4 U/g) was purchased from Sigma-Aldrich (St. Louis, MO, USA). Protamex and Flavourzyme was obtained from Novozymes (Bagsværd, Denmark), Papain LSG 100 from Enzybel (Waterloo, Belgium) and Corolase 2TS from AB Enzymes (Darmstadt, Germany). Analytical grade acetonitrile, trifluoroacetic acid, monosodium phosphate and molecular weight standards, i.e., bovine serum albumin, albumin from chicken egg white, carbonic anhydrase from bovine erythrocytes, lysozyme, cytochrome c from bovine heart, aprotinin from bovine lung, insulin chain B oxidized from bovine pancreas, angiotensin II human, bradykinin fragment 1-7, Val-Tyr-Val, and tryptophan were purchased from Sigma-Aldrich (St. Louis, MO, USA). Water used for HPLC was purified by deionization and 0.22 µm membrane filtration (MilliporeSigma, Burlington, MA, USA).

2.2. Raw materials

Protein-rich raw materials derived from chicken, turkey, salmon and mackerel were hydrolyzed by a selection of commercially available enzymes (see Table 1). The poultry raw materials (i.e., chicken mechanical deboning residue (CMDR), heat treated chicken mechanical deboning residue (hCMDR), chicken skin (CS), chicken bone (CB), turkey carcasses (TC) and turkey mechanical deboning residue (TMDR)) were supplied by a Norwegian slaughterhouse (Nortura, Hærland, Norway). Chicken fillets/muscle (CM) were purchased from a local grocery store in Ås, Norway. Salmon raw materials (i.e., heads (SH), bone (SB) and skin (SS)) were supplied by Nutrimar (Kverva, Norway). Mackerel raw materials (Ma) were supplied by Pelagia Tromsø (Tromsdalen, Norway). All samples were minced, packed in plastic bags and stored at −20 °C until further use.

Table 1
An overview of samples and hydrolysis reaction conditions.

Sample name[a]	Enzyme[b]	Enzyme loading (w/w)%[c]	Water (mL)[d]	Raw material (g)[e]	No. of samples[f]
CMDRA	Alcalase	1.5	1000	500	89
CMDRPa	Papain	1.5	1000	500	24
CMDRPr	Protamex	1.5	1000	500	36
hCMDRA	Alcalase	1.5	1000	500	23
hCMDRPa	Papain	1.5	1000	500	24
hCMDRPr	Protamex	1.5	1000	500	12
CMA	Alcalase	1.5	1000	500	87
CMPa	Papain	1.5	1000	500	23
CMPr	Protamex	1.5	1000	500	12
CSA	Alcalase	1.5	1000	500	22
CSPa	Papain	1.5	1000	500	24
CSPr	Protamex	1.5	1000	500	12
CBA	Alcalase	1.5	1000	500	12
CBPa	Papain	1.5	1000	500	12
CBPr	Protamex	1.5	1000	500	12
TCA	Alcalase	1	1000	500	22
TCC	Corolase 2TS	1	1000	500	24
TCF	Flavourzyme	1	1000	500	23
TMDRA	Alcalase	1	1000	500	24
TMDRC	Corolase 2TS	1	1000	500	24
TMDRF	Flavourzyme	1	1000	500	24
SHA	Alcalase	1	400	400	130
SSA	Alcalase	1	400	400	132
SBA	Alcalase	1	400	400	11
Ma	None	0	400	400	12
MaA	Alcalase	1	400	400	12
MaPa	Papain	1	400	400	11
MaF	Flavourzyme	1	400	400	12
All					885

[a] Raw materials are defined in chapter 2.2. The abbreviation for the enzymes used are added to the sample name.
[b] Alcalase, 2.4 U/g (A), Protamex (Pr), Flavourzyme (F), Papain LSG 100 (Pa) and Corolase 2TS (C).
[c] Enzyme loading relative to wet weight raw material.
[d] Water added to reaction mixture.
[e] Raw material loading.
[f] Number of samples in each enzyme-raw material group.

2.3. Enzymatic hydrolysis and sampling

All hydrolysis reactions were performed according to a previously published protocol using a Reactor-Ready™ jacketed reaction vessel (Radleys, Essex, United Kingdom) [16]. Water circulating through the vessel jacket was kept at 50 °C and was supplied using a JULABO circulator pump (JULABO GmbH, Seelbach, Germany). Raw materials and water were mixed in the ratios presented in Table 1. All reaction mixtures were thoroughly mixed and heated until the suspensions reached 50 ± 1 °C. This was followed by addition of 1-1.5% enzyme w/w to wet substrate weight. The reaction time, from addition of the enzyme, were 60 or 80 min. During the hydrolysis, aliquots of approximately 7 mL were collected at 11 or 12 time points (0.5, 2.5, 5, 7.5, 10, 15, 20, 30, 40, 50, 60 and 80 min, respectively). Many of the reactions were repeated multiple times, as seen in Table 1. After collecting the samples

from the reaction vessel, the enzyme was thermally inactivated before being allowed to cool to room temperature. The samples were then centrifuged to separate the mixtures into three phases: Solid, water and fat. The water phase was collected, and analytical samples of protein hydrolysate were prepared by filtration through Millex-HV PVDF 0.45 µm 33 mm filter (MilliporeSigma, Burlington, MA, USA).

