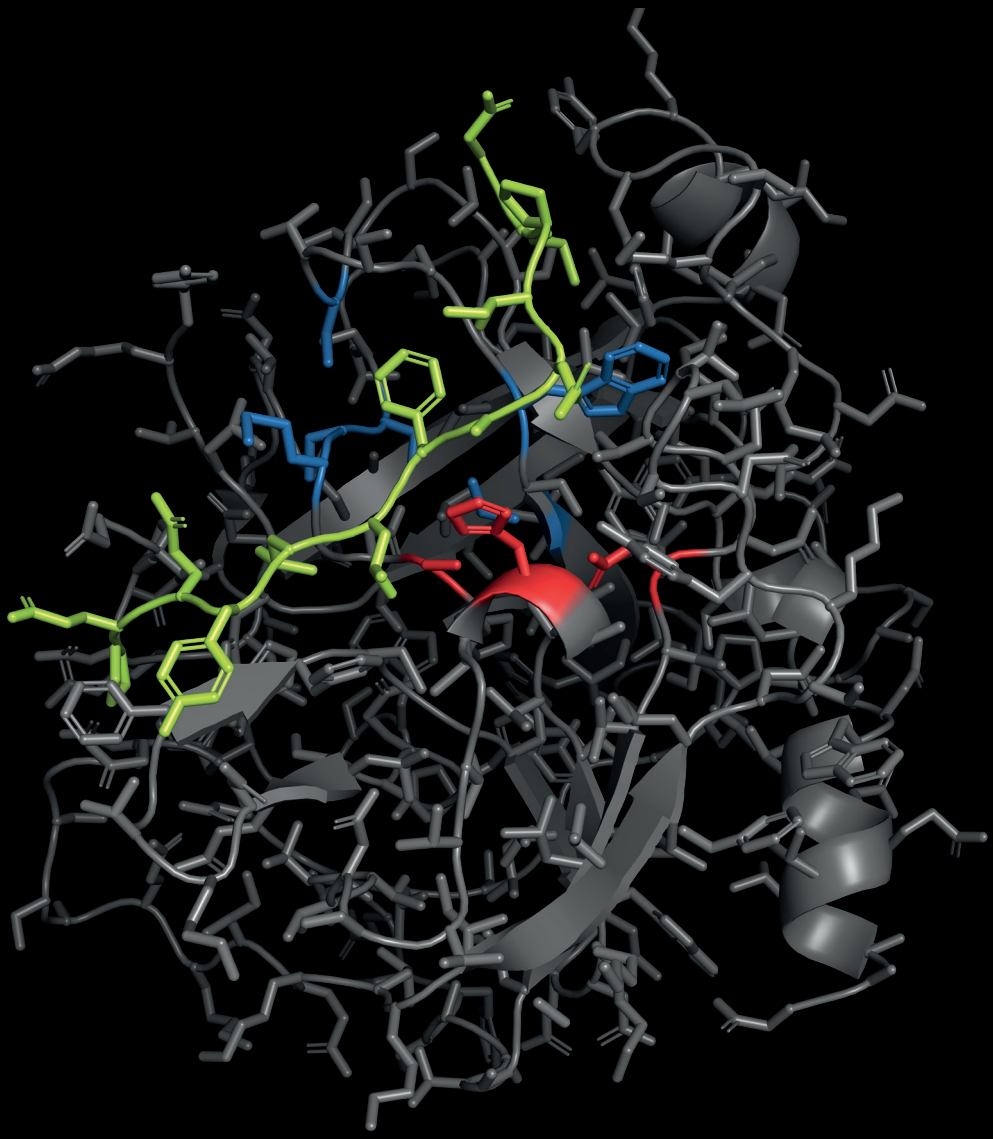
Robust, automated and quantitative peptide mapping

Compositional analysis of food protein hydrolysates and genetic variants, and gaining insight into the action of proteases



Gijs J. C. Vreeke

Propositions

1. From trypsin to chymotrypsin to pepsin, complexity increases in terms of specificity, preference and selectivity.
   (this thesis)

2. The analysis of combined known samples is the best predictor for the analysis quality of complex samples.
   (this thesis)

3. Documentation of data-processing in scientific articles is undervalued.

4. Sharing peer-reviewed research output via social media for professionals strongly enhances its reach.

5. Every human needs to learn how to perform cardiopulmonary resuscitation (CPR) in school.

6. A permanent job and income as requirement for renting a house is a violation of human rights.

Propositions belonging to the thesis, entitled
Robust, automated and quantitative peptide mapping
Compositional analysis of food protein hydrolysates and genetic variants, and gaining insight into the action of proteases

Gijs J.C. Vreeke
Wageningen, 23 January 2024

Robust, automated and quantitative peptide mapping

Compositional analysis of food protein hydrolysates and genetic variants, and gaining insight into the action of proteases

Gijs J.C. Vreeke

Thesis committee

Promotor
Prof. Dr J.-P. Vincken
Professor of Food Chemistry
Wageningen University & Research

Co-promotor
Dr P. A. Wierenga
Assistant professor, Laboratory of Food Chemistry
Wageningen University & Research

Other members
Prof. Dr T. Huppertz, Wageningen University & Research
Prof. Dr J. G. M. Janssen, Wageningen University & Research
Prof. Dr D. Dupont, INRAE-institut Agro, STLO, Rennes, France
Dr A. E. M. Janssen, Wageningen University & Research

This research was conducted under the auspices of VLAG Graduate School (Biobased, Biomolecular, Chemical, Food, and Nutrition sciences).

Robust, automated and quantitative peptide mapping

Compositional analysis of food protein hydrolysates and genetic variants, and gaining insight into the action of proteases

Gijs J.C. Vreeke

Thesis

Submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr A.P.J. Mol,
In the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Tuesday 23 January 2024
at 4 p.m. in the Omnia Auditorium.

Gijs J.C. Vreeke

Robust, automated and quantitative peptide mapping
Compositional analysis of food protein hydrolysates and genetic variants, and gaining insight into the action of proteases,
178 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2024)
With references, with summary in English

ISBN 978-94-6447-959-1
DOI https://doi.org/10.18174/641083

Abstract

Enzymatic protein hydrolysis is essential in the production of commercial hydrolysates and the digestion of food products. To thoroughly understand both processes, it is essential to identify the peptides formed and monitor their concentrations. Peptides can be analysed with LC-MS. The acquired data needs to be processed into reproducible and complete lists of peptides, to describe protease action, bio-active potential or genetic variant composition. Currently, such a data-processing method is not available. In this thesis, an automated method was developed for robust peptide mapping, according to guidelines set with a manual reference analysis of simple tryptic hydrolysates. Using these guidelines, a repeatability of 97 % was obtained for the analysis of abundant peptides in mixtures of these hydrolysates and full coverage of the amino acid sequence. All identified peptides were quantified absolute and label-free based on UV absorbance with comparable accuracy as obtained with reference peptides. The developed method was then applied to food hydrolysates with increasing complexity. First, complexity was induced by using extracts of different *Pisum sativum* cultivars as substrate, hydrolysed by a highly specific protease. Subsequently, complexity was induced by using a broadly specific protease (chymotrypsin) and finally a fully a-specific protease (pepsin) to hydrolyse milk protein isolates. The three studies led to unprecedented insights. For instance, the protein genetic variants expressed in yellow pea seeds were quantified and found to be surprisingly similar among cultivars. For digestive protease chymotrypsin, the specificity observed was much broader than previously reported. Furthermore, a proline residue hindered hydrolysis in certain sequence positions. For pepsin, which is naturally active in the stomach under dynamic pH conditions, relative hydrolysis rates of peptide bonds were generally independent of pH. For both digestive proteases, hydrolysis rates and kinetics widely varied, even for peptide bonds after similar amino acids. The developed method was shown to successfully fulfil the need for robust, automated and quantitative peptide mapping in food hydrolysates.

Table of contents

Chapter 1	General introduction	**1**
Chapter 2	A method to identify and quantify the complete peptide composition in protein hydrolysates	**21**
Chapter 3	Towards absolute quantification of protein genetic variants in *Pisum sativum* extracts	**47**
Chapter 4	The path of proteolysis by bovine chymotrypsin	**75**
Chapter 5	Quantitative peptide release kinetics to describe the effect of pH on pepsin preference	**101**
Chapter 6	General discussion	**137**
Summary		**167**
Acknowledgements		**171**
About the author		**175**

CHAPTER 1

General introduction

The breakdown of proteins into peptides is a vital process in the digestion of food and feed. This process is facilitated by digestive proteases in the stomach and small intestine. To better understand protein digestion, we need to unravel what these proteases actually do e.g. which bonds they break at which rate. This might be studied by following peptide release kinetics during *in vitro* digestion, which requires identification and quantification of the peptides formed at different stages of hydrolysis. The hydrolysates contain peptides of a few amino acids up to intact protein, in a diversity of concentrations and with various amino acid sequences. A methodology is required to identify and quantify all these peptides present. To follow peptide release kinetics, it is important that the peptide composition is reliably determined with high reproducibility and all peptides present are described by the results, but without false identifications. There are already methods that identify peptides, but these often result in low repeatability on peptide level (< 50 %), struggle with peptide quantification or require time-intensive manual data processing. In this thesis, a method is presented that meets all these criteria and enables untargeted identification and quantification of peptides, especially designed for characterising peptides in food protein hydrolysates.

INTRODUCTION TO THE METHODOLOGY

Characterisation of protein hydrolysates for food sciences without mass spectrometry

In the past, protein hydrolysates were often characterised using techniques that do not require mass spectrometry. To observe whether (intact) proteins were hydrolysed, typically SDS-PAGE was used. With this technique, the proteins are separated by gel electrophoresis and coloured with a (Coomassie) staining. Analysis of the bands at different timepoints shows the degradation or resistance of specific proteins during protein digestion [1, 2]. This technique is useful to evaluate whether protein hydrolysis occurred, but does generally not contain any information about the products. Only large peptide fragments can be identified with SDS-PAGE [3, 4]. Peptides can also be characterised for molecular weight using size exclusion chromatography (SEC). SEC is typically used to indicate differences in molecular weight distributions of peptides between samples [5-7].

Peptide identification with LC-MS

To analyse the amino acid sequence of peptides, nowadays mostly techniques involving mass spectrometry (MS) are used. With MS, the mass-over-charge (*m/z*) of the compounds of interest are measured before and after fragmentation. By fragmenting the peptide, the amino acids and their order can be confirmed. There are several choices to make in the experimental setup which affect the type of data obtained. Here, the most common setup is described that is typically used for analysing peptides. First, the peptides are separated with liquid chromatography (LC) using a C18 column and gradient of water and acetonitrile (ACN). Afterwards the peptides in the eluent are transferred from the solvent phase to the gas phase using a high voltage, also called electrospray ionisation (ESI) (for review on the principle of ESI see [8]). One peptide will give rise to many *m/z* signals in the MS spectra since (i) peptides are present in different charge states, (ii) at each charge state, a pattern of peaks is formed due to

carbon -and to a lower extent- nitrogen isotopes, (iii) part of the peptides have a neutral loss of water or ammonia, which can occur at every charge state [9, 10], and (iv) peptides could fragment during ionization, giving in-source fragments [11, 12]. All these signals end up in the mass spectrum acquired by the mass spectrometer. Generally, the mass spectrometers that are combined with LC can be divided in "scanning" MS techniques and "trapping" MS techniques. In a scanning technique, ions go through a quadrupole, prior to real-time analysis of the *m/z* with time-of-flight (TOF). In a trapping MS technique, ions are trapped and accumulated using electric fields before selective analysis of their masses. Examples of trapping MS techniques are the ion trap or Orbi-trap systems. To confirm the identity of the peptides, often the precursor ions are fragmented using collision induced dissociation (CID). This type of fragmentation breaks down the C-N amide bond and yields mostly b-fragments and y-fragments, which are the product ions containing the N-terminus and C-terminus of the peptide, respectively (**Figure 1.1**) [13]. In minor amounts, fragments with losses of NH_3 or H_2O are formed as well as a-ions (derived from b-fragments after release of CO_2). The CID fragmentation can be done for pre-selected targeted ions, for most intense MS ions or for all ions. The downside of fragmenting the most intense ions is that no MS/MS information of the less abundant peptides is obtained. Using MS^E, all ions eluting at the same time are fragmented, by using an energy ramp [14, 15]. The MS/MS fragments need to be matched with the precursor ions based on their chromatographic MS peak.

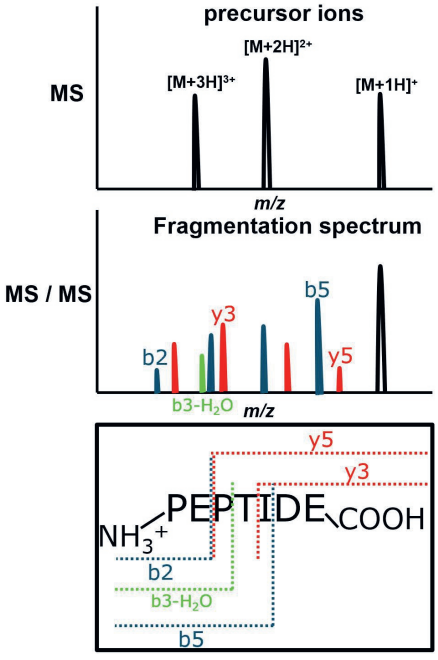


Figure 1.1. Illustration of confirming a peptide with MS/MS fragmentation. Precursor ions at multiple charge states are fragmented in a series of b-ions, y-ions and ions with a water of ammonium loss.

Current data-processing approaches

Several approaches can be used to process the acquired MS data (**Table 1.1**). At first, processing approaches can be categorised in targeted and untargeted approaches. In targeted approaches, the data is evaluated for the presence of a few peptides. These specific peptides can be targeted because of bio-functional properties [16, 17] or to confirm the authenticity of a product [18, 19]. The LC-MS data is searched directly for the theoretical masses of these target peptides to determine their presence or absence. A targeted approach cannot be used to characterise protein hydrolysates, since it is unknown which peptides are released. In an untargeted approach, all signals (above detection limit) are matched to a database with protein sequences. Untargeted approaches are typically evaluated using several objective parameters. Examples are the number of peptides and proteins identified, the percentage of peptides or proteins shared between replicate analyses of the same sample (repeatability) and to what extent the identified peptides cover the amino acid sequence of the proteins (amino acid sequence coverage). Based on performance on these parameters, we distinguish three (stereotype) data-processing styles: Proteomics, manual analysis and peptide analysis for food hydrolysates (**Table 1.1**).

- The goal in proteomics analysis is to determine the proteins present, with a high reproducibility on protein-level. These proteomics approaches are typically used to compare protein accumulation by different species [20], bacterial strains [21], cultivars [22] or organs [23, 24] and their changes in protein composition over time [25-27]. To identify the accumulated proteins, peptides need to be annotated that uniquely match a protein in a database. To identify as many proteins as possible, a large database of protein sequences is used and criteria on fragmentation spectra or signal intensity are relatively wide. Identification of as many proteins as possible is favoured over a high repeatability on peptide-level. The repeatability values reported on peptide-level are typically below 50 % [28, 29]. Short peptides are in most cases not of interest, since these can be matched to several proteins. Proteases with a clear specificity, as for instance trypsin, are used for hydrolysis and considered (only) in data-processing to limit the number of annotation options and limit thereby processing time. MS spectra and fragmentation spectra are matched to peptide sequences by an algorithm routine that considers the mass accuracy of the detector, likelihood of fragmentation and the intensities of fragments formed. In some cases, false positive identifications occur despite having a high score by the algorithm [30, 31].
- The goal in manual analysis is to determine the peptides present in a hydrolysate, with focus on a high repeatability and complete description of the hydrolysate on peptide-level. The manual analysis gives a robust list of peptides that can be used to study protease secondary specificity, for instance [32, 33]. The mass spectra of most dominant UV signals are manually matched with peptide masses from an *in silico* digest, considering the difference between the observed and expected mass, and observed and expected MS/MS fragments. These manual analyses are typically done on hydrolysates of one or two known, relatively small, proteins substrates. The focus is on identification of the most

abundant peptides and provide a complete overview of peptides (100 % amino acid sequence coverage). The intensive manual process of annotation limits the complexity of the samples that are feasible to analyse.

- The goal in peptide analysis for food hydrolysates is to determine the peptides present in a hydrolysate. For instance, studies characterised the peptides released upon protein digestion *in vitro* [2, 34-37] or *in vivo* [38, 39]. Other studies characterised peptides from various food sources for bio-functional properties [40-43] or techno-functional properties [44-47]. Typically, a few proteins (3-10) are dominant in these samples and their presence and sequences are known. The peptides formed a range from small to intact protein and can be formed by a specific or a-specific protease. The method needed to analyse these hydrolysates should annotate peptides with high repeatability leading to high amino acid sequence coverages. The complexity of the hydrolysates requires an automated approach. The method presented in this thesis aims to automate our in-house manual annotation procedure [48] and thereby allow the identification of peptides in food hydrolysates. The aim is to allow analysis of hydrolysates for food applications with similar robustness and evaluation of completeness as obtained with the manual analysis.

Table 1.1. Comparison of the typical proteomics data processing style, manual annotation and the method required to analyse food hydrolysates.

Proteomics analysis	**Method required to analyse food hydrolysates**	**Manual annotation**
Unknown substrates	Known substrates	Known substrates
Database with >100 proteins	< 10 protein sequences	1-2 protein sequences
Complex scoring with algorithms	Filter criteria based on MS/MS fragmentation	Manual spectra evaluation
Identification of peptides >7 AA	Identification of peptides >2 AA	Identification of peptides >2 AA
Fast and automated	Fast and automated	Slow and labour intensive
Fully specific analysis	Specific, semi-specific or a-specific analysis	Semi-specific
Low repeatability on peptide level, high repeatability on protein level	High repeatability on protein and peptide level	High repeatability on protein and peptide level

The steps in annotation of a peptide

The first step in untargeted annotation of a peptide is to match the precursor ion *m/z* with that of (theoretical) possible peptides, considering the mass accuracy of the mass spectrometer. This will result in one or more tentative options. To decide which annotation is correct, the MS/MS fragmentation spectra are compared with the theoretical possible fragments, for each annotation. This can be done manually, by a software that returns the fragments of each option (UNIFI) or by an algorithm which also considers intensities and calculates a score for the

Table 1.2. Studies using UNIFI software for peptide identification in various fields.

Goal	Type of analysis and specificity	Number of protein sequences	Post-translational modifications considered	Criteria set	First author and Reference
Study isomeric peptides for pharmaceutical industry with Cyclic IMS	Targeted analysis				Tomczyk *et al.*, 2021 [49]
Characterise peptides from Protamex and Flavourzyme in brewer's spent grain	Untargeted peptide mapping (a-specific)	16	0	Mass error (5 ppm); fragments (≥1)	Kriisa *et al.*, 2022 [50]
Compare peptide uptake by yeasts during wine fermentation	Untargeted peptide mapping (a-specific)	1 + *de novo* (2-4 AA)	0	Mass error (3 ppm for MS, 5 ppm for MS/MS); fragments (≥) ; low energy threshold (2000 counts), high energy threshold (200 counts)	Arju *et al.*, 2022 [51]
Characterise peptides during *in vitro* digestion of whey heated at different pH	Untargeted peptide mapping (a-specific)	7	Deamidation (N,Q), oxidation (M), pyroglutamic acid *N*-term (E, Q), phosphorylation (S, T, Y)	Fragments (≥10 % and ≥2)	Accardo *et al.*, 2022 [36]
Characterise peptides from endogenous plasmin activity during cheese maturation	Untargeted peptide mapping (a-specific)	4	Deamidation (N,Q), oxidation (M), Acetylation (protein *N*-term), phosphorylation (S, T, Y)	Mass error (5 ppm); fragments (≥1)	Xia *et al*, 2023 52]
Characterise peptides in *in vitro* digests of donkey milk processed with different treatments	Untargeted peptide mapping (a-specific)	13	Deamidation (N,Q), oxidation (single or double M, W), pyroglutamic acid *N*-term (E, Q), phosphorylation (S, T, Y)	-	Tedeschi *et al.*, 2023 [53]
Study Maillard glycation sites for α-LA peptides	Untargeted peptide mapping (specific)	1	Oxidation (M), Glycation with glucose, maltotriose, galacturonic acid (K)	Mass error (10 ppm)	Cardoso *et al.*2023 [54]

best match (Maxquant, used for proteomics). In this study, UNIFI software was used to annotate the peptides. This software makes all possible matches between precursor ions and the target protein sequence, within a certain mass error range. The software provides for each annotated peak the best match. In order to come to a repeatable and reliable list of identified peptides, criteria need to be set that are used in processing and to filter the list of tentative matches. Other studies which use UNIFI for annotation set typically thresholds for MS and MS/MS on mass error tolerance and minimum peak intensity and for MS/MS on the minimum number or percentage of fragments. An overview of studies using UNIFI is given in **Table 1.2**. None of these analyses are reported to be optimised using a manual annotation routine as reference.

A last important step to come to a list of reliable annotations is to recognize and remove in-source fragments. An in-source fragment cannot be distinguished by parent mass or MS/MS fragmentation from a (semi-specific) peptide. Kim *et al.* reported for a standard protein tryptic digest that ~57% of the unique 'peptide' entries were in-source fragments [11]. Picotti *et al.* reported an even higher percentage of in-source fragments (78 %) in the unique peptide list for a tryptic digest of five milk proteins [12]. For a complex sample (mouse brain lysate), the proportion of entries identified as in-source fragments was only 1 %. The in-source fragments had typically a 10 to 1,000 times lower intensity than the respective peptide [11]. In-source fragments do not seem to be an issue when the goal is to identify proteins as in proteomics because of two reasons. First, both the in-source fragment and peptide are attributed to the same protein. Secondly, the in-source fragments do not fully match the specificity of trypsin, and are therefore neglected anyway. However, in the analysis of protease release kinetics, it is important that in-source fragments are excluded from the results, since one could easily interpret the in-source fragments as peptides formed by a-specific protease activity.

Quantification of peptides

After identification of the peptides, it would be interesting to determine the concentration of the peptides, for food hydrolysates and proteomics, and use these to calculate protein concentrations (quantitative proteomics). To quantify proteins, several approaches exist (**Table 3**), which are elaborately reviewed [55-58]. Most quantification approaches use the intensity of the MS signal(s) of the peptide(s) for quantification. This is already complex since one peptide gives rise to many MS ions. Besides that, the intensity of an ion depends on the ionisation efficiency of the peptide and the type of mass spectrometer with settings applied (for ionisation, the quadrupole and detector). When analysing the same hydrolysate under constant analysis conditions, ion intensities can still be affected by ion-suppression, matrix effects, and day-to-day variance [59-61]. Typically in quantitative proteomics, MS signals are used to compare samples analysed in the same run (relative quantification, low accuracy) or the MS signals are compared with intensities of (isotopically labelled) reference peptides with a known concentration (absolute quantification, high accuracy). To quantify peptides in food hydrolysates, an alternative approach is needed since relative quantification is not accurate for individual peptides (≥ 30 % RSD [57]) and it is not feasible to synthesise a reference for each peptide in the hydrolysate. A good alternative to quantify all peptides in food hydrolysates

would be to use the UV absorbance of peptides. The benefit of UV-based quantification is that it is absolute and does not require labelling. The UV data can be acquired by coupling of a PDA-detector between LC and MS. The UV peak areas can be converted to an absolute concentration based on the law of Lambert-Beer, which requires the molar extinction coefficient of a peptide [4, 62]. Kuipers *et al.* determined the contributions of all amino acids and the peptide bond to the molar extinction at 214 nm of a peptide [63]. Based on these contributions, the molar extinction coefficient of any peptide sequence can be predicted based on its amino acid sequence. The mobile phase composition e.g. water or acetonitrile, did not affect the extinction coefficient of peptides [63]. The difference between the predicted and experimentally determined extinction coefficient was on average 12 %, calculated from data of 9 synthesised peptides by Kuipers *et al*. This makes UV-quantification more suitable for quantification of individual peptides than relative quantification approaches that use MS-intensity. UV-based peptide quantification has already been successfully applied in combination with size-exclusion chromatography [64, 65], ion-exchange chromatography [66] and with LC-MS by numerous research groups [67-72].

Table 1.3. Approaches for peptide and protein quantification.

Method	**Type of quantification**	**Approach**	**Label-free?**	**Examples**
Spectral counting	Relative quantification	Compare total peptide MS intensities between different samples in the same LC-MS run	Yes	PAI [73], emPAI [74], mSCI [75]
Chemical labelling	Relative quantification	Compare ion intensities in MS spectra for two samples mixed together, of which one is isotopically labelled. Isotopic labelling is performed after cell lysis	No	ICAT [76], iTRAQ [77], TMT [78]
Metabolic labelling	Relative quantification	Similar to previous but isotopic labelling is performed when growing the cell cultures.	No	SILAC [79]
Synthesised isotopically labelled reference peptide	Absolute quantification	Comparing MS intensity of a peptide in the digest with a (synthesised) isotopically labelled reference peptide	No	AQUA [80]
Absolute label free quantification	Relative peptide quantification but absolute protein quantification	Compare total MS intensity of peptide signals belonging to a protein with a calibration curve made by MS intensities of hydrolysates of different proteins at different known concentrations	Yes	IBAQ [81]
UV-based quantification	Absolute quantification	Convert UV absorbance at 214 nm to peptide concentrations based on predicted molar extinction coefficients	Yes	[63]

CONCEPTS OF PROTEIN HYDROLYSIS

In the digestive tract, proteases are present that cleave proteins into peptides and free amino acids, which are eventually small enough to cross the intestinal membrane and can be used as nutrients for the body. In the mouth, there is typically no hydrolysis of proteins, although the α-amylase can contribute to digestion by hydrolysing starch and thereby improving the accessibility to proteases later in the digestive tract [82]. In the stomach, pepsin hydrolyses proteins under acidic conditions [83]. In the small intestine, a mixture of proteases is released from the pancreas, named pancreatin, of which trypsin and chymotrypsin are the main endo-proteases [84]. Other (minor) proteases present are elastase [85, 86], kallikrein [87] and carboxypeptidases [84, 88, 89]. At last, the brush-border cells contain proteases [90, 91]. To describe and understand the activity of all these proteases, several concepts have been used. In this section, the most relevant concepts and parameters will be explained.

Degree of hydrolysis

A simple and widely used parameter to describe protein digestion is the degree of hydrolysis, which is the amount of amide bonds broken relative to the total amount of amide bonds. The degree of hydrolysis can be measured off-line using spectrophotometric assays as trinitrobenzenesulfonic acid (TNBS) or *o*-phthalaldehyde (OPA) [92, 93]. Alternatively, the degree of hydrolysis can be measured on-line with pH-stat. This titration device fixates the pH by titrating acid or base, to compensate for the pH change induced by breaking peptide bonds. At last, the degree of hydrolysis can be calculated from identified peptides and their concentrations, determined with LC-MS. The degree of hydrolysis is often used to compare the digestibility of different protein sources or evaluate different processing conditions or hydrolysis conditions [94-96].

The Linderstrøm-Lang theory

The second concept to describe protease activity is the affinity to intact protein, which affects the hydrolysis scenario. According to the Linderstrøm-Lang theory, protein hydrolysis depends on the enzyme's affinity for the initial cleavage site of the protein and the affinity to hydrolyse peptides, which results in two theoretical scenario's namely "one-by-one" or a "zipper" [97]. When the rate of protein denaturation is slow relative to the hydrolysis rate of intermediate peptides to small peptides (one-by-one scenario), the protease will first break down one protein into small peptides, before cleaving the next protein. This will lead to the presence of small peptides at relatively low degree of hydrolysis (**Figure 1.2**). In the zipper scenario, hydrolysis of the intact protein into intermediate peptides is much faster than the hydrolysis of intermediate peptides into small peptides. Determination of the hydrolysis scenario requires the measurement of the intact protein concentration and the degree of hydrolysis, both at several time points during hydrolysis. The hydrolysis scenario followed depends on the protease, the substrate (structure) [98] and conditions as pH [99] and protein concentration [100]. The affinity to intact protein influences the peptides and their concentrations during hydrolysis.

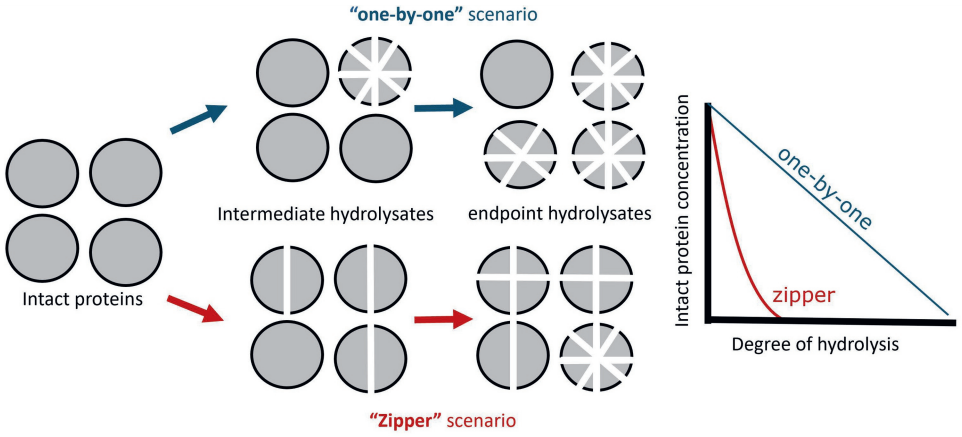


Figure 1.2. Schematic illustration of protein hydrolysis according to the Linderstrøm-Lang theory.

Hydrolysis mechanism, specificity and subsite model

Based on the amino acid residues facilitating the catalytic reaction of hydrolysis, proteases are categorised in mechanistic classes (**Table 4**). Bio-chemical studies unraveled the catalytic mechanism using synthetic substrates and confirmed these findings later by making X-ray crystal structures of protease-inhibitor complexes. To describe the interactions between enzyme (subsite) and substrate (binding site), Schechter and Berger introduced the subsite model [101] (**Figure 1.3**).

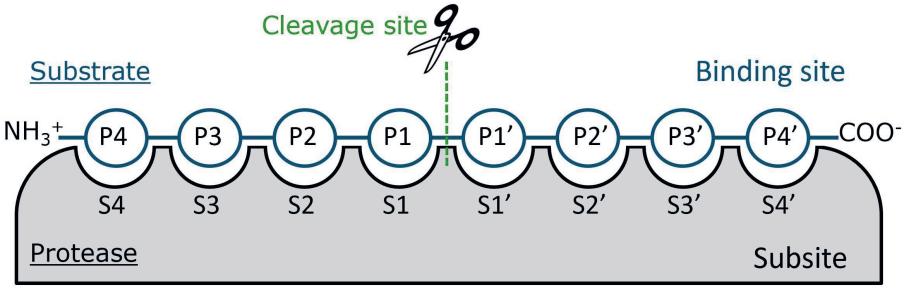


Figure 1.3. Illustration of the subsite model of Schechter and Berger [101], in which amino acid positions in the protein substrate (P4-P4') interact with the subsites on the protease (S4-S4').

In this model, the protease splits peptide bonds between the amino acid residue yielding the carboxyl group (P1), on the N-terminal protein side and the amino acid residue yielding the amino group (P1'), on the C-terminal protein side. The other binding site positions oriented to the protein N-terminus are named P2, P3, P4 etc., whereas the binding site positions oriented to the protein C-terminus are named P2', P3' and P4'. For proteases, the amide bonds that can be hydrolysed (specificity) depends on the amino acid interaction with the S1 specificity pocket [102]. Specificity is binary: a protease is able to hydrolyse a bond with a certain amino acid in P1 or not. Trypsin is specific for positively charged amino acid residues as lysine and arginine in

P1, since these are attracted by the negatively charged S1 specificity pocket. For chymotrypsin, the S1 specificity pocket is relatively deep and hydrophobic, resulting in a good fit for hydrophobic amino acids [103]. Some proteases, as for instance pepsin, facilitate hydrolysis with all amino acids in the P1 position and are called a-specific proteases.

Table 1.4. Proteases used in this thesis and their mechanism.

Mechanistic class	**Proteases**	**Source**	**Residues involved in catalytic mechanism**	**Typically reported specificity**	**Reference mechanism**
Serine proteases	Trypsin	Bovine	Ser195, His57, Asp102 (catalytic triad)	basic residues	[104]
	Chymotrypsin	Bovine	Ser195, His57, Asp 102, (catalytic triad)	aromatic residues or a-specific	[103]
	Bacillus licheniformis protease (BLP)	*E.coli*	Ser167, His47, Asp96, (catalytic triad)	acidic residues	[105]
Aspartic acid proteases	Pepsin	Porcine	Asp32, Asp215 (catalytic diad)	a-specific	[106]

Besides the interaction of the substrate with the S1 position, also neighbouring amino acids can interact with the subsite and promote or hinder hydrolysis. For bovine trypsin, it was observed that charged residues in the P2 or P2' position hinder hydrolysis, as well as a proline in P1' [33]. When amino acids in binding site positions other than P1 affect hydrolysis, the protease is said to also display a secondary specificity.

Maximum degree of hydrolysis and preference

The theoretical maximum degree of hydrolysis can be calculated with the specificity of a protease and the amino acid sequence of the substrate. In practice, this theoretical maximum is often not reached experimentally. This means that not all peptide bonds within the specificity are hydrolysed by the enzyme [107]. In some cases, these so-called "missed cleavages" could be caused by the enzyme's secondary specificity [108]. In other cases, the reason is not known (yet). For proteases that do not have a clearly defined specificity (a-specific proteases), the majority of bonds are not hydrolysed either [109, 110]. It was observed that bonds with some amino acids in the P1 position are hydrolysed more often than others. To describe the amino acids that are preferred, often the term "preference" is used [111-113]. In most cases, the preference is determined by analysis of peptides formed after (complete) hydrolysis. The peptide sequences are used to count how often the protease cleaved a peptide bond with a certain type of amino acid in the P1 position. Ideally, the preference should be corrected for the frequency of occurrence of the amino acids in the substrate sequence. When the concentrations of the peptides are determined, a preference value can be calculated,

which describes how much product is released with an amino acid in the P1 position relative to the expected product release when all amino acids would be similarly preferred.

Selectivity

Despite having a similar amino acid in the P1 position, peptide bonds can be hydrolysed at different rates. To describe hydrolysis kinetics at the level of individual peptide bonds, Butré *et al.* introduced "enzyme selectivity" [48]. Selectivity is a quantitative parameter to describe the hydrolysis rate of an individual cleavage site relative to the total hydrolysis rate. Determination of selectivity requires complete information on the peptides and their concentrations at several stages of hydrolysis (**Figure 1.4**).

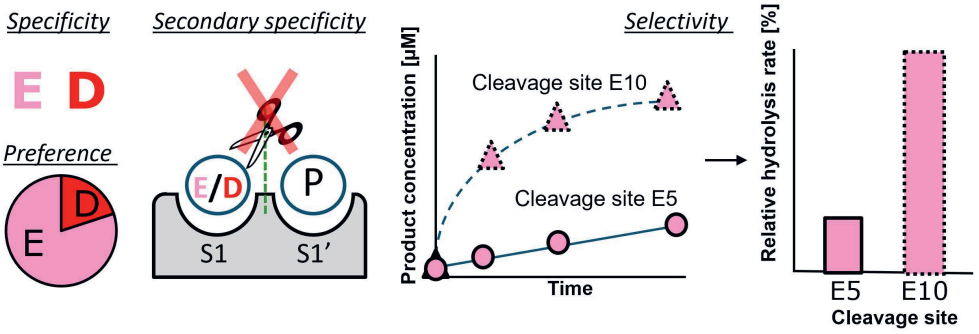


Figure 1.4. Illustration of the concepts "Specificity", "Preference", "Secondary specificity" and "Selectivity" and applied to *Bacillus licheniformis* protease. BLP has a specificity towards D and E, with a preference for E over D. BLP does not hydrolyse cleavage sites with a proline in the P1' position. Cleavage sites "E5" and "E10" are hydrolysed with different hydrolysis rates.

The introduction of selectivity brought a completely new level of detail in describing the hydrolysis process. The calculated selectivity values capture a lot of peptide data in a few quantitative parameters, which make it feasible to interpret and describe quantitative peptide release for different proteases, substrates and conditions. The selectivity of a protease-substrate combination is caused by the protease affinity to cleavage sites and the stochastic chance of hydrolysis, based on the distribution of cleavage sites over the protein [114]. Prior to this PhD project, selectivity had been used to describe the effect of pH and substrate concentration on hydrolysis by BLP [99, 115], to predict hydrolysis for bovine trypsin [33] and to compare hydrolysis by bovine, porcine and human trypsins [32].

Aim and outline

The aim of this thesis is to develop a methodology (approach) to identify and quantify the complete peptide composition in protein hydrolysates and to investigate its potential for a variety of applications. The automated peptide identification method should be able to process LC-MS data of complex hydrolysates in similar time as required for manual analysis of simple hydrolysates. In **Chapter 2** the current in-house peptide identification method [48] is automated and applied to mixtures of individual protein hydrolysates. The reproducibility, completeness and effect of coelution were described for single and mixed protein hydrolysates. In **Chapter 3**, the method is challenged by increasing substrate complexity. Hydrolysates of yellow pea extracts were analysed and used to calculate the protein genetic variant composition. First, purified fractions of legumins, vicilins and albumins were characterised to evaluate the performance of the method and to describe the challenge of converting peptide to protein concentrations. Later, the protein genetic variants were quantified for extracts of 8 different yellow pea cultivars. In **Chapter 4**, the method is applied to a broadly specific protease, chymotrypsin, to describe the path of hydrolysis. By analysis of peptide formation and degradation, the hydrolysis rates for individual cleavage sites were determined (selectivity). These hydrolysis rates were used to correlate the amino acids occupying the P4-P4' binding site positions to the preference and secondary specificity of chymotrypsin. In **Chapter 5**, the method is applied to study peptide release kinetics by an a-specific protease, porcine pepsin. The differences in relative hydrolysis rates were used to understand the effect of pH on pepsin activity. In the general discussion (**Chapter 6**), the reach of the method, limitations and future extensions to analyse even more complex hydrolysates are discussed, as well as the insights obtained on hydrolysis by chymotrypsin and pepsin.

References

1. Abrahamse, E., Thomassen, G. G. M., Renes, I. B., Wierenga, P. A., Hettinga, K. A. (**2022**). Gastrointestinal protein hydrolysis kinetics: Opportunities for further infant formula improvement. *Nutrients*, 14.
2. Dupont, D., Mandalari, G., Molle, D., Jardin, J., Léonil, J., Faulks, R. M., Wickham, M. S. J., Mills, E. N. C., Mackie, A. R. (**2010**). Comparative resistance of food proteins to adult and infant *in vitro* digestion models. *Molecular Nutrition and Food Research*, 54, 767-780.
3. De Laureto, P. P., Scaramella, E., Frigo, M., Wondrich, F. G., De Filippis, V., Zambonin, M., Fontana, A. (**1999**). Limited proteolysis of bovine α-lactalbumin: Isolation and characterization of protein domains. *Protein Science*, 8, 2290-2303.
4. Muñoz-Tamayo, R., de Groot, J., Bakx, E., Wierenga, P. A., Gruppen, H., Zwietering, M. H., Sijtsma, L. (**2011**). Hydrolysis of β-casein by the cell-envelope-located PI-type protease of *Lactococcus lactis*: A modelling approach. *International Dairy Journal*, 21, 755-762.
5. Leduc, A., Fournier, V., Henry, J. (**2020**). A standardized, innovative method to characterize the structure of aquatic protein hydrolysates. *Heliyon*, 6, e04170.
6. Cui, Q., Duan, Y., Zhang, M., Liang, S., Sun, Y., Cheng, J., Guo, M. (**2022**). Peptide profiles and antioxidant capacity of extensive hydrolysates of milk protein concentrate. *Journal of Dairy Science*.
7. Bak, K. H., Waehrens, S. S., Fu, Y., Chow, C. Y., Petersen, M. A., Ruiz-Carrascal, J., Bredie, W. L. P., Lametsch, R. (**2021**). Flavor characterization of animal hydrolysates and potential of glucosamine in flavor modulation. *Foods*, 10.
8. Wilm, M. (**2011**). Principles of Electrospray Ionization. *Molecular & Cellular Proteomics*, 10.
9. Sun, S., Yu, C., Qiao, Y., Lin, Y., Dong, G., Liu, C., Zhang, J., Zhang, Z., Cai, J., Zhang, H., Bu, D. (**2008**). Deriving the probabilities of water loss and ammonia loss for amino acids from tandem mass spectra. *Journal of Proteome Research*, 7, 202-208.
10. Neta, P., Pu, Q.-L., Kilpatrick, L., Yang, X., Stein, S. E. (**2007**). Dehydration versus deamination of N-terminal glutamine in collision-induced dissociation of protonated peptides. *Journal of the American Society for Mass Spectrometry*, 18, 27-36.
11. Kim, J.-S., Monroe, M. E., Camp, D. G., Smith, R. D., Qian, W.-J. (**2013**). In-Source fragmentation and the sources of partially tryptic peptides in shotgun proteomics. *Journal of Proteome Research*, 12, 910-916.
12. Picotti, P., Aebersold, R., Domon, B. (**2007**). The implications of proteolytic background for shotgun proteomics. *Molecular & Cellular Proteomics*, 6, 1589-1598.
13. Brodbelt, J. S. (**2016**). Ion activation methods for peptides and proteins. *Analytical Chemistry*, 88, 30-51.
14. Plumb, R. S., Johnson, K. A., Rainville, P., Smith, B. W., Wilson, I. D., Castro-Perez, J. M., Nicholson, J. K. (**2006**). UPLC/MSE; a new approach for generating molecular fragment information for biomarker structure elucidation. *Rapid Communications in Mass Spectrometry*, 20, 1989-1994.
15. Bond, N. J., Shliaha, P. V., Lilley, K. S., Gatto, L. (**2013**). Improving qualitative and quantitative performance for MSE-based label-free proteomics. *Journal of Proteome Research*, 12, 2340-2353.
16. Holder, A., Thienel, K., Klaiber, I., Pfannstiel, J., Weiss, J., Hinrichs, J. (**2014**). Quantification of bio- and techno-functional peptides in tryptic bovine micellar casein and β-casein hydrolysates. *Food Chemistry*, 158, 118-124.
17. Rutella, G. S., Solieri, L., Martini, S., Tagliazucchi, D. (**2016**). Release of the antihypertensive tripeptides valine-proline-proline and isoleucine-proline-proline from bovine milk caseins during in vitro gastrointestinal digestion. *Journal of Agricultural and Food Chemistry*, 64, 8509-8515.
18. Caira, S., Pinto, G., Nicolai, M. A., Chianese, L., Addeo, F. (**2016**). Simultaneously tracing the geographical origin and presence of bovine milk in Italian water buffalo Mozzarella cheese using MALDI-TOF data of casein signature peptides. *Analytical and Bioanalytical Chemistry*, 408, 5609-5621.