2.4. Size exclusion chromatography

SEC analysis was preformed according to Wubshet et al. [16]. The 2 mg/mL solutions of standards and the filtrates of the water phases collected from the hydrolysis were directly used as injection solutions without further modifications. The injection volumes were of 7 µL for the fish samples and 10 µL for the standards and poultry samples. Chromatographic separations of standards and samples were performed with an Agilent 1200 series instrument (Agilent Technologies, Santa Clara, CA, USA). Separation was performed at 25 °C using BioSep-SEC-s2000 (300 × 7.8 mm) columns from Phenomenex (Torrence, CA, USA). The mobile phase consisted of a mixture of acetonitrile and ultrapure water in a proportion 30:70 (v/v), containing 0.05% trifluoroacetic acid. Isocratic elution was carried out using a flow rate of 0.9 mL/min for 17.0 min. Between 17.0 and 17.1 min the mobile phase was changed to NaH_2PO_4 (0.10 M) and maintained until 20.0 min for column cleaning. Elution conditions were restored between minute 20.0 and 20.1 and the column was equilibrated for an additional 25 min. Chromatographic runs were controlled from OpenLAB CDS Rev. C. 01.07 (Agilent Technologies, Santa Clara, CA, USA). From chromatographic runs of both the standards and hydrolysate samples presented in Table 1, a UV trace of 214 nm was used. For the analytical standards, retention times were obtained from the automatic peak-picking algorithm of OpenLAB CDS. The average retention times from triplicate measurements of the standards were used to construct the calibration curves. Calibration data for one of the columns used including average retention times and standard deviation for all the analytical standers, are presented in the supporting information (SI) in Table S–1. Finally, M_w was calculated from the UV trace of a single chromatographic run for each of the hydrolysate samples. Calculations of the M_w were performed using PSS winGPC UniChrom V 8.00 (Polymer Standards Service, Mainz, Germany). The calculation from the software was based on a slicing method, similar to those previously used for analysis of protein hydrolysates [21].

2.5. FTIR spectroscopy

From each of the filtered protein hydrolysates, aliquots (5-7.5 µL) were deposited on 96-well IR-transparent Si-plates (Bruker, Billerica, MA, USA) and dried at room temperature for at least 30 min to form dry-films as described by Böcker et al. [15]. From each hydrolysate sample, five aliquots were deposited to allow for replicate measurements. FTIR measurements were performed using a High Throughput Screening eXTension (HTS-XT) unit coupled to a Tensor 27 spectrometer (Bruker, Billerica, MA, USA). The spectra were recorded in the region between 4000 and 400 cm^{-1} with a spectral resolution of 4 cm^{-1} and an aperture of 5.0 mm. For each spectrum, 40 interferograms were collected and averaged. Data acquisition was controlled using Opus v6.5 (Bruker, Billerica, MA, USA).

2.6. Data analysis

Pre-processing of FTIR spectra was performed using Savitzky-Golay 2nd derivative smoothing (window width 11 pt, 3rd order polynomial smoothing) followed by extended multiplicative signal correction (EMSC) with 2nd order polynomial correction with the mean spectrum as reference. This pre-processing approach was used to reduce the scattering effects in the spectra and suppress the effect of the varying thickness of the dry-films [22]. For all subsequent data analysis, the region from 1800 cm^{-1} to 700 cm^{-1} was used. The pre-processed FTIR spectra were further used for prediction of M_w. Two PLSR approaches were applied and compared: (1) Standard PLSR (henceforth denoted as the one-level PLSR model) and (2) a two-level modeling approach. In the first approach, PLSR was applied to the pre-processed FTIR spectra from all subgroups of enzymes and raw materials and the corresponding M_w values obtained from size exclusion chromatography. All the 885 FTIR spectra and subgroup datasets according to raw material origin and enzymes used (see Table 1) were modeled together. In the two-level modeling approach, spectra were first classified into 28 subgroups (as defined in Table 1) and then subjected to regression models tuned to each raw material and enzyme combination. Both the classification and regression models used PLSR for dimension reduction. In the classification a variant called canonical partial least squares (CPLS) was chosen for its ability to give simpler models and the response was a dummy coded representation of the subgroup [23]. PLS finds small subspaces in the high-dimensional FTIR data that span most of the covariation between spectra and response. For classification, the scores from the CPLS model were subjected to linear discriminant analysis (LDA) to obtain class memberships, while ordinary least squares regression was used on the scores when predicting M_w. All the modeling, classification and prediction were wrapped in a leave-one-out cross-validation to make sure that none of the spectra interfered with any of the models that it was to be predicted from, thus eliminating sources of information bleeding. All statistical analyses were performed using the MATLAB software (R2018a, The MathWorks, Inc., Natick, MA, USA).