19. Leni, G., Prandi, B., Varani, M., Faccini, A., Caligiani, A., Sforza, S. (**2020**). Peptide fingerprinting of *Hermetia illucens* and *Alphitobius diaperinus*: Identification of insect species-specific marker peptides for authentication in food and feed. *Food Chemistry*, 320, 126681.
20. Pan, Z., Fan, L., Zhong, Y., Guo, J., Dong, X., Xu, X., Wang, C., Su, Y. (**2023**). Quantitative proteomics reveals reduction in central carbon and energy metabolisms contributes to gentamicin resistance in *Staphylococcus aureus*. *Journal of Proteomics*, 277.
21. Putz, E. J., Fernandes, L. G. V., Sivasankaran, S. K., Bayles, D. O., Alt, D. P., Lippolis, J. D., Nally, J. E. (**2022**). Some like it hot, some like it cold; proteome comparison of *Leptospira borgpetersenii* serovar Hardjo strains propagated at different temperatures. *Journal of Proteomics*, 262.
22. Abdirad, S., Wu, Y., Ghorbanzadeh, Z., Tazangi, S. E., Amirkhani, A., Fitzhenry, M. J., Kazemi, M., Ghaffari, M. R., Koobaz, P., Zeinalabedini, M., Habibpourmehraban, F., Masoomi-Aladizgeh, F., Atwell, B. J., Mirzaei, M., Salekdeh, G. H., Haynes, P. A. (**2022**). Proteomic analysis of the meristematic root zone in contrasting genotypes reveals new insights in drought tolerance in rice. *Proteomics*, 22.
23. Kushner, I. K., Clair, G., Purvine, S. O., Lee, J.-Y., Adkins, J. N., Payne, S. H. (**2018**). Individual variability of protein expression in human tissues. *Journal of Proteome Research*, 17, 3914-3922.
24. Nissa, M. U., Pinto, N., Mukherjee, A., Reddy, P. J., Ghosh, B., Sun, Z., Ghantasala, S., Chetanya, C., Shenoy, S. V., Moritz, R. L., Goswami, M., Srivastava, S. (**2022**). Organ-based proteome and post-translational modification profiling of a widely cultivated tropical water fish, *Labeo rohita*. *Journal of Proteome Research*, 21, 420-437.
25. Duran, K., Magnin, J., America, A. H., Peng, M., Hilgers, R., de Vries, R. P., Baars, J. J., van Berkel, W. J., Kuyper, T. W., Kabel, M. A. (**2023**). The secretome of *Agaricus bisporus*: Temporal dynamics of plant polysaccharides and lignin degradation. *iScience*, 26.
26. Gazi, I., Reiding, K. R., Groeneveld, A., Bastiaans, J., Huppertz, T., Heck, A. J. R. (**2023**). Key changes in bovine milk immunoglobulin G during lactation: NeuAc sialylation is a hallmark of colostrum immunoglobulin G N-glycosylation. *Glycobiology*, 33, 115-125.
27. Santos, W. S., Montoni, F., Eichler, R. A. S., Arcos, S. S. S., Andreotti, D. Z., Kisaki, C. Y., Evangelista, K. B., Calacina, H. M., Lima, I. F., Soares, M. A. M., Gren, E. C. K., Carvalho, V. M., Ferro, E. S., Nishiyama-Jr, M. Y., Chen, Z., Iwai, L. K. (**2022**). Proteomic analysis reveals rattlesnake venom modulation of proteins associated with cardiac tissue damage in mouse hearts. *Journal of Proteomics*, 258.
28. Berg, M., Parbel, A., Pettersen, H., Fenyö, D., Björkesten, L. (**2006**). Reproducibility of LC-MS-based protein identification. *Journal of experimental botany*, 57, 1509-1514.
29. Delmotte, N., Lasaosa, M., Tholey, A., Heinzle, E., van Dorsselaer, A., Huber, C. G. (**2009**). Repeatability of peptide identifications in shotgun proteome analysis employing off-line two-dimensional chromatographic separations and ion-trap MS. *Journal of separation science*, 32, 1156-1164.
30. Chen, Y., Kwon, S. W., Kim, S. C., Zhao, Y. (**2005**). Integrated approach for manual evaluation of peptides identified by searching protein sequence databases with tandem mass spectra. *Journal of Proteome Research*, 4, 998-1005.
31. Chen, Y., Zhang, J., Xing, G., Zhao, Y. (**2009**). Mascot-derived false positive peptide identifications revealed by manual analysis of tandem mass spectra. *Journal of Proteome Research*, 8, 3141-3147.
32. Deng, Y., Gruppen, H., Wierenga, P. A. (**2018**). Comparison of protein hydrolysis catalyzed by bovine, porcine, and human trypsins. *Journal of Agricultural and Food Chemistry*, 66, 4219-4232.
33. Deng, Y., van der Veer, F., Sforza, S., Gruppen, H., Wierenga, P. A. (**2018**). Towards predicting protein hydrolysis by bovine trypsin. *Process Biochemistry*, 65, 81-92.
34. Sánchez-Rivera, L., Ménard, O., Recio, I., Dupont, D. (**2015**). Peptide mapping during dynamic gastric digestion of heated and unheated skimmed milk powder. *Food Research International*, 77, 132-139.
35. Sousa, R., Portmann, R., Dubois, S., Recio, I., Egger, L. (**2020**). Protein digestion of different protein sources using the INFOGEST static digestion model. *Food Research International*, 130.

36. Accardo, F., Leni, G., Tedeschi, T., Prandi, B., Sforza, S. (**2022**). Structural and chemical changes induced by temperature and pH hinder the digestibility of whey proteins. *Food Chemistry*, 387, 132884.
37. Accardo, F., Prandi, B., Terenziani, F., Tedeschi, T., Sforza, S. (**2023**). Evaluation of *in vitro* whey protein digestibility in a protein-catechins model system mimicking milk chocolate: Interaction with flavonoids does not hinder protein bioaccessibility. *Food Research International*, 169, 112888.
38. Egger, L., Ménard, O., Baumann, C., Duerr, D., Schlegel, P., Stoll, P., Vergères, G., Dupont, D., Portmann, R. (**2019**). Digestion of milk proteins: Comparing static and dynamic *in vitro* digestion systems with *in vivo* data. *Food Research International*, 118, 32-39.
39. Sayd, T., Dufour, C., Chambon, C., Buffière, C., Remond, D., Santé-Lhoutellier, V. (**2018**). Combined *in vivo* and *in silico* approaches for predicting the release of bioactive peptides from meat digestion. *Food Chemistry*, 249, 111-118.
40. Wongngam, W., Hamzeh, A., Tian, F., Roytrakul, S., Yongsawatdigul, J. (**2023**). Purification and molecular docking of angiotensin converting enzyme-inhibitory peptides derived from corn gluten meal hydrolysate and from in silico gastrointestinal digestion. *Process Biochemistry*.
41. Baba, W. N., Baby, B., Mudgil, P., Gan, C.-Y., Vijayan, R., Maqsood, S. (**2021**). Pepsin generated camel whey protein hydrolysates with potential antihypertensive properties: Identification and molecular docking of antihypertensive peptides. *LWT*, 143, 111135.
42. Ryan, J. T., Ross, R. P., Bolton, D., Fitzgerald, G. F., Stanton, C. (**2011**). Bioactive peptides from muscle sources: meat and fish. *Nutrients*, 3, 765-791.
43. Zamora-Sillero, J., Gharsallaoui, A., Prentice, C. (**2018**). Peptides from fish by-product protein hydrolysates and its functional properties: An overview. *Marine Biotechnology*, 20, 118-130.
44. Ryan, G., O'Regan, J., FitzGerald, R. J. (**2023**). Foaming and sensory properties of bovine milk protein isolate and its associated enzymatic hydrolysates. *International Dairy Journal*, 137, 105511.
45. García Arteaga, V., Apéstegui Guardia, M., Muranyi, I., Eisner, P., Schweiggert-Weisz, U. (**2020**). Effect of enzymatic hydrolysis on molecular weight distribution, techno-functional properties and sensory perception of pea protein isolates. *Innovative Food Science and Emerging Technologies*, 65.
46. da Silva Bambirra Alves, F. E., Carpiné, D., Teixeira, G. L., Goedert, A. C., de Paula Scheer, A., Ribani, R. H. (**2021**). Valorization of an abundant slaughterhouse by-product as a source of highly technofunctional and antioxidant protein hydrolysates. *Waste and Biomass Valorization*, 12, 263-279.
47. Chin, Y. L., Chai, K. F., Chen, W. N. (**2022**). Upcycling of brewers' spent grains via solid-state fermentation for the production of protein hydrolysates with antioxidant and techno-functional properties. *Food Chemistry: X*, 13, 100184.
48. Butré, C. I., Sforza, S., Gruppen, H., Wierenga, P. A. (**2014**). Introducing enzyme selectivity: A quantitative parameter to describe enzymatic protein hydrolysis. *Analytical and Bioanalytical Chemistry*, 406, 5827-5841.
49. Tomczyk, N., Giles, K., Richardson, K., Ujma, J., Palmer, M., Nielsen, P. K., Haselmann, K. F. (**2021**). Mapping isomeric peptides derived from biopharmaceuticals using high-resolution ion mobility mass spectrometry. *Analytical Chemistry*, 93, 16379-16384.
50. Kriisa, M., Taivosalo, A., Föste, M., Kütt, M.-L., Viirma, M., Priidik, R., Korzeniowska, M., Tian, Y., Laaksonen, O., Yang, B., Vilu, R. (**2022**). Effect of enzyme-assisted hydrolysis on brewer's spent grain protein solubilization - peptide composition and sensory properties. *Applied Food Research*, 2, 100108.
51. Arju, G., Berg, H. Y., Lints, T., Nisamedtinov, I. (**2022**). Methodology for analysis of peptide consumption by yeast during fermentation of enzymatic protein hydrolysate supplemented synthetic medium using UPLC-IMS-HRMS. *Fermentation*, 8, 145.
52. Xia, X., Arju, G., Taivosalo, A., Lints, T., Kriščiunaite, T., Vilu, R., Corrigan, B. M., Gai, N., Fenelon, M. A., Tobin, J. T., Kilcawley, K., Kelly, A. L., McSweeney, P. L. H., Sheehan, J. J. (**2023**). Effect of β-casein reduction and high heat treatment of micellar casein concentrate on proteolysis, texture and the

volatile profile of resultant Emmental cheese during ripening. *International Dairy Journal*, 138, 105540.

53. Tedeschi, T., Aspri, M., Loffi, C., Dellafiora, L., Galaverna, G., Papademas, P. (**2023**). Processing of raw donkey milk by pasteurisation and UV-C to produce freeze-dried milk powders: The effect on protein quality, digestibility and bioactive properties. *LWT*, 173, 114404.
54. Cardoso, H. B. (**2023**). Maillard reaction of milk proteins. Wageningen University.
55. Millán-Oropeza, A., Blein-Nicolas, M., Monnet, V., Zivy, M., Henry, C. (**2022**). Comparison of different label-free techniques for the semi-absolute quantification of protein abundance. *Proteomes*, 10.
56. Colaert, N., Vandekerckhove, J., Gevaert, K., Martens, L. (**2011**). A comparison of MS2-based label-free quantitative proteomic techniques with regards to accuracy and precision. *Proteomics*, 11, 1110-1113.
57. Schulze, W. X., Usadel, B. (**2010**). Quantitation in mass-spectrometry-based proteomics. *Annual review of plant biology*, 61, 491-516.
58. Ankney, J. A., Muneer, A., Chen, X. (**2016**). Relative and absolute quantitation in mass spectrometry–based proteomics. *Annual Review of Analytical Chemistry*, 11, 49-77.
59. Annesley, T. M. (**2003**). Ion suppression in mass spectrometry. *Clinical Chemistry*, 49, 1041-1044.
60. Collins, B. C., Hunter, C. L., Liu, Y., Schilling, B., Rosenberger, G., Bader, S. L., Chan, D. W., Gibson, B. W., Gingras, A.-C., Held, J. M., Hirayama-Kurogi, M., Hou, G., Krisp, C., Larsen, B., Lin, L., Liu, S., Molloy, M. P., Moritz, R. L., Ohtsuki, S., Schlapbach, R., Selevsek, N., Thomas, S. N., Tzeng, S.-C., Zhang, H., Aebersold, R. (**2017**). Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry. *Nature Communications*, 8, 291.
61. Bantscheff, M., Schirle, M., Sweetman, G., Rick, J., Kuster, B. (**2007**). Quantitative mass spectrometry in proteomics: a critical review. *Analytical and Bioanalytical Chemistry*, 389, 1017-1031.
62. Kosters, H. A., Wierenga, P. A., de Vries, R., Gruppen, H. (**2011**). Characteristics and effects of specific peptides on heat-induced aggregation of β-lactoglobulin. *Biomacromolecules*, 12, 2159-2170.
63. Kuipers, B. J. H., Gruppen, H. (**2007**). Prediction of molar extinction coefficients of proteins and peptides using UV absorption of the constituent amino acids at 214 nm to enable quantitative reverse phase high-performance liquid chromatography-mass spectrometry analysis. *Journal of Agricultural and Food Chemistry*, 55, 5445-5451.
64. Bodin, A., Framboisier, X., Alonso, D., Marc, I., Kapel, R. (**2015**). Size-exclusion HPLC as a sensitive and calibrationless method for complex peptide mixtures quantification. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences*, 1006, 71-79.
65. Beaubier, S., Pineda-Vadillo, C., Mesieres, O., Framboisier, X., Galet, O., Kapel, R. (**2023**). Improving the *in vitro* digestibility of rapeseed albumins resistant to gastrointestinal proteolysis while preserving the functional properties using enzymatic hydrolysis. *Food Chemistry*, 407, 135132.
66. Stanisavljević, N. S., Vukotić, G. N., Pastor, F. T., Sužnjević, D., Jovanović, Z. S., Strahinić, I. D., Fira, D. A., Radović, S. S. (**2015**). Antioxidant activity of pea protein hydrolysates produced by batch fermentation with lactic acid bacteria. *Archives of Biological Sciences*, 67, 1033-1042.
67. Møller, K. K., Rattray, F. P., Ardö, Y. (**2012**). Camel and bovine chymosin hydrolysis of bovine αS1- and β-caseins studied by comparative peptide mapping. *Journal of Agricultural and Food Chemistry*, 60, 11421-11432.
68. Fernández, A., Riera, F. (**2013**). β-Lactoglobulin tryptic digestion: A model approach for peptide release. *Biochemical Engineering Journal*, 70, 88-96.
69. Prandi, B., Faccini, A., Tedeschi, T., Cammerata, A., Sgrulletta, D., D'Egidio, M. G., Galaverna, G., Sforza, S. (**2014**). Qualitative and quantitative determination of peptides related to celiac disease in mixtures derived from different methods of simulated gastrointestinal digestion of wheat products. *Analytical and Bioanalytical Chemistry*, 406, 4765-4775.

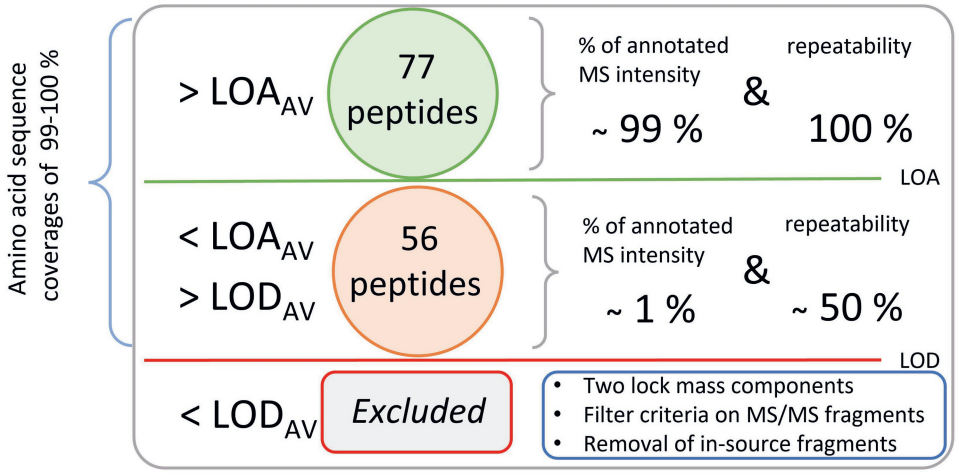
70. Rauh, V. M., Johansen, L. B., Ipsen, R., Paulsson, M., Larsen, L. B., Hammershøj, M. (**2014**). Plasmin activity in UHT milk: Relationship between proteolysis, age gelation, and bitterness. *Journal of Agricultural and Food Chemistry*, 62, 6852-6860.
71. Mao, Y., Krischke, M., Hengst, C., Kulozik, U. (**2018**). Comparison of the influence of pH on the selectivity of free and immobilized trypsin for β-lactoglobulin hydrolysis. *Food Chemistry*, 253, 194-202.
72. Rončević, T., Čikeš-Čulić, V., Maravić, A., Capanni, F., Gerdol, M., Pacor, S., Tossi, A., Giulianini, P. G., Pallavicini, A., Manfrin, C. (**2020**). Identification and functional characterization of the astacidin family of proline-rich host defence peptides (PcAst) from the red swamp crayfish (*Procambarus clarkii*, Girard 1852). *Developmental & Comparative Immunology*, 105, 103574.
73. Liu, H., Sadygov, R. G., Yates, J. R. (**2004**). A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Analytical Chemistry*, 76, 4193-4201.
74. Ishihama, Y., Oda, Y., Tabata, T., Sato, T., Nagasu, T., Rappsilber, J., Mann, M. (**2005**). Exponentially Modified Protein Abundance Index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Molecular & Cellular Proteomics*, 4, 1265-1272.
75. Sun, A., Zhang, J., Wang, C., Yang, D., Wei, H., Zhu, Y., Jiang, Y., He, F. (**2009**). Modified spectral count index (mSCI) for estimation of protein abundance by protein relative identification possibility (RIPpro): A new proteomic technological parameter. *Journal of Proteome Research*, 8, 4934-4942.
76. Sethuraman, M., McComb, M. E., Huang, H., Huang, S., Heibeck, T., Costello, C. E., Cohen, R. A. (**2004**). Isotope-Coded Affinity Tag (ICAT) approach to redox proteomics: Identification and quantitation of oxidant-Sensitive cysteine thiols in complex protein mixtures. *Journal of Proteome Research*, 3, 1228-1233.
77. Evans, C., Noirel, J., Ow, S. Y., Salim, M., Pereira-Medrano, A. G., Couto, N., Pandhal, J., Smith, D., Pham, T. K., Karunakaran, E., Zou, X., Biggs, C. A., Wright, P. C. (**2012**). An insight into iTRAQ: where do we stand now? *Analytical and Bioanalytical Chemistry*, 404, 1011-1027.
78. Dayon, L., Hainard, A., Licker, V., Turck, N., Kuhn, K., Hochstrasser, D. F., Burkhard, P. R., Sanchez, J. C. (**2008**). Relative quantification of proteins in human cerebrospinal fluids by MS/MS using 6-plex isobaric tags. *Analytical Chemistry*, 80, 2921-2931.
79. Ong, S.-E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., Mann, M. (**2002**). Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Molecular & Cellular Proteomics*, 1, 376-386.
80. Gerber, S. A., Rush, J., Stemman, O., Kirschner, M. W., Gygi, S. P. (**2003**). Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proceedings of the National Academy of Sciences*, 100, 6940-6945.
81. Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., Selbach, M. (**2011**). Global quantification of mammalian gene expression control. *Nature*, 473, 337-342.
82. Duodu, K. G., Nunes, A., Delgadillo, I., Parker, M. L., Mills, E. N. C., Belton, P. S., Taylor, J. R. N. (**2002**). Effect of grain structure and cooking on sorghum and maize *in vitro* protein digestibility. *Journal of Cereal Science*, 35, 161-174.
83. Malagelada, J.-R., Longstreth, G. F., Summerskill, W. H. J., Go, V. L. W. (**1976**). Measurement of gastric functions during digestion of ordinary solid meals in man. *Gastroenterology*, 70, 203-210.
84. Whitcomb, D. C., Lowe, M. E. (**2007**). Human pancreatic digestive enzymes. *Digestive Diseases and Sciences*, 52, 1-17.
85. Largman, C., Brodrick, J. W., Geokas, M. C. (**1976**). Purification and characterization of two human pancreatic elastases. *Biochemistry*, 15, 2491-2500.
86. Boros, E., Szabó, A., Zboray, K., Héja, D., Pál, G., Sahin-Tóth, M. (**2017**). Overlapping specificity of duplicated human pancreatic elastase 3 isoforms and archetypal porcine elastase 1 provides clues to evolution of digestive enzymes. *Journal of Biological Chemistry*, 292, 2690-2702.

87. Rinderknecht, H. (**1993**). Pancreatic secretory enzymes. *The pancreas: biology, pathobiology, and disease*, 219-251.
88. Vendrell, J., Avilés, F. X. (**1999**). Carboxypeptidases. In: Turk V, editor. Proteases New Perspectives. Basel: Birkhäuser Basel, p. 13-34.
89. Stahmann, M. A., Fruton, J. S., Bergmann, M. (**1946**). The specificity of carboxypeptidase. *Journal of Biological Chemistry*, 164, 753-760.
90. Hooton, D., Lentle, R., Monro, J., Wickham, M., Simpson, R. (**2015**). The secretion and action of brush border enzymes in the mammalian small intestine. In: Nilius B, Gudermann T, Jahn R, Lill R, Petersen OH, de Tombe PP, editors. Reviews of Physiology, Biochemistry and Pharmacology. Cham: Springer International Publishing, p. 59-118.
91. Ozorio, L., Mellinger-Silva, C., Cabral, L. M. C., Jardin, J., Boudry, G., Dupont, D. (**2020**). The influence of peptidases in intestinal brush border membranes on the absorption of oligopeptides from whey protein hydrolysate: An ex vivo study using an Ussing chamber. *Foods*, 9.
92. Adler-Nissen, J. (**1979**). Determination of the degree of hydrolysis of food protein hydrolysates by trinitrobenzenesulfonic acid. *Journal of Agricultural and Food Chemistry*, 27, 1256-1262.
93. Church, F. C., Swaisgood, H. E., Porter, D. H., Catignani, G. L. (**1983**). Spectrophotometric assay using *o*-Phthaldialdehyde for determination of proteolysis in milk and isolated milk proteins. *Journal of Dairy Science*, 66, 1219-1227.
94. Tang, J., Wichers, H. J., Hettinga, K. A. (**2022**). Heat-induced unfolding facilitates plant protein digestibility during *in vitro* static infant digestion. *Food Chemistry*, 375, 131878.
95. Hall, A. E., Moraru, C. I. (**2021**). Effect of High Pressure Processing and heat treatment on in vitro digestibility and trypsin inhibitor activity in lentil and faba bean protein concentrates. *LWT*, 152, 112342.
96. Laguna, L., Picouet, P., Guàrdia, M. D., Renard, C. M. G. C., Sarkar, A. (**2017**). In vitro gastrointestinal digestion of pea protein isolate as a function of pH, food matrices, autoclaving, high-pressure and re-heat treatments. *LWT*, 84, 511-519.
97. Adler-Nissen, J. (**1976**). Enzymic hydrolysis of proteins for increased solubility. *Journal of Agricultural and Food Chemistry*, 24, 1090-1093.
98. Dubois, V., Nedjar-Arroume, N., Guillochon, D. (**2005**). Influence of pH on the appearance of active peptides in the course of peptic hydrolysis of bovine haemoglobin. *Preparative Biochemistry and Biotechnology*, 35, 85-102.
99. Butré, C. I., Sforza, S., Wierenga, P. A., Gruppen, H. (**2015**). Determination of the influence of the pH of hydrolysis on enzyme selectivity of *Bacillus licheniformis* protease towards whey protein isolate. *International Dairy Journal*, 44, 44-53.
100. Deng, Y., Butré, C. I., Wierenga, P. A. (**2018**). Influence of substrate concentration on the extent of protein enzymatic hydrolysis. *International Dairy Journal*, 86, 39-48.
101. Schechter, I., Berger, A. (**1967**). On the size of the active site in proteases. I. Papain. *Biochemical and Biophysical Research Communications*, 27, 157-162.
102. Hedstrom, L. (**2002**). Serine protease mechanism and specificity. *Chemical Reviews*, 102, 4501-4524.
103. Blow, D. M. (**1976**). Structure and mechanism of chymotrypsin. *Accounts of Chemical Research*, 9, 145-152.
104. Radisky, E. S., Lee, J. M., Lu, C.-J. K., Koshland, D. E. (**2006**). Insights into the serine protease mechanism from atomic resolution structures of trypsin reaction intermediates. *Proceedings of the National Academy of Sciences*, 103, 6835-6840.
105. Ye, W., Wang, H., Ma, Y., Luo, X., Zhang, W., Wang, J., Wang, X. (**2013**). Characterization of the glutamate-specific endopeptidase from *Bacillus licheniformis* expressed in *Escherichia coli*. *Journal of Biotechnology*, 168, 40-45.
106. Northrop, D. B. (**2001**). Follow the protons: A low-barrier hydrogen bond unifies the mechanisms of the aspartic proteases. *Accounts of Chemical Research*, 34, 790-797.

107. Siepen, J. A., Keevil, E.-J., Knight, D., Hubbard, S. J. (**2007**). Prediction of missed cleavage sites in tryptic peptides aids protein identification in proteomics. *Journal of Proteome Research*, 6, 399-408.
108. Gershon, P. D. (**2014**). Cleaved and missed sites for trypsin, Lys-C, and Lys-N can be predicted with high confidence on the basis of sequence context. *Journal of Proteome Research*, 13, 702-709.
109. Powers, J. C., Harley, A. D., Myers, D. V. (**1977**). Subsite specificity of porcine pepsin. In: Tang J, editor. Acid Proteases: Structure, Function, and Biology. New York, NY: Springer US, p. 141-157.
110. Hamuro, Y., Coales, S. J., Molnar, K. S., Tuske, S. J., Morrow, J. A. (**2008**). Specificity of immobilized porcine pepsin in H/D exchange compatible conditions. *Rapid Communications in Mass Spectrometry*, 22, 1041-1046.
111. Diamond, S. L. (**2007**). Methods for mapping protease specificity. *Current Opinion in Chemical Biology*, 11, 46-51.
112. Debela, M., Magdolen, V., Schechter, N., Valachova, M., Lottspeich, F., Craik, C. S., Choe, Y., Bode, W., Goettig, P. (**2006**). Specificity profiling of seven human tissue kallikreins reveals individual subsite preferences. *Journal of Biological Chemistry*, 281, 25678-25688.
113. Talanian, R. V., Quinlan, C., Trautz, S., Hackett, M. C., Mankovich, J. A., Banach, D., Ghayur, T., Brady, K. D., Wong, W. W. (**1997**). Substrate specificities of Caspase family proteases. *Journal of Biological Chemistry*, 272, 9677-9682.
114. Butré, C. I. (**2014**). Introducing enzyme selectivity as a quantitative parameter to describe the effects of substrate concentration on protein hydrolysis. Wageningen, the Netherlands. PhD thesis Wageningen University & Research.
115. Butré, C. I., Sforza, S., Gruppen, H., Wierenga, P. A. (**2014**). Determination of the influence of substrate concentration on enzyme selectivity using whey protein isolate and *Bacillus licheniformis* protease. *Journal of Agricultural and Food Chemistry*, 62, 10230-10239.

CHAPTER 2

A method to identify and quantify the complete peptide composition in protein hydrolysates



Vreeke, G. J. C., Lubbers, W., Vincken, J.-P., Wierenga, P. A. (**2022**). A method to identify and quantify the complete peptide composition in protein hydrolysates. *Analytica Chimica Acta*, *1201*, 339616. https://doi.org/10.1016/j.aca.2022.339616

Abstract

Automated approaches from proteomics are used to characterise peptides for food applications and in protein digests. Peptide annotations and confidence in these annotations are then based on the fragment spectra. Low reproducibility in repeat analyses has been reported even for annotations with high confidence. When analysing protein hydrolysates (in food) it is important to determine criteria that yield highly reproducible annotations. This study provides a structured approach to determine these criteria. Tryptic hydrolysates of α-lactalbumin, β-lactoglobulin and β-casein were analysed manually and automatically, using an UPLC-PDA-MS method for untargeted identification and absolute label-free quantification of peptides. A lock mass with two components was introduced resulting in an average mass error of 1 ppm. Processing filters were set to ensure reliable annotations based on MS/MS fragmentation, while maintaining maximum amount of information. Peptides in the individual hydrolysates with an MS intensity above the limit of annotation represented 99% of total MS intensity and were 100 % consistently annotated between four replicates. Amino acid and peptide sequence coverages for the individual protein hydrolysates were 99-100 % and 89-95 %, respectively. Mixing the hydrolysates resulted in a loss of 11% of the peptide annotations above the LOA and lower reproducibility (97 %) for the remaining annotations, as well as more co-eluting peptides. Calculated concentrations of co-eluting peptides in mixed hydrolysates varied 37 ±21 % from the value for single hydrolysates. The proposed approach allows complete description of peptide composition with highly repeatable annotations and quantification of peptides even in mixed hydrolysates.

Introduction

Identification of peptides present in enzymatic protein hydrolysates using UPLC-MS is essential in a large variety of scientific disciplines and industrial research [1]. In recent years, this is done often with data processing software originating from proteomics. In traditional proteomics, the goal is to identify the proteins that were in the original sample based on unique peptides identified. This does not necessarily require identification of all the peptides that originate from that protein. In contrast to proteomics, the goal for food applications and digestion studies, is to identify the formed peptides when the proteins in the original sample are known. In such studies, often the presence of peptides is compared in a set of different samples. It is important to know the level of confidence in the presence of individual peptides as reported after automated annotation. It is also important to know how many of the total hydrolysate is included in the analysis. Therefore, the completeness of the analysis should also be evaluated using different parameters than used in proteomics. For instance, a parameter should be used to identify if any peptides were lost during sample preparation (i.e. check for mass balance).

In this study, we test and optimise a method for automated identification and absolute label-free quantification of peptides using a non-targeted UPLC-PDA-MS approach. The aim was to propose a structured approach for data processing and reporting on completeness of peptide analysis.

Peptide Identification with mass spectrometry

To characterise the peptide composition in a hydrolysate, an untargeted approach is required. In this approach, all *m/z* signals detected in the mass spectra should be included in the analysis and then converted to a list of identified peptides. The steps in this process are to (1) separate the signals from the noise, and (2) to attribute peptide sequences to the signals. The dilemma in separating the signals from the noise is that with a high noise threshold peptides with a low intensity are not identified. With a low noise threshold non peptide related MS signals are also included in the analysis. In the process of attributing the included *m/z* signals to peptides, multiple challenges occur:

- One peptide in the sample can result in multiple *m/z* peaks in the spectrum: Peptides typically (i) occur in different charge states, (ii) can be present as adducts, or (iii) can be present as in-source fragments [2].
- To link the *m/z* values to peptide sequences a list of potential peptide masses could be generated based on the primary amino acid sequence of the substrate and protease specificity. This requires choices on whether to include (i) peptides that do not fall within the protease specificity, (ii) peptides with missed-cleavages, (iii) peptides with modified AA residues or (iv) peptides that originate from protein impurities.
- In addition, to link the *m/z* value to a peptide sequence, a certain mass error should be taken into account. The number of matches is highly dependent on the mass error [3, 4]. If the mass error is set too strictly, peptides may not be included in the final list. If the mass error is set too widely, there is a chance of incorrect identification of the *m/z* value.
- The last challenge is that in some cases, multiple peptide sequences can be matched to an *m/z* signal within the mass accuracy. This is for instance the case for isobaric peptide sequences. To decide what is the correct peptide that should be assigned to a *m/z* value, often the fragmentation spectra are used. The MS/MS fragments are decisive to confirm the positive identification of a peptide.

Key parameters for peptide identification

Despite the dilemmas listed above, many people publish lists of peptides annotated in complex mixtures. To come to the list of peptides, several approaches are used in practice to deal with (1) mass accuracy and (2) MS/MS fragmentation. The mass error used in peptide identification is often reported without explanation how the set value was chosen. In some cases, the choice is made based on the type and settings of the mass spectrometer [5], or based on the observed distribution of mass deviations [6].

To confirm the identity of the peptide, fragmentation spectra need to be analysed. A choice is made on how many of the possible fragments need to be identified to confirm the identification of the peptide. Although this choice is crucial, there is no general consensus on the (absolute or relative) number of fragments that is required for confirmation. Clearly, the number of required MS/MS fragments for confirmation depends on the number of options within the mass error. To distinguish between tryptic peptides originating from a single substrate does not require as many identified fragments as for *de novo* sequencing of peptides [7]. Many standardised algorithms are used in the field of proteomics to automatically identify

peptides based on MS/MS fragments as for instance MASCOT [8], SEQUEST [9], or Andromeda [10] with (incorporated) scoring functions [11, 12]. In literature, different studies using the same algorithm often do not apply the same threshold scores [13]. The score of a peptide annotation is often linked to a certain confidence level. However, even for annotations above the threshold score, still (only) 32 - 45 % of the identified peptides were repeatably annotated in all replicate injections [14, 15]. The question is how one could define a parameter to describe the confidence in the repeatability of the annotation, without the need to analyse multiple replicates.

Peptide quantification

To quantify compounds in mass spectrometry, typically the MS intensity of the ions is used. This intensity is known to vary because of ion-suppression, matrix effects, variation in charge states and day-to-day differences in absolute intensity [16-18]. Ideally, in the targeted MS approach, the MS intensity is corrected for these variations by using isotopically labelled standards, preferably with correction based on a standard addition to a reference sample [19]. In the untargeted approach, it is impossible to have isotopically labelled standards for each peptide, since beforehand it is not known which peptide are present. To avoid the need for (isotopically labelled) standards, Butré *et al.* have developed in recent years an approach for absolute label-free quantification of peptides based on UV absorbance [20]. The approach uses the predicted molar extinction coefficient of each peptide based on Kuipers *et al.* to convert UV peak areas to absolute peptide concentrations [21]. This quantification method was successfully applied in the past to for example determine differences in peptide release kinetics by bovine, human and porcine trypsin [22] and to quantify complex peptide mixtures with size-exclusion high-performance liquid chromatography [23]. In complex mixtures, UV peaks of eluting peptides are not always baseline separated. In some cases the individual UV peaks cannot be separately integrated, so that one UV peak should be divided over multiple peptides. It was previously suggested that this could be done using the ratio of MS intensities of the co-eluting peptides. Considering that peptides with similar retention times have more or less similar chemical properties, it was considered that ionisation efficiencies would be comparable as well [20].

Parameters to evaluate the completeness of analysis

When peptides are studied in food sciences, mostly a list or table of annotations is reported, e.g. [24] without any parameters describing the completeness of the analysis. In some of these cases, a plot is provided in which the identified peptides are mapped against the sequences of the initial protein substrates, e.g. [25]. These plots may aid the reader in evaluating the completeness, but do not give a value that describes the completeness. In other cases [26, 27], the protein sequence coverage, known from the field of proteomics [28], is reported. This parameter describes how many amino acids from the parental protein sequence were identified in at least one of the peptides. This parameter is purely based on unique amino acids, and therefore renamed to amino acid sequence coverage by Butré *et al.* [20]. They further introduced the quantitative parameters “peptide sequence coverage” and the “molar sequence coverage”, to describe the completeness of the peptide identification and of the peptide quantification respectively [20].

To test the reproducibility and completeness of automated annotation, three single protein hydrolysates were analysed. A set of criteria was developed to optimize completeness and validity of annotations. In addition, based on replicate analyses an objective parameter was defined to distinguish the annotations with high reproducibility and low reproducibility.

Materials & methods

Protein isolates, protease and chemicals

α-lactalbumin (α-LA) was obtained from Davisco Foods International. Inc. (Le Sueur, MN, USA). The α-LA was treated with ethylenediaminetetraacetic acid (EDTA) to remove the calcium ions attached to protein, as described by Deng *et al*. [29]. β-lactoglobulin (β-LG, L0130), β-casein (β-cas, C6905), bovine trypsin (EC 3.4.21.4, T1426) and aprotinin from bovine lung (A6279) were purchased from Sigma-Aldrich (St. Louis, MO, USA). Leucine enkephalin (Leu-enk, L9133), and Insulin (XI5500) were obtained from Sigma-Aldrich (St. Louis, MO, USA) Angiotensin was obtained from Alfa Aesar (Karlsruhe, Germany). The bovine trypsin had a protein content of~80 % of which 100 % was bovine trypsin, based on UV_{214nm} area analysis with UPLC-PDA-MS. According to the supplier, the trypsin activity was ≥ 10,000 BAEE units mg^{-1} protein. The bovine trypsin was treated with N-tosyl-L-phenylalanyl chloromethyl ketone (TPCK) to inactivate any chymotrypsin activity (≤ 0.1 % BTEE units mg^{-1} protein). The aprotinin solution contained 2.3 mg mL^{-1} aprotinin based on previous UPLC-MS results [30]. All other chemicals were purchased of analytical grade and purchased from Sigma or Merck.

Enzymatic hydrolysis of proteins

α-LA, β-LG and β-cas were each dissolved in 10 mL Millipore water at 1 % [weight powder / volume]. The pH was adjusted to pH 8.0 and the solutions were equilibrated for 0.5 h at 37 °C. Trypsin was dissolved (10 mg powder mL^{-1}) in Millipore water and added to the equilibrated solutions at an enzyme to substrate ratio of 1:100 [w/w]. Protein hydrolysis was performed in duplicate for 2 h in a pH-stat (Metrohm, Herisau, Switzerland) with a 0.2 M NaOH solution to keep the pH constant. The volume of added NaOH was used to calculate the degree of hydrolysis (DH) with equation 1 [31],

$$DH_{stat}[\%] = V_b \times N_b \times \frac{1}{\alpha} \times \frac{1}{m_p} \times \frac{1}{h_{tot}} \times 100\ \% \quad \text{(Eq. 1)}$$

where V_b [mL] is the volume of added NaOH; N_b [mol L^{-1}] is the normality of NaOH; α is the average degree of dissociation of the α-NH group (1/α = 1.3 at 37 °C and pH 8 [32]; m_p [g] is the amount of protein in solution; h_{tot} [mmol g^{-1}] is the number of peptide bonds per gram of protein. Trypsin was inactivated by addition of 15 µL aprotinin / mL hydrolysate, afterwards the samples were stored at -20 °C.

Sample preparation

The protein hydrolysates were incubated for 2 h in 10 mM dithiothreitol (DTT) and 50 mM Tris-HCl buffer at pH 8.0 at a protein concentration of 0.5 % [w/v], to reduce disulphide bonds. After incubation, the individual protein hydrolysates were mixed in mixtures of two substrates (α-LA

+ β-LG, α-LA + β-cas, β-LG + β-cas) and a mixture of three substrates (α-LA + β-LG + β-cas). The individual hydrolysates were diluted [1:2] with 0.15 % TFA [v/v] in MQ and the mixtures of two substrates were diluted [2:1] with 0.3 % TFA [v/v] in MQ water. 2 µL of 5 % [v/v] TFA was added to 100 µL of the mixture with three substrates. The final molar protein concentrations are shown in (**Table 1**). Afterwards, the samples were centrifuged (10 minutes, 14,000 x *g*, 20 °C) and the supernatants were injected in four replicates on the UPLC-MS.