3. Results and discussion

3.1. Average molecular weight and FTIR profiling

A total of 885 hydrolysate samples were prepared in the current study, collected at different time points during enzymatic protein hydrolysis of a variety of raw materials and enzymes (see Table 1). All samples were analyzed by SEC and corresponding M_w were calculated. For a given protein hydrolysate, M_w is a highly descriptive molecular weight distribution parameter that can serve as a measure of the extent of hydrolysis. The relationship between hydrolysis time and M_w for two

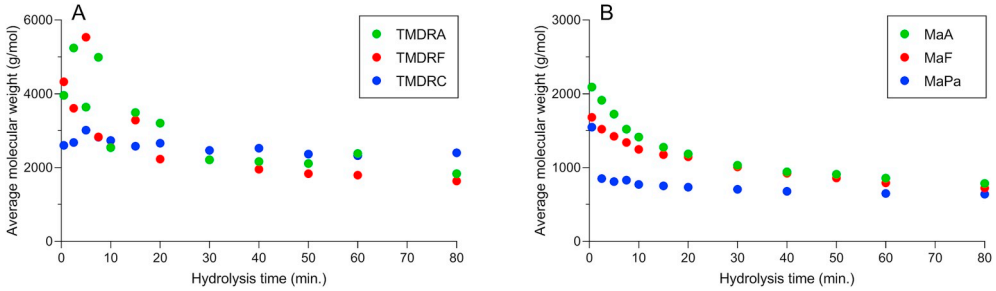

Fig. 1. M_w plotted against the hydrolysis time for protein hydrolysates obtained using four enzymes and two different raw materials from A) Turkey and B) Mackerel. Raw materials and enzyme abbreviations are defined in chapter 2.2. and Table 1.

raw materials (turkey and mackerel) is seen in Fig. 1. The maximum, minimum and average M_w for all raw material and enzyme combinations are displayed in SI, Table S–2.

A large variation in M_w between samples prepared from different raw materials and enzymes was observed. The general trend was a fast reduction in M_w at the start of the process followed by a slower change with increasing hydrolysis time, as shown in Fig. 1. Similar observations have been reported previously for hydrolysis reactions using different complex substrates and enzymes [16,21]. For many of the reactions, significant fluctuations were also observed in the first 20-30 min as shown in Fig. 1. It is reasonable to explain this observation by the substantial heterogeneity of the reaction mixtures, as raw materials were only ground to a semi-homogenous mass to simulate industry-relevant pretreatment before hydrolysis [24].

The turkey hydrolysates generally contained larger amounts of longer peptides at the beginning of the process, as compared to almost all of the other raw materials. The difference could be explained by differences in amount of structural proteins between the two poultry species. Turkey contains more collagenous compounds than chicken, and all proteases used have relatively higher specificity for peptide bonds prevalent within myofibrillar proteins. Another possible explanation is the formation of relatively higher amounts of "virtual intermediate peptides" after liberation of peptides within the collagen-rich turkey hydrolysates. Virtual intermediate peptides are aggregates of cleaved peptides, limiting the accessibility of the specific peptide bonds for peptide hydrolysis [25].

Depending on both the raw material quality and the enzyme used, enzymatic liberation of peptides and free amino acids during an enzymatic hydrolysis process of complex protein-rich substrates will result in large variations in the water phase composition. This is clearly observed in the dry-film FTIR spectra of all protein hydrolysates, especially in the spectral region 1800-1300 cm^{-1}. Fig. 2 displays the second derivative FTIR spectra representing hydrolysis time-series using two enzymes on two raw materials. The most important bands for describing these variations are marked, and include the NH_3^+ deformation (1516 cm^{-1}), the COO^- stretching (1400 cm^{-1}), the amide I (~1650 cm^{-1}), and the amide II (~1550 cm^{-1}) bands [15,16]. Additional examples of second derivative FTIR time-series of chicken and salmon hydrolysis reactions are provided in SI Fig. S-1.

A few general trends can be deduced from the second derivative FTIR spectra in Fig. 2. The dominating variation seen in the amide I and amide II bands are related to changes in the corresponding secondary structures of the proteins. As the larger proteins are broken down into smaller peptide fragments, the amide I and II changes and simplifies accordingly (i.e., there are fewer peaks in the second derivative spectra after 80 min than in the beginning of the hydrolysis). These changes are most pronounced in hydrolysis of the turkey samples due to the complex protein composition of this raw material. Another observation is the differences seen between the enzymes used. For Alcalase, large changes are observed in all the four bands marked in Fig. 2, while for Flavourzyme, more dominant changes are seen in the signals from the C- and N-terminals. These developments can be explained by the fact that Alcalase contains mostly endo-peptidases, while Flavourzyme, which mainly contains exo-peptidases, releases more free amino acids into the water phase [26].

3.2. Multivariate calibrations (PLSR)

The hydrolysates analyzed in the current study were produced using a variety of raw materials and enzymes. This resulted in hydrolysates with large variations in composition and size distribution, as illustrated in the previous section. Due to the extensive spectral variations seen in the FTIR spectra of the hydrolysates, two multivariate regression approaches were used and compared to establish the relationship between the FTIR spectra and the corresponding M_w values: (1) A one-level PLSR model and (2) a two-level PLSR model. In the one-level PLSR approach, the different main subgroups of Table 1 were combined in one single PLSR model. The results are provided in Table 2, and as shown in the table, PLSR models with moderate to high cross-validated coefficients of determination (R^2) were obtained. The exception was the turkey protein hydrolysates where the R^2 value was 0.455. The PLSR results of Table 2 can also be presented in measured vs. predicted plots, as shown in Fig. 3. Here the raw material groups are color-coded. Fig. 3 reveals that for the one-level PLSR approach, the prediction of lower and higher M_w values were less accurate than the region between approx. 1500-4000 g mol^{-1}. The most challenging samples, as also indicated in the previous section, are the hydrolysates originating from turkey.

The results from merging all the data and constructing a one-level PLSR model are shown in Fig. 3 and Table 2. As judged by R^2 and the root mean square error of cross-validation (RMSECV), the model is only moderately good. Thus, a two-level PLSR model was utilized to see if any improvements could be achieved. In the two-level modeling approach, the spectra were first classified into 28 subgroups of raw materials + enzymes as defined in Table 1 and then subjected to

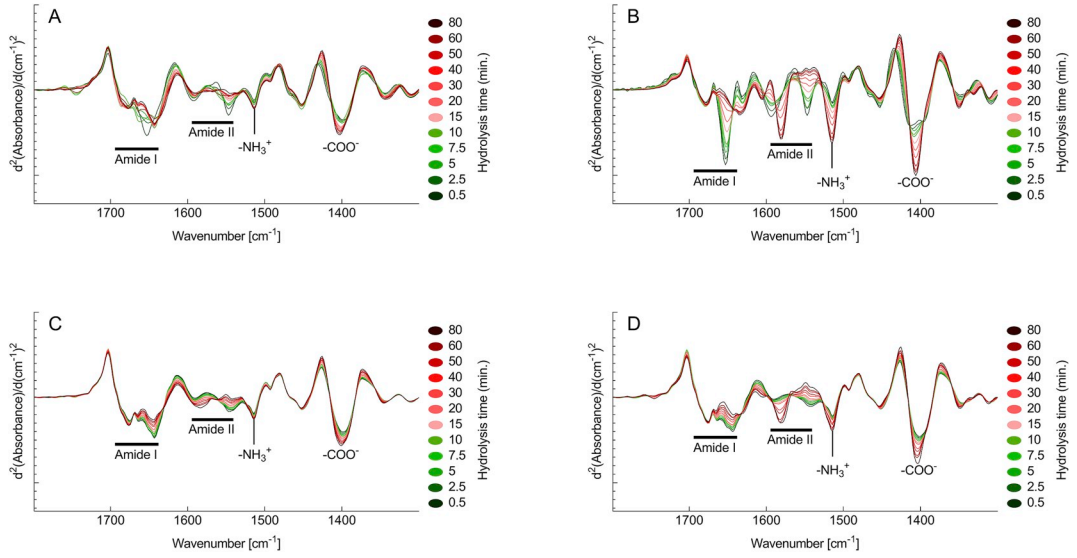

Fig. 2. Second derivative FTIR spectra (1800-1300 cm^{-1}) of hydrolysate time-series from hydrolysis reactions using two different enzymes and two different raw materials. A and B) Turkey mechanical debone residue. C and D) Mackerel. A and C) Alcalase. B and D) Flavourzyme.

Table 2
One-level PLSR model results for different groups of protein hydrolysis samples.

One-level PLSR model	No. of samples	R^2	RMSECV (g mol^{-1})
All	885	0.834	446
Poultry	565	0.817	449
Chicken	424	0.929	285
Turkey	141	0.455	667
Fish	320	0.934	298
Salmon	273	0.926	299
Mackerel	47	0.992	57

regression models tuned to each raw material + enzyme subgroup. Both the classification and regression models used PLS for dimension reduction. A score plot showing the two first PLS components of the classification step is provided in Fig. 4. It can be seen from the figure, using the two first components of the classification model, that 81.25% of the samples are correctly classified into subgroups. The score plot provides a very good illustration of the raw material effect on the FTIR spectra. In the plot, there is a main separation between fish raw materials (lower (i.e., mackerel) and left (i.e., salmon) part of the plot) and poultry materials (right and upper part of the plot). The chicken raw materials are found along the whole length of the second component, whereas most of the turkey raw materials are found in the middle part of the plot. There are also significant overlaps between some of the raw material subgroups, and a total of 24 components were needed for correctly classifying 884 of the 885 samples in the current sample set (data not shown). Since this is a supervised classification, it is also interesting to note that the effect of the hydrolysis time, which is one of the major parameters contributing to protein size differences, is virtually absent in the score plot.

The results of the two-level PLSR model are provided in Table 3 and Fig. 5. The table reveals that for all groups, there is a considerable improvement in regression results compared to Table 2. The exception is the mackerel data, which is very well modeled also using the one-level approach. A similar trend is shown in Fig. 5 where large improvements are observed for all raw material groups, especially in the M_w regions that were more challenging using the one-level approach. This shows that using the two-level approach is a feasible tool for quantifying generic features from highly detailed spectroscopic measurements of samples of different origin. However, even when using the two-level approach, the estimation errors for prediction of M_w higher than 4000 g mol^{-1} are still high, especially for the turkey hydrolysates.

As previously suggested, the error in prediction of higher M_w may be largely related to inaccuracies of the reference analysis itself. This is not only due to the SEC column exclusion range, but also partly due to the detection method used. In this study a UV detector was used measuring peptide bonds at 214 nm, which is the most common detection method and wavelength used for this purpose. However, there are some