Reverse phase ultra-high performance liquid chromatography (RP-UPLC)

The hydrolysates were analysed on a Waters H-class Acquity UPLC system (Milford, MA, USA). Peptide separation was done using a BEH C18 column (1.7 µm, 2.1 x 150 mm, Waters) that was coupled to a Waters Acquity UPLC PDA detector. The mobile phase consisted of a gradient of two solutions. Eluent A was UPLC-Grade water with 1 % [v/v] acetonitrile (ACN) and 0.1 % [v/v] trifluoroacetic acid (TFA) and eluent B was ACN with 0.1 % [v/v] TFA. 4 µL of the supernatant was injected into the column thermostated at 30 °C. The peptides were separated using the following elution profile: 0-2 min isocratic on 3 % B; 2-10 min linear gradient from 3-22 % B; 10-16 min linear gradient 22-30 % B; 16-21 min linear gradient 30-100 % B; 21-26 min isocratic on 100% B; 26-28 min linear gradient 100-3 % B and 28-32 min isocratic on 3 % B. During the first two minutes of isocratic elution (1.3 column volumes), the flow was directed to the waste to protect the MS and avoid any influence of remaining salt or unbound material on the MS or UV signals. The flow rate was set on 350 µL min^{-1}. Detection was performed using a PDA, which scanned the absorbance at the fixed wavelength of 214 nm at 1.2 nm resolution and 40 datapoints s^{-1}.

Electrospray ionisation time of flight mass spectrometry (ESI-Q-TOF-MS)

Mass spectra were obtained by an online Waters Synapt G2-Si high definition mass spectrometer coupled to the RP-UPLC, equipped with a z-spray electrospray ionization source, a hybrid quadrupole and an orthogonal time-of-flight analyser. The capillary voltage was set to 3 kV with the source operation in positive ion mode and the source temperature at 150 °C. The sample cone was operated at 35 V and nitrogen was used as desolvation gas (500°C, 800 L h^{-1}) and cone gas (200 L h^{-1}). Full scan MS and MS/MS data were acquired between 200 and 3000 m/z with a scan time of 0.3 seconds in resolution mode (V-mode) using an MSe method. MSe is a data-independent approach, were all precursor ions present in the MS at a given time were fragmented simultaneously. The trap collision energy was set at 4 V in single MS mode and ramped from 20 to 45 V in MS/MS mode. Prior to the analysis, the system was calibrated using sodium iodide, which was accepted when the average mass error on the calibrant peaks was below 2 ppm. Online lock mass data were acquired as a separate trace using Waters LockSpray at a set lockspray capillary voltage of 3.0 kV and at a sample infusion rate of 20 µL min^{-1}. Three peptides were evaluated as lock mass components: Leucine-enkephaline, $[M+H]^+$: 556.276575 *m/z*, Angiotensin II, $[M+2H]^{2+}$: 523.774534 *m/z* $[M+H]^+$: 1046.541791 *m/z* and Insulin $[M+3H]^{3+}$: 1910.876843 *m/z*. The optimised lock mass solution contained 0.4 µM Leu-Enk and 0.7 µM insulin dissolved in 50% [v/v] methanol containing 3% [v/v] acetic acid and 0.4% [v/v] diethylamine in UPLC grade water.

The development of data processing

The data processing method was developed by following the proposed steps (**Figure 2.1**). To set up the method in UNIFI, the α-LA hydrolysate was processed with the default method, without lock mass, without filters on MS/MS fragmentation or in-source fragments. The lock mass was optimised by analysis of the β-LG hydrolysate using different (combinations of) lock mass compounds. A concentration series of the α-LA hydrolysate was analysed with the optimised lock mass compounds to determine the LOD and LOA. The data, of the individual α-LA, β-LG and β-cas hydrolysates were obtained with the optimised double lock mass, and processed manually (1 replicate / protein) and automatically (4 replicates / protein). The mixtures of the proteins were processed only automatically (4 replicates / protein).

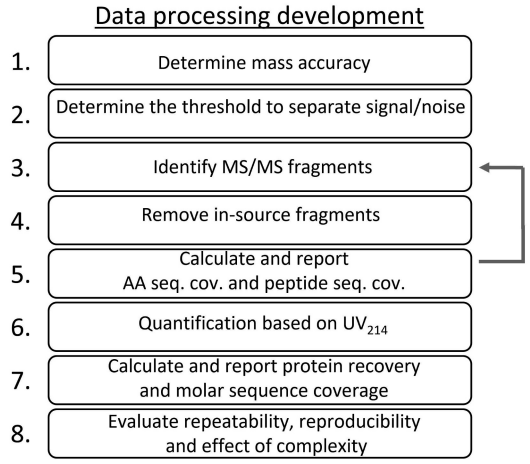


Figure 2.1. The proposed steps for the development of an automated UPLC-PDA-MS data processing method.

Peptide identification manually

Analysis of the mass spectrometry data of the individual protein hydrolysates was done manually in MassLynx software version 4.2. Manual annotation was performed similarly as in previous studies of our group [20, 30]. The *m/z* signals in the spectra were linked to possible peptides, based on the primary amino acid sequence of the substrate of interest. The peptides from other proteins than the main protein were not considered in manual annotation. Two AA modifications were taken into account: Methionine oxidation (+16 Da) for α-LA and serine phosphorylation (+80 Da per phosphoserine) for β-cas. The maximum allowed mass error was 100 ppm. To confirm the tentative annotations, the MS/MS spectra were used to identify b- and y-fragments. Mass spectrum deconvolution was used to extract the intact protein mass with the MaxEnt function in MassLynx.

Automated peptide identification

Automated peptide annotation was performed using the peptide mapping package in UNIFI software version 1.8. The amino acid sequences of α-LA, β-LG (variant A and B) and β-cas were

inserted and processed with trypsin as enzyme specificity on the semi-digest option. The semi-digest option included peptides that matched trypsin's specificity on at least the N- or C-terminal side. The included variable modifications were oxidation of the methionine (up to 1 per peptide) and serine phosphorylation (up to 5 per peptide). First, a default peak processing method was used with the default UNIFI settings without lock mass correction. In the default method, all signals were processed with a minimum signal intensity of 1000 detector counts in both the MS and MS/MS chromatograms. The maximum acceptable mass error was set at 100 ppm in the MS and 20 ppm in the MS/MS. For the final peak processing method, the minimum signal intensity was changed to 250 detector counts for the MS and 75 detector counts for the MS/MS, which corresponded with 3x the noise in the MS and MS/MS spectra. The maximum acceptable mass error was decreased from 100 ppm to 10 ppm. After peak processing, the match between *m/z* signals and a peptide sequence (peptide-spectrum match) was done with the algorithm incorporated in UNIFI. The algorithm returned for all precursor ions a potential peptide annotation (when possible). In case of multiple tentative annotations within the mass error, the peptide was matched to the annotation with most identified MS/MS fragments. This list still contains many entries of which some are not considered sufficiently reliable (e.g. if only one fragment was identified in a 6 AA peptide). These were removed by applying a filter selection constructed by the user. All annotations that were not confirmed with (at least) 2 b/y fragments were excluded. Peptides were included when 2 b/y fragments were identified (relevant for peptides with 2-6 amino acids) and when more than 15 % of the possible b/y fragments were identified (relevant for peptide with 7-16 amino acids) or when at least 5 b/y fragments were identified (relevant for peptides ≥ 17 amino acids). The peptide-length dependence of the fragmentation criteria were set based on the peptides that could theoretically be formed using a-specific hydrolysis. Listing these peptides showed that with these settings there was a negligible chance to have isobaric peptides that still meet the requirements. In addition, an additional mass error restriction of 5 ppm was set for peptides eluting within 15.00 minutes after the injection. The filter also removed in-source fragments recognised by UNIFI and annotations with a H_2O or NH_3 adduct. Annotations with a methionine oxidation were only included when originating from α-LA, while annotations with serine phosphorylation were only included when originating from β-cas. In-source fragments that were not recognised by the UNIFI software as such, were removed using PeptQuant, an in-house developed script in Matlab v2018b. Annotations were considered as in-source fragments if the parent peptide and potential in-source fragment eluted at a similar retention time and the in-source fragment included the same sequence as the peptide and the in-source fragment had a lower parent ion MS intensity than the peptide. In case a unique peptide was annotated twice, the peptide with the lowest MS intensity was removed. The presence of intact protein was manually evaluated and if present, added to the UNIFI output.

Peptide quantification

Peptides were quantified based on UV absorbance at 214 nm. The UV peaks between 1 to 20 [min] were integrated using the peak integration option in Masslynx. The peak integration was performed using a peak to peak baseline noise ratio of 500, a peak width of 0.28 min and a

baseline increase of 1 % (all values determined manually). The UV peaks corresponding to tris, DTT and aprotinin were excluded. The list of UV peak areas and retention times was coupled to the filtered UNIFI output using PeptQuant. The coupling was based on the start and end retention time [min] of the integrated UV peak and the retention time of the annotated peptide [min], taking into account the retention time offset between UV and MS (0.08 min). If multiple peptides were linked to the same UV peak, the UV peak area was divided over the co-eluting peptides based on their total ion count and molar extinction coefficient ε (Equation 2).

$$A_{214,i}\ [\mu\text{AU} \cdot \text{min}] = \left(\frac{\varepsilon_{214,i} \cdot MS_{tic,i}}{\sum \varepsilon_{214} \cdot MS_{tic}} \right) \times A_{214,tot} \qquad \text{(Eq. 2)}$$

where $A_{214,i}$ [µAU·min] is the UV peak area at 214 nm assigned to co-eluting peptide i, $A_{214,tot}$ [µAU·min] is the total UV peak area at 214 nm, $\varepsilon_{214,i}$ [L Mol^{-1} cm^{-1}] is the molar extinction coefficient at 214 nm and $MS_{tic,i}$ [counts] is the total ion count for co-eluting peptide i.

The concentration of each peptide, $C_{peptide}$ [µM], was calculated with Equation (3).

$$C_{peptide}\ [\mu M] = \frac{A_{214} \cdot Q}{\varepsilon_{214} \cdot l \cdot V_{inj} \cdot k_{cell}} \qquad \text{(Eq. 3)}$$

where A_{214} [µAU min] is the UV peak area at 214 nm, V_{inj} [µL] is the volume of sample injected, Q [µL min^{-1}] is the flow rate and l [cm] is the path length of the UV cell, which is 1 cm according to the manufacturer. The molar extinction coefficient ε_{214} [L Mol^{-1} cm^{-1}] for each peptide was calculated according to Kuipers et al. [21]. The cell constant, k_{cell} for the UV detector was 0.78. The k_{cell} was determined with a concentration series of α-LA and angiotensin II, with known concentrations. Corrected for protein content, purity and dilution during hydrolysis, the expected protein concentrations were 86 µM for α-LA, 78 µM for β-LG and 50 µM for β-cas. Equation (3) was also used to calculate the expected total UV based on the starting protein concentrations. The molar extinction coefficients ε_{214} [L Mol^{-1} cm^{-1}] of the hydrolysates were corrected for the degree of hydrolysis, resulting in a coefficient of 294,089 L Mol^{-1} cm^{-1} for α-LA, 281,944 L Mol^{-1} cm^{-1} for β-LG and 412,089 L Mol^{-1} cm^{-1} for β-cas.

Limits of detection, annotation and quantification

The limit of detection (LOD) and limit of annotation (LOA) of peptides were determined using a dilution series of the α-LA hydrolysate with a hydrolysate concentration from 0.00005 to 5 g L^{-1}. The MS intensity was reported for the highest dilution in which a peptide was respectively detected or annotated. The limit of detection (LOD) was defined as the lowest MS intensity of a peptide at which the precursor ion was recognised as signal in UNIFI. To be detected as a signal, the datapoints in the spectra had to form a recognisable (Gaussian) peak shape and the MS peak height had to be above the minimum detector count threshold in the MS (> 250 counts). The limit of annotation (LOA) was defined as the lowest MS intensity for a peptide to be annotated and meet the criteria on MS/MS fragmentation as stated in the UNIFI filters. The individual LOD and LOA of the peptides were averaged to determine the general LOD and LOA for this method. The average LOD was used in this study to differentiate signals from the noise and the average

LOA was used to differentiate abundant from non-abundant annotations. The limit of quantification (LOQ) was defined as 10 x the standard deviation of the noise in the UV chromatogram and was determined to be $3x10^1$ µAU · min. Since the peptides had large differences in molar extinction coefficient (from $2 \cdot 10^3$ to $3 \cdot 10^5$ L Mol^{-1} cm^{-1}), the LOQ was not expressed in µM for individual peptides.

Reproducibility

The individual hydrolysates and mixed hydrolysates were injected in four replicates. The repeatability of automatically annotated peptides was expressed as the percentage of unique peptides that were annotated similarly in all 4 replicates. The repeatability was calculated for peptides above the average LOA and for peptides between the LOD and LOA. The standard deviations over the total UV area, annotated UV area and absolute peptide concentrations were calculated based on the individual α-LA hydrolysate. To calculate the error on the concentration, annotations were used that were annotated similarly in all four replicates.

Tools to assess the completeness of peptide annotation and quantification

The completeness of the peptide analyses was evaluated by calculating the amino acid sequence coverages, peptide sequence coverages, protein recoveries and molar sequence coverages, as previously introduced by Butré *et al* [20]. The amino acid sequence coverage, also used in proteomics [28], was calculated by dividing the number of unique amino acids annotated in the peptides by the total number of amino acids in the protein sequence (Equation 4).

$$Amino\ acid\ sequence\ coverage\ [\%] = \frac{\#\ unique\ annotated\ amino\ acids}{\#\ amino\ acids\ in\ protein\ sequence} \cdot 100\ \% \qquad \text{(Eq. 4)}$$

When a peptide is annotated, other peptides should be present that cover the amino acids directly before and after this peptide. When this is not the case, the amino acids that should be covered form a 'missing' sequence. Moreover, a certain unique amino acid could be covered by multiple peptides. A 100 % amino acid sequence coverage does therefore not necessarily imply that all peptides in the hydrolysate are identified. To include both aspects in the sequence coverage, the peptide sequence coverage was calculated. This was calculated by dividing the number of unique annotated peptides by the number of expected peptides (Equation 5).

$$Peptide\ sequence\ coverage\ [\%] = \frac{\#\ AA\ (annotated\ peptides)}{\#\ AA\ (annotated\ peptides) + \#\ AA\ (missing\ peptides)} \cdot 100\ \%$$

To assess the completeness of quantification, the concentration of the peptides has to be considered. Based on the law of mass conservation, all the amino acids [µM] in the initial substrate should end up after hydrolysis as free amino acids, peptides or remaining intact protein. The protein recovery was calculated to assess to what extent the measured average AA concentrations matched the injected protein concentration (Equation 6).

$$Protein\ recovery\ [\%] = \left(\frac{\left(\frac{\sum C_n}{\#AA_{protein}} \right)}{C_0} \right) \cdot 100\ \% \quad \text{(Eq. 6)}$$

where C_n [µM] is the concentration of each individual AA (n) in the protein sequence, and $\#AA_{protein}$ is the number of amino acids in the initial protein and C_0 [µM] is the initially injected protein concentration. At last, the molar sequence coverage was calculated, which considers that certain parts of the protein sequence might be over-quantified whereas other regions are quantified with a lower concentration compared to the expected concentration.

The molar sequence coverage represents to what extent the peptides that cover an amino acid in the protein sequence are quantified relative to the injected molar concentration [µM] (Equation 7).

$$Molar\ sequence\ coverage\ [\%] = \left(1 - \frac{\sqrt{\frac{\sum(C_n - C_0)^2}{(\#AA_{protein} - 1}}}{C_0}\right) \cdot 100\ \% \qquad \text{(Eq. 7)}$$

where C_n [µM] is the concentration of each individual AA (n) in the protein sequence, C_0 [µM] is the initially injected protein concentration and $\#AA_{protein}$ is the number of amino acids in the initial protein.

Results & Discussion

Characterisation of the starting protein isolates and hydrolysates

The protein isolates of α-LA, β-LG and β-cas had a protein content [w protein / w DM] of 93 %, 96 % and 90 % and a protein purity of 90 %, 100 % and 90 % respectively (**Table 2.1**). The remaining 10 % of protein in the α-LA isolate was identified as β-LG with UPLC-MS. The remaining proteins in the β-cas had masses between 25 and 35 kDa. Analysis of the intact proteins showed that β-LG was equally present as genetic variant A or B. Literature indicates that the methionine residue [M90] in α-LA is prone to oxidation. Uniprot indicated that the serine residues [S15,S17,S18,S19,S35] in β-cas were phosphorylated, which was confirmed by the intact protein mass in the MS. The protein isolates of α-LA, β-LG and β-cas were hydrolysed with bovine trypsin and reached a $DH_{stat,max}$ of respectively 5.6 % (±0.1 %), 7.7 (±0.2 %), and 6.2 % (±0.1 %) (**Figure 2.2**). The $DH_{stat,max}$ values were in line with previous results under the same conditions [30].

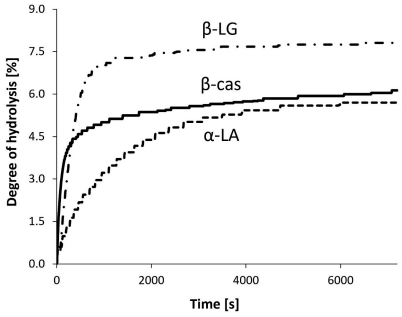


Figure 2.2. Degree of hydrolysis (DH_{stat}) versus time of 1% α-LA (- - -), β-LG (— ·) and β-cas (—) hydrolysed with bovine trypsin.

Table 2.1. Characteristics of the protein material as used as starting material for the hydrolysis.

Protein and Uniprot code[1]	N-factor[1] [g of protein / g N]	Protein content [w/w]	Purity[2] [%]	Protein loss with hydrolysis [3] [w/w]	Injected hydrolysate concentration [µM]	Molecular weight[1]	ε_{214} [L Mol^{-1} cm^{-1}]	#AA [1]	#CS	#Possible specific peptides	DH_{Stat} [%]
α-LA (P00711)	6.25	93 %	90 %	7.8 %	86 µM	14,186	300,395	123	13	105	5.6 ±0.1 %
β-LG (P02754)	6.29	96 %	100 %	7.8 %	78 µM	18,367 (A[5]) 18,281 (B[5])	293,410 (A[5]) 293,362 (B[5])	162	18	136	7.7 ±0.2 %
β-cas (P02666)	6.39	90 %	90 %	7.8 %	50 µM	23,983[4]	423,992	209	15	190	6.2 ±0.1 %
Trypsin (P00760)	5.97	80 %	100 %		-	-	-	-	-		

α-LA: α-Lactalbumin, β-cas: β-casein; β-LG: β-Lactoglobulin.

[1] From Uniprot (http://www.uniprot.org).

[2] Reported previously in [30].

[3] This protein loss factor corrects for (1) sampling during hydrolysis (2) Addition of trypsin inhibitor after hydrolysis and (3) pH adjustment in the pH-stat.

[4] The molecular weight of β-casein takes into account five phosphorylated serine residues, as identified with RP-UPLC-MS.

[5] The A and B indicate the genetic variant of β-lactoglobulin.

Manual peptide identification

The manual annotation of peptides in individual protein hydrolysates of α-LA, β-LG and β-cas yielded 27, 39, and 24 unique annotated peptides respectively. Of these 90 peptides, 67 peptides resulted from specific hydrolysis for trypsin and 23 peptides from semi-specific hydrolysis, i.e. either the peptide bond on the C- or on the N-terminal side that was hydrolysed did not match trypsin specificity. The methionine residue in α-LA was present in both the oxidised and in the non-oxidised form. The serine residues (S15, S17, S18, S19 and S35) in β-cas were always phosphorylated. The traditional amino acid sequence coverage was 100 % for all three substrates. The peptide sequence coverages, which take into account peptides that should be present based on the other formed peptides present, were respectively 91 %, 97 %, and, 97 % for the α-LA, β-LG and β-cas, respectively. These sequence coverage values were comparable to previous sequence coverages of studies of Butré and Deng, who used the same manual approach [20, 30].

Automated peptide annotation with UNIFI - Default run

In the **default** analysis of the α-LA hydrolysate 2034 unique *m/z* signals were identified. Of these *m/z* values, 843 were not matched to a peptide sequences, 279 were recognised by UNIFI as in-source fragments and 912 were tentatively matched to peptide sequences. Among these 912 annotations were 56 peptide annotations with losses of H_2O or NH_4 and 157 non-unique annotations. The remaining 699 unique peptide annotations had an absolute average mass error of 47 ±31 ppm and an average MS/MS fragment recovery of 4 ±14 %. Of the 27 manually annotated α-LA peptides, 26 peptides were also identified in the default analysis. The number of annotations was clearly higher in the default automated analysis than in the manual analysis. The question arises or all the (new) annotations should be considered valid and how to create confidence in the identified peptides.

Evaluation of different lock mass components

The 26 manually confirmed α-LA peptides in the **default** analysis had an average absolute mass error of 12 ± 6 ppm and showed a negative dependency with mass with a slope of -0.005 ppm Da^{-1} (**Figure 2.3**). Therefore, to ensure that large peptides were included in the analysis a high mass error threshold (100 ppm) was used. At the same time, the high mass error threshold would result in multiple tentative peptides that could be matched with a parent ion mass within the mass error, and potentially result in wrong annotations. To reduce the increase in mass error with increasing peptide mass, different lock mass combinations were evaluated using a β-LG hydrolysate. Without a lock mass, the peptides in the β-LG hydrolysate yielded an average absolute mass error of 5.1 ppm, with a maximum of 12.1 ppm and a slope in the mass residuals of -0.0052 ppm Da^{-1} (**Table 2.2**). Analysis of the same hydrolysate with lock mass yielded an average absolute mass error of 5.1 ppm for Insulin [3+], 2.0 ppm for Angiotensin [1+] and 2.0 ppm for LeuEnk [1+]. The insulin [3+] was not effective as a lock mass, probably because the *m/z* was higher than the majority of the peptides. The average mass error was efficiently decreased when Angiotensin [1+] or LeuEnk [1+] was used, but the mass error still showed a dependency with increasing mass, respectively -0.0017 ppm Da^{-1} for Angiotensin and -0.0015 ppm Da^{-1} for LeuEnk. Therefore the lock mass processing was performed with two components. Processing

the same data with LeuEnk [1+] and Insulin [3+] decreased the average absolute mass error to 1.4 ppm and reduced the slope to -0.0007 ppm Da^{-1}.

Table 2.2. The mass error of β-LG peptides analysed with different lock mass components.

Lock mass [charge state]	Maximum observed mass error (absolute, ppm)	Average mass error (absolute, ppm)	Trendline slope (ppm Da^{-1})
No lock mass	12.1	5.1	-0.0052
Insulin [3+]	7.8	5.1	-0.0013
Angiotensin [1+] & Insulin [3+]	5.5	2.1	-0.0001
LeuEnk [1+]	7.4	2.0	-0.0015
Angiotensin [1+]	7.3	2.0	-0.0017
LeuEnk [1+] & Angiotensin [1+]	7.3	1.8	-0.0011
LeuEnk [1+] & Insulin [3+]	4.8	1.4	-0.0007
LeuEnk [1+] & Insulin [3+] (with diethylamine)	4.8	1.2	-0.0008

It was observed that the insulin was mainly present in the [M+4] [M+5] and [M+6] state in the lockspray spectrum, whereas the charge state of interest [M+3] comprised only ~ 0.2 % [MS intensity] of the mass spectrum's signal intensity. Therefore, the charge state was altered by changing the solvent conditions and the addition of diethylamine. The relative abundance of [M+3] increased from 0.2 % to 90 % of the mass spectrum's signal intensity (**Annex 2.1+2.2**). This change in solvent composition yielded a final average mass error of 1.2 ± 1.1 ppm for the β-LG hydrolysate. Processing of the other samples showed a comparable average mass error of respectively 1.0 ±0.9 ppm for α-LA and 1.5 ±1.9 ppm for β-cas.

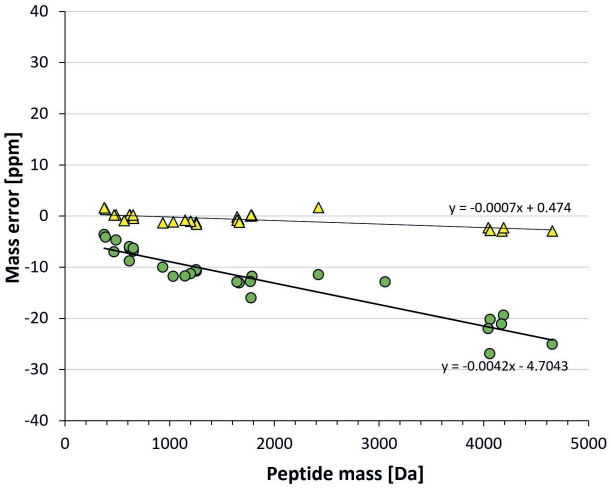


Figure 2.3. The mass error [ppm] plotted for manually confirmed α-LA peptides as function of the peptide mass [Da]. The green dots (●)represent α-LA peptides in the default run without lock mass, the yellow triangles (△) represent α-LA peptides analysed with the lock mass combination of LeuEnk [1+] and Insulin [3+] with diethylamine.

For the manually confirmed α-LA peptides, the slope was reduced from -0.005 to -0.0007 ppm Da^{-1} using the optimised lock mass combination of Leu-Enk with Insulin (**Figure 2.3**). Based on this, the mass error threshold was set at 10 ppm for the analyses with double lock mass in further sections.

Separate signals from the noise based on the LOD / LOA

The limit of detection (LOD) was on average $1.6 \cdot 10^5 \pm 1.7 \cdot 10^5$ counts, and was used to filter the signals from the noise. The limit of annotation (LOA) was on average $2.4 \cdot 10^6 \pm 3.5 \cdot 10^6$ counts which is ~15x higher than the average LOD (**Figure 2.4**). Without applying cut-offs for the LOD or LOA, analysis of the α-LA hydrolysate with double lock mass yielded 288 unique annotated α-LA peptides, 389 β-LG peptides and 343 β-cas peptides. The MS intensities of 599 of these annotations were below the LOD and were therefore excluded, leaving 421 peptides with MS intensities above LOD. It was observed that for these remaining 421 peptides, part of the annotations were not confirmed with a sufficient number of b/y fragments. Therefore, a filter was introduced to include only annotations with sufficient identified b/y fragments. Of the total MS intensity above the LOA, 90.0 % was attributed to peptides that passed the applied filter, which resulted in the identification of 73 peptides, of which 43 from α-LA, 21 from β-LG peptides and 9 from β-cas.

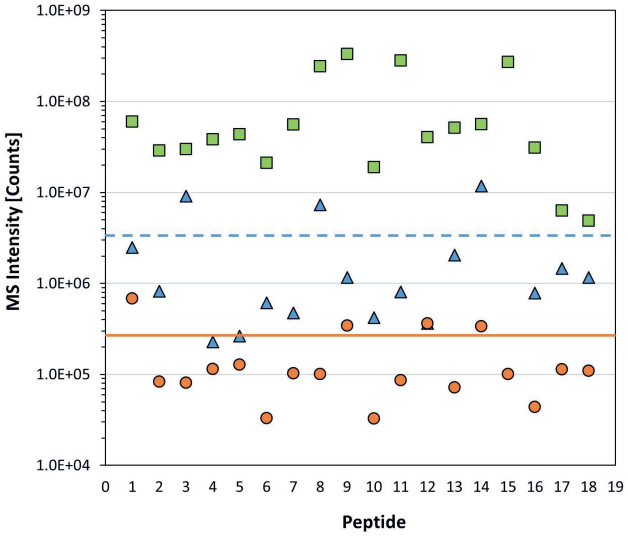


Figure 2.4. MS intensities of α-LA peptides in the highest dilution at which the MS signals were detected (●) and the peptides were annotated with MS/MS fragments (▲). The MS intensities as in the individual α-LA hydrolysate were also indicated (■). The average LOD (—) and average LOA (- -) were calculated based on the average of the individual peptides.

Removal of in-source fragments

For some peptides, formed in-source fragments were recognised as such by UNIFI or were excluded since the MS intensity was below the LOD. In other cases, these fragments were

incorrectly annotated as unique peptide. From the 73 annotated peptides in the α-LA hydrolysate, 6 in-source fragments (~10 % of total) were incorrectly identified as peptide by UNIFI and therefore removed in PeptQuant.

Peptide identification in the individual hydrolysates

The analysis of the individual α-LA, β-LG and β-cas hydrolysates using the proposed data-processing method yielded in total 77 peptides above the LOA, (26 of α-LA, 29 of β-LG and 22 of β-cas) and 56 peptides between the LOD and LOA (14 α-LA, 32 β-LG and 10 β-cas). The total MS intensity of these peptides was described for 99 ± 0.6 % by peptides above the LOA and 1 ± 0.6 % by the peptides between the LOD and LOA. 97 % of the manually identified peptides were also identified with the automated annotation. The automated analysis identified in total 4 additional peptides above the LOA and 40 additional peptides below the LOA that had not been found in the manual analysis.

Repeatability of peptide identification in the individual hydrolysates

The 77 peptides identified above the LOA were consistently annotated in the four replicates with a repeatability of 100 % (**Figure 2.5**). For the 56 peptides between the LOD and LOA, 50 % of the peptides was annotated consistently in all four replicates. The LOA could therefore be used as an MS intensity threshold to describe the confident and repeatable part of the annotations. The repeatability of peptides below the LOA implies that a peptide could be annotated below the LOA, but that the inclusion or exclusion is not as consistent as for peptides above the LOA. The repeatability of peptide identification in this work (100 % > LOA, 50 % < LOA) is higher than the repeatability in peptides for proteomics purposes. In work of Tabb and co-workers, a typical repeatability of 35-60% was described between two technical replicates [33].

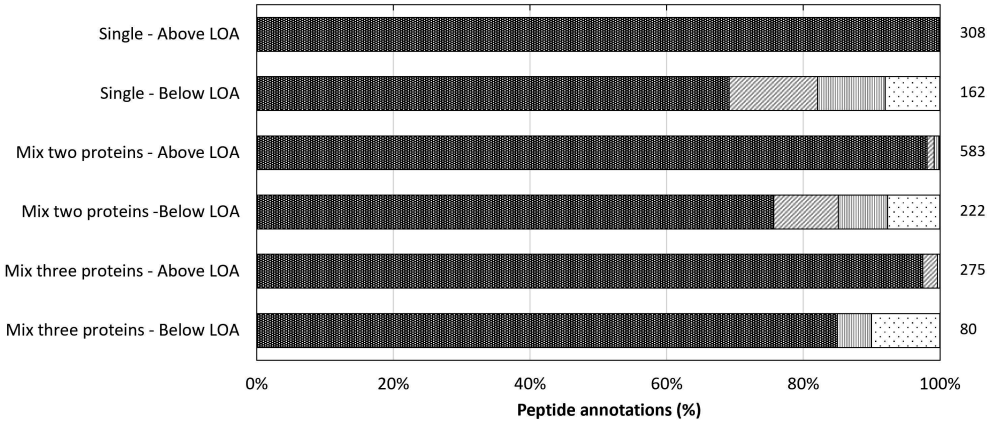


Figure 2.5. Repeatability of peptide identifications in the individual hydrolysate and mixtures. The percentage (%/%) of α-LA, β-LG and β-cas annotations in 4/4 replicates (▩), 3/4 replicates (▨), 2/4 replicates (▥) or 1/4 replicates (□). The total number of annotations is given.

Completeness of peptide identification in the individual hydrolysates

Using the automated annotation, the amino acid sequence coverages were 100 ±0 % for α-LA and β-LG; and 99 ±0 % for β-cas for the individual hydrolysates (**Table 2.3**). The peptide sequence coverages were 91.4 ±1 % for α-LA, 95.4 ±2 % for β-LG and 88.8 ±1 % for β-cas. The peptide sequence coverages of α-LA and β-LG were in line with the manual peptide sequence coverage (91.0 for α-LA & 97.0 % for β-LG). The peptide sequence coverage of the automated β-cas analysis (88.8 ±1) was lower than that with the manual analysis (97 %). This is probably due to 4 'missing' peptides, that were expected based on the 10 peptides that were annotated additionally to the manually annotated peptides. The variation in annotation of peptides between replicates did not result in substantial variation in the amino acid and peptide sequence coverages between the replicates. In the analysis of the individual α-LA hydrolysate, also β-LG was identified, with an amino acid sequence coverage of 98 ±0 % and a peptide sequence coverage of 88 ±2 %. This gives a first insight that the identification of peptides in a mixture of two proteins works. The amino acid sequence coverages in the analysis of the individual hydrolysates were higher than the typical amino acid sequence coverages reported in proteomics, which were typically below 50 % [34, 35].

Table 2.3. Sequence coverages of α-LA, β-LG and β-cas in the individual hydrolysates and the mixtures. The standard deviation is calculated over the four replicate injections for the hydrolysates analysed with the automated peptide annotation method.

Protein	Sample	AA sequence coverage [%]	Peptide sequence coverage [%]	Protein recovery [%]	Molar sequence coverage [%]
α-LA	Manual	100 %	91 %	103 %	80 %
	Automated	100 ±0 %	91 ± 1 %	101 ± 4 %	81 ± 2 %
	+ β-LG	100 ±0 %	93 ± 1 %	100 ± 1 %	75 ± 2 %
	+ β-cas	100 ±0 %	91 ± 1 %	110 ± 4 %	77 ± 2 %
	+ β-LG + β-cas	100 ±0 %	92 ± 3 %	106 ± 11 %	76 ± 8 %
β-LG	Manual	100 %	97 %	98 %	64 %
	Automated	100 ±0 %	95 ± 2 %	101 ± 4 %	77 ± 1 %
	+ α-LA	100 ±0 %	95 ± 2 %	101 ± 1 %	67 ± 1 %
	+ β-cas	100 ±0 %	88 ± 1 %	114 ± 3 %	63 ± 2 %
	+ α-LA + β-cas	100 ±0 %	96 ± 1 %	122 ± 11 %	67 ± 5 %
β-cas	Manual	100 %	97 %	99 %	62 %
	Automated	99 ±0 %	89 ± 1 %	94 ± 1 %	56 ± 0 %
	+ α-LA	99 ±0 %	88 ± 1 %	76 ± 4 %	63 ± 1 %
	+ β-LG	99 ±0 %	88 ± 1 %	85 ± 4 %	52 ± 3 %
	+ α-LA + β-LG	99 ±0 %	89 ± 1 %	82 ± 5 %	67 ± 6 %

Reproducibility of peptide quantification

The identified peptides were quantified with the corresponding UV areas at 214 nm and the predicted molar extinction coefficients. Over the four replicates of the individual α-LA hydrolysate, the average relative standard deviation of an integrated UV peak was 5%. The relative standard deviation of the total UV area in a chromatogram was 6.4 %. The relative standard deviation of the calculated peptide concentrations was 3.7 % for the absolute concentration of peptides above the LOA and 10.2 % for that of peptides between the LOA and

LOD. The relative standard deviations are comparable to those obtained with quantification techniques that require metabolic or chemical labelling (< 10 % RSD), and lower than those obtained with (relative) label-free quantification approaches in proteomics (10-30 % RSD) [36].

Quantification of the peptides in the individual hydrolysates

The total UV area in the chromatograms of α-LA, β-LG and β-cas were respectively 105 ±7 %, 104 ±4 % and 112 ±1 % of the expected UV. The UV area attributed to the automated analysed peptides was 100 ±5 % of the expected amount of UV for the individual α-LA hydrolysate, 99 ±4 % for β-LG and 99 ±2 % for β-cas. These values indicate that the amount of UV included in the analysis was in line with the expected amount based on protein concentrations. The protein recoveries yielded comparable values for the manual and automated analysis (**Table 2.3**). Peptide losses due to insolubility during sample preparation or peptide instability during the analysis seem to be neglectable for these hydrolysates from relatively 'clean' protein isolates. The protein recovery values do not indicate whether certain regions of the parental protein were over-quantified or under-estimated by the peptide composition. Therefore, the concentration of the amino acids in the peptides were plotted against the protein sequence (**Figure 2.6**). To describe the completeness of this plot with a quantitative parameter, the molar sequence coverages were calculated, which were 80 %, 64 % and 62 % for α-LA, β-LG and β-cas using the manual input and 81 ±2 %, 77 ±1 % and 56 ±0 % for β-cas using the automated input (**Table 2.3**). The molar sequence coverages for α-LA and β-LG were in line with reported molar sequence coverages reported for BLP hydrolysates, (70 ±10 % by Butré *et al*.) and for tryptic hydrolysates, (79 ±6 % by Deng *et al*.) [20, 30]. The molar sequence coverage for β-cas was lower than previously reported values, but was in the study of Deng *et al.* also mentioned to be below the average [30].

Peptide identification in the hydrolysate mixtures

In the mixtures there were no peptides annotated above the LOA that were not identified in the individual hydrolysates. However, the reverse is not true. Of the 308 peptide annotations above the LOA combined in the four individual hydrolysate replicates, 95 % were similarly annotated in the mixtures of two proteins, and 89 % were similarly annotated in the mixture of three proteins (**Figure 2.7**). Peptide (14-16) of α-LA was one of the peptides above the LOA that disappeared upon mixing with β-LG and with β-LG + β-cas. In this particular case, the parent ion *m/z* was above the LOD in all samples, but the minimum requirement of 2 b/y fragments was not met. This could probably be a result of the co-elution with peptide (139-141) of β-LG. For the 162 peptide annotations below the LOA in the individual hydrolysates, 64 % were similarly annotated in the mixtures of two proteins and 47 % were similarly annotated in the mixture of three proteins. Mixing the hydrolysates resulted in a substantial loss of annotations between the LOD and LOA. The remaining peptides after mixing between the LOD and LOA, showed a relative improvement in the repeatability between replicates from 69 % in the individual hydrolysates to 76 % for the mixtures with two proteins to 85 % for the mixtures with three proteins (**Figure 2.5**). The repeatability of the annotated peptides above the LOA decreased from 100 % in the individual hydrolysates to 98 % for the mixture with two proteins and 97 % for the

mixture with three proteins (**Figure 2.5**). For peptides above the LOA, the variation introduced by mixing three proteins (11 %) was slightly higher than the variation between replicates (3 %). For peptides between the LOD and LOA, the variation introduced by mixing three proteins (53 %) was in line with the variation between replicates (28 %). The repeatability of peptide annotations between the LOD and the LOA is in line with repeatability in proteomics studies [33].

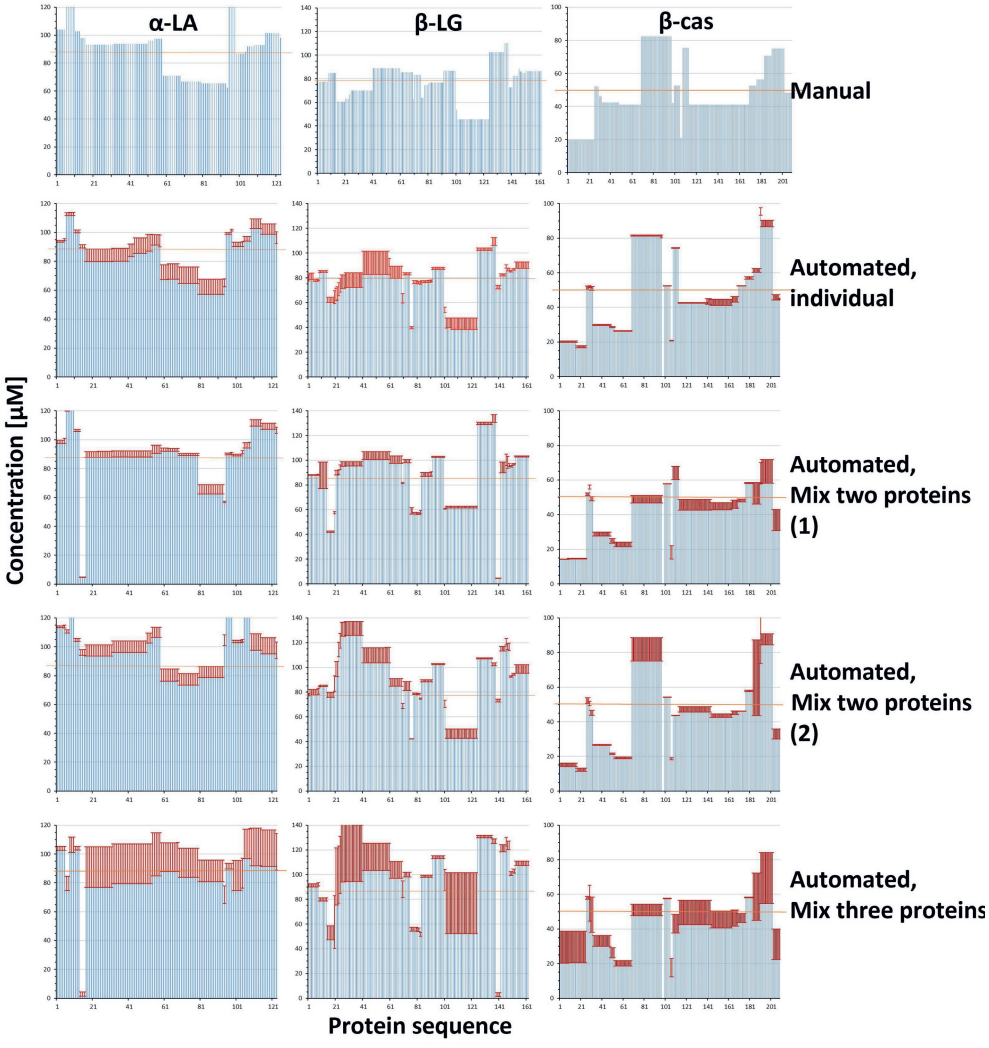


Figure 2.6. Concentration of all amino acids C_n for α-LA (left), β-LG (middle) and β-cas (right) of the manual analysis (row 1), individual protein hydrolysates analysed automatically (row 2), the 1:1 mixtures: α-LA in α-LA + β-LG (row 3, left), β-LG in α-LA + β-LG (row 3, middle), β-cas in α-LA + β-cas (row 3, right), α-LA in α-LA + β-cas (row 4, left), β-LG in β-LG + β-cas (row 4, middle), and β-cas in β-LG + β-cas (row 4, right), and the mixture of α-LA, β-LG and β-cas (row 5). The standard deviation shown was calculated over the four injections analysed with the automated annotation method. The orange line indicates the initial protein concentration in µM.