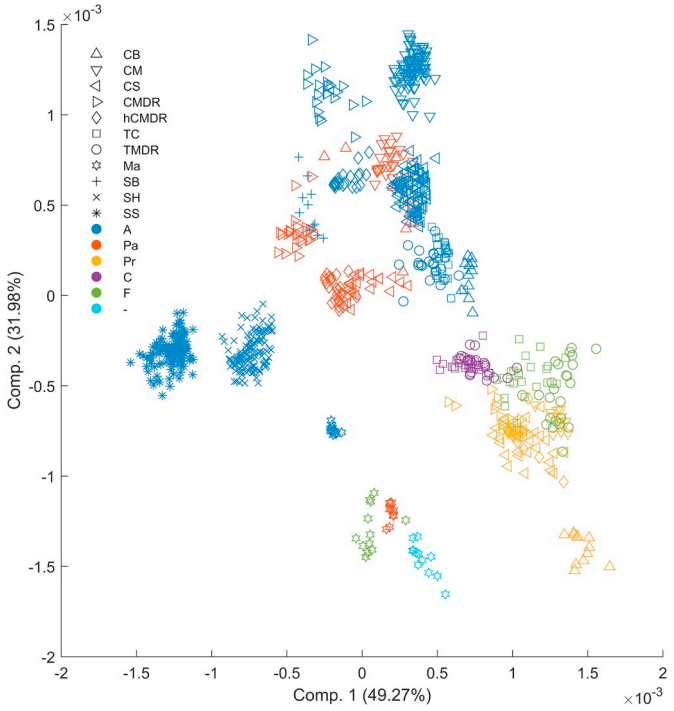

Fig. 4. Score plot with the two first components of the PLSR classification. The shape of the points denote raw material origin, whereas enzymes are denoted by color. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 3
Two-level PLSR model results for different groups of protein hydrolysis samples.

Two-level PLSR model	No. of samples	R^2	RMSECV (g mol^{-1})
All	885	0.944	260
Poultry	565	0.926	286
Chicken	424	0.989	114
Turkey	141	0.651	534
Fish	320	0.970	201
Salmon	273	0.962	217
Mackerel	47	0.995	44

limitations to this method as free amino acids are barely detected at this wavelength, while proteins and peptides are detected by absorption contributions from both peptide bonds and side-groups [27,28]. This, together with poor retention of larger peptides and protein fragments, will result in scaling errors, which in turn will affect how well the

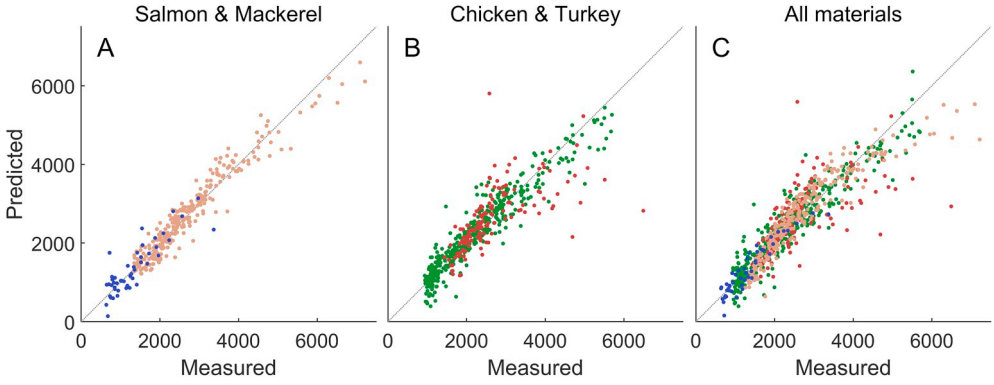

Fig. 3. One-level PLSR results, (blue) mackerel, (orange) salmon, (green) chicken and (red) turkey. A) 320 Salmon and mackerel samples. B) 565 Chicken and turkey samples. C) All 885 samples. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

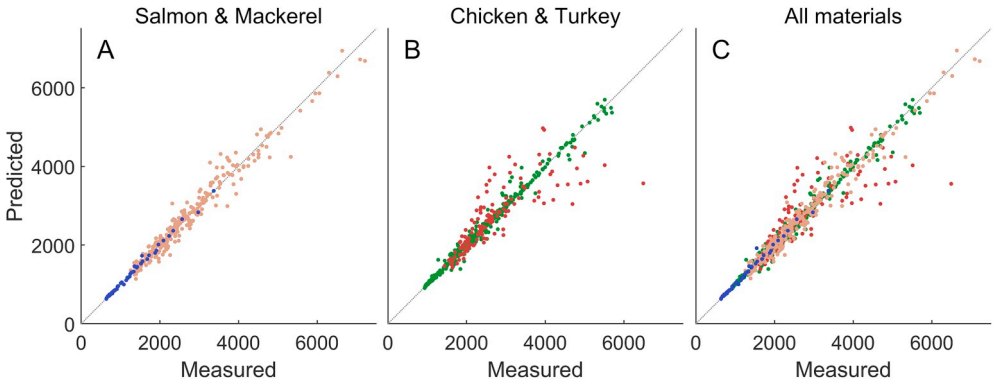

Fig. 5. Two-level PLSR results, (blue) mackerel, (orange) salmon, (green) chicken and (red) turkey. A) 320 Salmon and mackerel samples. B) 565 Chicken and turkey samples. C) All 885 samples. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

calculated M_w reflects the actual molecular weight distribution in the sample.

Another potential source of error affecting the relationship between the FTIR spectra and M_w values, is the inherent levels of chemical detail reflected in the FTIR spectra. When the hydrolysates have high molecular weights, larger protein and peptide fragments having a secondary structure will dominate the FTIR spectra, particularly in the amide I and the amide II regions. At lower molecular weights, where the enzymes have broken down some of these larger fragments, the spectral features related to secondary structures will be less pronounced. In addition, the raw materials used in the current study have different complexity levels. Fish raw materials generally have less complex protein composition than poultry raw materials. This will lead to less complexity in the amide I and amide II regions, as illustrated in Fig. 2, which in turn influences the possibility of making an adequate protein size calibration covering many different raw materials. The complexity differences in the FTIR amide bands related to secondary structures between subgroups also serve as a very good illustration of why the two-level PLSR outperforms the one-level PLSR approach in the current study. The regression coefficients of the PLSR models provide a good support for the effects of raw material on prediction accuracy. The regression coefficients of two PLSR models (i.e., chicken muscle hydrolyzed with Alcalase and turkey mechanical debone residue hydrolyzed with Flavourzyme) are shown in Fig. 6. Chicken muscle hydrolysate samples are expected to contain the least complex protein composition of the two raw materials and the COO^- stretching band (around 1400 cm^{-1}) is therefore a major feature of the regression coefficients. For the turkey mechanical debone residue, on the other hand, features in the amide region are more dominant. The regression coefficients of the other combinations of raw material and enzyme are presented in SI, Fig. S-2.

The use of two-level or hierarchical modeling is usually related to three conditions: (1) The reference value of interest relates differently to the recorded spectra in different subgroups of the sample material, thus leading to a higher prediction accuracy if a spectrum is predicted with the correct local model. (2) The spectral signals that are used for classifying into subgroups are stable and possibly distinct from the signals used for the reference predictions. (3) The precision of the classification into subgroups is high and/or the consequence of a wrong classification is low because similar subgroups have similar prediction models. If the data being analyzed does not adhere to these conditions, a hierarchical approach may have little value or be negative for the overall prediction accuracy. The hydrolysates analyzed in this study fit perfectly to all three conditions. The dominant variation in the spectra is due to the differences between subgroups (condition 2), while this source of variation is small inside each subgroup and across closely related subgroups (conditions 1 and 3).

One usually imagines a hierarchical model as a structure where samples enter at the top, getting classified through one or more levels

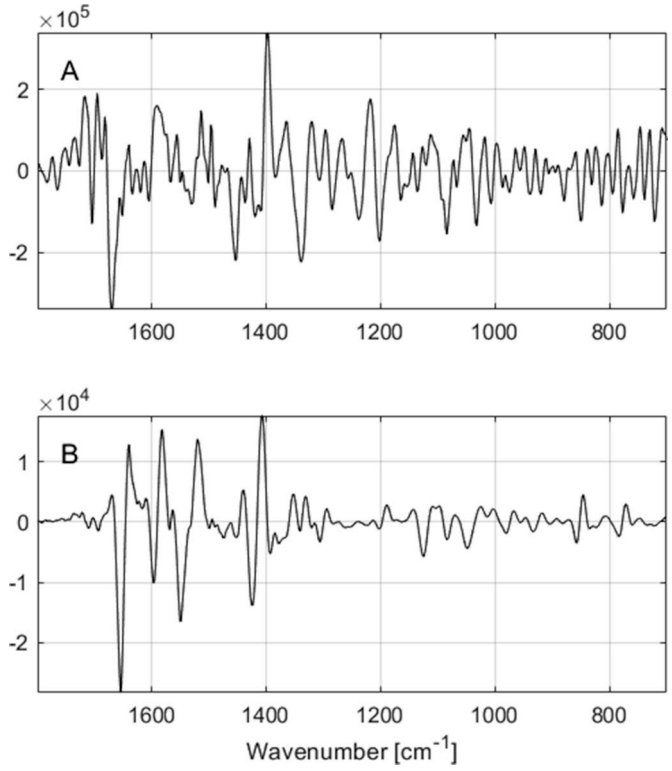

Fig. 6. Regression coefficients of the PLSR models of all hydrolysate time-series produced using: A) Chicken muscle and Alcalase. B) Turkey mechanical debone residue and Flavourzyme.

on their way down, while predictions (or classifications) fall out of the bottom of the model. In the two-level PLS model, the upper level is thus the global subgroup classifier (CPLS + LDA), while the lower level consists of a set of PLSR models that return predicted M_w values. The upper level in the approach uses known subgroups as classes. In cases where such subgroups follow the above-mentioned conditions, they are well suited for this type of modeling. If the subgroups were unknown, they might still be possible to infer from the structure of the spectra, e.g. using the clustering approach of HC-PLS. More complex relations between subgroups with different structures in the signals might benefit from more levels in the modeling, e.g. like the Hot PLS approach. In all the hierarchical approaches, a balance between the local adaptiveness of the model, i.e., how finely split the subgroups are, and the size of the subgroups, i.e., how robust and precise the prediction models are, must be found. In this case, optimization for this was not performed, but a good balance was achieved by the sheer number of samples and the

design of the enzyme-raw material combinations that were chosen. This two-level approach may also be further developed to handle unspecified hydrolysate samples (with regard to enzyme and raw material used to hydrolysate). For this, an unsupervised classification system is needed to group samples with similarities. An example of this has been presented by Perez-Guaita D. et al. [19].

It is important to note that full cross-validation was used for all regression models in the present study. As the sample size of the different local regression models in the hierarchical approach varied from 11 samples to 132 samples, the cross-validation approach was the only validation allowing to appropriately compare modeling results. Segmented cross-validation leaving out sets of replicates corresponding to each material-enzyme-time combination was tested, but these results were closely comparable to the ones presented here (< 1% difference in R^2, data not shown). In future work, when larger subgroups are present, it will be essential to also employ proper test set validation of the local regression models.