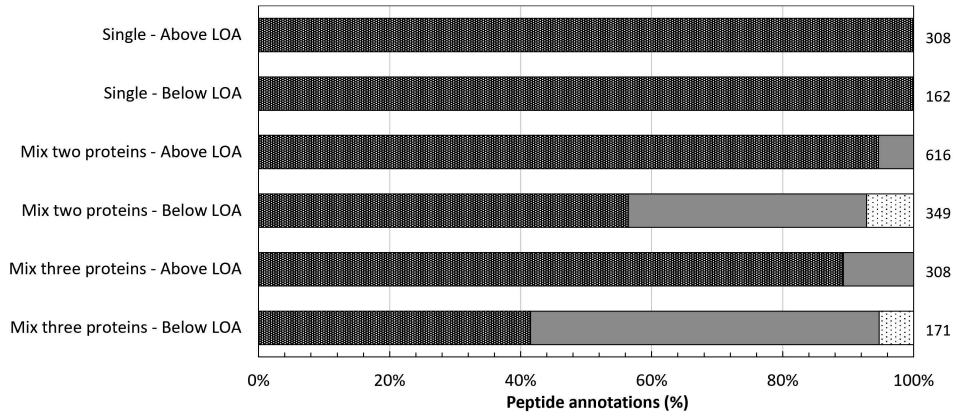


Figure 2.7. The effect of mixing on peptide identifications. The percentage (%/%) of α-LA, β-LG and β-cas annotations in the mixed hydrolysates annotated similarly as in the individual hydrolysates (▩), missing in the mixed hydrolysates (▢) or appearing in the mixed hydrolysates (▢) The total number of annotations is given.

In the end, the amino acid sequence coverages were identical for the mixed hydrolysates and the individual hydrolysates (**Table 2.3**). The peptide sequence coverages did not change significantly upon mixing, except for β-LG in the mixture with β-cas.

The effect of co-elution on peptide quantification in the mixed hydrolysates

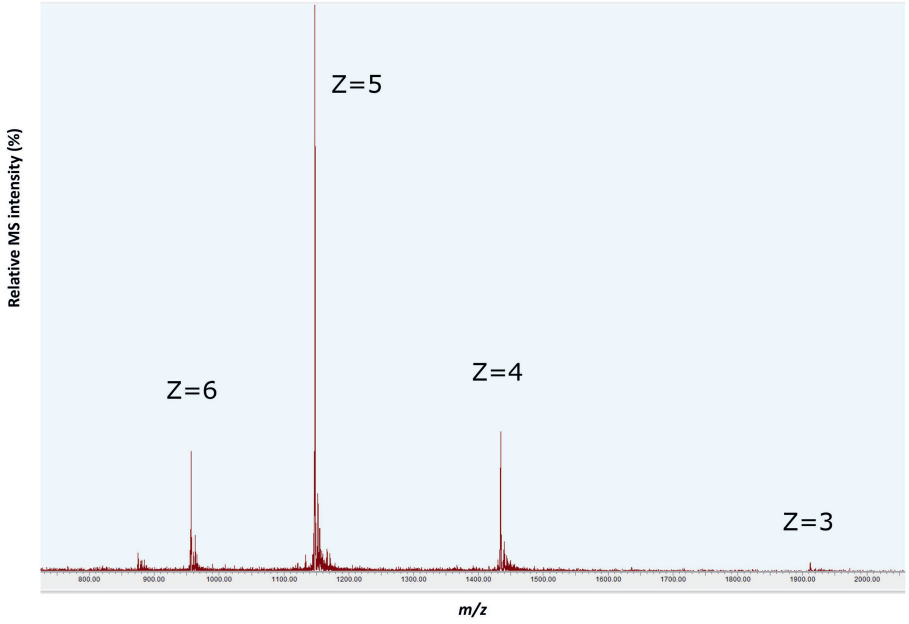
The individual protein hydrolysates were mixed to evaluate the effect of co-elution on the quantification of individual peptides (**Annex 2.3**). The molar sequence coverage for mixed hydrolysates was on average 4 ±8 % lower than for the individual hydrolysates. The average standard deviation in molar sequence coverage over the four replicates increased upon mixing from ±2 % to ±8 %. In mixtures of two hydrolysates, on average ~40 % of the total annotated UV area was linked to co-eluting peptides. This value increased to 57 % for the mixture with three hydrolysates. For 21 co-eluting peptides, the effect of co-elution on quantification was analysed by comparing the UV and MS signals as well as calculated concentrations in the individual and mixed hydrolysate. The total UV areas for each peptide in the individual differed on average 7 % from that in the mixed hydrolysate. For some co-eluting peptides there was no ion suppression (0%), while for others there was (max 48 %, average 21 %). When ion suppression occurred, it affected all co-eluting peptides at that RT similarly. To calculate the concentration of co-eluting peptides, the UV area is divided over the peptides assuming that the peptides have a more or less similar ionisation efficiency (i.e. MS intensity per amount of peptide). At short RT the ionisation efficiency is typically lower ($1 \cdot 10^5$ Counts μM^{-1}) than at higher RT ($1 \cdot 10^7$ Counts μM^{-1}) (**Annex 2.4**). However, for each set of co-eluting peptides, the ionisation efficiencies differed maximally with a factor 2-3 between each set of co-eluting peptides. In the chromatogram maximum variation in ionisation efficiency of a factor 5 to 8 were observed at close retention times, although these were not co-eluting (**Annex 2.4**). The differences in ionisation efficiency of co-eluting peptides resulted in a difference in calculated concentration of on average 37 ± 21

% compared to the concentrations in the individual hydrolysates. This error in calculated concentration of co-eluting peptides was substantially larger than the (relative) standard deviation over concentrations in replicate injections (6.3 %) for the studied peptides.

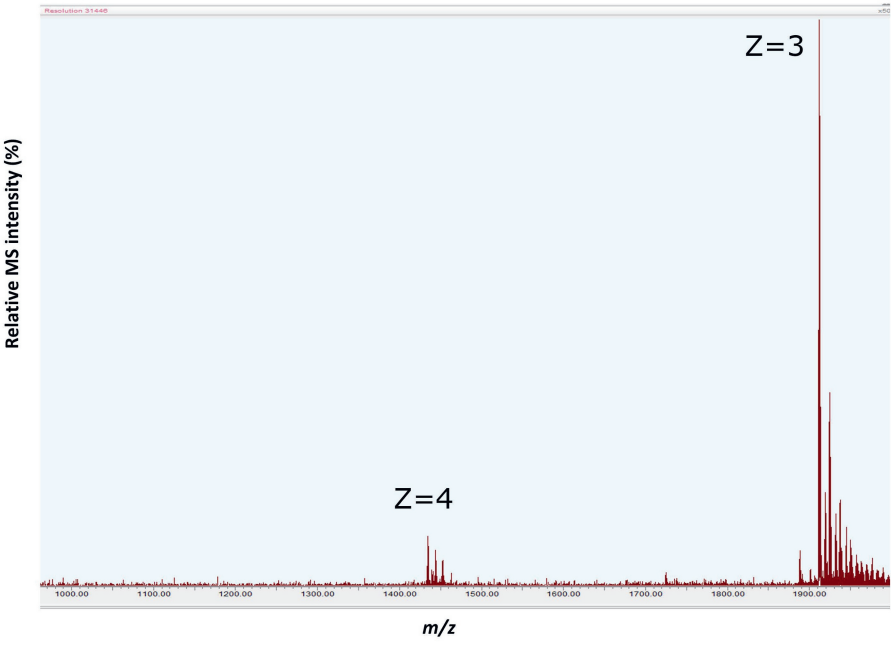
Conclusion

A method was evaluated for reproducible automated annotation and absolute quantification of peptides. It was shown that using the LOA a distinction could be made between peptides with 100% repeatability in single hydrolysates (99 % of the MS intensity assigned to peptides) and those with lower repeatability (50 %, and ~1 % of the MS intensity). For peptides above the LOA, mixing the hydrolysates resulted in an 11 % loss of identified peptides and a 3 % decrease in repeatability. The increased number of co-eluting peptides due to mixing had minor effects on amino acid, peptide or molar sequence coverage. However, calculated concentrations of individual co-eluting peptides in mixed systems varied on average 37%. The proposed approach enables automation of the hydrolysate compositional analysis while maintaining confidence in the repeatability of peptide annotations and completeness of the analysis. In addition, it opens up new possibilities for future research towards more complex protein hydrolysates.

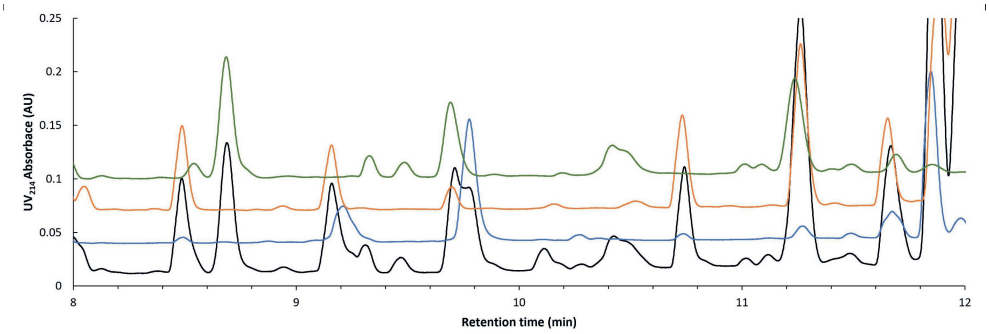
Annexes chapter 2



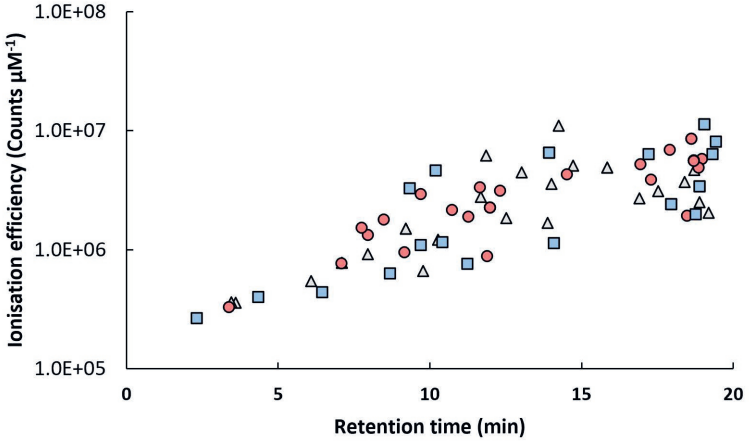
Annex 2.1. Lock mass spectrum of insulin without addition of diethylamine.



Annex 2.2. Lock mass spectrum of insulin with addition of diethylamine.



Annex 2.3. $UV_{214,nm}$ chromatograms of α-LA (—), β-LG (—), β-cas (—) and the mixture of α-LA, β-LG and β-cas (—) from 8-12 minutes.



Annex 2.4. Ionisation efficiencies of peptides above the LOA in the individual hydrolysates of α-LA (△), β-LG (○) and β-cas (□).

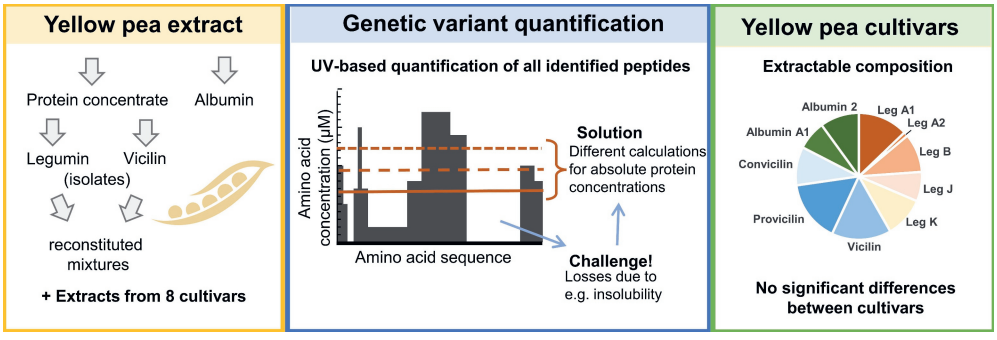
References

1. Mamone, G., Picariello, G., Caira, S., Addeo, F., Ferranti, P. (**2009**). Analysis of food proteins and peptides by mass spectrometry-based techniques. *Journal of Chromatography A*, 1216, 7130-7142.
2. Kim, J.-S., Monroe, M. E., Camp, D. G., Smith, R. D., Qian, W.-J. (**2013**). In-Source fragmentation and the sources of partially tryptic peptides in shotgun proteomics. *Journal of Proteome Research*, 12, 910-916.
3. Zubarev, R. A., Håkansson, P., Sundqvist, B. (**1996**). Accuracy requirements for peptide characterization by monoisotopic molecular mass measurements. *Analytical Chemistry*, 68, 4060-4063.
4. Kind, T., Fiehn, O. (**2006**). Metabolomic database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinformatics*, 7, 234.
5. Bristow, A. W. T., Webb, K. S. (**2003**). Intercomparison study on accurate mass measurement of small molecules in mass spectrometry. *Journal of the American Society for Mass Spectrometry*, 14, 1086-1098.
6. Zubarev, R., Mann, M. (**2007**). On the proper use of mass accuracy in proteomics. *Molecular & Cellular Proteomics*, 6, 377-381.
7. Frank, A. M., Savitski, M. M., Nielsen, M. L., Zubarev, R. A., Pevzner, P. A. (**2007**). De novo peptide sequencing and identification with precision mass spectrometry. *Journal of Proteome Research*, 6, 114-123.
8. Perkins, D. N., Pappin, D. J., Creasy, D. M., Cottrell, J. S. (**1999**). Probability ‐ based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis: An International Journal*, 20, 3551-3567.
9. Eng, J. K., McCormack, A. L., Yates, J. R. (**1994**). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, 5, 976-989.
10. Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., Mann, M. (**2011**). Andromeda: a peptide search engine integrated into the MaxQuant environment. *Journal of Proteome Research*, 10, 1794-1805.
11. Keller, A., Nesvizhskii, A. I., Kolker, E., Aebersold, R. (**2002**). Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical Chemistry*, 74, 5383-5392.
12. Koenig, T., Menze, B. H., Kirchner, M., Monigatti, F., Parker, K. C., Patterson, T., Steen, J. J., Hamprecht, F. A., Steen, H. (**2008**). Robust prediction of the MASCOT score for an improved quality assessment in mass spectrometric proteomics. *Journal of Proteome Research*, 7, 3708-3717.
13. Cooper, B. (**2011**). The problem with peptide presumption and low Mascot scoring. *Journal of Proteome Research*, 10, 1432-1435.
14. Berg, M., Parbel, A., Pettersen, H., Fenyö, D., Björkesten, L. (**2006**). Reproducibility of LC-MS-based protein identification. *Journal of experimental botany*, 57, 1509-1514.
15. Delmotte, N., Lasaosa, M., Tholey, A., Heinzle, E., van Dorsselaer, A., Huber, C. G. (**2009**). Repeatability of peptide identifications in shotgun proteome analysis employing off ‐ line two ‐ dimensional chromatographic separations and ion-trap MS. *Journal of separation science*, 32, 1156-1164.
16. Annesley, T. M. (**2003**). Ion suppression in mass spectrometry. *Clinical Chemistry*, 49, 1041-1044.
17. Collins, B. C., Hunter, C. L., Liu, Y., Schilling, B., Rosenberger, G., Bader, S. L., Chan, D. W., Gibson, B. W., Gingras, A.-C., Held, J. M., Hirayama-Kurogi, M., Hou, G., Krisp, C., Larsen, B., Lin, L., Liu, S., Molloy, M. P., Moritz, R. L., Ohtsuki, S., Schlapbach, R., Selevsek, N., Thomas, S. N., Tzeng, S.-C., Zhang, H., Aebersold, R. (**2017**). Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry. *Nature Communications*, 8, 291.
18. Bantscheff, M., Schirle, M., Sweetman, G., Rick, J., Kuster, B. (**2007**). Quantitative mass spectrometry in proteomics: a critical review. *Analytical and Bioanalytical Chemistry*, 389, 1017-1031.

19. Brun, V., Dupuis, A., Adrait, A., Marcellin, M., Thomas, D., Court, M., Vandenesch, F., Garin, J. (**2007**). Isotope-labeled protein standards: Toward absolute quantitative proteomics. *Molecular & Cellular Proteomics*, 6, 2139-2149.
20. Butré, C. I., Sforza, S., Gruppen, H., Wierenga, P. A. (**2014**). Introducing enzyme selectivity: A quantitative parameter to describe enzymatic protein hydrolysis. *Analytical and Bioanalytical Chemistry*, 406, 5827-5841.
21. Kuipers, B. J. H., Gruppen, H. (**2007**). Prediction of molar extinction coefficients of proteins and peptides using UV absorption of the constituent amino acids at 214 nm to enable quantitative reverse phase high-performance liquid chromatography-mass spectrometry analysis. *Journal of Agricultural and Food Chemistry*, 55, 5445-5451.
22. Deng, Y., Gruppen, H., Wierenga, P. A. (**2018**). Comparison of protein hydrolysis catalyzed by bovine, porcine, and human trypsins. *Journal of Agricultural and Food Chemistry*, 66, 4219-4232.
23. Bodin, A., Framboisier, X., Alonso, D., Marc, I., Kapel, R. (**2015**). Size-exclusion HPLC as a sensitive and calibrationless method for complex peptide mixtures quantification. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences*, 1006, 71-79.
24. Nongonierma, A. B., Mazzocchi, C., Paolella, S., FitzGerald, R. J. (**2017**). Release of dipeptidyl peptidase IV (DPP-IV) inhibitory peptides from milk protein isolate (MPI) during enzymatic hydrolysis. *Food Research International*, 94, 79-89.
25. Dupont, D., Mandalari, G., Molle, D., Jardin, J., Léonil, J., Faulks, R. M., Wickham, M. S. J., Mills, E. N. C., Mackie, A. R. (**2010**). Comparative resistance of food proteins to adult and infant *in vitro* digestion models. *Molecular Nutrition and Food Research*, 54, 767-780.
26. Rajendran, S. R. C. K., Mason, B., Udenigwe, C. C. (**2016**). Peptidomics of peptic digest of selected potato tuber proteins: post-translational modifications and limited cleavage specificity. *Journal of Agricultural and Food Chemistry*, 64, 2432-2437.
27. Gu, S., Chen, N., Zhou, Y., Zhao, C., Zhan, L., Qu, L., Cao, C., Han, L., Deng, X., Ding, T., Song, C., Ding, Y. (**2018**). A rapid solid-phase extraction combined with liquid chromatography-tandem mass spectrometry for simultaneous screening of multiple allergens in chocolates. *Food Control*, 84, 89-96.
28. Meyer, B., Papasotiriou, D. G., Karas, M. (**2011**). 100% protein sequence coverage: a modern form of surrealism in proteomics. *Amino Acids*, 41, 291-310.
29. Deng, Y., Butré, C. I., Wierenga, P. A. (**2018**). Influence of substrate concentration on the extent of protein enzymatic hydrolysis. *International Dairy Journal*, 86, 39-48.
30. Deng, Y., van der Veer, F., Sforza, S., Gruppen, H., Wierenga, P. A. (**2018**). Towards predicting protein hydrolysis by bovine trypsin. *Process Biochemistry*, 65, 81-92.
31. Adler-Nissen, J. (**1986**). Enzymic hydrolysis of food proteins. London, UK: Elsevier Applied Science Publishers.
32. Butré, C. I., Wierenga, P. A., Gruppen, H. (**2014**). Influence of water availability on the enzymatic hydrolysis of proteins. *Process Biochemistry*, 49, 1903-1912.
33. Tabb, D. L., Vega-Montoto, L., Rudnick, P. A., Variyath, A. M., Ham, A.-J. L., Bunk, D. M., Kilpatrick, L. E., Billheimer, D. D., Blackman, R. K., Cardasis, H. L. (**2010**). Repeatability and reproducibility in proteomic identifications by liquid chromatography− tandem mass spectrometry. *Journal of Proteome Research*, 9, 761-776.
34. Garza, S., Moini, M. (**2006**). Analysis of complex protein mixtures with improved sequence coverage using (CE−MS/MS)n. *Analytical Chemistry*, 78, 7309-7316.
35. Krug, K., Carpy, A., Behrends, G., Matic, K., Soares, N. C., Macek, B. (**2013**). Deep coverage of the escherichia coli proteome enables the assessment of false discovery rates in simple proteogenomic experiments. *Molecular and Cellular Proteomics*, 12, 3420-3430.
36. Schulze, W. X., Usadel, B. (**2010**). Quantitation in mass-spectrometry-based proteomics. *Annual review of plant biology*, 61, 491-516.

CHAPTER 3

Towards absolute quantification of protein genetic variants in *Pisum sativum* extracts



Vreeke, G. J. C., Meijers, M. G. J., Vincken, J.-P., Wierenga, P. A. (**2023**). Towards absolute quantification of protein genetic variants in *Pisum sativum* extracts. *Analytical Biochemistry*, *665*, 115048. https://doi.org/10.1016/j.ab.2023.115048

Abstract

In recent years, several studies have used proteomics approaches to characterize genetic variant profiles of agricultural raw materials. In such studies, the challenge is the quantification of the individual protein variants. In this study a novel UPLC-PDA-MS method with absolute and label-free UV-based peptide quantification was applied to quantify the genetic variants of legumin, vicilin and albumins in pea extracts. The aim was to investigate the applicability of this method and to identify challenges in determining protein concentration from the measured peptide concentrations. Analysis of the protein mass balance showed significant losses of proteins in extraction (37 %) and of peptides in further sample preparation (69 %). The challenge in calculating the extractable individual protein concentrations was how to deal with these insoluble peptides. The quantification approach using average amino acid concentrations in each position of the sequence showed most reproducible results and allowed comparison of the genetic protein composition of 8 different cultivars. The extractable protein composition (μM/μM) was remarkably similar for all cultivar extracts and consisted of legumins A1 (12.8 ±1.2 %), A2 (1.1 ±0.4 %), B (9.9 ±1.6 %), J (7.5 ±1.0 %) and K (10.3 ±2.1%), vicilin (15.2 ±1.7 %), provicilin (15.7 ±2.5 %), convicilin (9.8 ±0.8 %), albumin A1 (7.4 ±2.0 %), albumin 2 (10.0 ±1.5 %) and protease inhibitor (0.4 ±0.4 %).

Introduction

The protein composition of protein concentrates describes the different classes of protein e.g. globulins and albumins, or specific types of globulins e.g. legumin and vicilin present in the sample. In leguminosa seeds, the most abundant proteins are legumins and vicilins, e.g. in pea (*Pisum sativum*) isolates these proteins together represent approximately 53 ± 7 % (w/w) of the total protein content [1, 2]. The amount and relative ratio of the different types of proteins can be identified using electrophoresis techniques, such as SDS-PAGE [3]. Each type of protein, however, can be present in different genetic variants, which cannot be identified with this technique. For pea, the first reports about the existence of different legumin genetic variants are based on genomic data and date back 30 years [4, 5]. Different genetic variants can be identified with proteomics techniques, which are already widely applied on agricultural raw materials such as milk and soy [6, 7]. In that field there is quite some discussion on how the mass spectrometry data can be used to determine the absolute amount of individual proteins present (quantification) [8]. By combining mass spectrometric data with UV signals of individual peptides, we have shown that peptides can be accurately quantified in protein hydrolysates using UPLC equipped with a photodiode-array detector (PDA) and MS (UPLC-PDA-MS) [9, 10] (**Chapter 2**). The aim of this study is to illustrate the use of this method to obtain an absolute, label-free quantification of the protein composition of complex pea protein samples at genotype-level.

To identify which proteins are present in a sample, (plant) proteins are hydrolysed, the formed peptides are separated by (2D) gel-electrophoresis or liquid chromatography (LC) and

analysed using mass-spectrometry (MS). This is similar in proteomics and in the approach described in this study. To quantify the proteins, first the peptides need to be quantified, which could be done via several strategies [11]. When the concentrations of the peptides are determined, they need to be converted to a concentration of a protein genetic variant. The quantification of peptides was initially done relatively to each other, using the MS peak intensities of the HPLC chromatogram, for genetic variants in bovine milk [12]. In this case, the intensity of a proteotypic peptide, which uniquely represented a protein and was (reproducibly) formed at high yield [13], was used to compare protein abundances. The downside of relative quantification is that individual MS intensities are highly affected by ion-suppression, matrix effects and day-to-day variation [14-16]. Therefore, later, absolute peptide concentrations were determined by comparing the MS peak intensity of the respective peptide with the intensity of an isotopically labelled reference peptide with similar ionisation properties [17]. The downside of using labelled reference peptides is that it can be costly and laborious. Due to the limited number of reference peptides, only one or a few peptides per protein are quantified. A benefit of this approach is that it has a lower relative standard deviation (RSD) on the calculated protein concentration (<10 %) than the MS intensity based approaches (>10 %) [18]. An alternative approach to quantification based on MS signals, is the quantification of peptides using the absolute UV absorbance [10]. For the absolute quantification of each peptide this method uses the UV peak area at 214 nm and the molar extinction coefficient predicted using the method of Kuipers *et al*. [9] (**Chapter 2**). The benefit of this technique is that it has the accuracy of quantification techniques with labelling (~6 % RSD for peptide concentrations in replicate injections [10]), but does not require chemical or isotopic labelling. A downside of applying UV-based quantification in complex digests, is that for coeluting peptides, the quantification is considerably less accurate (**Chapter 2**).

Converting the peptide concentrations to a concentration of a protein genetic variant introduces three challenges.

- Enzymatic hydrolysis of a substrate does not always yield the same peptides at the same concentrations, when different hydrolysis conditions and incubation times are used [19, 20]. Variations in digestion methods resulted in large relative standard deviations of 102 % up to 1305 % in quantification of individual peptides [20]. This could give problems when only one or a few (isotopically labelled reference) peptides are used to quantify a protein. Similarly, more than one peptide (sequence) could be released including the same amino acids of the original protein sequence. Therefore, one should sum peptide concentrations that cover the same amino acids in the protein sequence. Both issues could be solved when all peptides are quantified and used to determine the protein concentration, as for instance done with UV-based quantification or MS intensity based approaches as exponentially modified protein abundance index (emPAI) and intensity based absolute quantification (IBAQ) [21-23].
- A second challenge is how to deal with peptides that are not unique to one genetic protein variant [18, 24]. In proteomics analyses, these non-unique peptide sequences are typically excluded from the analysis. The loss of information by excluding these non-unique peptides

leads to an underestimation of the protein concentration. The impact will depend on the proportion of non-unique peptide sequences, affected by the similarity between protein sequences.

- Underestimation could also result from peptide losses during sample preparation [25, 26] or from intrinsic instability [19, 27]. As a result, not all peptides that are formed during enzymatic hydrolysis are included in the analysis. Small molecules as free amino acids, di-peptides and (some) tri-peptides will also be excluded from the analysis since they are not detected in the typical RP-HPLC methods used [28]. The challenge is how to deal with this missing information in calculating protein concentrations.

To estimate the impact of these challenges on the protein quantification it is necessary to have knowledge on the mass balance *e.g.* how much of the initial protein(s) is included in the analysis. This is generally not considered in quantitative proteomics, but could easily be checked by analysis of the protein content before and after sample preparation and centrifugation, or by analysis of the amount of UV in the chromatograms compared with the expected amount based on protein content and composition. In this study the completeness of the analysis is evaluated in detail using the amino acid sequence coverage, UV recovery, matched UV, protein recovery and molar sequence coverage plot as described previously by Butre *et al*. [10] and in **Chapter 2**.

For pea, genetic variants were only quantified relatively in protein extracts using the volume percentage from 2-D electrophoresis [29]. Bourgeois *et al.* showed a 2-D map with 626 Coomassie-blue spots, of which 124 were analysed using MS-techniques (Maldi-TOF MS and LC-MS/MS). Altogether, 156 polypeptides were identified, belonging to 55 different proteins. This number can be an overestimation, since precursors, post-translationally cleaved proteins and proteins with modifications were reported as different proteins. Vicilin, convicilin, legumin, and albumin families represented 56 % of the polypeptides identified [29]. Several genetic variants were identified, for example for pea legumin, three genetic variants were reported that were reviewed in Uniprot: legumin A2, J and K. In total, out of 11 genes that were described for pea legumin the primary structures of 5 different legumin genetic variants are reported in Uniprot: legumin A1, A2 (former LEG2), J, K and B (Uniprot) [30]. Legumin A1 and A2 are closely related with a sequence similarity of 97.5 %. Legumin B is more similar to J and K (**Table 3.1**; Uniprot).

This study aims at quantifying the protein composition of pea extracts at genetic variant-level, using the recently developed UPLC-PDA-MS method in **Chapter 2**. The method will be used to calculate protein concentrations using the concentrations of all peptides, based on UV absorbance. The applicability and challenges to quantify proteins will be investigated. The method will be tested on purified pea legumin, vicilin and albumin fractions and afterwards applied to characterise the genetic protein composition of 8 pea cultivars.

Table 3.1. Protein genetic variants in *Pisum sativum* reported on Uniprot used in peptide annotation screening.

Protein	Uniprot code	Molecular weight (Da)[1]	Sequence similarity (%)[2]		Identified?
Legumin A1	P02857	56605	-		Yes
Legumin J	P05692	54587	40.1	Leg A1	Yes
Legumin K	P05693	39800	66.7	Leg J	Yes
Legumin B	P14594	38990	47.2	Leg K	Yes
Legumin A2	P15838	56968	97.5	Leg A1	Yes
Vicilin	P13918	49341	-		Yes
Provicilin B	P02854	44878	74.1	Vic	Yes
Convicilin A	P13915	63932	-		Yes
Albumin 2	P08688	25621	-		Yes
Albumin-1 A	P62926	11234	-		Yes
IBBB	P56679	7864	-		Yes
PIP20	Q41015	20862	-		Yes
Albumin-1 B	P62927	11274	96.9	Alb-1 A	No
Albumin-1 C	P62928	11234	90.8	Alb-1 A	No
Albumin-1 D	P62929	11238	96.9	Alb-1 A	No
Albumin-1 F	P62931	11235	96.9	Alb-1 A	No
Provicilin A	P02855	31540	47.9	Vic	No
Convicilin B	P13919	43275	50.9	Con A	No

[1] Signal peptide and (post translational) modifications were not taken into account.

[2] The sequence similarity gives the amino acid sequence similarity of a protein and the most similar protein higher in the list for peptide annotation.

Materials & methods

Yellow peas (*Pisum sativum* Leguminosae) were purchased from Alimex Europe B.V. (Sint-Kruis, Belgium). The following pea variants (*Pisum sativum* Leguminosae) were provided by the Centrum voor Genetische Bronnen Nederland (CGN, Wageningen, The Netherlands): Lente Krombek (CGN02949), Vroegste Gele Krombek (CGN02950), Venlosche Lage (CGN02962), Belinda (CGN10266), Miranda (CGN10292), Paloma (CGN10296), Flavandra (CGN13290), Montana (CGN24055). *Bacillus licheniformis* protease (BLP) was provided by Novozymes (Bagsvaerd, Denmark). BLP is a serine protease which is able to hydrolyse bonds at the C-terminal of aspartic acid (D) and glutamic acid (E) residues. Previous studies observed 1000x faster hydrolysis after glutamic acid (E) than aspartic acid (D) residues [10, 31]. The BLP powder was further treated to remove insoluble material as described by [32]. In short, a suspension of BLP was made and centrifuged for 10 min (4000 *x g*, 25°C). The supernatant was dialyzed with a 12-14 kDa membrane against 150 mM NaCl and subsequently against demineralised water. Afterwards, the retentate was frozen and freeze-dried. The freeze-dried BLP had a protein content of 60 % (w/w, as is) and an activity of 3.9 AU mg^{-1} min^{-2} according to analysis by Deng *et al*. [33]. SDS-PAGE marker, gels, sample buffer and running buffer were purchased from Bio-Rad Laboratories (Hercules, CA, USA). Coomassie blue stain was purchased form Expedeon (San Diego, CA, USA). A glycoprotein staining kit was purchased from Thermo Scientific (Waltham, MA, USA). Sep-Pak C18 6 cc Vac Cartridges (WAT043395) were purchased from Waters (Milford, MA, USA). All other chemicals were of analytical grade and purchased from either Merck (Darmstadt, Germany), Sigma-Aldrich (St. Louis, MO, USA) or Acros Organics (Geel, Belgium). All water was demineralised (conductivity of 2 µS cm^{-1}) or obtained from a Milli-Q system (Millipore, Billerica, MA, USA; conductivity of 0.5 µS cm^{-1}).

Protein isolation and fractionation from yellow pea

Preparation of pea protein concentrate

Pea protein concentrate (PPC) was prepared by alkaline extraction followed by iso-electric precipitation, as described by O'Kane [34], with minor alterations. Whole frozen peas (Alimex) were broken with a pin mill (LV 15M Condux-Werk, Wolfgang bei Hanau, Germany) and subsequently milled (ZPS50 impact mill, Hosokawa-Alpine, Augsburg, Germany). The pea flour (10 %, w/w) was suspended in Milli-Q water (MQ). The suspension was adjusted to pH 8.0, followed by centrifugation (17,000 x *g*, 4 °C, 20 min). The supernatant was collected and adjusted to pH 4.5, followed by centrifugation (17,000 x *g*, 4 °C, 20 min). The pellet was recovered and suspended in MQ at a final concentration of 10 % (w/w, wet pellet) and adjusted to pH 8.0. The obtained solution was centrifuged (17,000 x *g*, 4 °C, 20 min), and the resulting supernatant was frozen (PPC_{-20}), freeze-dried and named pea protein concentrate (PPC). Prior to all centrifugation steps, suspensions and solutions were kept at 4 °C and the set pH while being stirred for minimally 2 hours.

Legumin and vicilin fractionation from PPC

The PPC_{-20} was further fractionated to obtain pea legumin fraction (PLF) and pea vicilin fraction (PVF) as described by O'Kane et al. with alterations [34]. The solution was adjusted to pH 8.0 with NaOH, and stirred for 1 hour at 4 °C. The solution was subsequently diluted 1:1 with a McIlvaine buffer of pH 4.8, to a final concentration of 200 mM disodium phosphate and 100 mM citric acid containing 200 mM NaCl. The sample was stirred at 4 °C for minimally 2 hours, followed by centrifugation (17,000 x *g*, 4 °C, 20 min). The obtained supernatant containing the pea vicilin was filtered using an ultrafiltration system with a 5 kDa membrane (Hydrosart Ultrafilter, Sartorius AG, Frankfurt, Germany). The liquid removed during ultrafiltration was replenished by MQ. The PVF was frozen and freeze-dried. The legumin-rich pellet was resuspended in 20.0 mM Tris-HCl buffer, pH 8.0, (buffer A) at a final concentration of approximately 10 g L^{-1}. The solution was stirred for minimally 2 hours, prior to centrifugation (17,000 x *g*, 4 °C, 20 min). The obtained supernatant was filtered over a glass fiber pre-filter (13400-142-K, Sartorius) with a Whatman filter paper (black ribbon, 589/1, GE Healthcare, Uppsala, Sweden). The filtrate was applied onto a Source 15Q column (Fineline, Pfizer Manufacturing, Freiburg, Germany) coupled to an ÄKTA explorer system (GE Healthcare). Elution was similar to the method as described by O'Kane *et al.* and fractions were collected [34]. The fractions rich in legumin were pooled and filtered using an ultrafiltration system with a 5 kDa membrane (Hydrosart Ultrafilter, Sartorius AG). The liquid removed during ultrafiltration was replenished by MQ. The PLF was frozen and freeze-dried.

Preparation of pea albumin fraction

The pea albumin fraction (PAF) was isolated by grinding approximately 500 g of yellow peas (Alimex) using a centrifugal mill (Retsch ZM 200, Haan, Germany). The flour was suspended at (20 %, w/w) in MQ. The pH of the suspension was adjusted to 8.0. Afterwards the sample was centrifuged (38,400 x *g*, 15 min., 20 °C) and the obtained supernatant was adjusted to pH 4.5. The dispersion at pH 4.5 was centrifuged (38,400 x *g*, 15 min., 20 °C) and the supernatant was dialysed using an ultra-filtration system with a 10 kDa cut-off, whilst stored on ice. The retentate was subsequently frozen, freeze-dried and labelled pea albumin fraction (PAF). Prior to centrifugation, the samples were stirred for 3 hours at room temperature and the pH of the samples was checked regularly and adjusted to the desired pH if necessary.

Preparation of cultivar extracts

Protein was extracted from eight different pea varieties. From two varieties (Miranda and Montana) seeds were included from two different harvest years. Approximately 8 – 10 g of peas were ground using a centrifugal mill (Retsch ZM 200, Haan, Germany). The flour was suspended (20 %, w/w) in MQ containing 2 % SDS. The pH of the suspension was adjusted to 8.0, and the samples were stirred for 3 hours at room temperature (RT). The pH of the samples was checked regularly and adjusted if necessary. Afterwards the samples were centrifuged (38,400 x *g*, 15 min., 20 °C). The supernatants were dialysed against demineralised water using slide-a-lyzers (Thermo Scientific) with a 10 kDa cut-off and subsequently frozen and freeze-dried (**Table 3.2**).

Table 3.2. List of protein extracts and samples including their code and Centrum voor Genetische Bronnen Nederland (CGN) number, if applicable.

Code	Extracts from CGN peas	CGN-number	Code	Samples non CGN peas
LKE	Lente Krombek	CGN02949	YPE	yellow pea extract
VGKE	Vroegste Gele Krombek	CGN02950	PPC	pea protein concentrate
VLE	Venlosche Lage	CGN02962	PLF	pea legumin fraction
BelE	Belinda	CGN10266	PVF	pea vicilin fraction
Mir89E	Miranda 1989	CGN10292	PAF	pea albumin fraction
Mir20E	Miranda 2020	CGN10292		
PalE	Paloma	CGN10296		
FlaE	Flavandra	CGN13290		
Mon20E	Montana 2020	CGN24055		
Mon06E	Montana 2006	CGN24055		

Total nitrogen content

The total nitrogen content was determined in triplicate using the Dumas method (Flash EA 1112 N analyzer, Thermo Scientific), according to manufacturer's protocol. Methionine was used as standard for the nitrogen quantification. For the pea protein extracts, PPC, PLF and PVF a nitrogen conversion factor of 5.4 was used. This was calculated from the average nitrogen conversion factor of the following pea protein genetic variants: legumin A (P02857, Uniprot Database), legumin J (P05692, Uniprot Database), legumin A2 (P15838, Uniprot Database), legumin K (P05693, Uniprot Database), legumin B (P14594, Uniprot Database), and vicilin (P13918, Uniprot Database) [35]. For the PAF a nitrogen conversion factor of 6.22 was used, assuming only albumin 2 (P08688, Uniprot Database) to be present in the sample. This protein content of the samples was calculated assuming all nitrogen originated from protein. The signal peptide was not included in any of the sequences used. In addition it was assumed that there were no post-translational modifications to the proteins.

The protein recovery for extracts, concentrate and isolated fractions was calculated according to **equation 1**:

$$Protein\ recovery_{Dumas} = \frac{Protein\ in\ sample\ (g)}{Protein\ in\ flour\ (g)} \times 100\ \% \qquad \text{(Eq. 1)}$$

Protein composition using SDS-PAGE

The protein composition of the samples was determined using SDS-PAGE in the presence and absence of a reducing agent. The samples were diluted to 3 g L^{-1} and analysed according to the manufacturer's protocol. The samples were applied to gels (any kD™, Mini-protean TGX precast protein gels, Bio-Rad Laboratories), and separated on a Miniprotean II system (Bio-Rad Laboratories). The proteins were visualized by staining with Coomassie blue stain (InstantBlue, Expedeon). The gels were scanned and analysed using a densitometer (GS-900™, Bio-rad

laboratories) and Image Lab software (Bio-Rad laboratories). Under reducing conditions the following bands were annotated: ~93 kDa lipoxygenase [36], ~70 kDa convicilin, ~50 kDa vicilin, ~38-40 kDa legumin acidic polypeptide, ~33 and 30 kDa vicilin αβ and βγ fragments [34], ~ 26 kDa albumin 2 [37], ~19-22 kDa legumin basic polypeptide, and ~19, 16 and 13.5 kDa vicilin α, β and γ fragments [34]. Legumin basic polypeptides and vicilin fragments were differentiated from one another, by comparing the gels under reducing and non-reducing conditions. Under non-reducing conditions legumin was present as a monomer consisting of an acidic and basic polypeptide chain, therefore bands of ~57 – 62 kDa were ascribed to legumin [34]. The intensity of all unidentified bands was summed and the total was referred to as "other proteins". The relative protein composition was determined from the optical density (OD), by dividing to OD of the protein of interest by the total OD in a lane. The relative composition under reducing and non-reducing conditions was averaged. SDS-PAGE analysis was also performed on the PPC, PLF, PVF, PAF, YPE after dithiothreitol (DTT) incubation and after the TFA addition on the supernatant (with and without SPE). The SDS-PAGE analysis was performed under reducing conditions as described above.