The results of this study clearly illustrate the potential of using FTIR for quantifying protein sizes in a range of different protein hydrolysates. The study also provides a feasible solution for building a generic calibration for protein sizes in hydrolysates. The approach of hierarchical modeling is also expected to be a potential solution in other FTIR approaches where the aim is to quantify a generic component in different raw materials (e.g. fatty acid quantification across microbial strains). As the use of dry-film FTIR for automated high-throughput analysis including automated sample handling and robotics is gaining increasing attention, a commercial system for protein size estimations in enzymatic protein hydrolysis industry could thus be expected when proper technical developments are made. A tool for protein size estimation would potentially also find applications in a range of different fields, including reaction kinetics, *in vitro* protein digestion, protein production by fermentation, and characterization of protein and peptide compositions of dairy products.

4. Conclusion

In the present study, we have shown that M_w of protein hydrolysates can be predicted with high accuracy using FTIR spectroscopy. The best result was obtained using a hierarchical PLSR approach where FTIR spectra of the protein hydrolysates were classified according to raw material type and enzyme prior to local modeling. This shows that prediction of protein sizes in protein hydrolysates can be achieved for a range of different raw materials using a single mathematical model. The results therefore demonstrate the potential of using FTIR for monitoring protein sizes during enzymatic protein hydrolysis in industrial settings, while also paving the way for measurements of protein sizes in other applications.

CRediT authorship contribution statement

Kenneth Aase Kristoffersen: Writing - original draft, Data curation, Conceptualization, Visualization, Investigation, Formal analysis. **Kristian Hovde Liland:** Writing - original draft, Software, Methodology, Conceptualization, Visualization, Investigation. **Ulrike Böcker:** Data curation, Conceptualization, Writing - review & editing. **Sileshi Gizachew Wubshet:** Conceptualization, Writing - review & editing. **Diana Lindberg:** Writing - review & editing. **Svein Jarle Horn:** Writing - review & editing. **Nils Kristian Afseth:** Writing - original draft, Writing - review & editing, Conceptualization, Investigation, Methodology.

Acknowledgement

Ria Das, Ornella Fleury, Rasmus Karstad and Marte Ryen Dalsnes are acknowledged for excellent technical assistance. Financial support from the Norwegian Fund for Research Fees for Agricultural Products, Norway through the projects “FoodSMaCK” (no. 262308) and “SunnMat” (no. 262300), and from the Norwegian Research Council, Norway through the project “Notably” (no. 280709) is greatly acknowledged. Internal funding from Nofima through the project “PepTek” is also greatly acknowledged.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.talanta.2019.06.084.