Detection of glycosylated protein using periodic acid – Schiff's reagent staining

The presence of glycosylated proteins was determined under non-reducing conditions. All samples were dissolved at approximately 2 g L^{-1} protein in MQ containing 2 % SDS. Horseradish peroxidase and soybean trypsin inhibitor were used as a positive and negative control, respectively. The controls were dissolved at 2 g L^{-1} protein in MQ. The proteins were separated using SDS-PAGE as described above. The gels were stained using a glycoprotein staining kit containing periodic acid – Schiff's reagent according to the manufacturer's protocol (Thermo Scientific).

Enzymatic protein hydrolysis

The freeze-dried protein extracts were dissolved at 1.0 % (w/v) in 10 mL milli-pore water, adjusted to pH 8.0 and equilibrated for 30 min at 40 °C. The freeze-dried BLP was dissolved at 0.05 mg μL^{-1} of which 30 µL was added to start hydrolysis. The enzymatic hydrolysis was performed in duplicate for 2 h in a pH-stat device (Metrohm, Herisau, Switzerland). This device was used to keep the pH constant by titration of 0.2 M NaOH. Samples of 200 µL were taken before addition of the enzyme and after 10, 30 and 120 minutes of hydrolysis. The enzymatic hydrolysis was stopped by lowering the pH by addition of 20 µL mL^{-1} hydrolysate of 5 M HCl and changing the pH back after 10 min with 20 µL mL^{-1} hydrolysate of 5 M NaOH. Afterwards, the samples were stored frozen at -20 °C. The degree of hydrolysis was calculated according to **equation 2**.

$$DH_{stat}[\%] = V_b \times N_b \times \frac{1}{\alpha} \times \frac{1}{m_p} \times \frac{1}{h_{tot}} \times 100\ \% \qquad \text{(Eq. 2)}$$

where V_b [mL] is the volume of added NaOH; N_b [mol L^{-1}] is the normality of NaOH; α is the average degree of dissociation of the α-NH group (1/α=1.257 at 40 °C and pH 8 [32]); m_p [g] is

the amount of protein in solution; h_{tot} [mmol g^{-1}] is the number of peptide bonds per gram of protein. h_{tot} [mmol g^{-1}] was calculated using the protein composition from SDS-PAGE to be 8.69 for PPC and the extracts, 8.74 for PLF, 8.68 for PVF and 8.77 for PAF.

Sample preparation for RP-UHPLC-MS

The hydrolysates were diluted (1:1) with a 100 mM Tris-HCl buffer at pH 8.0, containing 20 mM DTT and incubated for minimally 2 hours at RT to reduce the disulphide bonds. Afterwards, part of the incubated sample was further processed with solid phase extraction (see section below) and part was used as is. The incubated PLF and PVF (one replicate) were mixed in ratios of 90:10, 75:25, 50:50, 25:75, 10:90 (v/v). The samples and mixtures were acidified by addition of 40 µL of 10 % TFA per mL incubated hydrolysate, which lowered the pH to 1-2. The samples were centrifuged for 10 min at 14,000 *x g* prior to injection. The PLF, PVF, PAF, PPC (hydrolysis in duplicate) and mixtures of PLF and PVF (originating from one hydrolysis) were injected twice. The hydrolysates of the different cultivars (hydrolysis in duplicate) were injected once.

Solid Phase Extraction (SPE)

Solid phase extraction (SPE) was performed using Sep-Pak C18 columns according to the manufacturer's protocol (Waters). The Sep-Pak C18 columns were washed 3 times with 1 mL 50 % acetonitrile in MQ and subsequently 3 times with 1 mL MQ, prior to loading the sample. Approximately 1 mL of the samples was loaded onto the columns. The impurities in the samples were removed by washing 3 times with 1 mL MQ and afterwards 3 times with 1 mL 3 % acetonitrile in MQ. The peptides were removed from the column by washing with 1 mL 50 % acetonitrile in MQ. The acetonitrile solution was evaporated under a N_2-flow. The dried samples were solubilized in 250 µL MQ using ultrasonication for 10 minutes.

Reverse phase ultra-high performance liquid chromatography (RP-UPLC)

The hydrolysates were analysed on the Acquity Premier UPLC equipped with a PDA. A gradient was applied of two mobile phases: Eluent A, containing UPLC-grade water with 1% acetonitrile (ACN) + 0.1% TFA and eluent B, containing ACN with 1% water + 0.1% TFA. The gradient was 0-2 min isocratic on 3 % B; 2-10 min linear gradient from 3-22 % B; 10-16 min linear gradient 22-30 % B; 16-21 min linear gradient 30-100 % B; 21-26 min isocratic on 100% B; 26-28 min linear gradient 100-3 % B and 28-32 min isocratic on 3 % B. The peptides were separated on the Acquity Premier peptide column, BEH C18 2,1*150 300 A 1.7 µm, with a flowrate of 350 µL min^{-1}. The injected volume was 4 µL. The PDA was used to scan the UV absorbance at fixed wavelength of 214 nm at 1.2 nm resolution and 40 scans/second.

Electro spray ionisation time of flight mass spectrometry (ESI-Q-TOF-MS)

The mass spectra (50-3000 *m/z*) were collected with the Select Series Cyclic IMS operating in time-of-flight and V-mode (Waters, Milford, MA, USA). The peptides were ionized in the electrospray ionization source with a capillary voltage of 2.5 kV and a source temperature of 150 °C. The sample cone was operated at 40 V and nitrogen was used as desolvation gas (500 °C, 800

L h^{-1}) and cone gas (200 L h^{-1}). Online lock mass data were acquired by infusing 10 µL min^{-1} of 50 pg $µL^{-1}$ Leucine-Enkephalin via the Waters LockSpray at a capillary voltage of 2.7 kV. The quadrupole was operated using the automatic quad profile. The collision energy applied in the trap was 6 V for the MS, and ramped up in the MS^E method from 28 to 56 V for MS/MS. The cyclic ion mobility cell was not used in this experiment. The collision energy in the transfer was 4 V. Prior to analysis, the TOF-analyzer was calibrated up to 4000 *m/z* using sodium iodide.

Data processing - Peptide identification

Identification of the peptides was performed in UNIFI software version 1.8 according to the suggested settings in **Chapter 2**. The amino acid sequences of the proteins, so without the signal peptide, in **Table 3.1** were used in UNIFI. All different genetic variants were inserted as unique proteins. Post-translational modifications as oxidation, glycosylation and phosphorylation were not reported for these proteins on Uniprot and therefore not included in this analysis. First, a BLP specific analysis was performed on PPC, PLF, PVF, PAF, PPC and YPE with all potential proteins to evaluate their presence. Protein variants that showed no unique peptides were excluded (albumin variants B, C, D, F, provicilin A and convicilin B). This was done to reduce the number of non-unique sequences. Afterwards, a semi-specific analysis was performed on all samples. The semi-specific analysis included peptides of which either the peptide bond on the C- or on the N-terminal side that was hydrolysed did not match the specificity of BLP (assumed specificity for glutamic acid + aspartic acid). The semi-specific analysis method is essential for good coverage of proteins that naturally occur as linked poly-peptide fragments as vicilin. The protein variants were inserted in the same order as **Table 3.1**.

The processing parameters were set based on the guideline in **Chapter 2**. In the peak detection, all *m/z* signals with an intensity above 1000 detector counts were processed, and all MS/MS signals with an intensity of more than 250 detector counts were processed. Peptides were annotated with a maximal acceptable mass error of 10 ppm. After processing, peptides were excluded that did not meet the criteria for MS/MS fragmentation as set in **Chapter 2**. The average limit of detection was determined for the Select Series Cyclic IMS with a dilution series of a tryptic hydrolysate of α-LA with concentrations 2.5 mg L^{-1} to 5 g L^{-1}. The average MS intensity in the lowest dilution at which the parent ion *m/z* was detected was 1.6 * 10^4 Counts (LOD-lowest level of detection). Peptide annotations were excluded when the MS intensity was below the limit of detection, considering that at this intensity no clear MS/MS spectrum is acquired. In-source fragments, recognised in UNIFI or in PeptQuant, and adducts from water or ammonium were also removed (In this case all peptides annotated > LOD were included in analysis). As shown before a small part of these peptides may not be reproducibly identified in repeated analyses **Chapter 2**. In the current study, all samples were injected in duplicate, minimizing possible errors in obtained quantification.

Data processing - Peptide quantification

The absolute concentration of peptides was measured by analysis of the UV absorbance at 214 nm and the molar extinction coefficient of the particular peptide, predicted from Kuipers *et al.*, 2007 [9]. The UV peak areas were integrated in Masslynx version 4.2, with the integration settings as described in **Chapter 2**. The UV peak areas that were originating from the Tris and DTT were removed from the list. The list of peak areas was coupled to the peptide list, taking into account the time offset between PDA and MS (0.08 min) using PeptQuant, an in-house developed Matlab script. The concentration of the peptide was calculated with **equation 3**.

$$C_{peptide\ [\mu M]} = \frac{A_{214} \cdot Q}{\varepsilon_{214} \cdot l \cdot V_{inj} \cdot k_{cell}} \quad \text{(Eq. 3)}$$

where A_{214} [µAU·min] is the UV peak area at 214 nm, V_{inj} [µL] is the volume of sample injected, Q [µL min^{-1}] is the flow rate and l [cm] is the path length of the UV cell, which is 1 cm according to the manufacturer. The molar extinction coefficient ε_{214} [L Mol^{-1} cm^{-2}] for each peptide was calculated according to Kuipers *et al.* [9]. The cell constant, k_{cell} for the UV detector was 0.78 (**Chapter 2**). In case multiple peptides were assigned to the same UV peak, the corresponding area was divided based on the MS intensities and molar extinction coefficients of both peptides **Chapter 2**.

Data processing - Protein quantification

The peptide concentrations were used to calculate the concentration of each protein (variant). To do this, first the concentration of each amino acid occupying a unique position of the protein sequence (unique amino acid) was calculated by summation of the peptide concentrations containing that unique amino acid. The concentration of each unique amino acid was plotted against the sequence of the protein (see results section for an example). If all peptides were completely included in the analysis, the unique amino acid concentrations would be identical for each amino acid in the protein sequence and equal to the initial protein concentration. Typically, variations were observed in the calculated concentrations of unique amino acids. Therefore, three different calculation methods (I-III) were used to calculate the concentration of a protein:

I Concentrations of all unique amino acids were averaged.
II All quantified concentrations of unique amino acids > 0 µM were averaged.
III Averaging the concentrations of all unique amino acids with C>average of II.

Tools to analyse the completeness of peptide identification and quantification

The completeness of peptide identification was analysed by calculating the amino acid sequence coverage, also known as protein sequence coverage in proteomics [38]. This parameter describes how many of the unique amino acids in a certain protein are covered in at least one of the identified peptides (**Equation 4**).

$$Amino\ acid\ sequence\ coverage\ [\%] = \frac{\#\ unique\ annotated\ amino\ acids}{\#\ amino\ acids\ in\ protein\ sequence} \cdot 100\ \%$$ (Eq. 4)

The completeness of quantification was roughly estimated by the UV recovery, which was calculated by dividing the expected amount of UV by the total UV in the chromatogram. The expected amount of UV area was calculated with **equation 3**, with a correction of the molar extinction coefficient for broken peptide bonds during hydrolysis and the protein concentrations based on protein content and the estimated composition based on SDS-PAGE. To assess the completeness of quantification of each unique amino acid, plots were made of the sum of the absolute peptide concentration involving a certain unique amino acid residue.

Results & discussion

Pea flours and derived fractions: protein content, composition and losses during extraction

The protein contents of the prepared pea flours were: 17.6 ± 1.4 % (w/w, on sample as is, **Table 3.3**). The protein extracts had protein contents of 58.9 ± 1.0 % (w/w, on sample as is). The protein contents of the pea protein concentrate (PPC), pea legumin fraction (PLF), pea vicilin fraction (PVF) and pea albumin fraction (PAF) were higher than those of the crude protein extracts from the different cultivars: 71.6 – 87.3 % (w/w, on sample as is, **Table 3.3**). For the different cultivars, the extracted protein represented 58 – 60 % of the total amount of protein in the original sample (**Table 3.3**). Other authors report similar extractabilities using a Tris-HCl buffer at pH 8.0 (60 ± 7 % [1]). The first challenge in analysis of the protein composition is that ~40 % of the proteins in the pea flour are actually not extracted and therefore not analysed. PAS-staining SDS-PAGE gels did not indicate any glycosylated proteins in the extracts, concentrate and fractions (**Annex 3.1**). The protein composition of the extracts as analysed by SDS-PAGE stained with Coomassie blue stain showed small differences in the legumin:vicilin ratio (L:V) 28:72 – 41:59 (w/w, **Table 3.3, Annex 3.3**), but no other differences in presence or absence of specific proteins (**Figure 3.1**). The L:V ratio in PPC, PLF and PVF were 28:72, 94:6 and 6:94 (w/w), respectively. Comparable L:V ratios were obtained using size-exclusion chromatography (UV_{214}): PPC 44:56, PLF 100:0, PVF 16:84 (results not shown).

BLP hydrolysis

The BLP hydrolysis of PPC, PLF, PVF reached a degree of hydrolysis of respectively 6.7 ± 0.2 %, 6.7 ± 0.2 % and 7.6 ± 0.3 %. The hydrolysis of PAF yielded a lower degree of hydrolysis of 4.1 ± 0.2 %. For the extracts the final DH was 6.9 ± 0.5 %, which indicated that all extracts were hydrolysed to the same extent. The obtained degrees of hydrolysis were 64-85 % of the expected value based on the percentage glutamic acid residues in the protein sequences (10.5 % for legumin A1, 10.2 % vicilin and 4.8 % albumin 2). This was in line with the hydrolysis efficiencies observed by Butré *et al.* for BLP with dairy proteins [10]. SDS-PAGE of the protein hydrolysates showed the presence of intact protein after digestion (**Annex 3.4**).

Table 3.3. Yield (%), total protein recovery and legumin:vicilin ratios of protein extracts, concentrate and fractions. Protein content (% w/w, on sample "as is") including standard deviations of pea flours and resulting extracts, concentrate and fractions.

Code	Yield (%)[1]	Total protein recovery (%)[2]	Protein content pea flour (%, w/w "as is")	Protein content extract / concentrate / fraction (%, w/w "as is")	Legumin:vicilin ratio (w/w)
PPC	11	46	17.3 ± 0.2	71.6 ± 0.5	28:72
PLF	2	10	17.3 ± 0.2	84.2 ± 0.9	94:6
PVF	5	21	17.3 ± 0.2	72.0 ± 1.4	6:94
PAF	1	5	17.3 ± 0.2	87.3 ± 0.5	38:62
BelE	22	67	19.3 ± 0.2	59.3 ± 0.7	31:69
FlaE	22	70	18.8 ± 0.1	59.5 ± 0.4	36:64
Mon20E	17	62	16.1 ± 0.5	58.4 ± 0.5	28:72
Mir20E	18	63	16.2 ≤ 0.0	58.6 ± 0.6	33:67
LKE	19	63	18.5 ± 0.2	60.3 ± 0.6	38:62
PalE	18	61	18.1 ± 0.3	60.2 ± 0.3	31:69
VLE	23	70	18.8 ± 0.2	58.5 ± 0.4	31:69
Mon06E	16	60	15.8 ± 0.4	58.9 ± 0.4	31:69
YPE	14	50	17.3 ± 0.2	57.3 ± 0.4	31:69
Mir89E	16	59	15.7 ± 0.1	57.4 ± 1.0	31:69
VGKE	23	70	19.6 ± 0.2	59.9 ± 0.9	41:59

[1] Powder (g) / flour (g) * 100.

[2] Protein in sample (g) / protein in flour (g) * 100.

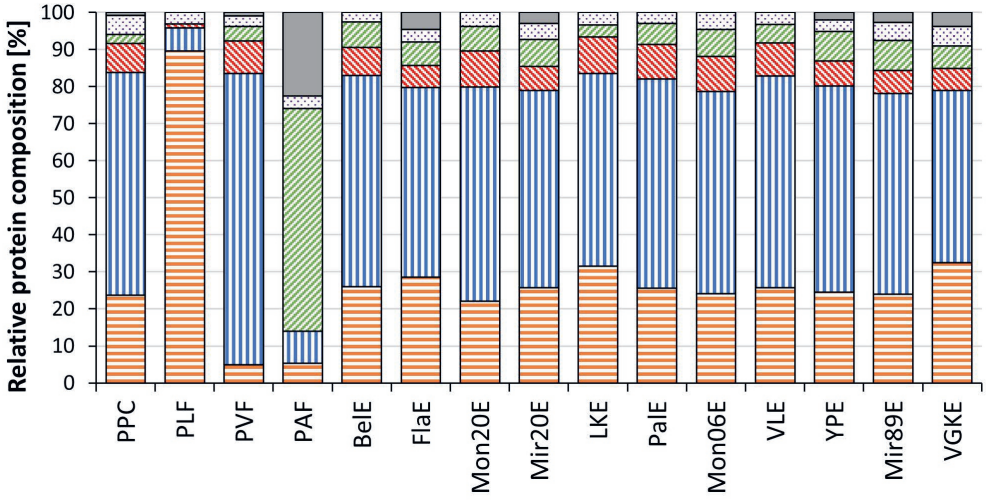


Figure 3.1. Relative protein composition (w/w, %) based on densitometry of SDS-PAGE gels showing, legumin (▭), vicilin (▯), albumin (▨), convicilin (▩), lipoxygenase (▨) and other proteins (▭). Values are an average of the results under reducing and non-reducing conditions, with an average standard deviation 1 % and a maximum standard deviation of 6 %.

Quantification of the protein genetic variants and challenges concerned

Peptide losses during sample preparation

The second challenge was the loss of peptides during sample preparation. The total UV peak areas in the chromatograms of the PPC, PLF, PVF and PAF were 68.8 ±1 %, 75.5 ±6 %, 70.8 ±1 %, and 38.0 ±1.0 % of the expected amounts of UV based on protein content and composition, respectively (**Table 3.4**). The chromatograms of the extracts of different cultivars represented 31 ± 5 % of the expected total UV. This means that only part of the extracted protein was included in the analysis. In our previous study on tryptic digests of milk protein isolates, the observed amount of UV absorbance was equal to the expected amount **Chapter 2**.

Table 3.4. Amount of integrated UV_{214} area in chromatograms, mean and standard deviations over replicates.

	Expected[1] UV area (·10 ^{5} AU x min)	**Total UV (·10 ^{5} AU x min)**	**Annotated UV peak area (·10 ^{5} AU x min)**	**Total UV / Expected UV (%)**	**Relative Annotated UV area (%)**	**UV recovery: Annotated area / expected area (%)**
PPC	3.7	2.5 ± 0.0	1.8 ±0.4	69 ± 0.6	71 ± 15	49 ± 11
PLF	4.9	3.7 ± 0.3	3.1 ±0.3	76 ± 6.2	85 ± 7	64 ± 7
PVF	3.7	2.6 ± 0.0	2.2 ±0.1	71 ± 0.4	83 ± 2	59 ± 1
PAF	4.6	1.8 ± 0.1	1.0 ±0.1	38 ± 1.0	58 ± 5	22 ± 2
PLF:PVF_90:10	4.8	3.4 ± 0.0	2.9 ±0.1	71 ± 0.4	86 ± 1	60 ± 1
PLF:PVF_75:25	4.6	3.1 ± 0.1	2.4 ±0.1	68 ± 1.9	77 ± 2	52 ± 3
PLF:PVF_50:50	4.3	2.9 ± 0.0	2.2 ±0.0	67 ± 0.2	78 ± 1	52 ± 1
PLF:PVF_25:75	4.0	2.5 ± 0.0	1.9 ±0.1	64 ± 0.2	77 ± 2	49 ± 1
PLF:PVF_10:90	3.8	2.3 ± 0.0	1.7 ±0.0	60 ± 0.4	76 ± 2	46 ± 1
BelE	3.2	0.9 ± 0.0	0.7 ± 0.0	27 ± 0.1	75 ± 0	20 ± 0
FlaE	3.1	1.3 ± 0.0	1.0 ± 0.0	42 ± 0.1	76 ± 0	32 ± 0
Mon20E	3.1	1.0 ± 0.1	0.8 ± 0.1	33 ± 2.9	75 ± 1	24 ± 2
Mir20E	3.0	0.9 ± 0.0	0.7 ± 0.0	30 ± 0.5	73 ± 1	22 ± 1
LKE	3.2	0.9 ± 0.1	0.8 ± 0.0	29 ± 1.5	82 ± 1	24 ± 1
PalE	3.2	0.9 ± 0.0	0.7 ± 0.0	26 ± 0.7	76 ± 2	20 ± 0
VLE	3.1	1.0 ± 0.1	0.8 ± 0.0	33 ± 3.0	79 ± 1	26 ± 3
Mon06E	3.1	1.0 ± 0.1	0.8 ± 0.0	33 ± 2.9	77 ± 4	25 ± 1
YPE	3.0	0.7± 0.0	0.6 ± 0.0	24 ± 0.6	78 ± 1	19 ± 1
Mir89E	3.0	0.9 ± 0.1	0.6 ± 0.1	29 ± 4.6	74 ± 0	21 ± 3
VGKE	3.1	1.0 ± 0.0	0.8 ± 0.1	32 ± 1.2	81 ± 2	26 ± 2

[1]Expected amount was calculated with protein content, estimated molar extinction coefficient based on (SDS-PAGE) protein composition and correction for degree of hydrolysis.

Typically, in (quantitative) proteomics studies, the recovery of injected protein material is not described. An exception, Wang *et al.* reported also low recoveries of 18-60 % for plant protein extracts (barley leaves), dependent on the sample preparation procedure for proteomics analysis [39]. The low UV recoveries observed in the current study were attributed to insoluble

aggregates formed when changing the pH to eluent conditions, which were visible as turbidity and then removed by centrifugation. To try and avoid this problem, samples were also prepared by applying solid phase extraction at the pH 8 (after reduction of the disulfide bonds), but the SPE treatment did not improve the UV recovery. The observed UV peak areas ranged between 14 - 53 % of the expected UV absorbances. The UPLC-MS data of the same samples with and without SPE treatment did not show changes in *m/z* peaks and ion intensities. Therefore, further analyses were all performed on the dataset without SPE treatment.

Peptide identification in the PLF, PVF, PAF, PPC and YPE

Of the UV peaks that were present in the chromatograms, on average 77 ±8 % was attributed to peptides (**Table 3.4**). The highest matched UV was observed for PLF (91 %) and the minimum was observed for PAF (54 %). The matched UV for the extracts varied between 73 - 82 %. UV areas that were not matched with peptide sequences were mostly from remaining intact proteins and phenolic compounds. The number of identified peptides was 301 ±8 in the PLF, 293 ±7 in the PVF, 98 ±9 in the PAF, 264 ±37 in the PPC and 186±9 in the YPE. For 78 ±3 % of these peptides the MS intensity was above the average limit of annotation, which was previously used to indicate annotations with high repeatability in **Chapter 2**. To be as complete as possible for this study we also included the peptide annotation < limit of annotation (LOA, but these were all still >limit of detection (LOD) and confirmed with sufficient MS/MS fragments). Between replicate injections of PLF, PVF and PAF, 86 ±2 % of all the peptides were annotated similarly between replicates. Between duplicate hydrolyses of the same fractions, 85 ±3 % of the peptides were annotated similarly. This means that the variation in peptide composition between hydrolysates of replicate digestions, did not exceed the variation between replicate measurements of the same hydrolysate. The repeatability for duplicate injections in this study was higher (86 %) than repeatability at peptide level observed in proteomics studies, where typically 35-79% were similarly annotated between duplicate injections [40-43].

Completeness of peptide identification

The peptides identified for each protein genotype were visualised against the sequence of the protein, as illustrated with legumin A1 in PLF (**Figure 3.2**). In some cases, part of the protein sequence was annotated in multiple peptides. For example, amino acid Leucine on position 1 for legumin A1 was present in peptides 1-3 and 1-9. In other cases, part of the sequence was not covered by any of the annotated peptides, e.g. legumin A1: 362-397. The protein sequences of the most abundant proteins in the fractions were covered with high amino acid sequence coverages, respectively 91 ±4 % for legumin A1 in the PLF, 80 ±3 % for vicilin in the PVF and 97 ±1 % for albumin 2 in the PAF (**Table 3.5**). Substantial sequence coverages for the legumin variants A2, B, J and K (28 - 69 %) confirmed that legumin was present in different genetic variants. The peptides identified in PVF yielded amino acid sequence coverages of 80 ±3 % for vicilin, 62 ±4 % for provicilin and 53 ±3 % for convicilin. For the two protease inhibitors that were included in the analysis the coverages were relatively low in all purified fractions (≤ 20 %). The amino acid sequence coverages observed in this study were lower than coverages reported in previous studies with hydrolysates of 1 to 3 milk proteins (amino acid sequence coverages of 99-

100 %) [10] (**Chapter 2**), analysed with the same procedure. The amino acid sequence coverages were logically affected in the pea hydrolysates by the observed peptide losses in sample preparation. Furthermore, sample complexity could be relevant (higher number of peptides in similar injection volume), leading to lower concentrations of individual peptides. For instance, the coverage of legumin A1 in YPE (34 ±3 %), is ~ 1/3 of the coverage of legumin A1 in the purified legumin fraction from the same pea cultivar (PLF, 91 ±4 %).

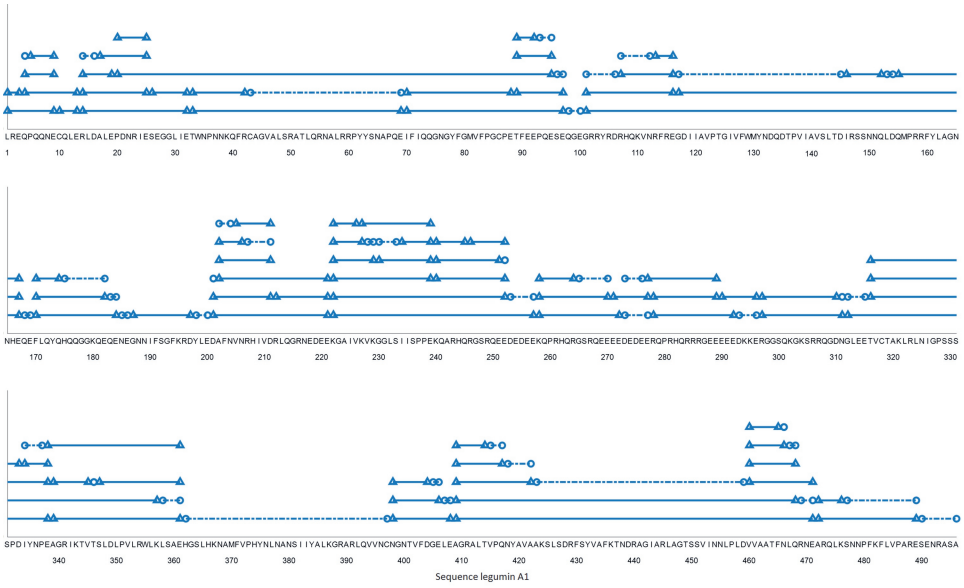


Figure 3.2. Peptides of legumin A1 identified in PLF, visualised against the protein sequence of legumin A1. Dotted lines indicate missing peptides.

Table 3.5. Amino acid sequence coverages (%) + standard deviation of pea proteins identified in PPC, PLF, PVF, PAF and YPE.

Protein	PPC	PLF	PVF	PAF	YPE
Legumin A1	53 ± 21	91 ± 4	47 ± 10	3 ± 1	34 ± 3
Legumin A2	14 ± 2	28 ± 2	9 ± 4	3 ± 0	2 ± 2
Legumin B	41 ± 8	66 ± 3	39 ± 11	2 ± 1	29 ± 1
Legumin J	30 ± 4	38 ± 10	13 ± 3	2 ± 1	24 ± 3
Legumin K	41 ± 8	69 ± 8	34 ± 7	2 ± 1	25 ± 4
Vicilin	71 ± 9	56 ± 5	80 ± 3	1 ± 0	55 ± 2
Provicilin	48 ± 24	48 ± 9	62 ± 4	14 ± 16	27 ± 0
Convicilin	50 ± 2	44 ± 4	53 ± 3	3 ± 2	35 ± 2
Albumin A1	46 ± 0	12 ± 2	71 ± 0	67 ± 6	46 ± 0
Albumin 2	47 ± 12	16 ± 15	36 ± 6	97 ± 1	31 ± 3
IBBB	0 ± 0	0 ± 0	5 ± 0	3 ± 0	3 ± 0
PIP20	5 ± 0	10 ± 8	20 ± 12	13 ± 16	3 ± 0

Peptide and protein quantification in the purified fractions

All peptides in the PLF, PVF and PAF were quantified based on their UV absorbance and predicted molar extinction coefficient, which yielded a wide variety of individual peptide concentrations. For instance, the different peptides originating from legumin A1 in PLF had concentrations ranging between 0.4 nM (for peptide 258-264) to 55 µM (for peptide 26-32). This implies that using one of these quantified peptides as reference for the protein concentration, as done often with isotopically labelling, would yield very different protein concentrations dependent on the reference peptide chosen. Therefore, to tackle this challenge, we choose to sum all peptide concentrations to determine the protein concentration. To visualise this way of protein quantification, the amino acid concentrations for each position of the sequence were plotted against the amino acid sequence of each protein (**Figure 3.3**). For legumin A1 in PLF, the standard deviation of the observed amino acid concentration in four replicates was 4%. The average amino acid concentration was 25 ±3 µM over the four replicates. In absence of losses during sample preparation, the retrieved amino acid concentration would be similar to the (molar) concentration of protein before hydrolysis. The composition of the PLF, PVF and PAF were determined using the average amino acid concentration as indication for the protein concentrations (calculation I) (**Figure 3.4**). With this calculation, minor differences were observed between replicate analyses and major differences in the relative protein composition for the different fractions were observed. The PLF consisted for 87 ±5 % of the legumins, the PVF had 62 ±1 % of vicilin, provicilin and convicilin and the PAF had 72 ±5 % of albumin 2.

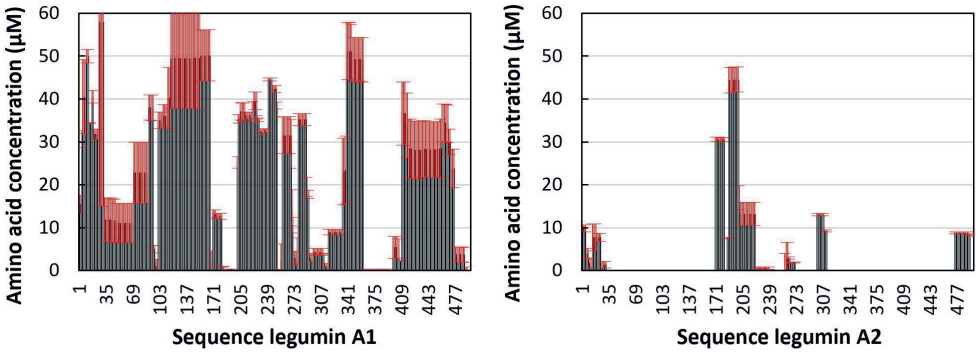


Figure 3.3. The unique amino acid concentration (µM) + standard deviation (µM, in red) for legumin A1 (left) and A2 (right) in PLF, calculated with the peptide concentrations including that respective amino acid.

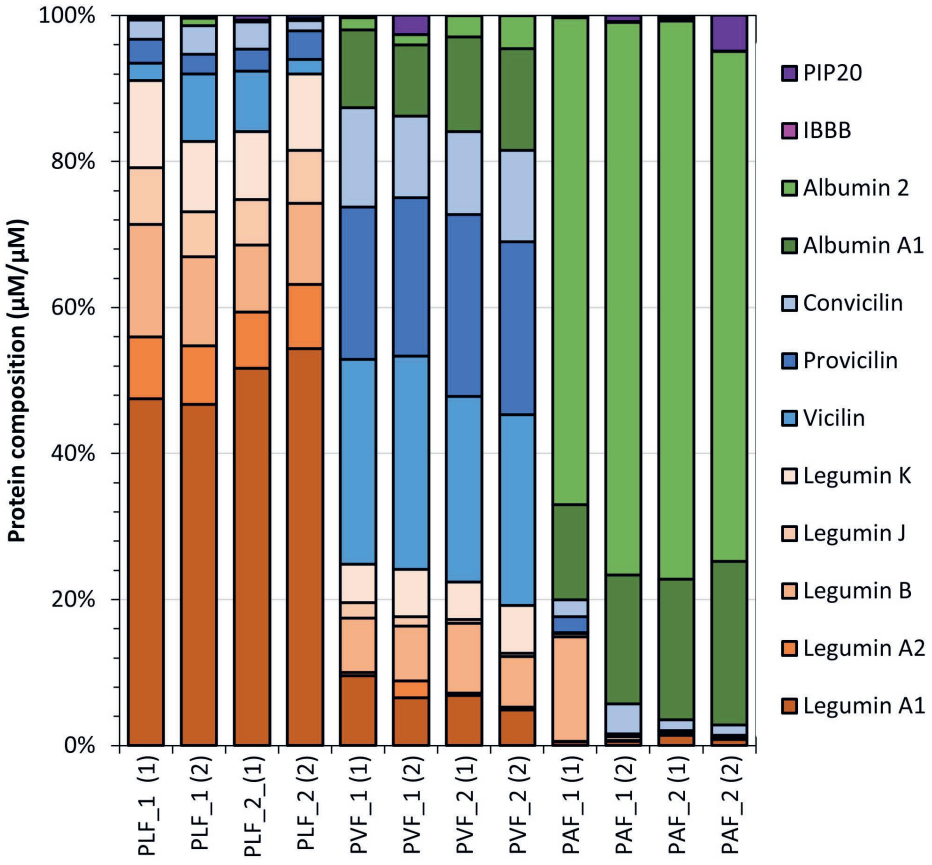


Figure 3.4. Composition µM/µM (%/%) of the PLF, PVF and PAF determined with calculation I. Reproducibility was tested in duplicate hydrolyses (first number) and duplicate injections (second number).

Correcting protein concentrations for missing peptide information

Generic peptide sequences

The third challenge was that some peptide sequences can occur in several protein variants. This could lead to overestimation of one variant, and underestimation of the other variant in protein quantification. For the 134 peptides annotated to legumin A1 or A2 in PLF, only 1 peptide could also originate from the sequence of legumin B, J or K. Out of these 133 peptides unique for legumin A, 29 were unique for genetic variant A1, 20 were unique for genetic variant A2 and 75 peptides could originate from either legumin A1 or A2. Since these 75 peptides were -in the software- now attributed to legumin A1, the amino acid sequence coverage for A2 had large parts of the sequence where no peptides were attributed to (**Figure 3.3**). Therefore, the concentrations of proteins should ideally be calculated with peptides that were unique for that genetic variant.

How to tackle missing peptide information in calculated protein concentrations

Two additional calculations were evaluated to transform the peptide concentrations into an absolute protein concentration, taking into account the unique sequences per variant (calculation II) and possibly low recovery of part of the sequence (calculation III). Based on protein content and the SDS-PAGE composition, 67 µM of legumins were expected to be present in the PLF and 57 µM of vicilins would be present in the PVF (**Table 3.6**). These values are higher than what was calculated from averaging the amino acid concentrations of quantified peptides; legumins in PLF (43 ±3 µM) and vicilins in PVF (20 ±1 µM). A correction for non-unique sequences, which was done by calculating the concentrations with only peptides that were unique for a certain variant, overestimated the concentration of the minor legumin variants (A2, B, J, K) in the PVF. When the protein concentrations were calculated with all amino acids that were above the average, which was done to exclude the sequences of the protein that were quantified clearly lower than the maximum in the plot, the observed concentration of legumins was almost 2x the (maximum) expected concentration.

Table 3.6. Illustrating the effect of different approaches to convert peptide to protein concentrations (µM) in PLF and PVF based on UPLC-PDA-MS (calculation I-III) and protein composition based on SDS-PAGE.

	SDS-PAGE		**Calculation I (Av)[1]**		**Calculation II (Av > 0)[2]**		**Calculation III (Av>Av)[3]**	
	PLF	**PVF**	**PLF**	**PVF**	**PLF**	**PVF**	**PLF**	**PVF**
Legumin A1	-	-	24.7 ±3.2	2.8 ±0.7	27.1 ±1.3	3.1 ±1.3	41.3 ±3.5	16.3 ±14.2
Legumins (A1, A2, B, J, K)	66.7	3.2	43.0 ±2.8	9.2 ±0.9	67.0 ±4.9	28.3 ±6.4	116.0 ±7.9	27.2 ±2.2
Vicilin	-	-	2.8 ±2.0	11.1 ±0.6	4.8 ±3.3	13.8 ±0.6	12.9 ±3.2	23.1 ±14.1
Vicilins (Vicilin + provicilin)	5.2	57.4	4.4 ±1.9	20.4 ±0.9	8.1 ±3.5	28.9 ±2.7	67.6 ±10.4	51.8 ±2.3

[1] Concentrations of all unique amino acids were averaged.

[2] All quantified concentrations of unique amino acids > 0 µM were averaged.

[3] Averaging the concentrations of all unique amino acids with C > average of II.

Similarly, an unrealistic amount of legumins was observed in the PVF. This error seems to originate from parts of the protein sequence of legumins that were quantified at concentrations far above the (observed) average amino acid concentrations. This was for instance the case for sequences of legumin B: 43-50, 83-90 and 103-107 and legumin A2: 170-182, 187-200. For these peptides the identification might be incorrect. For the LegA2 peptides 170-182 and 187-200 the identifications were confirmed with 19 and 21 MS/MS fragments, respectively, which excludes the possibility of having alternative annotations. For legumin B 103-107, with sequence KEEED, an isobaric alternative assignment would be provicilin DKEEE of 307-311. Leg B peptides 43-50 and 83-90 did not have alternative assignments either. These flaws in the analysis of the peptides, in combination with the losses during sample preparation will affect the absolute

concentrations calculated. However, it is expected that the current flaws will be similar for all samples and have thereby a minor effect on the conclusions on relative differences in genetic composition.

Genetic composition of mixtures of PLF and PVF

The fractions PLF and PVF were mixed in different ratios to evaluate the robustness in the genetic protein composition. For the majority of the calculated amino acid concentrations, the height changed gradually with the amount of the protein in the mixture and the overall the pattern of the plot remained similar (**Figure 3.5**). This means that the majority of the peptides were identified and quantified consistently. Exceptions were for example the legumin B sequences, 83-90 and 103-107 mentioned in the previous section: For these, the maximum concentrations were not observed in the PLF. Regardless of the calculation used, the analysed composition of the 50:50 mixture was a good representative of the purified PLF and PVF fractions. Since the composition of the purified fractions was most reproducible when determined with calculation I, so without correction for the missing peptide information, this calculation was also used to determine the composition of the different pea cultivars.

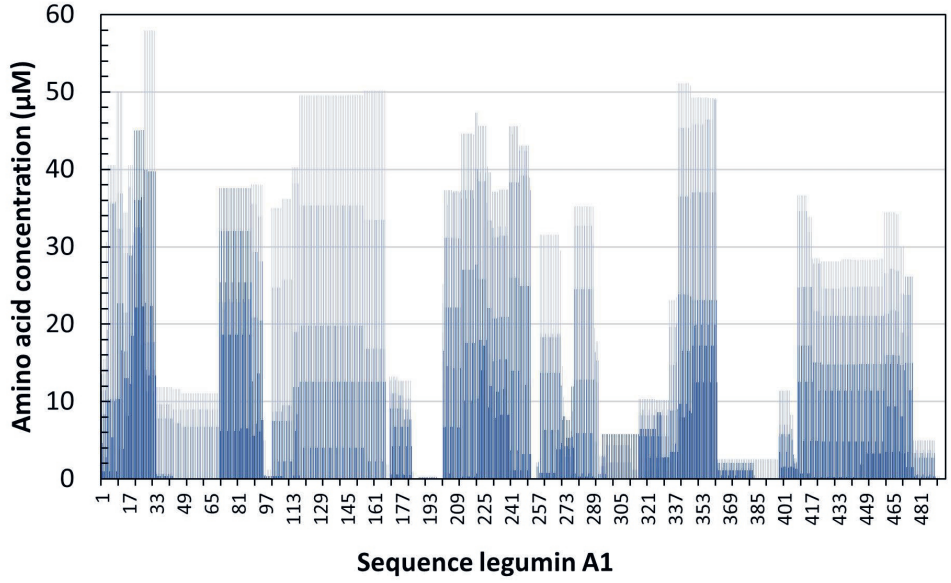


Figure 3.5. Amino acid concentration (µM) observed for legumin A1 for different PLF:PVF ratios: 100:0 ( ), 90:10 ( ), 75:25 ( ) 50:50 ( ) 25:75 ( ), 10:90 ( ) and 0:100 ( ).

Variation in genetic composition of different pea cultivars

For the pea cultivar extracts, approximately 20 ±5 % of the total protein in the peas was analysed by RP-UHPLC-MS, based on the measured UV absorbance, amount of injected sample, and protein contents of the extracts and flours. In total, 15 ±4 % of the total protein in the different pea cultivars was annotated. The protein composition of the analysed part of the different pea cultivar extracts showed no significant differences (**Figure 3.6**).

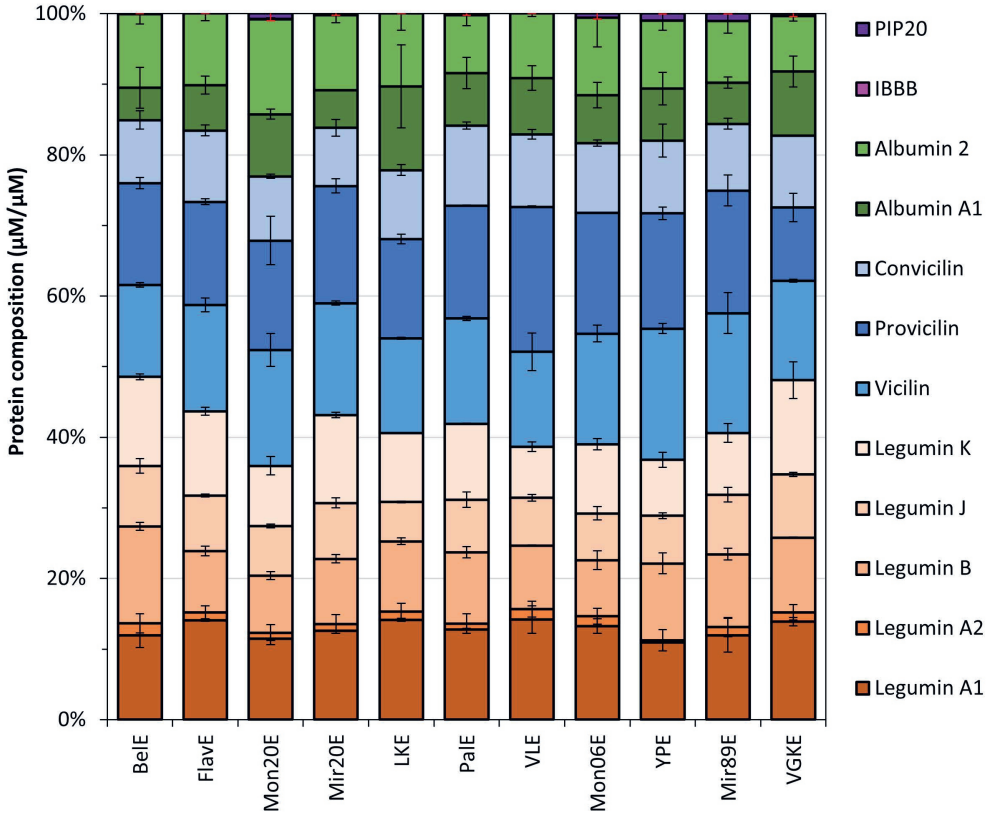


Figure 3.6. Absolute protein composition µM/µM (%/%) on genetic-variant level for extracts of different pea cultivars determined with calculation I.

Based on the calculation using all amino acids the cultivars were composed of 12.8 ±1.2 % legumin A1, 1.1 ±0.4 % legumin A2, 9.9 ±1.6 % legumin B, 7.5 ±1.0 % legumin J, 10.3 ±2.1 % legumin K, 15.2 ±1.7 % vicilin, 15.7 ±2.5 % provicilin, 9.8 ±0.8 % convicilin, 7.4 ±2.0 % albumin A1, 10.0 ±1.5 % albumin 2 and 0.4 ±0.4 % PIP20. The L:V ratio found in these samples was 57:43 % (µM μM^{-1}), whereas approximately 70 % of the samples described in literature and measured with SDS-PAGE has a ratio ranging between 17:83 – 38:62 (w/w %) [1, 2, 44-46]. The ratio determined with UPLC-PDA-MS reflected the differences in recovery of legumins (65 %) compared with vicilin (36 %) in the purified fractions. Legumin A1, A2, B, J and K were present in all samples, and legumin A (A1 + A2) was most abundant, 14.1 ±1.4 %. The other legumin genetic variants occurred in relatively similar quantities in each pea cultivar and the composition (n/n %) between the genetic variants was comparable. Besides IBBB, all the protein genetic variants considered in this study were found in all pea cultivars in similar quantities. A recent study by Burstin *et al.* provided insights into the pea genome [47]. Our study shows that the genes responsible for the production of the proteins considered in this study are expressed in all cultivars in similar quantities.

Conclusion

In this study, a new way of protein quantification was illustrated, in which all peptides in the analysis were used to quantify protein genetic variants. Using UV quantification, we were not limited to a small number of reference peptides, enabling the quantification of all peptides, normally only achieved with MS intensity-based quantification. Analysis of the protein mass balance showed losses during sample preparation and protein extraction. This allowed to describe how much of the original pea protein was described by the analysis. With the approach taken, the high impact of wrong annotations on calculated protein concentrations was identified. Without correcting for the peptide losses, differences in composition were still reproducibly determined for fractions of legumins, vicilins and albumins, as well as their mixtures. For the pea protein extracts from different cultivars this method showed that all considered protein genetic variants were present in similar amounts.

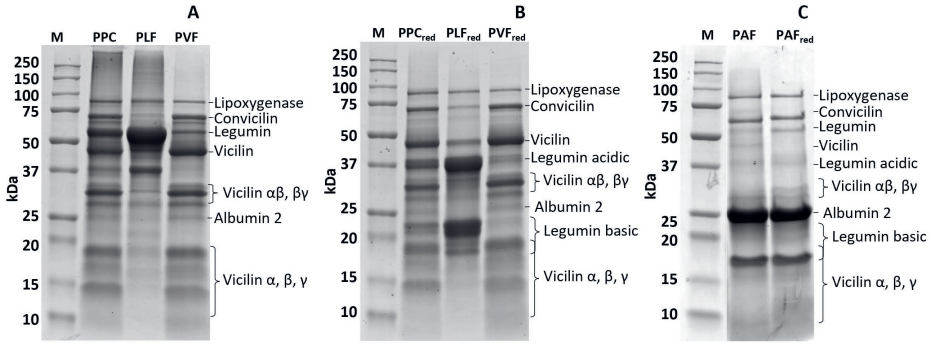
References

1. Barać, M. B., Čabrilo, S., Pešić, M. B., Stanojević, S. P., Žilić, S., Maćej, O., Ristić, N. (**2010**). Profile and functional properties of seed proteins from six pea (*Pisum sativum*) genotypes. *International Journal of Molecular Sciences*, 11, 4973-4990.
2. Tzitzikas, E. N., Vincken, J.-P., de Groot, J., Gruppen, H., Visser, R. G. F. (**2006**). Genetic variation in pea seed globulin composition. *Journal of Agricultural and Food Chemistry*, 54, 425-433.
3. García Arteaga, V., Kraus, S., Schott, M., Muranyi, I., Schweiggert-Weisz, U., Eisner, P. (**2021**). Screening of twelve pea (*Pisum sativum* L.) cultivars and their isolates focusing on the protein characterization, functionality, and sensory profiles. *Foods*, 10.
4. Shewry, P. R., Casey, R. (**1999**). Seed proteins. Seed proteins: Springer, p. 1-10.
5. Thompson, A. J., Bown, D., Yaish, S., Gatehouse, J. A. (**1991**). Differential expression of seed storage protein genes in the pea legJ subfamily; Sequence of gene legK. *Biochemie und Physiologie der Pflanzen*, 187, 1-12.
6. Agregán, R., Echegaray, N., López-Pedrouso, M., Kharabsheh, R., Franco, D., Lorenzo, J. M. (**2021**). Proteomic advances in milk and dairy products. *Molecules*, 26.
7. Houston, N. L., Lee, D.-G., Stevenson, S. E., Ladics, G. S., Bannon, G. A., McClain, S., Privalle, L., Stagg, N., Herouet-Guicheney, C., MacIntosh, S. C. (**2011**). Quantitation of soybean allergens using tandem mass spectrometry. *Journal of Proteome Research*, 10, 763-773.
8. Colaert, N., Vandekerckhove, J., Martens, L., Gevaert, K. (**2011**). A case study on the comparison of different software tools for automated quantification of peptides. Gel-free proteomics: Springer, p. 373-398.
9. Kuipers, B. J. H., Gruppen, H. (**2007**). Prediction of molar extinction coefficients of proteins and peptides using UV absorption of the constituent amino acids at 214 nm to enable quantitative reverse phase high-performance liquid chromatography-mass spectrometry analysis. *Journal of Agricultural and Food Chemistry*, 55, 5445-5451.
10. Butré, C. I., Sforza, S., Gruppen, H., Wierenga, P. A. (**2014**). Introducing enzyme selectivity: A quantitative parameter to describe enzymatic protein hydrolysis. *Analytical and Bioanalytical Chemistry*, 406, 5827-5841.
11. Lau, K. W., Jones, A. R., Swainston, N., Siepen, J. A., Hubbard, S. J. (**2007**). Capture and analysis of quantitative proteomic data. *Proteomics,* 7, 2787-2799.
12. Jensen, H., Poulsen, N., Andersen, K., Hammershøj, M., Poulsen, H., Larsen, L. (**2012**). Distinct composition of bovine milk from Jersey and Holstein-Friesian cows with good, poor, or noncoagulation properties as reflected in protein genetic variants and isoforms. *Journal of Dairy Science*, 95, 6905-6917.
13. Keerthikumar, S., Mathivanan, S. (**2017**). Proteotypic Peptides and Their Applications. In: Keerthikumar S, Mathivanan S, editors. Proteome Bioinformatics. New York, NY: Springer New York, p. 101-107.
14. Annesley, T. M. (**2003**). Ion suppression in mass spectrometry. *Clinical Chemistry*, 49, 1041-1044.
15. Trufelli, H., Palma, P., Famiglini, G., Cappiello, A. (**2011**). An overview of matrix effects in liquid chromatography–mass spectrometry. *Mass Spectrometry Reviews*, 30, 491-509.
16. Piehowski, P. D., Petyuk, V. A., Orton, D. J., Xie, F., Moore, R. J., Ramirez-Restrepo, M., Engel, A., Lieberman, A. P., Albin, R. L., Camp, D. G. (**2013**). Sources of technical variability in quantitative LC–MS proteomics: human brain tissue sample analysis. *Journal of Proteome Research*, 12, 2128-2137.
17. Bär, C., Mathis, D., Neuhaus, P., Dürr, D., Bisig, W., Egger, L., Portmann, R. (**2019**). Protein profile of dairy products: Simultaneous quantification of twenty bovine milk proteins. *International Dairy Journal*, 97, 167-175.
18. Schulze, W. X., Usadel, B. (**2010**). Quantitation in mass-spectrometry-based proteomics. *Annual review of plant biology*, 61, 491-516.

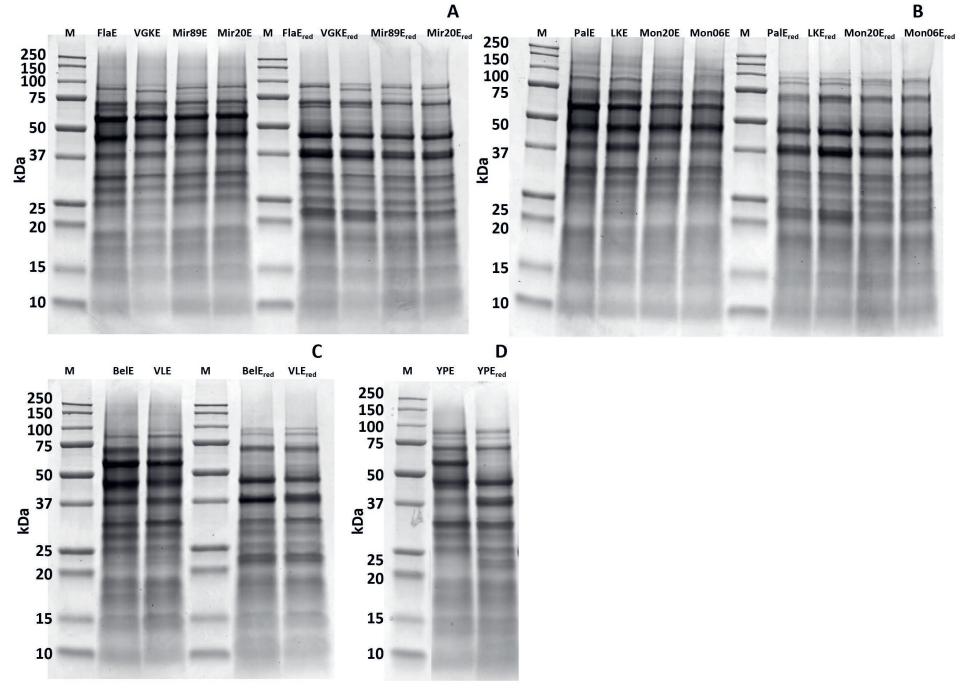
19. Butré, C. I., Buhler, S., Sforza, S., Gruppen, H., Wierenga, P. A. (**2015**). Spontaneous, non-enzymatic breakdown of peptides during enzymatic protein hydrolysis. *Biochimica et Biophysica Acta - Proteins and Proteomics*, 1854, 987-994.
20. Lowenthal, M. S., Liang, Y., Phinney, K. W., Stein, S. E. (**2013**). Quantitative bottom-up proteomics depends on digestion conditions. *Analytical Chemistry*, 86, 551-558.
21. Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., Selbach, M. (**2011**). Global quantification of mammalian gene expression control. *Nature*, 473, 337-342.
22. Battisti, I., Ebinezer, L. B., Lomolino, G., Masi, A., Arrigoni, G. (**2021**). Protein profile of commercial soybean milks analyzed by label-free quantitative proteomics. *Food Chemistry*, 352, 129299.
23. Ishihama, Y., Oda, Y., Tabata, T., Sato, T., Nagasu, T., Rappsilber, J., Mann, M. (**2005**). Exponentially Modified Protein Abundance Index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Molecular & Cellular Proteomics*, 4, 1265-1272.
24. Blein-Nicolas, M., Xu, H., Vienne, D., Giraud, C., Huet, S., Zivy, M. (**2012**). Including shared peptides for estimating protein abundances: A significant improvement for quantitative proteomics. *Proteomics*, 12, 2797-2801.
25. Kraut, A., Marcellin, M., Adrait, A., Kuhn, L., Louwagie, M., Kieffer-Jaquinod, S., Lebert, D., Masselon, C. D., Dupuis, A., Bruley, C., Jaquinod, M., Garin, J., Gallagher-Gambarelli, M. (**2009**). Peptide storage: Are you getting the best return on your investment? Defining optimal storage conditions for proteomics samples. *Journal of Proteome Research*, 8, 3778-3785.
26. Zapadka, K. L., Becher, F. J., Gomes Dos Santos, A. L., Jackson, S. E. (**2017**). Factors affecting the physical stability (aggregation) of peptide therapeutics. *Interface focus*, 7, 20170030-20170030.
27. Planyavsky, M., Huber, M. L., Staller, N. A., Müller, A. C., Bennett, K. L. (**2015**). A longitudinal proteomic assessment of peptide degradation and loss under acidic storage conditions. *Analytical Biochemistry*, 473, 11-13.
28. Butré, C. I. (**2014**). Introducing enzyme selectivity as a quantitative parameter to describe the effects of substrate concentration on protein hydrolysis [PhD thesis Wageningen University]. Wageningen. Wageningen University.
29. Bourgeois, M., Jacquin, F., Savois, V., Sommerer, N., Labas, V., Henry, C., Burstin, J. (**2009**). Dissecting the proteome of pea mature seeds reveals the phenotypic plasticity of seed protein composition. *Proteomics*, 9, 254-271.
30. Casey, R., Domoney, C. (**1999**). Pea Globulins. In: Shewry PR, Casey R, editors. Seed Proteins. Dordrecht: Springer Netherlands, p. 171-208.
31. Breddam, K., Meldal, M. (**1992**). Substrate preferences of glutamic acid-specific endopeptidases assessed by synthetic peptide substrates based on intramolecular fluorescence quenching. *European Journal of Biochemistry*, 206, 103-107.
32. Butré, C. I., Wierenga, P. A., Gruppen, H. (**2014**). Influence of water availability on the enzymatic hydrolysis of proteins. *Process Biochemistry*, 49, 1903-1912.
33. Deng, Y., Butré, C. I., Wierenga, P. A. (**2018**). Influence of substrate concentration on the extent of protein enzymatic hydrolysis. *International Dairy Journal*, 86, 39-48.
34. O'Kane, F. E., Happe, R. P., Vereijken, J. M., Gruppen, H., van Boekel, M. A. J. S. (**2004**). Characterization of pea vicilin. 1. denoting convicilin as the α-Subunit of the pisum vicilin family. *Journal of Agricultural and Food Chemistry*, 52, 3141-3148.
35. UniProtKB. http://www.uniprot.org/. *04/06/2020*.
36. Szymanowska, U., Jakubczyk, A., Baraniak, B., Kur, A. (**2009**). Characterisation of lipoxygenase from pea seeds (*Pisum sativum* var. Telephone L.). *Food Chemistry*, 116, 906-910.
37. Higgins, T. J. V., Beach, L. R., Spencer, D., Chandler, P. M., Randall, P. J., Blagrove, R. J., Kortt, A. A., Guthrie, R. E. (**1987**). cDNA and protein sequence of a major pea seed albumin (PA 2 : Mr≈26 000). *Plant Molecular Biology*, 8, 37-45.
38. Meyer, B., Papasotiriou, D. G., Karas, M. (**2011**). 100% protein sequence coverage: a modern form of surrealism in proteomics. *Amino Acids*, 41, 291-310.

39. Wang, W.-Q., Jensen, O. N., Møller, I. M., Hebelstrup, K. H., Rogowska-Wrzesinska, A. (**2018**). Evaluation of sample preparation methods for mass spectrometry-based proteomic analysis of barley leaves. *Plant methods*, 14, 1-13.
40. Tabb, D. L., Vega-Montoto, L., Rudnick, P. A., Variyath, A. M., Ham, A.-J. L., Bunk, D. M., Kilpatrick, L. E., Billheimer, D. D., Blackman, R. K., Cardasis, H. L. (**2010**). Repeatability and reproducibility in proteomic identifications by liquid chromatography- tandem mass spectrometry. *Journal of Proteome Research*, 9, 761-776.
41. Berg, M., Parbel, A., Pettersen, H., Fenyö, D., Björkesten, L. (**2006**). Reproducibility of LC-MS-based protein identification. *Journal of experimental botany*, 57, 1509-1514.
42. Delmotte, N., Lasaosa, M., Tholey, A., Heinzle, E., van Dorsselaer, A., Huber, C. G. (**2009**). Repeatability of peptide identifications in shotgun proteome analysis employing off-line two-dimensional chromatographic separations and ion-trap MS. *Journal of separation science*, 32, 1156-1164.
43. Tsou, C. C., Tsai, C. F., Teo, G. C., Chen, Y. J., Nesvizhskii, A. I. (**2016**). Untargeted, spectral library-free analysis of data-independent acquisition proteomics data generated using Orbitrap mass spectrometers. *Proteomics*, 16, 2257-2271.
44. Casey, R., Sharman, J. E., Wright, D. J., Bacon, J. R., Guldager, P. (**1982**). Quantitative variability in *Pisum* seed globulins: its assessment and significance. *Plant Foods for Human Nutrition*, 31, 333-346.
45. Gueguen, J., Barbot, J. (**1988**). Quantitative and qualitative variability of pea (*Pisum sativum* L.) protein composition. *Journal of the Science of Food and Agriculture*, 42, 209-224.
46. Schroeder, H. E. (**1982**). Quantitative studies on the cotyledonary proteins in the genus *Pisum*. *Journal of the Science of Food and Agriculture*, 33, 623-633.
47. Burstin, J., Kreplak, J., Macas, J., Lichtenzveig, J. (**2020**). Pisum sativum (Pea). *Trends in Genetics*, 36, 312-313.

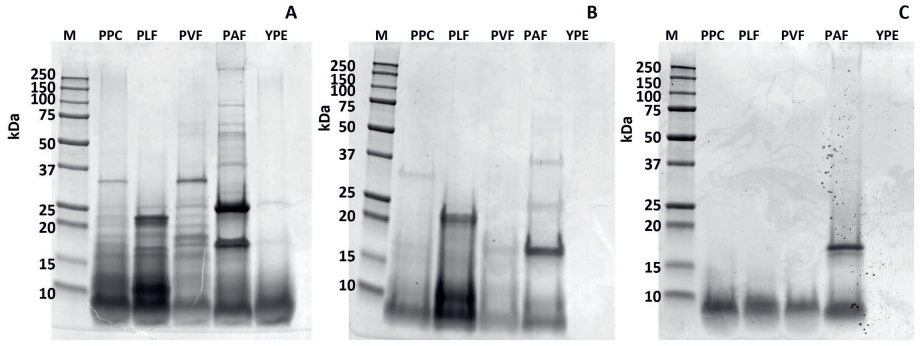
Annexes chapter 3



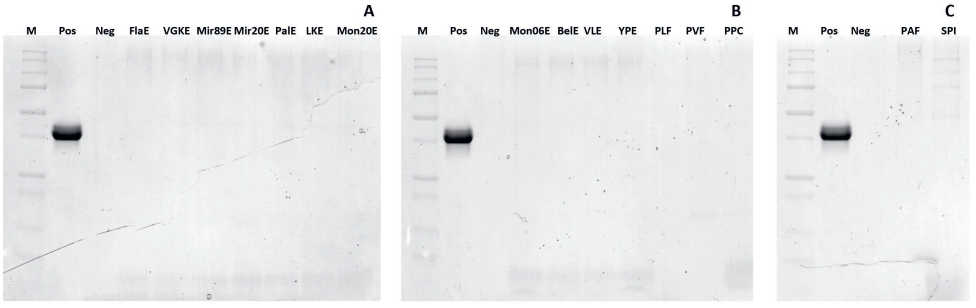
Annex 3.1. SDS-PAGE gels stained with Coomassie showing a marker (M), PPC, PLF, PVF and PAF under non-reducing and reducing (red) conditions, annotated bands indicated.



Annex 3.2. SDS-PAGE gels stained with Coomassie showing a marker (M), FlaE, VGKE, Mir89E, Mir20E, PalE, LKE, Mon20E, Mon06E, BelE, VLE and YPE under non-reducing and reducing (red) conditions.



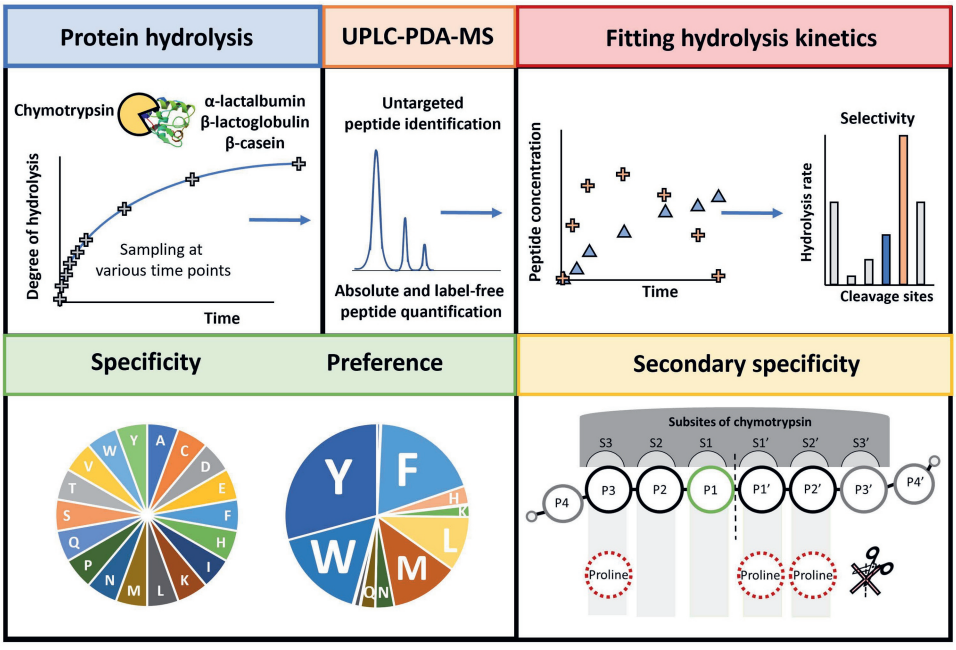
Annex 3.3. SDS-PAGE gels stained with Coomassie showing a marker (M), PPC, PLF, PVF and YPE after enzymatic hydrolysis and DTT incubation (A). After enzymatic hydrolysis, DTT incubation and addition of TFA (supernatant, B). After enzymatic hydrolysis, DTT incubation, SPE, and addition of TFA (supernatant, C).



Annex 3.4. SDS-PAGE gels stained with Periodic acid–Schiff (PAS) showing a marker (M), positive control (Pos), negative control (Neg), FlaE, VGKE, Mir89E, Mir20E, PalE, LKE, Mon20E, Mon06E, BelE, VLE, YPE, PLF, PVF, PPC, PAF and Soy Protein Isolate (SPI).

CHAPTER 4

The path of proteolysis by bovine chymotrypsin



Vreeke, G. J. C., Vincken, J.-P., Wierenga, P. A. (**2023**). The path of proteolysis by bovine chymotrypsin. *Food Research International*, *165*, 112485.
https://doi.org/10.1016/j.foodres.2023.112485

Abstract

Chymotrypsin is one of the major proteases in intestinal protein digestion. Observations about the type of bonds that are hydrolysed (specificity and preference) were in the past derived from the peptide composition after digestion or hydrolysis rates of synthetic peptides. In this study, the path of hydrolysis by bovine chymotrypsin, *i.e* formation and degradation of peptides, were described for α-lactalbumin, β-lactoglobulin and β-casein. The peptide compositions, determined with UPLC-PDA-MS at different time points were used to determine the digestion kinetics for individual cleavage sites. It was evaluated how statements on (secondary) specificity from literature were reflected in the release kinetics of peptides. β-Lactoglobulin reached the highest degree of hydrolysis (10.9 ±0.1 %) and was hydrolysed fastest (28 ±1 $mM_{peptide\ bonds}/s/mM_{enzyme}$), regardless of its globular (tertiary) structure. Chymotrypsin showed a preference towards aromatic amino acids, methionine and leucine, but was also tolerant to other amino acids. For the cleavage sites within this preference, ~73% of the cleavage sites were hydrolysed with high or intermediate selectivity. For the missed cleavages within the preference, 45 % was explained by hindrance of proline, which affected hydrolysis only when in positions P3, P1' or P2'. No clear indication (based on primary structure) was found to explain the other missed cleavages. A few cleavage sites were hydrolysed extremely efficient in α-lactalbumin (F9, F31, W104) and β-casein (W143, L163, F190). This study gave unique and quantitative insight in peptide formation and degradation by chymotrypsin in the digestion of proteins. The approach used showed potential to explore the path of hydrolysis for other proteases with less defined specificity.

Introduction

Chymotrypsin is commonly known for the involvement in intestinal protein digestion in many animals species [1-3] and humans [4]. Nowadays, the protease is also used in the production of bio-active peptides [5, 6] and as alternative to trypsin for proteomics [7]. In the 20th century, many studies in biochemistry have been performed to unravel the hydrolysis mechanism of chymotrypsin [8, 9]. In these studies, parameters for hydrolysis kinetics as k-cat were determined for synthetic (amide) substrates to find the relationship between amino acids around the cleavage site in the substrate and protease activity. These findings about the mechanism were (later) confirmed with X-ray crystallography. These efforts led to several hypotheses about the effect of amino acids in the different binding positions of the protease. Recent advances in mass spectrometry and peptide quantification enabled us to actually measure the hydrolysis rates of cleavage sites during proteolysis instead of analysing synthetic model substrates. The aim of this study is to describe the path of hydrolysis for three milk proteins with varying structures. Current statements on specificity from synthetic substrates will be evaluated during actual hydrolysis of intact proteins.

Prior knowledge about chymotrypsin's mechanism, specificity and preference

Chymotrypsin is a serine protease, in which three amino acid residues (Ser195, His57, Asp102) form a catalytic triad that facilitates the hydrolysis of the peptide bond via a two-step

mechanism. First, a temporary bond (acylation) is made between the peptide bond in the substrate and the serine residue in the catalytic site. Second, the serine residue is de-acylated by releasing the amide part and subsequently the carboxylic acid product. The type of peptide bonds that can be hydrolysed (protease specificity), depends on whether the amino acid in the P1 position of the substrate fits the S1 pocket of the enzyme. Chymotrypsin consists of a relatively deep S1 pocket, in which the amino acid residue in the P1 position can make hydrophobic interactions with for instance Ser190, Cys191, Cys220, Val213, Trp215 and Tyr228 [8, 9]. Considering these interactions, it is generally believed that chymotrypsin has a specificity for hydrophobic and aromatic residues in the P1 position. But, the exact reported specificity is broader (**Table 4.1**). For instance, occurrences of hydrolysis by chymotrypsin were described after methionine, glycine or arginine residues. These observations seem to imply that the S1 pocket of chymotrypsin could accommodate a wider range of amino acid residues. Besides protease specificity, studies report that not all amino acids are hydrolysed with the same likelihood, often described with the term "preference". In literature, the preferred amino acids are mostly phenylalanine (F), tryptophan (W) and tyrosine (Y) (**Table 4.1**).

Table 4.1 Reported specificity and preference of chymotrypsin.

Assumed specificity / statement about specificity	**Preference**	**Source**
Aromatic and hydrophobic residues		[10]
F, H, L, M, W, Y	F, W, Y > H, L, M	[11]
A, F, G, H, I, M, S, T, V, W, Y	F, W, Y > M, I, S, V, H, G, A, T	[12]
F, L, M, W, Y	W > Y > F > M > L	[13]
F, K, L, M, R, W	F, W, Y > L, M > K, R	[14]
Correlates with hydrophobicity		[15]
F, L, M, N, Q, W, Y	W > Y > F > M > L > Q, N	[16]
F, Y, L		MEROPS database [17]
F, H, L, M, W, Y	F, W, Y > H, L, M	ExPASy peptide cutter

Observations of secondary specificity

Several studies indicated that not all theoretical cleavage sites within the assumed preference were hydrolysed by chymotrypsin. For example, Galvão *et al.* described that only 56 % of the theoretical cleavage sites were hydrolysed in whey proteins [18]. Deng *et al.* observed that 54 % of the theoretical maximum degree of hydrolysis ($DH_{max,\ theo}$) was reached for chymotrypsin hydrolysis of apo alpha-lactalbumin [19]. In a proteomics study by Giansanti *et al*. a similar observation was done. They reported that ~80 % of the peptides identified in chymotrypsin digests of *E. coli* lysate still contained 1-4 intact cleavage sites [7]. Possibly, these observations could be explained by neighbouring amino acids that hinder the hydrolysis of the cleavage sites, also called "secondary specificity".

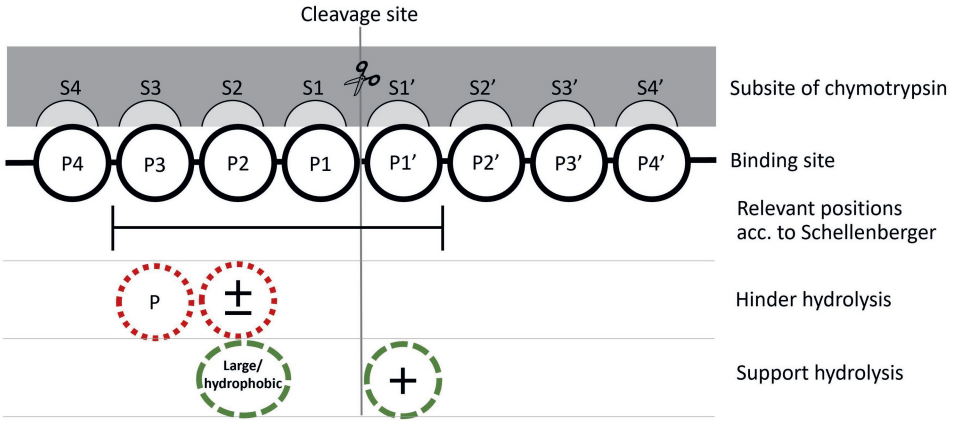


Figure 4.1. Illustration of the subsite model of Schechter and Berger with influence of binding site positions on chymotrypsin hydrolysis. + and - indicate positively charged and negatively charged residues, P indicates proline.

The concept of secondary specificity was introduced with the subsite model of Schechter and Berger [20] (**Figure 4.1**). Schellenberger *et al.* described in 1989 that chymotrypsin hydrolysis was influenced by the amino acids in the P3-P1' positions [21]. A study from 1972 using synthetic substrates, observed that a proline residue in the P3 position hindered hydrolysis [22]. In 1991, Schellenberger *et al*. performed a Quantitative Structure-Activity Relationship (QSAR) study on chymotrypsin hydrolysis of synthetic substrates tested in literature [23]. The model showed preference for large residues in the P2 position, no effect of proline in the P2 position and confirmed the earlier observed hindrance by proline in the P3 position. Hedstrom reported a slight preference for hydrophobic amino acid residues in the P2 position and a preference for positively charged residues in the P1' position [15]. Another study, focussing on the influence of the amino acid in the P1' position in relation to that in the P1 position, also described the positive contribution of lysine and arginine in the P1' position [11]. The presence of a charged residue in the P2 position would negatively affect the probability of hydrolysis [11].

Besides the primary substrate structure, a study of Wright suggested a conformational specificity, in which the tertiary structure of the amino acids in the P3 and P4 position would influence the hydrolysis [24]. Similarly, research of Vorob'ev *et al.* suggested that some peptide bonds in proteolysis have to be demasked before other bonds become accessible to chymotrypsin [16].

Peptide quantification and enzyme selectivity

The previous studies on chymotrypsin determined hydrolysis kinetics for synthetic substrates or analysed the peptides formed after proteolysis. For the latter, the identified peptides were used to determine the preference by analysing the amino acids in all cleavage sites at which hydrolysis was observed, sometimes corrected for frequency of occurrence. Typically, the concentrations of the peptides formed were not measured and considered. In this manuscript, all individual

peptides (at each timepoint) will be absolutely and label-free quantified, based on UV absorbance and predicted molar extinction coefficient according to Kuipers *et al*. [25]. The peptide composition at different time points allows us to determine the hydrolysis rate of each individual cleavage site in relation to the total hydrolysis rate, previously introduced as protease selectivity [26]. The determination of the selectivity makes it possible to see whether previous observations about secondary specificity based on synthetic substrates are also valid for real protein substrates. Ultimately, the aim is to evaluate whether the path of proteolysis of any protein by chymotrypsin, and the respective peptides released, can be predicted.

Materials & Methods

Materials

α-Lactalbumin was obtained from Davisco Foods International Inc. (Le Sueur, MN, USA). The protein isolate was further treated into apo α-Lactalbumin (α-LA) by removal of the calcium ions with EDTA, similarly done as described by Deng *et al.* [19]. β-Lactoglobulin (β-LG, L0130), β-casein (β-cas, C6905), bovine α-chymotrypsin (C-3142) and aprotinin from bovine lung (A6279) were purchased from Sigma-Aldrich (St. Louis, MO, USA). The protein contents of the protein samples were 93 % (w/w) for α-LA, 96 % (w/w) for β-LG and 90 % for β-cas and 86 % (w/w) for chymotrypsin (based on Dumas) (**Table 4.2**). The protein purities were 90 % for α-LA, 100 % for β-LG and 90 % for β-cas. The other proteins in α-LA were β-lactoglobulin and bovine serum albumin. The other proteins in β-cas were other type of caseins. The chymotrypsin did not contain any trypsin, according to SDS-PAGE and UPLC-PDA-MS measurements. According to the supplier, the chymotrypsin had an activity of ≥ 40 N-benzoyl-L-tyrosine-ethyl-ester units mg^{-1} protein and was treated with Nα-Tosyl-L-lysine chloromethyl ketone hydrochloride (TLCK) to inactivate residual trypsin activity. Aprotinin was present at 2.3 mg mL^{-1}, based on previously determinations with RP-UPLC [27].

Table 4.2. Protein materials used in this study with DH_{max} for chymotrypsin hydrolysis.

Protein	Uniprot code	N-factor[1] [g of protein / g N]	Protein content [w/w]	Protein purity	h_{tot}[2] [mmol/g protein]	F, Y, W in sequence (%)	DH_{max} experimental at E:S ratio 1:25
α-LA	P00711	6.25	93 %	90 %	8.6	9.8 %	8.4 ±0.3 %
β-LG	P02752	6.29	96 %	100 %	8.8	6.2 %	10.9 ±0.1 %
β-cas	P02666	6.39	90 %	90 %	8.7	6.7 %	8.2 ±0.3 %
CT	P00766	5.99	86 %	100 %			

[1] From Uniprot (http://www.uniprot.org)

[2] h_{tot} is the amount of peptide bonds per gram of protein calculated with the amino acid sequence and molecular weight of the protein reported in Uniprot.

Methods

Protein hydrolysis by chymotrypsin

Hydrolysis was performed for 3 hours with a pH-stat device (Metrohm, Herisau, Switzerland) equipped with 0.2 M NaOH. α-LA, β-LG and β-cas were dissolved at 1% (w/v) in Millipore water. Afterwards, the pH was adjusted to 8.0 and the solutions were equilibrated for 30 min at 37 °C. Hydrolysis was performed at enzyme to substrate ratios of 1:100 and 1:25, both in duplicate. Samples (200 µL) were taken before and during the hydrolysis at various time points, after which aprotinin was immediately added to inactivate the chymotrypsin. To the samples with an E:S ratio of 1:100, 6 µL of aprotinin was added and to the samples at E:S ratio of 1:25, 24 µL of aprotinin was added. The inactivation of chymotrypsin activity was confirmed by addition of aprotinin to a (ongoing) hydrolysis in the pH-stat after 1 min, in which the increase in DH stopped at moment of addition. Hydrolyses without enzyme were performed to exclude potential consumption of NaOH due to CO_2 dissolution. The degree of hydrolysis was calculated using equation 1.

$$DH_{stat}[\%] = V_b \times N_b \times \frac{1}{\alpha} \times \frac{1}{m_p} \times \frac{1}{h_{tot}} \times 100\ \% \qquad \text{(Eq. 1)}$$

where V_b [mL] is the volume of added NaOH; N_b [mol/L] is the normality of NaOH; α is the average degree of dissociation of the α-NH group (1/α=1.3 at 37 °C and pH 8 [28]); m_p [g] is the amount of protein in solution; h_{tot} [mmol/g] is the number of peptide bonds per gram of protein. The DH in time was fitted with second order reaction kinetics to determine the overall hydrolysis rate constant (k_h) (equation 2). The k_h was used to calculate the turnover number [$mM_{peptidebonds}$ s^{-1} mM_{EZ}^{-1}], which was the amount of peptide bonds in solution hydrolysed by an equal molar amount of enzyme each second.

$$DH_{fit}[\%] = DH_{max,fit} - DH_{max,fit} \,/\, (1 + k_h \times DH_{max,fit} \times t) \quad \text{(Eq. 2)}$$

Sample preparation

The protein hydrolysates were mixed (1:1) [v/v] with 20 mM dithiothreitol (DTT) and 100 mM Tris-HCl buffer at pH 8.0 and incubated for 2 hours to reduce disulphide bonds. Afterwards, the intact protein samples were diluted 10x in eluent A and the hydrolysates were diluted 3x [v/v] in eluent A. The samples were centrifuged (10 minutes, 14,000 x g, 20 °C) and the supernatant was injected on the UPLC-MS. The endpoint hydrolysate of β-LG was injected three times to estimate reproducibility.

Reverse phase ultra-high performance liquid chromatography (RP-UPLC)

The peptides were analysed on an Acquity Premier UPLC coupled to a Photodiode Array (PDA) detector, both from Waters (Milford, MA, USA). A gradient was used of 1 % ACN + 0.1 % TFA in UPLC-grade water (eluent A) and 1 % UPLC-grade water + 0.1 % TFA in ACN (eluent B). The gradient was 0-2 min isocratic on 3 % B; 2-10 min linear gradient from 3-22 % B; 10-16 min linear

gradient 22-30 % B; 16-21 min linear gradient 30-100 % B; 21-26 min isocratic on 100 % B; 26-28 min linear gradient 100-3 % B and 28-32 min isocratic on 3 % B. The column was the Acquity Premier peptide column BEH C18 2.1 mm *150 mm 300 Å 1.7 µm, thermostated at 30 °C. The applied flow was 350 µL/min and the volume injected was 4 µL.

Electro spray ionisation time of flight mass spectrometry

The mass spectra were acquired from 50-3000 *m/z* with the Select Series Cyclic IMS (Waters). The machine was operated in Time-of-flight and V-mode without using the ion mobility dimension. Ionisation was done with the electrospray ionisation source at capillary voltage of 2.5 kV and a temperature of 150 °C. The sample cone was operated at 40 V and nitrogen was used as desolvation gas (500 °C, 800 L/h) and cone gas (200 L/h). Lock mass data was acquired by injection of 50 pg/µL Leucine-Enkephalin at 10 µL/min, at a capillary voltage of 2.7 kV. An automatic quadrupole profile was applied. Fragmentation was done in the trap-cell by changing the voltage for MS (6 V) into a MS^E ramp from 28 V to 56 V.