References

[1] U. Böcker, R. Ofstad, H.C. Bertram, B. Egelandsdal, A. Kohler, Salt-induced changes in pork myofibrillar tissue investigated by FT-IR microspectroscopy and light microscopy, J. Agric. Food Chem. 54 (18) (2006) 6733–6740.
[2] P. Gelfand, R.J. Smith, E. Stavitski, D.R. Borchelt, L.M. Miller, Characterization of protein structural changes in living cells using time-lapsed FTIR imaging, Anal. Chem. 87 (12) (2015) 6025–6031.
[3] A. Barth, The infrared absorption of amino acid side chains, Prog. Biophys. Mol. Biol. 74 (3) (2000) 141–173.
[4] N. Perisic, N.K. Afseth, R. Ofstad, A. Kohler, Monitoring protein structural changes and hydration in bovine meat tissue due to salt substitutes by fourier transform infrared (FTIR) microspectroscopy, J. Agric. Food Chem. 59 (18) (2011) 10052–10061.
[5] P.V. Andersen, E. Veiseth-Kent, J.P. Wold, Analyzing pH-induced changes in a myofibril model system with vibrational and fluorescence spectroscopy, Meat Sci. 125 (2017) 1–9.
[6] S.H. Arabi, B. Aghelnejad, C. Schwieger, A. Meister, A. Kerth, D. Hinderberger, Serum albumin hydrogels in broad pH and temperature ranges: Characterization of their self-assembled structures and nanoscopic and macroscopic properties, Biomater. Sci. 6 (3) (2018) 478–492.
[7] G. Martra, C. Deiana, Y. Sakhno, I. Barberis, M. Fabbiani, M. Pazzi, M. Vincenti, The Formation and self-assembly of long prebiotic oligomers produced by the condensation of unactivated amino acids on oxide surfaces, Angew. Chem. Int. Ed. 53 (18) (2014) 4671–4674.
[8] H.-F. Okabayashi, H.-H. Kanbe, C.J. O'Connor, The role of an L-leucine residue on the conformations of glycyl-L-leucine oligomers and its N- or C-terminal dependence: Infrared absorption and Raman scattering studies, Eur. Biophys. J. 45 (1) (2016) 23–34.
[9] C. Ruckebusch, L. Duponchel, J.P. Huvenne, P. Legrand, N. Nedjar-Arroume, B. Lignot, P. Dhulster, D. Guillochon, Hydrolysis of hemoglobin surveyed by infrared spectroscopy II. Progress predicted by chemometrics, Anal. Chim. Acta 396 (2–3) (1999) 241–251.
[10] C. Ruckebusch, L. Duponchel, J.P. Huvenne, Degree of hydrolysis from mid-infrared spectra, Anal. Chim. Acta 446 (1–2) (2001) 257–268.
[11] C. Ruckebusch, B. Sombret, R. Froidevaux, J.P. Huvenne, On-line mid-infrared spectroscopic data and chemometrics for the monitoring of an enzymatic hydrolysis, Appl. Spectrosc. 55 (12) (2001) 1610–1617.
[12] G. Guler, E. Dzafic, M.M. Vorob'ev, V. Vogel, W. Mantele, Real time observation of proteolysis with Fourier transform infrared (FT-IR) and UV-circular dichroism spectroscopy: Watching a protease eat a protein, Spectrochim. Acta Mol. Biomol. Spectrosc. 79 (1) (2011) 104–111.
[13] G. Guler, M.M. Vorob'ev, V. Vogel, W. Mantele, Proteolytically-induced changes of secondary structural protein conformation of bovine serum albumin monitored by Fourier transform infrared (FT-IR) and UV-circular dichroism spectroscopy, Spectrochim. Acta Mol. Biomol. Spectrosc. 161 (2016) 8–18.
[14] N.A. Poulsen, C.E. Eskildsen, M. Akkerman, L.B. Johansen, M.S. Hansen, P.W. Hansen, T. Skov, L.B. Larsen, Predicting hydrolysis of whey protein by mid-infrared spectroscopy, Int. Dairy J. 61 (2016) 44–50.
[15] U. Böcker, S.G. Wubshet, D. Lindberg, N.K. Afseth, Fourier-transform infrared spectroscopy for characterization of protein chain reductions in enzymatic reactions, Analyst 142 (15) (2017) 2812–2818.
[16] S.G. Wubshet, I. Mage, U. Böcker, D. Lindberg, S.H. Knutsen, A. Rieder, D.A. Rodriguez, N.K. Afseth, FTIR as a rapid tool for monitoring molecular weight distribution during enzymatic protein hydrolysis of food processing by-products, Anal. Methods 9 (29) (2017) 4247–4254.
[17] K. Tøndel, U.G. Indahl, A.B. Gjuvsland, J.O. Vik, P. Hunter, S.W. Omholt, H. Martens, Hierarchical Cluster-based Partial Least Squares Regression (HC-PLSR) is an efficient tool for metamodelling of nonlinear dynamic models, BMC Syst. Biol. 5 (1) (2011) 90.
[18] K.H. Liland, A. Kohler, V. Shapaval, Hot PLS—a framework for hierarchically ordered taxonomic classification by partial least squares, Chemometr. Intell. Lab. Syst. 138 (2014) 41–47.
[19] D. Perez-Guaita, J. Kuligowski, G. Quintás, S. Garrigues, M.d.l. Guardia, Modified locally weighted—partial least squares regression improving clinical predictions from infrared spectra of human serum samples, Talanta 107 (2013) 368–375.
[20] O. Preisner, R. Guiomar, J. Machado, J.C. Menezes, J.A. Lopes, Application of fourier transform infrared spectroscopy and chemometrics for differentiation of Salmonella enterica serovar enteritidis phage types, Appl. Environ. Microbiol. 76 (11) (2010) 3538–3544.
[21] D.S.T. Hsieh, C. Lin, E.R. Lang, N. Catsimpoolas, C.K. Rha, Molecular-weight

distribution of soybean globulin peptides produced by peptic hydrolysis, Cereal Chem. 56 (4) (1979) 227–231.
[22] N.K. Afseth, A. Kohler, Extended multiplicative signal correction in vibrational spectroscopy, a tutorial, Chemometr. Intell. Lab. Syst. 117 (2012) 92–99.
[23] U.G. Indahl, K.H. Liland, T. Naes, Canonical partial least squares—a unified PLS approach to classification and regression problems, J. Chemom. 23 (9) (2009) 495–504.
[24] M.C. Archer, J.O. Ragnarsson, S.R. Tannenbaum, D.I.C. Wang, Enzymatic solubilization of an insoluble substrate, fish protein concentrate: Process and kinetic considerations, Biotechnol. Bioeng. 15 (1) (1973) 181–196.
[25] R. Muñoz-Tamayo, J. de Groot, P.A. Wierenga, H. Gruppen, M.H. Zwietering, L. Sijtsma, Modeling peptide formation during the hydrolysis of β-casein by Lactococcus lactis, Process Biochem. 47 (1) (2012) 83–93.
[26] M. Merz, W. Claaßen, D. Appel, P. Berends, S. Rabe, I. Blank, T. Stressler, L. Fischer, Characterization of commercially available peptidases in respect of the production of protein hydrolysates with defined compositions using a three-step methodology, J. Mol. Catal. B Enzym. 127 (2016) 1–10.
[27] P. Hong, S. Koza, E.S.P. Bouvier, A review size-exclusion chromatography for the analysis of protein biotherapeutics and there aggregates, J. Liq. Chromatogr. Relat. Technol. 35 (20) (2012) 2923–2950.
[28] G. Brusotti, E. Calleri, R. Colombo, G. Massolini, F. Rinaldi, C. Temporini, Advances on size exclusion chromatography and applications on the analysis of protein biopharmaceuticals and protein aggregates: A mini review, Chromatographia 81 (1) (2018) 3–23.