Peptide identification

Peptides were identified using the recently developed guidelines and methodology of Vreeke *et al.* with UNIFI software version 1.8 (**Chapter 2**). Processing was done semi-specific with a specificity of F, Y, W, M, H, L and (fully) a-specific. The semi-specific processing option identified peptides that matched chymotrypsin's specificity on at least the N- or C-terminal side, whereas the a-specific digest considered potential hydrolysis of all peptide bonds. The a-specific method was used in the end for all analyses. The threshold set in peak detection was 1000 counts for MS peaks and 250 counts for MS/MS peaks. There was no maximum set to the number of included MS and MS/MS signals. The threshold on mass error was 10 ppm for the MS and 20 ppm for MS/MS fragments. The requirements on number of MS/MS fragments to include a peptide were similar as described in (**Chapter 2**). The limit of detection (LOD) and limit of annotation (LOA) were determined by injection of a concentration series (2.5 mg/L to 5 g/L) of a tryptic α-LA hydrolysate. For this mass spectrometer under the applied conditions, the average LOD for a peptide was $1.6 \cdot 10^4$ Counts and the average LOA was $2.1 \cdot 10^5$ Counts. Peptides consisting of more than 41 amino acids and intact proteins were not detected in the a-specific analysis, and therefore obtained from the semi-specific processing result. Adducts and in-source fragments that were recognised in UNIFI were excluded from the analysis. In-source fragments that were not recognised as such were removed using PeptQuant, an in-house written script in Matlab v2018b. Annotations were considered as in-source fragment if the peptide and potential in-source fragment eluted at the same retention time, the in-source fragment included the same sequence as the peptide and the in-source fragment had a lower MS intensity than the peptide.

Peptide quantification

Quantification of the peptides was done using the UV absorbance at 214 nm. The UV peak areas were integrated in Masslynx software version 4.2 using the parameters as described in **Chapter 2**. The areas were converted into absolute peptide concentrations, $C_{peptide}$ [µM], using the molar

extinction coefficients, predicted based on Kuipers *et al.* [25] and the calculation of Butré *et al*. (equation 3).

$$C_{peptide}\ [\mu M] = \frac{A_{214} \cdot Q}{\varepsilon_{214} \cdot l \cdot V_{inj} \cdot k_{cell}} \qquad \text{(Eq. 3)}$$

where A_{214} [µAU min] is the UV peak area at 214 nm, V_{inj} [µL] is the volume of sample injected, Q [µL min^{-1}] is the flow rate, ε_{214} [L Mol^{-1} cm^{-1}] the predicted molar extinction coefficient of the peptide and l [cm] is the path length of the UV cell, which is 1 cm according to the manufacturer. The cell constant, k_{cell} for the UV detector was 0.78 (**Chapter 2**).

Completeness of analysis

The completeness of the peptide identification was evaluated with the amino acid sequence coverage and the peptide sequence coverage (equation 4 & 5), the completeness of quantification was evaluated with the protein recovery and the molar sequence coverage (equation 6 & 7). All these parameters were introduced by Butré *et al.* [26].

$$Amino\ acid\ sequence\ coverage\ [\%] = \frac{\#\ unique\ annotated\ amino\ acids}{\#\ amino\ acids\ in\ protein\ sequence} \cdot 100\ \% \qquad \text{(Eq. 4)}$$

$$Peptide\ sequence\ coverage\ [\%] = \frac{\#\ AA\ (annotated\ peptides)}{\#\ AA\ (annotated\ peptides) + \#\ AA\ (missing\ peptides)} \cdot 100\ \%$$

$$Protein\ recovery\ [\%] = \left(\frac{\left(\frac{\sum C_n}{\#AA_{protein}} \right)}{C_0} \right) \cdot 100\ \% \quad \text{(Eq. 6)}$$

where C_n [µM] is the concentration of each individual AA (n) in the protein sequence, and $\#AA_{protein}$ is the number of amino acids in the initial protein and C_0 [µM] is the initially injected protein concentration.

$$Molar\ sequence\ coverage\ [\%] = \left(1 - \frac{\sqrt{\frac{\sum (C_n - C_0)^2}{(\#AA_{protein} - 1}}}{C_0} \right) \cdot 100\ \% \qquad \text{(Eq. 7)}$$

where C_n [µM] is the concentration of each individual AA (n) in the protein sequence, C_0 [µM] is the initially injected protein concentration and $\#AA_{protein}$ is the number of amino acids in the initial protein.

Description of the endpoint hydrolysate with protease specificity and preference

Protease specificity is defined as the type of amino acid on the carboxylic side next to which the enzyme can hydrolyse peptide bonds. In literature, this is typically determined by analysis of the

start and end amino acid residues of the peptides formed after hydrolysis. For comparison, this was done similarly in this study. However, the specificity does not give any information on the probability that cleavage occurred or the extent of hydrolysis. Therefore, protease preference was calculated. This was defined as the proportion to which hydrolysis occurs compared to the theoretical probability based on the frequency of occurrence of the amino acids (equation 8). When a type of amino acid residue has a preference of 1, it is hydrolysed to the same extent as expected with the frequency of occurrence.

$$Preference\ [-] = \frac{\left(\frac{\sum C_{endpoint,i}}{\sum C_{endpoint,total}}\right)}{\left(\frac{\#AA_i}{\#AA_{tot}-1}\right)} \qquad \text{(Eq. 8)}$$

where $C_{endpoint,i}$ [µM] is the concentration cleavage site products formed after amino acids of type i in the endpoint hydrolysate, $C_{endpoint,total}$ [µM] is the total amount of cleavage site products formed in the endpoint hydrolysate, $\#AA_i$ is the frequency of occurrence of amino acid residues of type i in the sequence of the substrate, except the C-terminal amino acid. The $\#AA_{tot}$ is the total number of amino acid residues in the sequence.

The concentration of cleavage site products $C_{t,j}$ for cleavage site j at timepoint t was derived from the peptide concentrations (equation 9).

$$C_{t,j}[\mu M] = \sum\{C_{peptide}[x-y] \mid j = x-1 \cup j = y\} \qquad \text{(Eq. 9)}$$

where $C_{t,j}$ [µM] equals the sum of all peptide concentrations ($C_{peptide}$) with sequence x-y, which are released after amino acid j or which end by amino acid j.

Determination of the hydrolysis rate constants

For each cleavage site, the formation of cleavages site products in time was fitted using second order reaction kinetics (equation 10) and the sequential (second order) reaction kinetics (equation 11).

$$C_{j,t}[\mu M] = 2(C_0 - C_0\ /(1 + k_{j,app} \times t \times C_0)) \qquad \text{(Eq. 10)}$$

$$C_{j,t}[\mu M] = 2(C_0 - \frac{C_0 \times \left(\frac{k_2}{k_2-k_1}\right)}{1+C_0 \times k_1 \times t} + \frac{C_0 \times \left(\frac{k_1}{k_1-k_2}\right)}{1+C_0 \times k_2 \times t}) \qquad \text{(Eq. 11)}$$

where $C_{j,t}$ [µM] is the concentration cleavage site products for cleavage site j at timepoint t, C_0 [µM] is the (expected) protein concentration and $k_{j,app}$ [s^{-1}] is the apparent hydrolysis rate constant for cleavage site j. Both the C_0 value and the apparent hydrolysis rate constant were fitted. The C_0 was fitted because for some cleavage sites the experimental plateau value was lower than C_0. If in these cases the theoretical C_0 would be used, the $k_{j,app}$ would be

underestimated. For some cleavage sites, product formation started in a later stage of the hydrolysis and the data could better be described by sequential kinetics. The k_2 of the sequential fit was used when the R^2 improved by 0.10 in comparison with the regular second order fit and the R^2 of the sequential fit was > 0.50.

The apparent hydrolysis rate constants k, or in case of sequential kinetics k_2, were divided by the mass of enzyme used in mg (equation 12) and used to calculate the selectivity (equation 13). Cleavage sites were excluded in the selectivity analysis when cleavage site products were observed in (only) one timepoint, in only one of the duplicates hydrolyses or at concentrations below 1 µM.

$$k_i\ [s^{-1}mg^{-1}] = \frac{k_{i,app}}{m_E} \qquad \text{(Eq. 12)}$$

$$Selectivity\ [\%] = \frac{k_i \times C_0}{\sum(k_i \times C_0)} \times 100\% \qquad \text{(Eq. 13)}$$

where $k_{i,app}$ [s^{-1}] is the apparent hydrolysis rate constant calculated with the second order, or sequential order fit, m_E is the mass of enzyme in mg. The C_0 [µM] was derived from the fit.

Clustering of cleavage sites and subsite analysis

The cleavages sites were manually clustered into high selectivity sites (HSS), intermediate selectivity sites (ISS), low selectivity sites (LSS) and no hydrolysis (NH) by analysis of the cleavage site product formation in time. HSS cleavage sites showed formation of > 50 µM of cleavage site products within 3 minutes of hydrolysis. ISS cleavage sites formed > 50 µM of cleavage site products, but not within 3 minutes of hydrolysis. LSS showed product formation in multiple timepoints but at concentrations < 50 µM. Cleavages sites were categorised as no hydrolysis when product formation was < 1 µM, or only observed in one of the duplicates or at only one of the timepoints. After clustering, the amino acids in the P4-P4' position of the cleavage sites were analysed in relation to their relative occurrence to the four selectivity clusters. It was taken into consideration that the number of cleavage sites in each cluster was different and whether the amino acid in the P1 position was preferred or not. For each rule, it was also tested how often the inverse was correct. To have as many observations as possible the separate analyses of α-LA, β-LG and β-cas were combined.

Results & Discussion

Description of the hydrolysis

Based on the frequency of occurrence of the aromatic residues (F, Y, W) a DH_{max} was expected of 9.8 % for α-LA, 6.2 % for β-LG and 6.7 % for β-cas, respectively. Hydrolysis of 1% α-LA, β-LG and β-cas with chymotrypsin reached degree of hydrolysis values of 6.5 ±0.1 %, 8.7 ±0 % and 5.7 ±0.4% at an E:S ratio of 1:100. These values did neither match with the DH_{max} values expected based on aromatic residues, nor with the order of the DH_{max} values of the substrates. The

assumed preference and amino acid composition did not solely explain the observed degree of hydrolyses. To confirm that the degree of hydrolysis reached was the absolute maximum, the experiment was repeated at higher E:S ratio (1:25). For all substrates a significant increase in DH_{max} was observed, to 8.4 ±0.3 %, 10.9 ±0.1 % and 8.2 ±0.3 %, respectively (**Figure 4.2**). For chymotrypsin, the increase in $DH_{max, exp}$ with increasing E:S ratio was previously also reported at hydrolysis of α-LA at low (0.1% w/v) substrate concentration [19]. Similar shifts in DH_{max} were also observed for trypsin [19], and pepsin [29] and suggested to be the result of the formation of inhibitory peptides [19, 30]. Further experiments on peptide release were all done using the samples at an E:S ratio of 1:25, to minimise the effect of inhibition. The turnover rate of chymotrypsin was slightly higher for β-LG (28 ±1 mM s^{-1} mM^{-1}) than for β-cas and α-LA (17 ±0 and 20 ±4 mM s^{-1} mM^{-1}, respectively). This gave a first impression that the activity of chymotrypsin is not affected by protein structure. The turnover rate was in line with the activity provided by the supplier (≥17 mM s^{-1} mM^{-1}). During the hydrolysis, samples were taken at different time points to determine the peptides present and their concentrations.

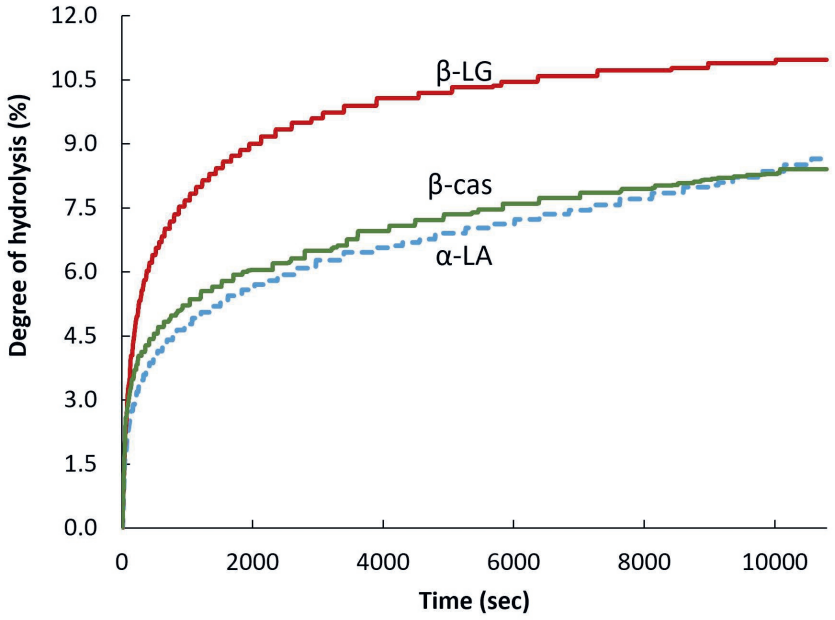


Figure 4.2. Degree of hydrolysis in time for 1% α-LA (- - -), β-LG (——) and β-cas (——) with bovine chymotrypsin at an E:S ratio of 1:25.

Peptide identification

Peptide identification was performed with semi-specific and a-specific data-processing in UNIFI, to check for possible differences between both processing options. Almost all peptides in the semi-specific processing were also identified with a-specific processing. One peptide sequence was identified with semi-specific processing, which exceeded the maximum length for a-specific

processing (α-LA 61-104). The a-specific processing yielded a few fully a-specific peptides as for instance β-LG 71-76 (K)IIAEKT (confirmed with 6 MS/MS fragments). The corresponding *m/z* value was assigned to β-LG 1-6 LIVTQT (3 MS/MS fragments) with semi-specific processing. For this example, the a-specific annotation was clearly a better match to the MS/MS spectrum. The a-specific analysis yielded ~10 % of in-source fragments that were not correctly recognised as such by the software, but were annotated as unique peptide. All these annotations were removed automatically in further analyses, based on sequence and retention time. In the endpoint hydrolysate of β-LG, the 61 unique peptides were identified in three replicate injections. Each replicate contained on average 54±2 peptides. Of the 61 unique entries, 82 % had a MS intensity above the average limit of annotation and 80% was identified in all three replicates. In total, the hydrolysates of α-LA, β-LG and β-cas consisted of 78, 169 and 141 peptides, respectively, in the various time points.

Peptide quantification

All peptides were quantified based on UV absorbance and their predicted molar extinction coefficient. On average, ~94 ±1 % of the UV peak area present in the α-LA chromatograms was assigned to annotated peptides. For β-LG and β-cas, these values were lower, 84 ±3 % and 73 ±7 %, respectively. The peptide concentrations during hydrolysis varied between 0.06 µM to 304 µM, with an average relative standard deviation of 5.7 % on individual peptide concentrations.

Evaluation of completeness of analysis

The identified peptides of α-LA covered the amino acid sequence for 100 %, for all time points. The amino acid sequence coverages for β-LG and β-cas were on average 97 ±5% and 90 ±8%, respectively (**Table 4.3**). Parts of the protein sequence that were not covered by peptides were for example, sequence 106-122 of β-LG and sequence 7-42 for β-cas. The chromatograms were investigated manually with extracted ion chromatograms for the expected masses, but this did not lead to the identification of these sequences. The average peptide sequence coverages obtained were 96 ±2 % for α-LA, 92 ±5 % for β-LG and 86 ±5 % for β-cas hydrolysates. These coverages were similar to those reported previously for tryptic hydrolysates of the same substrates (**Chapter 2**). For α-LA, the average amino acid concentration in a certain position on the sequence was 94 ±10 % of the injected protein concentration. This resulted in a molar sequence coverage of 80 ±7 %. For β-LG and β-cas, the average retrieved amino acid concentrations were lower than the injected concentrations, resulting in (average) protein recoveries of 82 ±14 % for β-LG and 71 ±13 % for β-cas. As a result, the molar sequence coverages were also lower than observed in the past. For β-LG, this was mainly due to low coverage of sequence region 106-122. For β-cas, sequence region 194-209 was quantified ~2x the expected concentration, while other part of the sequence were quantified below the injected (protein) concentration.

Table 4.3. Amino acid sequence coverages, peptide sequence coverages, protein recoveries and molar sequence coverages for the chymotrypsin hydrolysates of α-LA, β-LG and β-cas.

	α-LA		β-LG		β-cas	
	Average + stdev [%]	**Min-Max [%-%]**	**Average + stdev [%]**	**Min-Max [%-%]**	**Average + stdev [%]**	**Min-Max [%-%]**
Amino acid sequence coverage (%)	100 ±0	100	97 ±5	86 - 100	90 ±8	83 - 100
Peptide sequence coverage (%)	96 ±2	92 - 99	92 ±5	80 - 98	86 ±5	79 - 95
Protein recovery (%)	94 ±10	70 - 110	82 ±14	58 - 111	71 ±13	55 - 105
Molar sequence coverages (%)	80 ±7	65 - 89	63 ±14	36 - 81	34 ±14	9 - 62

Chymotrypsin's specificity and preference

The specificity of chymotrypsin was determined based on the peptides present in the endpoint hydrolysates. Peptides were released by hydrolysis of α-LA peptide bonds after the aromatic amino acids (phenylalanine, tyrosine and tryptophan) but also after asparagine, glutamic acid, leucine, lysine, histidine, methionine, serine, threonine, and valine residues. Besides these amino acids, hydrolysis of β-LG showed peptides released after alanine, aspartic acid, cysteine, glutamine, isoleucine and proline. Based on these observations, chymotrypsin could better be classified as a-specific protease than as specific protease. Within the broad specificity, chymotrypsin showed a preference to hydrolyse cleavage sites with phenylalanine, tryptophan and tyrosine in the P1 position. To describe this preference quantitatively, the amount of product released after a certain AA was compared to the amount that one would expect for completely random enzymatic hydrolysis (all AA similarly preferred). For phenylalanine, 5.2 (±2.0) times more hydrolysis products were released than one would expect based on the presence of phenylalanine in the sequences of α-LA, β-LG and β-cas (**Figure 4.3**). Tryptophan and tyrosine had preference values of 4.5 ±2.2 and 8.0 ±1.8. Other amino acids that were preferred by chymotrypsin in the P1 position were methionine and leucine with preferences of 3.3 ±0.2 and 2.7 ±0.5. The residues histidine, lysine, glutamine and asparagine had a preference around 1, but the individual preference values for each substrate varied substantially (e.g. hydrolysis of these residues was not observed for all substrates.) The order of the preference values matched reported preferences in literature (**Table 4.1**). In order to see whether cleavage sites with similar amino acid in the P1 position, were hydrolysed at similar rate, the kinetics of each individual cleavage site had to be determined.

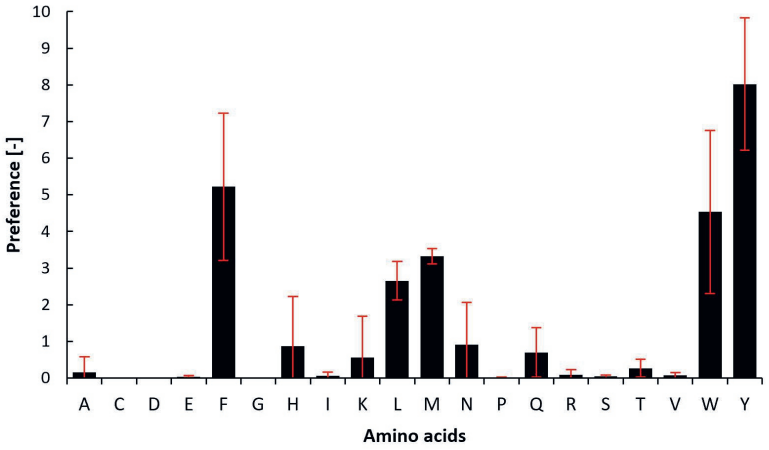


Figure 4.3. Preference of chymotrypsin for the amino acid residue in the P1 position. The average and standard deviation are shown of the preference values determined against α-LA, β-LG and β-cas.

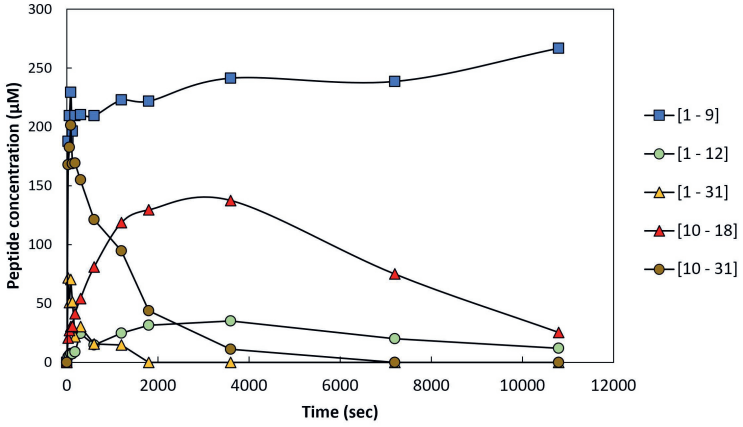


Figure 4.4. The absolute concentration in time of a few α-LA peptides to illustrate peptide formation and degradation by chymotrypsin.

Hydrolysis rate kinetics of individual cleavage sites

By analysis of the peptides and their concentrations at various timepoints (**Figure 4.4**), the hydrolysis rate constants were calculated for each individual cleavage site. A few cleavage sites in α-LA (F9, F31, and W104) and β-cas (W143, L163 and F190) were hydrolysed extremely fast. For these, >90 % of the cleavage site was hydrolysed in the 30 s^{-1} hydrolysate sample. As a result, all intact proteins were hydrolysed at 90 s^{-1} for α-LA and at 30 s^{-1} for β-cas. The degree of hydrolysis calculated from the peptide composition was for the first time points higher than the degree of hydrolysis based on pH-stat. Theoretically, both should be similar. A possible explanation is that after inactivation, hydrolysis of these cleavage sites continued. Possibly, chymotrypsin had a higher affinity towards these few peptide bonds in α-LA and β-cas than towards the inhibitor (aprotinin). Another hypothesis could be that the enzyme-aprotinin

complex dissociated (partly) during sample preparation for LC-MS, leading to a continuation of the hydrolysis of these sites. The fast depletion of intact protein was also observed in a previous study for α-LA hydrolysed with chymotrypsin [19]. Another study also observed that F31 and W104 in α-LA were hydrolysed relatively early by chymotrypsin compared to other cleavage sites [31].

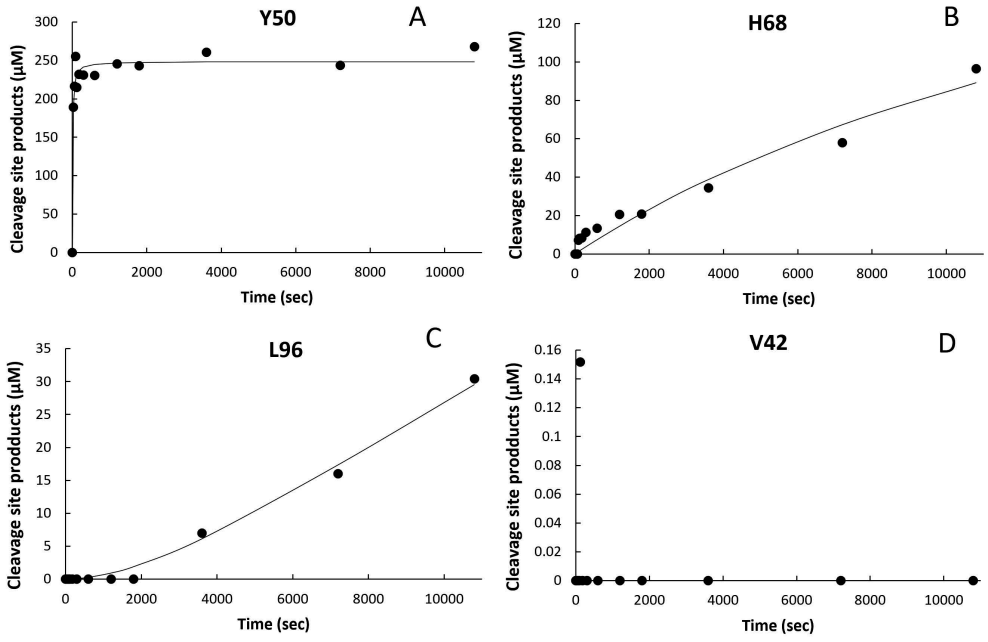


Figure 4.5. Release of cleavage site products in time for Y50 (A), H68 (B), L96 (C) and V42 (D) upon chymotrypsin hydrolysis of α-LA. The examples illustrating cleavage sites in the HSS cluster (A), ISS cluster (B), LSS cluster (C) and NH cluster (D). Markers are experimental data, lines are from second-order fit (A,B) and sequential second-order fit (C).

For β-cas, it has been reported that certain peptide bonds have to be hydrolysed first (de-masked) by chymotrypsin before remaining cleavage sites become accessible [16]. For the cleavage sites, the data were best described by fitting the hydrolysis rate constant and C_0 as variables, since in many cases the maximum concentration (µM) reached was lower than the initial protein concentration. Several cleavage sites were hydrolysed from the start of the incubation and were fully hydrolysed within a few minutes (**Figure 4.5A**). For these high selectivity sites (HSS), hydrolysis is efficient and seems independent of the hydrolysis of other peptide bonds. For intermediate selectivity sites (ISS), cleavage site products were released from the start of the hydrolysis and no clear plateau concentration was reached within 3 hours (**Figure 4.5B**), indicating a much lower rate of hydrolysis by chymotrypsin. For some cleavage sites, the hydrolysis products were not observed in the first few timepoints, but started to release at a

later stage of the hydrolysis (**Figure 4.5C**). After the lag time the release of cleavage site products increased to substantial concentrations. It seems that the hydrolysis of these cleavage sites is efficient, but requires other peptide bonds to be hydrolysed in advance (de-masking). For these cleavage sites, the sequential kinetics did better describe the data than the normal second order kinetics. For all hydrolysed and non-hydrolysed cleavage sites, the amino acids in the P4-P4' positions were analysed to evaluate whether the primary structure can be used to predict the hydrolysis efficiency of a peptide bond during proteolysis.

Chymotrypsin selectivity and the effect of the P1 position

Analysis of the hydrolysis rate constant of each cleavage site showed differences in the rate of hydrolysis of bonds with similar amino acid in the P1 position (**Figure 4.6-4.8**). For example, some cleavage sites with phenylalanine (F) in the P1 position in β-cas were not hydrolysed (F33, F87, F157 and F205), and some at slow rate, ($9 \pm 1 \cdot 10^{-4}$ $s^{-1} \cdot mg^{-1}_{enzyme}$; F52) up to a hydrolysis rate of 8.3 $s^{-1} \cdot mg^{-1}_{enzyme}$ (F190). Furthermore, it was observed that cleavage sites with non-aromatic amino acids could be HSS. This was for example the case for cleavage sites H32 and M90 in α-LA and D98 and A139 in β-LG (**Annex 4.1**).

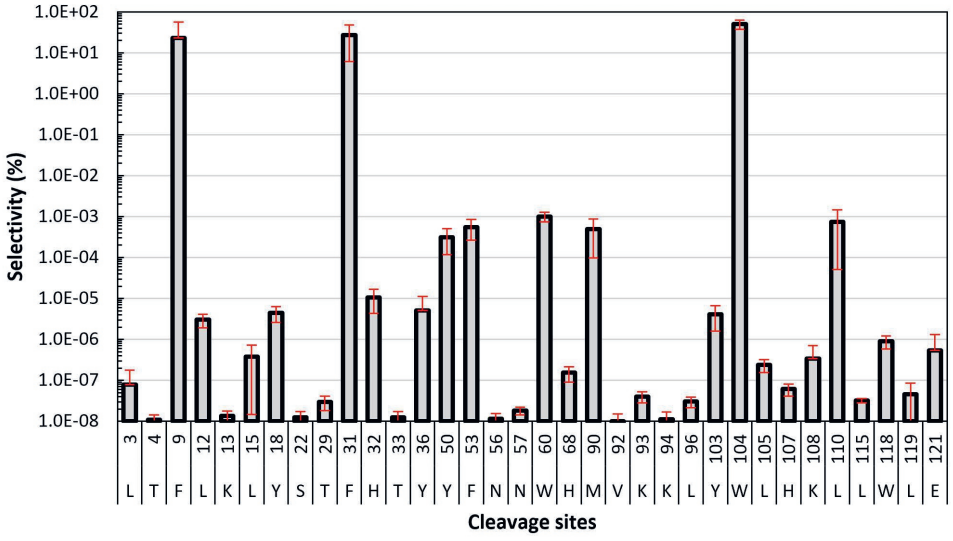


Figure 4.6. Selectivity (%) of chymotrypsin for hydrolysed cleavage sites of α-LA, plotted on logarithmic scale.

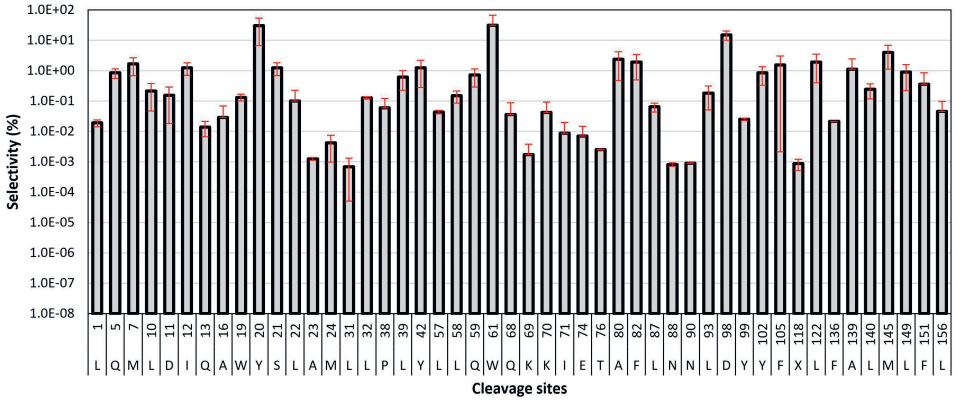


Figure 4.7. Selectivity (%) of chymotrypsin for hydrolysed cleavage sites of β-LG, plotted on logarithmic scale.

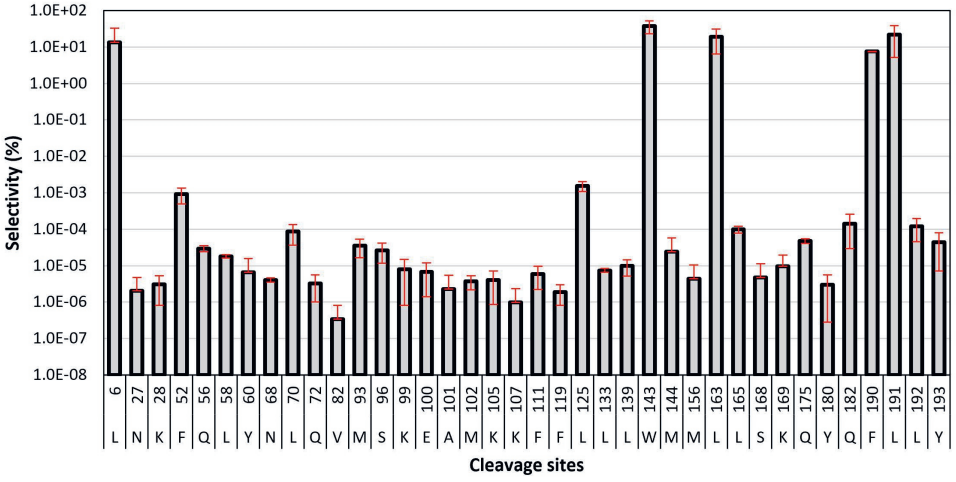


Figure 4.8 Selectivity (%) of chymotrypsin for hydrolysed cleavage sites of β-cas, plotted on logarithmic scale.

In order to predict the extent of hydrolysis, it should be considered how often cleavages sites with a certain AA in the P1 position are hydrolysed. For the cleavage sites in α-LA, β-LG and β-cas with phenylalanine, tyrosine or tryptophan in the P1 position, hydrolysis was observed for 28 out of 36 occurrences (**Figure 4.9**). For methionine, chymotrypsin hydrolysed 8 out of 11 occurrences, despite the absence of an aromatic ring. Leucine was hydrolysed less frequently (49 %) and at relatively slower hydrolysis rates than the other preferred residues. This suggests that leucine is less attractive for chymotrypsin to be positioned in the S1 binding pocket, or, that the effect of surrounding amino acids seem more dominant than for phenylalanine, tyrosine, tryptophan and methionine. Cleavage sites with non-preferred amino acids were hydrolysed, but only in less than 30 % of the occurrences and at hydrolysis rates < $1{\cdot}10^{3}$ s^{-1}. Possibly, the low

hydrolysis rates of cleavage sites with preferred amino acids in the P1 position could be explained by amino acids neighbouring the P1 position (secondary specificity).

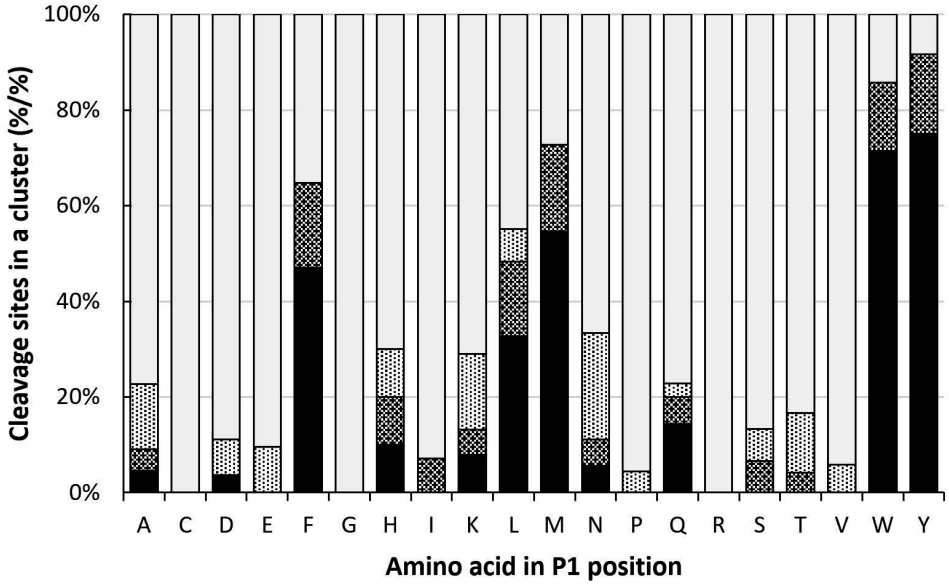


Figure 4.9. Occurrence (%/%) of cleavage sites of α-LA, β-LG and β-cas within the clusters containing the high selectivity sites (HSS) (■), intermediate selectivity sites (ISS) (▩), low selectivity sites (LSS) (▨) or not hydrolysed (NH) (□), categorised based on the amino acid in the P1 position.

The effect of neighbouring amino acids: Proline

Literature suggested that proline could hinder chymotrypsin hydrolysis when present around the cleavage site. In the sequences of α-LA, β-LG and β-cas, proline was present in each binding site position 45 times. The cleavage sites with proline in the position P4, P2, P3' and P4' were similarly distributed in the HSS/ISS/LSS/NH clusters as one would expect based on the relative number of cleavage sites in each cluster (**Table 4.4**). Therefore, a proline in these positions does not seem to influence the hydrolysis rate. On the contrary, a proline in the P3 position did never yield a cleavage site in HSS or ISS, despite that for 7 of these cleavage sites, a preferred amino acid was in the P1 position. For cleavage sites with a proline in the P1' and P2' positions, 1 and 2 cleavage sites were respectively clustered as HSS. This was also 2.5 times and 5 times less than expected. Cleavage sites that had a phenylalanine, tyrosine, tryptophan or methionine in the P1 position and that were classified as LSS or NH (11 occurrences), 8 had a proline in the P3, P1' or P2' position. For leucine, 10 out of 29 cleavage sites in LSS or NH had a proline in the P3, P1' or P2' position. The proline rule explained 73 % of the missed cleavages for phenylalanine, tyrosine, tryptophan and methionine, but only 34 % of the missed cleavages for leucine.

Table 4.4. Clustering of cleavage sites with proline at different positions of the cleavage site.

		P4	P3	P2	P1	P1'	P2'	P3'	P4'	Total for all cleavages sites
Occurrence P in position		45	45	45	45	45	45	45	45	
Occurrence with FYWML in the P1 position		7	7	13	0	16	8	12	9	
Occurence of a cleavage site in a cluster, when proline occurs in that respective position.	HSS	6	0	9	0	2	1	8	7	59
	ISS	2	0	4	0	2	2	4	3	29
	LSS	2	1	2	2	0	1	2	3	35
	NH	35	44	30	43	41	41	31	32	375
Proline in position, observed / expected[1]	HSS	1.1	0.0	1.7	0.0	0.4	0.2	1.5	1.3	
	ISS	0.8	0.0	1.5	0.0	0.8	0.8	1.5	1.1	
	LSS	0.6	0.3	0.6	0.6	0.0	0.3	0.6	0.9	
	NH	1.0	1.3	0.9	1.3	1.2	1.2	0.9	0.9	

[1] The ratio observed/expected was calculated by dividing the number of cleavage sites in a specific cluster with proline in a certain position by the expected number of cleavage sites for that category based on the distribution of all cleavage sites.

The effect of neighbouring amino acids: Charged amino acids

A study from Hedstrom suggested that a lysine or arginine residue in the P1' position would enhance hydrolysis. A lysine or arginine in the P1' occurred 46 times, of which 9 in combination with a preferred AA residue in the P1 position. Of the 46 cleavage sites with a positive residue in the P1' position, 10 were HSS, 6 were ISS, 11 were LSS and 19 were not hydrolysed (**Table 4.5**). The number of cleavage sites in the respective clusters were 1.8x (HSS), 2.2x (ISS), 3.4x (LSS) more than expected. The trend as described by Hedstrom was confirmed, but was not as dominant as the observations with proline. A study from Keil suggested that a (positively or negatively) charged residue in the P2 position would hinder hydrolysis. Out of the 115 occurrences, of which 22 with a preferred residue in the P1 position, 9 cleavage sites were classified as HSS and 6 as ISS. The occurrences were respectively 0.7 and 0.9 times the expected number of cleavage site in HSS and ISS and thereby within the expectation. The effect of charged amino acid residues on the P1' position and P2 position were not the main reason behind low hydrolysis rates.

The effect of neighbouring amino acids: Size and hydrophobicity

For the P3 position, the large residues alanine and arginine would fit best according to the QSAR model of Schellenberger. Cleavage sites with these residues in the P3 position were 1.7 and 1.1 times the expectation in HSS and ISS, respectively. These values were not significantly different from the values in other positions, considering the number of observations. Previous studies described for the P2 position a good fit when either the residues leucine and valine, or, (other) hydrophobic residues were in the P2 position. From our dataset, both did not show considerable effects. Although these observations were done with synthetic substrates, their effect in proteolysis seems minor.

Table 4.5. Clustering of cleavage sites with charged residues at different positions of the cleavage site.

		P4	P3	P2	P1	P1'	P2'	P3'	P4'
Occurence K or R in position		45	45	46	46	46	45	45	45
Occurence with FYWML in P1 position		6	18	9	0	9	9	8	12
Occurence K, R, E or D in position		113	114	115	115	115	113	112	112
Occurence with FYWML in P1		21	39	22	0	23	25	23	28
K of R in position, cleavage site in	HSS	2	8	5	3	10	7	6	9
	ISS	1	6	3	2	6	5	3	3
	LSS	5	3	4	6	11	5	5	1
	NH	37	28	35	35	19	28	31	32
K of R in position, expected/observed	HSS	0.4	1.5	0.9	0.6	1.8	1.3	1.1	1.7
	ISS	0.4	2.3	1.1	0.7	2.2	1.9	1.1	1.1
	LSS	1.6	0.9	1.2	1.9	3.4	1.6	1.6	0.3
	NH	1.1	0.8	1.0	1.0	0.5	0.8	0.9	0.9
K,R,D,E in position, cleavage site in	HSS	5	17	9	4	15	11	16	15
	ISS	4	13	6	2	8	8	5	7
	LSS	9	5	7	12	18	8	8	6
	NH	95	80	94	97	75	86	83	84
K, R, D, E in position, expected/observed	HSS	0.4	1.2	0.7	0.3	1.1	0.8	1.2	1.1
	ISS	0.6	1.9	0.9	0.3	1.2	1.2	0.8	1.1
	LSS	1.1	0.6	0.9	1.5	2.2	1.0	1.0	0.8
	NH	1.1	0.9	1.1	1.1	0.9	1.0	1.0	1.0

Conclusion

In this study, the aim was to describe the path of hydrolysis with chymotrypsin. Digestion kinetics for individual cleavage sites allowed us to relate subsite composition with chymotrypsin's activity. Although chymotrypsin is generally considered to have a specificity for aromatic and hydrophobic residues, this study showed that it is able to hydrolyse peptide bonds after almost all amino acid residues. A preference was observed for phenylalanine, tyrosine, tryptophan and methionine (hydrolysed in 78 % of the cleavage sites) and to a lower extent leucine (hydrolysed in 49 % of the cleavage sites). For cleavage sites with these preferred amino acids in the P1 position, still a substantial number of cleavage sites were not hydrolysed. The negative effect of proline around the cleavage site was found to be position dependent. A proline residue in the P3, P1' or P2' position hindered chymotrypsin hydrolysis, whereas a proline in other positions had no effect. Hindrance by proline explained 45 % of the missed-cleavages for cleavage sites within the preference. Charge or hydrophobic amino acids as neighbouring amino acids in the cleavage sites did not show major effects. The approach taken gave fundamental insight in chymotrypsin and is promising to apply to other digestive or novel (commercial) proteases.

References

1. Gorrill, A., Thomas, J. (**1967**). Trypsin, chymotrypsin, and total proteolytic activity of pancreas, pancreatic juice, and intestinal contents from the bovine. *Analytical Biochemistry*, 19, 211-225.
2. Gorrill, A. D. L., Friend, D. W. (**1970**). Pancreas size and trypsin and chymotrypsin activities in the pancreas and intestinal contents of pigs from birth to 5 weeks of age. *Canadian Journal of Physiology and Pharmacology*, 48, 745-750.
3. Guyonnet, V., Tłuscik, F., Long, P. L., Polanowski, A., Travis, J. (**1999**). Purification and partial characterization of the pancreatic proteolytic enzymes trypsin, chymotrypsin and elastase from the chicken. *Journal of Chromatography A*, 852, 217-225.
4. Feinstein, G., Hofstein, R., Koifmann, J., Sokolovsky, M. (**1974**). Human pancreatic proteolytic enzymes and protein inhibitors: isolation and molecular properties. *European Journal of Biochemistry*, 43, 569-581.
5. Srinivas, S., Prakash, V. (**2010**). Bioactive peptides from bovine milk α-casein: Isolation, characterization and multifunctional properties. *International Journal of Peptide Research and Therapeutics*, 16, 7-15.
6. Kamath, V., Niketh, S., Chandrashekar, A., Rajini, P. S. (**2007**). Chymotryptic hydrolysates of α-kafirin, the storage protein of sorghum (*Sorghum bicolor*) exhibited angiotensin converting enzyme inhibitory activity. *Food Chemistry*, 100, 306-311.
7. Giansanti, P., Tsiatsiani, L., Low, T. Y., Heck, A. J. (**2016**). Six alternative proteases for mass spectrometry–based proteomics beyond trypsin. *Nature Protocols*, 11, 993-1006.
8. Blow, D. M. (**1976**). Structure and mechanism of chymotrypsin. *Accounts of Chemical Research*, 9, 145-152.
9. Steitz, T. A., Hendekson, R., Blow, D. M. (**1969**). Structure of crystalline α-chymotrypsin: III. Crystallographic studies of substrates and inhibitors bound to the active site of α-chymotrypsin. *Journal of Molecular Biology*, 46, 337-348.
10. Adler-Nissen, J. (**1986**). Enzymic hydrolysis of food proteins. London, UK: Elsevier Applied Science Publishers.
11. Keil, B. (**1992**). Specificity of proteolysis: Springer-Verlag Berlin Heidelberg, New York, USA.
12. Burrell, M. M. (**1993**). Methods in Molecular Biology, Vol 16 Enzymes of Molecular Biology. Humana, p.
13. Lin, I. C., Sookkheo, B., Phutrakul, S., Chen, S. T., Tseng, M. J., Wang, K. T. (**1999**). Combinatorial peptide library for probing the selectivity of the s-1 subsite of proteases. *Journal of the Chinese Chemical Society*, 46, 147-152.
14. Hudáky, P., Kaslik, Gyula, Venekei, István, Gráf, L. (**1999**). The differential specificity of chymotrypsin A and B is determined by amino acid 226. *European Journal of Biochemistry*, 259, 528-533.
15. Hedstrom, L. (**2002**). Serine protease mechanism and specificity. *Chemical Reviews*, 102, 4501-4524.
16. Vorob'ev, M. M. (**2013**). Quantification of two-step proteolysis model with consecutive demasking and hydrolysis of peptide bonds using casein hydrolysis by chymotrypsin. *Biochemical Engineering Journal*, 74, 60-68.
17. Rawlings, N. D., Waller, M., Barrett, A. J., Bateman, A. (**2014**). MEROPS: The database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Research*, 42, D503-D509.
18. Galvão, C. M. A., Souza Silva, A. F., Custódio, M. F., Monti, R., Giordano, R. d. L. C. (**2001**). Controlled hydrolysis of cheese whey proteins using trypsin and α-chymotrypsin. *Applied Biochemistry and Biotechnology*, 91, 761.
19. Deng, Y., Butré, C. I., Wierenga, P. A. (**2018**). Influence of substrate concentration on the extent of protein enzymatic hydrolysis. *International Dairy Journal*, 86, 39-48.
20. Schechter, I., Berger, A. (**1967**). On the size of the active site in proteases. I. Papain. *Biochemical and Biophysical Research Communications*, 27, 157-162.

21. Schellenberger, V., Schellenberger, U., Mitin, Y. V., Jakubke, H. D. (**1990**). Characterization of the S′ -subsite specificity of bovine pancreatic α-chymotrypsin via acyl transfer to added nucleophiles. *European Journal of Biochemistry*, 187, 163-167.
22. Segal, D. M. (**1972**). Kinetic investigation of the crystallographically deduced binding subsites of bovine chymotrypsin Aγ. *Biochemistry*, 11, 349-356.
23. Schellenberger, V., Braune, K., Hofmann, H. J., Jakubke, H. D. (**1991**). The specificity of chymotrypsin: a statistical analysis of hydrolysis data. *European Journal of Biochemistry*, 199, 623-636.
24. Wright, H. T. (**1977**). Secondary and conformational specificities of trypsin and chymotrypsin. *European Journal of Biochemistry*, 73, 567-578.
25. Kuipers, B. J. H., Gruppen, H. (**2007**). Prediction of molar extinction coefficients of proteins and peptides using UV absorption of the constituent amino acids at 214 nm to enable quantitative reverse phase high-performance liquid chromatography-mass spectrometry analysis. *Journal of Agricultural and Food Chemistry*, 55, 5445-5451.
26. Butré, C. I., Sforza, S., Gruppen, H., Wierenga, P. A. (**2014**). Introducing enzyme selectivity: A quantitative parameter to describe enzymatic protein hydrolysis. *Analytical and Bioanalytical Chemistry*, 406, 5827-5841.
27. Deng, Y., van der Veer, F., Sforza, S., Gruppen, H., Wierenga, P. A. (**2018**). Towards predicting protein hydrolysis by bovine trypsin. *Process Biochemistry*, 65, 81-92.
28. Butré, C. I., Wierenga, P. A., Gruppen, H. (**2014**). Influence of water availability on the enzymatic hydrolysis of proteins. *Process Biochemistry*, 49, 1903-1912.
29. Ruan, C.-Q., Chi, Y.-J., Zhang, R.-D. (**2010**). Kinetics of hydrolysis of egg white protein by pepsin. *Czech journal of food sciences*, 28, 355-363.
30. Loveday, S. M., Peram, M. R., Singh, H., Ye, A., Jameson, G. B. (**2014**). Digestive diversity and kinetic intrigue among heated and unheated β-lactoglobulin species. *Food & Function*, 5, 2783-2791.
31. Catiau, L., Delval-Dubois, V., Guillochon, D., Nedjar-Arroume, N. (**2011**). Characterization and identification of a chymotryptic hydrolysate of alpha-Lactalbumin stimulating cholecystokinin release in STC-1 cells. *Applied Biochemistry and Biotechnology*, 165, 1264-1273.

Annex 4.1 Hydrolysis rate constants and selectivity of bovine chymotrypsin against cleavage sites hydrolysed in α-LA, β-LG and β-cas, clustered in HSS/ISS/LSS.

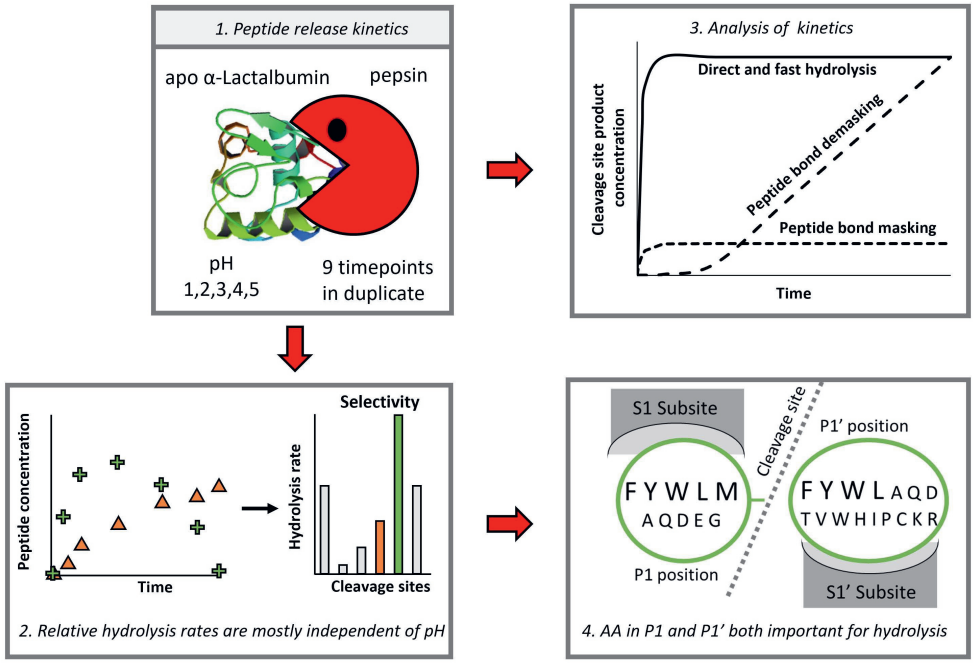
Cluster	Protein	CS	K * C0	K * C0 stdev	Sel. [%]	Sel. stdev	P4	P3	P2	P1	P1'	P2'	P3'	P4'
HSS	α-LA	F9	4.0E+03	6E+03	2.3E+01	3E+01	C	E	V	F	R	E	L	K
HSS	α-LA	L12	6.7E-04	4E-04	3.0E-06	1E-06	F	R	E	L	K	D	L	K
HSS	α-LA	F31	6.3E+03	6E+03	2.7E+01	2E+01	C	T	T	F	H	T	S	G
HSS	α-LA	H32	2.4E-03	2E-03	1.1E-05	6E-06	T	T	F	H	T	S	G	Y
HSS	α-LA	Y36	1.3E-03	2E-03	5.1E-06	6E-06	T	S	G	Y	D	T	Q	A
HSS	α-LA	Y50	7.2E-02	6E-02	3.1E-04	2E-04	S	T	E	Y	G	L	F	Q
HSS	α-LA	F53	1.3E-01	9E-02	5.6E-04	3E-04	Y	G	L	F	Q	I	N	N
HSS	α-LA	W60	2.2E-01	1E-01	1.0E-03	3E-04	N	K	I	W	C	K	D	D
HSS	α-LA	M90	1.1E-01	1E-01	4.9E-04	4E-04	D	D	I	M	C	V	K	K
HSS	α-LA	Y103	9.5E-04	8E-04	4.1E-06	3E-06	G	I	N	Y	W	L	A	H
HSS	α-LA	W104	1.1E+04	5E+03	5.0E+01	1E+01	I	N	Y	W	L	A	H	K
HSS	α-LA	L110	1.8E-01	2E-01	7.5E-04	7E-04	H	K	A	L	C	S	E	K
ISS	α-LA	L3	1.4E-05	2E-05	7.9E-08	1E-07		E	Q	L	T	K	C	E
ISS	α-LA	L15	9.0E-05	1E-04	3.8E-07	4E-07	L	K	D	L	K	G	Y	G
ISS	α-LA	Y18	1.0E-03	6E-04	4.5E-06	2E-06	L	K	G	Y	G	G	V	S
ISS	α-LA	H68	3.4E-05	2E-05	1.5E-07	6E-08	Q	N	P	H	S	S	N	I
ISS	α-LA	L105	5.3E-05	3E-05	2.4E-07	8E-08	N	Y	W	L	A	H	K	A
ISS	α-LA	K108	6.2E-05	6E-05	3.4E-07	4E-07	L	A	H	K	A	L	C	S
ISS	α-LA	W118	2.0E-04	1E-04	8.9E-07	3E-07	L	D	Q	W	L	C	E	K
ISS	α-LA	L119	8.6E-06	6E-06	4.6E-08	4E-08	D	Q	W	L	C	E	K	L
LSS	α-LA	T4	2.4E-06	1E-06	1.1E-08	3E-09	E	Q	L	T	K	C	E	V
LSS	α-LA	K13	3.0E-06	2E-06	1.4E-08	4E-09	R	E	L	K	D	L	K	G
LSS	α-LA	S22	2.8E-06	2E-06	1.2E-08	5E-09	G	G	V	S	L	P	E	W
LSS	α-LA	T29	6.7E-06	4E-06	3.0E-08	1E-08	W	V	C	T	T	F	H	T
LSS	α-LA	T33	2.8E-06	2E-06	1.2E-08	5E-09	T	F	H	T	S	G	Y	D
LSS	α-LA	N56	2.6E-06	2E-06	1.1E-08	4E-09	F	Q	I	N	N	K	I	W
LSS	α-LA	N57	4.0E-06	2E-06	1.8E-08	4E-09	Q	I	N	N	K	I	W	C
LSS	α-LA	V92	1.1E-06	2E-06	6.3E-09	9E-09	I	M	C	V	K	K	I	L
LSS	α-LA	K93	8.9E-06	5E-06	4.0E-08	1E-08	M	C	V	K	K	I	L	D
LSS	α-LA	K94	2.6E-06	2E-06	1.1E-08	6E-09	C	V	K	K	I	L	D	K
LSS	α-LA	L96	6.7E-06	4E-06	3.0E-08	9E-09	K	K	I	L	D	K	V	G
LSS	α-LA	H107	1.4E-05	8E-06	6.1E-08	2E-08	W	L	A	H	K	A	L	C
LSS	α-LA	L115	6.8E-06	9E-07	3.2E-08	4E-09	S	E	K	L	D	Q	W	L
LSS	α-LA	E121	9.5E-05	1E-04	5.5E-07	8E-07	W	L	C	E	K	L		
HSS	β-LG	Q5	3.7E-03	1E-03	4.2E-01	6E-01	I	V	T	Q	T	M	K	G
HSS	β-LG	M7	1.8E-02	9E-04	2.4E+00	3E+00	T	Q	T	M	K	G	L	D
HSS	β-LG	L10	2.0E-03	2E-03	7.8E-02	1E-01	M	K	G	L	D	I	Q	K
HSS	β-LG	W19	1.5E-03	5E-04	1.6E-01	2E-01	A	G	T	W	Y	S	L	A
HSS	β-LG	Y20	2.3E-01	3E-02	3.5E+01	5E+01	G	T	W	Y	S	L	A	M
HSS	β-LG	L32	3.2E-03	2E-03	2.9E-01	4E-01	I	S	L	L	D	A	Q	S
HSS	β-LG	L39	1.7E-02	2E-02	6.9E-01	1E+00	S	A	P	L	R	V	Y	V
HSS	β-LG	Y42	2.0E-02	4E-03	3.2E+00	5E+00	L	R	V	Y	V	E	E	L
HSS	β-LG	Q59	2.6E-03	0E+00	3.7E-01	5E-01	I	L	L	Q	K	W	E	N
HSS	β-LG	W61	1.4E-01	2E-01	1.7E+00	2E+00	L	Q	K	W	E	N	X	E
HSS	β-LG	F82	2.9E-02	6E-03	4.6E+00	7E+00	P	A	V	F	K	I	D	A
HSS	β-LG	L93	3.5E-03	3E-04	5.2E-01	7E-01	N	K	V	L	V	L	D	T
HSS	β-LG	D98	1.6E-03	1E-03	8.2E-02	1E-01	L	D	T	D	Y	K	K	Y
HSS	β-LG	Y99	6.3E-04	3E-04	6.2E-02	9E-02	D	T	D	Y	K	K	Y	L
HSS	β-LG	Y102	1.5E-02	3E-04	2.1E+00	3E+00	Y	K	K	Y	L	L	F	C
HSS	β-LG	F105	1.3E-02	6E-03	2.4E+00	3E+00	Y	L	L	F	C	M	E	N
HSS	β-LG	L122	1.6E-02	2E-03	2.5E+00	4E+00	C	Q	C	L	V	R	T	P

Cluster	Protein	CS	K * C0	K * C0 stdev	Sel. [%]	Sel. stdev	P4	P3	P2	P1	P1'	P2'	P3'	P4'
HSS	β-LG	F136	5.2E-04	2E-04	4.9E-02	7E-02	L	E	K	F	D	K	A	L
HSS	β-LG	A139	1.5E-03	1E-03	3.5E-01	5E-01	F	D	K	A	L	K	A	L
HSS	β-LG	L140	4.3E-03	5E-04	5.5E-01	8E-01	D	K	A	L	K	A	L	P
HSS	β-LG	M145	6.4E-02	9E-03	9.9E+00	1E+01	A	L	P	M	H	I	R	L
HSS	β-LG	L149	1.1E-02	2E-03	1.7E+00	2E+00	H	I	R	L	S	F	N	P
ISS	β-LG	L1	1.6E-04	8E-05	1.4E-02	2E-02				L	I	V	T	Q
ISS	β-LG	I12	4.6E-04	4E-05	6.0E-02	9E-02	G	L	D	I	Q	K	V	A
ISS	β-LG	Q13	8.3E-05	5E-05	6.4E-03	9E-03	L	D	I	Q	K	V	A	G
ISS	β-LG	S21	4.6E-04	4E-05	6.0E-02	9E-02	T	W	Y	S	L	A	M	A
ISS	β-LG	L22	2.0E-04	2E-04	4.5E-02	6E-02	W	Y	S	L	A	M	A	A
ISS	β-LG	M24	7.2E-05	5E-05	4.7E-03	7E-03	S	L	A	M	A	A	S	D
ISS	β-LG	L57	4.3E-04	2E-04	4.2E-02	6E-02	L	E	I	L	L	Q	K	W
ISS	β-LG	L58	4.1E-04	3E-04	2.8E-02	4E-02	E	I	L	L	Q	K	W	E
ISS	β-LG	I71	4.5E-05	3E-05	9.2E-03	1E-02	Q	K	K	I	I	A	E	K
ISS	β-LG	T76	5.7E-05	2E-05	6.0E-03	8E-03	A	E	K	T	K	I	P	A
ISS	β-LG	F151	1.4E-04	2E-04	3.9E-02	6E-02	R	L	S	F	N	P	T	Q
LSS	β-LG	D11	5.7E-05	1E-05	9.4E-03	1E-02	K	G	L	D	I	Q	K	V
LSS	β-LG	A16	2.5E-05	3E-05	5.9E-04	8E-04	Q	K	V	A	G	T	W	Y
LSS	β-LG	A23	4.3E-06	1E-06	5.0E-04	7E-04	Y	S	L	A	M	A	A	S
LSS	β-LG	L31	5.5E-06	1E-07	7.9E-04	1E-03	D	I	S	L	L	D	A	Q
LSS	β-LG	P38	9.8E-05	8E-05	5.3E-03	8E-03	Q	S	A	P	L	R	V	Y
LSS	β-LG	Q68	2.3E-05	2E-05	5.5E-03	8E-03	E	C	A	Q	K	K	I	I
LSS	β-LG	K69	1.9E-05	7E-06	3.5E-03	5E-03	C	A	Q	K	K	I	I	A
LSS	β-LG	K70	7.1E-05	3E-05	1.3E-02	2E-02	A	Q	K	K	I	I	A	E
LSS	β-LG	E74	6.9E-06	2E-06	1.2E-03	2E-03	I	I	A	E	K	T	K	I
LSS	β-LG	A80	1.6E-03	1E-03	3.3E-01	5E-01	K	I	P	A	V	F	K	I
LSS	β-LG	N90	2.7E-05	2E-05	2.2E-03	3E-03	L	N	E	N	K	V	L	V
LSS	β-LG	X118	7.2E-06	2E-06	8.6E-04	1E-03	Q	S	L	X	C	Q	C	L
LSS	β-LG	L156	5.6E-05	1E-05	9.2E-03	1E-02	P	T	Q	L	E	E	Q	C
HSS	β-cas	L6	6.0E+02	9E+02	1.2E+00	2E+00	L	E	E	L	N	V	P	G
HSS	β-cas	K28	1.0E-03	9E-04	1.6E-06	2E-06	R	I	N	K	K	I	E	K
HSS	β-cas	F52	2.5E-01	1E-02	2.6E-04	4E-04	I	H	P	F	A	Q	T	Q
HSS	β-cas	Q56	3.6E-03	1E-03	4.6E-06	6E-06	A	Q	T	Q	S	L	V	Y
HSS	β-cas	L58	4.2E-03	2E-03	5.3E-06	7E-06	T	Q	S	L	V	Y	P	F
HSS	β-cas	N68	1.6E-03	9E-04	2.2E-06	3E-06	P	I	P	N	S	L	P	Q
HSS	β-cas	L70	4.3E-03	3E-03	6.4E-06	9E-06	P	N	S	L	P	Q	N	I
HSS	β-cas	Q72	9.6E-04	8E-04	1.5E-06	2E-06	S	L	P	Q	N	I	P	P
HSS	β-cas	M93	1.8E-02	1E-03	1.7E-05	2E-05	P	E	V	M	G	V	S	K
HSS	β-cas	M102	7.5E-04	5E-04	1.1E-06	2E-06	K	E	A	M	A	P	K	H
HSS	β-cas	K105	1.9E-03	2E-03	3.2E-06	4E-06	M	A	P	K	H	K	E	M
HSS	β-cas	L125	8.8E-01	1E-01	9.9E-04	1E-03	S	Q	S	L	T	L	T	D
HSS	β-cas	L133	2.6E-03	1E-03	3.3E-06	5E-06	V	E	N	L	H	L	P	L
HSS	β-cas	L139	5.0E-03	4E-04	5.3E-06	7E-06	P	L	P	L	L	Q	S	W
HSS	β-cas	W143	1.7E+04	3E+03	2.0E+01	3E+01	L	Q	S	W	M	H	Q	P
HSS	β-cas	M144	3.2E-04	4E-04	6.0E-07	8E-07	Q	S	W	M	H	Q	P	H
HSS	β-cas	L163	7.0E+03	7E+03	1.2E+01	2E+01	Q	S	V	L	S	L	S	Q
HSS	β-cas	L165	2.2E-02	7E-03	2.7E-05	4E-05	V	L	S	L	S	Q	S	K
HSS	β-cas	K169	1.1E-03	1E-03	1.9E-06	3E-06	S	Q	S	K	V	L	P	V
HSS	β-cas	Q175	3.5E-02	1E-02	4.4E-05	6E-05	P	V	P	Q	K	A	V	P
HSS	β-cas	Y180	1.8E-03	2E-03	3.1E-06	4E-06	A	V	P	Y	P	Q	R	D
HSS	β-cas	F190	5.6E+03	3E+03	7.5E+00	1E+01	I	Q	A	F	L	L	Y	Q
HSS	β-cas	L191	3.4E+03	7E+02	2.9E+00	4E+00	Q	A	F	L	L	Y	Q	E
HSS	β-cas	L192	5.6E-03	1E-03	4.7E-06	6E-06	A	F	L	L	Y	Q	E	P

Cluster	Protein	CS	K * C0	K * C0 stdev	Sel. [%]	Sel. stdev	P4	P3	P2	P1	P1'	P2'	P3'	P4'
HSS	β-cas	Y193	8.6E-03	3E-03	6.5E-06	9E-06	F	L	L	Y	Q	E	P	V
ISS	β-cas	N27	2.5E-04	3E-04	4.4E-07	6E-07	T	R	I	N	K	K	I	E
ISS	β-cas	Y60	5.4E-04	8E-04	1.1E-06	2E-06	S	L	V	Y	P	F	P	G
ISS	β-cas	K99	6.4E-04	6E-04	1.0E-06	1E-06	S	K	V	K	E	A	M	A
ISS	β-cas	A101	2.9E-04	3E-04	5.1E-07	7E-07	V	K	E	A	M	A	P	K
ISS	β-cas	F111	8.7E-04	8E-04	1.5E-06	2E-06	E	M	P	F	P	K	Y	P
ISS	β-cas	F119	1.1E-03	9E-04	1.7E-06	2E-06	V	E	P	F	T	E	S	Q
ISS	β-cas	M156	2.9E-04	4E-04	5.6E-07	8E-07	P	T	V	M	F	P	P	Q
ISS	β-cas	S168	3.1E-04	4E-04	6.1E-07	9E-07	L	S	Q	S	K	V	L	P
ISS	β-cas	Q182	6.7E-03	5E-03	1.0E-05	1E-05	P	Y	P	Q	R	D	M	P
LSS	β-cas	E42	3.6E+03	6E+02	4.1E+00	6E+00	Q	Q	T	E	D	E	L	Q
LSS	β-cas	D43	4.2E+03	1E+03	3.3E+00	4E+00	Q	T	E	D	E	L	Q	D
LSS	β-cas	V82	7.3E-05	9E-05	1.4E-07	2E-07	Q	T	P	V	V	V	P	P
LSS	β-cas	S96	1.7E-05	5E-06	2.1E-08	3E-08	M	G	V	S	K	V	K	E
LSS	β-cas	E100	3.1E-04	2E-04	4.8E-07	7E-07	K	V	K	E	A	M	A	P
LSS	β-cas	K107	7.3E-05	1E-04	1.4E-07	2E-07	P	K	H	K	E	M	P	F
LSS	β-cas	N132	9.9E-02	5E-02	1.3E-04	2E-04	D	V	E	N	L	H	L	P
LSS	β-cas	P138	9.7E-02	5E-02	1.3E-04	2E-04	L	P	L	P	L	L	Q	S

CHAPTER 5

Quantitative peptide release kinetics to describe the effect of pH on pepsin preference



Vreeke, G. J. C., Vincken, J.-P., Wierenga, P. A. (**2023**). Quantitative peptide release kinetics to describe the effect of pH on pepsin preference. *Process Biochemistry*, *134*, 351-362. https://doi.org/10.1016/j.procbio.2023.10.021

Abstract

The preference of pepsin to hydrolyse certain peptide bonds is typically determined by counting peptides after hydrolysis, without considering concentrations and kinetics. In this study, peptide release was quantified to describe proteolysis by pepsin. The aim was to investigate whether pH affects individual hydrolysis rates of peptide bonds. α-Lactalbumin was hydrolysed by porcine pepsin systematically at pH 1 to 5 and peptides were identified and quantified with UPLC-PDA-MS at eight time points. Apparent pH-based differences in specificity were caused by differences in total hydrolysis rate but the relative hydrolysis rates of cleavage sites were generally independent of pH. Previous statements of pepsin preference for amino acids in the P3-P3' positions withstand when considering the hydrolysis rates of cleavage sites and were pH independent. Despite the a-specificity of pepsin, many bonds were not or slowly hydrolysed, some cleavage sites became more accessible during hydrolysis (demasking) and some became less accessible (masking).

Introduction

Protein digestion *in vivo* can be simulated using *in vitro* digestion protocols that mimic the stages of the digestive tract [1]. For *in vitro* studies, the pH of the gastric phase is often fixed at a certain pH, which differs between studies. During *in vivo* gastric digestion, the pH in the stomach increases with food intake and decreases gradually during the digestion [2]. Pepsin is known to possess proteolytic activity up to pH 5 and stability is assumed up to pH 7.5 [3, 4]. However, it is not clear how pH affects the formation and degradation of individual peptides by pepsin, which is relevant to understand digestion kinetics. One study suggested that the preference of pepsin is dependent on pH [5], while other studies did not observe any distinctive effect of pH on peptide release by pepsin [6, 7]. However, in none of these studies the concentrations of the peptides were considered and neither their formation kinetics. The first aim of this study is to understand how pH affects peptide release kinetics during pepsin hydrolysis, by comparison of the individual hydrolysis rates for (all) peptide bonds at five different pH conditions. The second aim is to understand how pepsin hydrolyses a protein. Peptide release kinetics will be correlated with the type of amino acids occupying the binding site positions (P3-P3').

What factors influence peptide release kinetics?

Several factors influence the peptides present and their concentrations during enzymatic protein hydrolysis:

- At first, the folding state of the substrate could affect the accessibility of bonds for the enzyme. For pepsin, it has frequently been reported that globular substrates, as for instance β-LG, have a low susceptibility to be hydrolysed [8, 9]. Besides that, (change in) folding state could alter the hydrolysis scenario from "one-by-one" to "zipper" according to the Linderstrøm-Lang theory. Both scenario's yield different peptides in intermediate stages of hydrolysis [10, 11].

- Secondly, the enzyme activity e.g. the amount of bonds hydrolysed per minute, determines the extent of hydrolysis after a certain incubation time. The activity of the protease depends on the pH and temperature of the hydrolysis.
- Thirdly, proteases could hydrolyse only certain types of bonds (specificity) and could have a preference within this specificity for amino acids in the binding site positions.
- Fourthly, individual cleavage sites could be hydrolysed with different rates. For instance, Deng *et al.* reported 3 log differences in hydrolysis rates by bovine trypsin, all being trypsin specific cleavage sites [12].

At last, peptide bonds could be inaccessible in the initial substrate, and become accessible after hydrolysis of other bonds. This will give a delay in the onset of hydrolysis of a certain cleavage site. Vorob'ev described this phenomenon as "demasking" [13, 14]. When peptides during the hydrolysis form aggregates, the opposite can occur and cleavage sites become less accessible during the process. This was described as "secondary masking" [15].

The effect of the hydrolysis mechanism on peptide release

According to the Linderstrøm-Lang theory, protein hydrolysis could be described by a "one-by-one" or a "zipper" scenario, depending on the rate of the first cleavage and the hydrolysis rate of intermediate peptides to small peptides [16]. A (partial) shift in the scenario could change the concentrations of intact protein, intermediate peptides and small peptides at early time points. For instance, concentrations of (bio-active) peptides were altered by unfolding haemoglobin with urea before pepsin hydrolysis [10]. A change in the hydrolysis scenario will affect peptide release kinetics. To focus on the effect of pH on the enzyme preference of pepsin for primary structure, one substrate with a secondary and tertiary structure similar at each pH was used for this study. α-Lactalbumin was brought into the apo form by depletion of calcium ions to have a constant molten globule structure independent of the pH [17, 18].

The effect of pH on rate and extent of hydrolysis

Alternating the pH affects the protonation state of the (acidic) amino acid residues, both on the substrate binding site as well as the enzyme subsite [19]. Prior studies on the effect of pH on pepsin activity focused generally on the total hydrolysis rate and the maximum degree of hydrolysis (DH_{max}) or formed peptides after digestion. Results on the effect of pH on the DH_{max} are deviating. In work of Salelles *et al.*, the degree of hydrolysis of egg white proteins was more than 2x higher at pH 2 (DH_{max} ~12 %) than that at pH 4 (DH_{max} ~5 %) after two hours of digestion [20]. A study with whey protein isolate compared digestion at static pH (pH 3) with digestion at a dynamic pH profile, which yielded comparable DH_{max} (3-4 %) [21]. In a study of Miralles *et al.*, similar peptides were released under dynamic and static conditions, but dynamic conditions showed a stepwise increase in peptide numbers, whereas static conditions yielded many peptides at early stage of digestion [22].

Pepsin mechanism, specificity, preference and the effect of pH

Pepsin is classified as an aspartic acid protease. Two aspartic acid residues are involved in the mechanism of hydrolysis, of which one is protonated (Asp32, pKa 5.02) and one deprotonated (Asp215, pKa 1.57) [23, 24]. Pepsin does not have a clearly defined specificity. Many efforts have been done to estimate the probability of hydrolysis based on the amino acids surrounding the cleavage site (binding site) [7, 25, 26]. For example, Hamuro and co-workers studied which peptide bonds were hydrolysed for 39 proteins [7]. In this study, pepsin showed a preference to hydrolyse peptide bonds after phenylalanine, leucine and methionine, hydrolysed respectively in 46, 44 and 35 % of the cases. Hydrolysis was observed in < 0.5 % of the occurrences after lysine, arginine, histidine and proline residues. A proline in the P2 position or a positively charged residue in the P3 position both negatively affected the probability of hydrolysis [7]. It is not clear how these cleavage probabilities are affected by the pH. Lockridge suggested that pepsin has a stronger preference towards favourable amino acids at pH 1.3 and hydrolyses non-specific above pH 2 [5]. However, Palashoff did not report distinctive differences in preference at pH 1.0, 2.5 and 4.0 [6]. Palashoff identified considerably less peptides at pH 4.0 than at pH 1.0 and 2.5, which was expected to be a result of lower (total) enzymatic activity.

Challenges in describing peptide release kinetics for a-specific protein hydrolysis

Already at one pH condition, many peptides are formed during the hydrolysis. To allow comparison of the different conditions, this information needs to be captured into a number of kinetic parameters. Therefore, Butré *et al.* introduced “enzyme selectivity” [27]. This quantitative parameter describes the hydrolysis rate constants of an individual cleavage site relative to the total hydrolysis rate, and is calculated based on the peptide composition at multiple time points during hydrolysis. Determination of selectivity helps to understand -and ultimately- predict protein hydrolysis. Previously it was used successfully to correlate amino acids in the binding site positions with hydrolysis efficiency, for instance for bovine trypsin and chymotrypsin [12] (**Chapter 4**). In this study, the same approach is applied on pepsin, a non-specific protease. To allow analysis of complex samples, our in-house data-processing method was recently automated (**Chapter 2**). The analysis of peptides formed by a-specific hydrolysis could be challenging because of the numerous tentative sequences that match each *m/z* signal. First, different combinations of amino acids have similar masses (isobaric). For these, distinctive MS/MS fragments need to identify the order of the amino acid residues. Secondly, amino acid stretches could occur more than once in the protein sequence(s). This could give difficulties in assigning the peptide to the right part of the substrate sequence. Third, in-source fragments could be difficult to recognize as such in complex analysis [28]. At last, every annotation software has limitations to the minimum and maximum peptide length. In our previous study, a methodology was evaluated for automated peptide identification and quantification of tryptic digests (**Chapter 2**). The same method will be applied to characterise peptides resulting from a-specific hydrolysis. Subsequently, the method will be used to understand the effect of pH on the peptide release kinetics in hydrolysis of apo α-LA.

Materials and Methods

Materials

Alpha-Lactalbumin was obtained from Davisco Foods International Inc. (Le Sueur, MN, USA). A treatment with EDTA was used to remove remaining calcium ions attached to the protein, as described by Deng *et al.* [29]. The protein content of the treated apo alpha-lactalbumin powder (α-LA) was 93 % (w/w) based on Dumas. Of the protein present, ~90 % was alpha-lactalbumin and ~10 % was beta-lactoglobulin (**Chapter 2**). Pepsin from porcine stomach mucosa (V1959) was obtained from Promega Corporation (Madison, WI, USA). The pepsin powder contained 89 % protein based on Dumas. All other chemicals were purchased either from Sigma-Aldrich (St. Louis, MO, USA) or Merck (Darmstadt, Germany).

Circular dichroism spectroscopy

The secondary and tertiary protein structures were analysed by circular dichroism (CD) measurements with the spectropolarimeter Jasco J-815 (Jasco Corp., Tokio, Japan). Apo α-LA was dissolved at 1 mg mL^{-1} (near-UV) and 0.2 mg mL^{-1} (far-UV) in millipore water and the pH was altered by addition of HCl to the desired pH. In addition, α-LA before EDTA treatment and α-LA in 10 mM $CaCl_2$ solution were analysed at neutral pH. For each sample 300 µL were transferred in a quartz cuvette with a path length of 1 mm, heated and kept at 37 °C during the measurement. Spectra were recorded as average of 10 scans at a scan speed of 50 nm min^{-1} with 2 nm bandwidth at near-UV (350-250 nm) and at the far-UV (260-190 nm). The relative content of secondary structure elements was calculated using software package CD-fit.

Table 5.1. pH dependent conditions used during hydrolysis of 1% α-LA with pepsin.

pH	Molarity HCl	Alpha at 37 °C	Amount of 1M NaOH added for inactivation (µL)	Amount of 2 M HCl added before freezing (µL)	$DH_{stat,max}$ (%)
1.0	0.5	-0.992	10	0	9.8 ±1.9
2.0	0.2	- 0.93	5	0	10.1 ±0.1
3.0	0.2	- 0.56	5	0	11.2 ±0.2
4.0	0.1	- 0.11	4	1	8.7 ±5.7
5.0	0.01	- 0.009	4	1	1.5 ±2.1

Hydrolysis in pH-stat

α-LA was dissolved in Millipore water at 1 % (weight/volume). Solutions were heated to 37 °C, adapted to the pH of the hydrolysis (**Table 5.1**) and equilibrated for 30 min. Hydrolysis was performed in duplicate for 2 hours in a pH-stat system (Metrohm, Herisau, Switzerland). The pH-stat was used to keep the pH static during hydrolysis by titrating HCl (see **table 5.1** for the molarity of added HCl). Porcine pepsin was dissolved at 50 mg mL^{-1} in Millipore water, of which 20 µL was added to the protein solution, corresponding to 1 mg pepsin / solution and an enzyme to substrate ratio of 1:100. During the hydrolysis samples were taken (200 µL) before pepsin addition, after 30 s, 60 s, 120 s, 180 s, 600 s, 1800 s, 3600 s and 7200 s of hydrolysis. The pepsin

was inactivated by changing the pH above 7.5 by addition of NaOH (**Table 5.1**). The amount titrated acid was used to calculate the degree of hydrolysis (**Equation 1**).

$$DH_{stat}[\%] = V_b \times N_b \times \frac{1}{\alpha} \times \frac{1}{m_p} \times \frac{1}{h_{tot}} \times 100\ \% \qquad \text{(Eq. 1)}$$

where V_b [mL] is the volume of added NaOH; N_b [mol L^{-1}] is the molarity of NaOH; α is the average degree of dissociation of the α-amino and α-carboxylic groups in the protein, calculated with **equation 2** [30]; m_p [g] is the amount of protein in solution; h_{tot} [mmol g^{-1}] is the number of peptide bonds per gram of protein (8.6 mmol g^{-1} for α-LA).

$$\alpha = \frac{1}{1+10^{(pKa^{NH_2}-pH)}} - \frac{1}{1+10^{(pH-pKa^{COOH})}} \qquad \text{(Eq. 2)}$$

where pH is the pH of hydrolysis, pKa^{NH2} is the average pKa of the α-amino group (7.48 at 310.15 K, calculated based on Butré *et al*. [31]), pKa^{COOH} is the average pKa of the α-carboxylic group, 3.10 according to Margot *et al.* [30].

Sample preparation for UPLC-PDA-MS

The hydrolysates (10 mg mL^{-1}) were mixed 1:1 (v/v) with a 100 mM Tris-HCl buffer pH 8 containing 20 mM DTT and incubated for 2 hours to reduce the disulphide bonds. Afterwards, the intact protein sample was diluted 10x and the hydrolysates were 5x diluted in eluent A. The diluted hydrolysates were centrifuged (10 min, 14,000 x *g*, 20 °C) and the supernatant was injected on the UPLC-PDA-MS.

Reverse phase ultra-high performance liquid chromatography (RP-UPLC)

The hydrolysates (4 µL) were injected on an Acquity Premier UPLC coupled to a Photodiode Array (PDA) detector, both from Waters (Milford, MA, USA). The column, thermostated at 30 °C, was an Acquity Premier peptide column BEH C18 2.1 mm *150 mm 300 Å 1.7 µm. A gradient was used of 1% ACN + 0.1 % TFA in UPLC-grade water (eluent A) and 1% UPLC-grade water + 0.1 % TFA in ACN (eluent B), as previously described (**Chapter 3**).

Electro spray ionisation time of flight mass spectrometry

Mass spectra were acquired with the Select Series Cyclic IMS (Waters, Milford, MA, USA). The instrument was operated in positive V-mode, without use of ion mobility. The peptides were ionised with the electrospray ionisation source at capillary voltage of 2.5 kV and a temperature of 150 °C. The sample cone was operated at 40 V and nitrogen was used as desolvation gas (500 °C, 800 L h^{-1}) and cone gas (200 L h^{-1}). The quadrupole was operated with a manual profile, focusing on 400, 500 and 600 *m/z*. Lock mass data were acquired by injection of 50 pg μL^{-1} Leucine-Enkephalin at 10 µL min^{-1}, at a capillary voltage of 2.7 kV. Fragmentation was done in the trap-cell with an MS^E ramp from 28 V to 56 V. The Time of Flight analyser was calibrated with NaI up to 4000 *m/z*. Mass spectra were collected between 50 and 3000 *m/z.*

Data processing: Peptide annotation in UNIFI

Analysis of the mass spectra was done in UNIFI version 1.8 according to **Chapter 2**. The data was processed with the sequences of α-LA (uniprot code: P00711) and β-LG A + B (uniprot code: P02754). Methionine oxidation [+16 Da] was included as modification. Processing was initially done non-specifically, with annotated peptide sequences of 3 to 40 amino acids. The thresholds in peak detection were 1000 counts for MS peaks and 250 counts for MS/MS peaks. There was no maximum number of included MS and MS/MS signals. The thresholds on mass error were 10 ppm for the MS and 20 ppm for MS/MS fragments. The MS/MS fragments were attributed to MS signals based on chromatographic peak shape. The list of tentative peptide annotations from the UNIFI software was filtered with requirements on MS/MS fragments, similarly as described in **Chapter 2**. All annotations with less than two MS/MS fragments were excluded from the analysis. Moreover, peptides were only included when more than 15 % of the possible b/y fragments were identified (relevant for peptide with 7-16 amino acids) or when at least 5 b/y fragments were identified (relevant for peptide with ≥ 17 amino acids). Adducts, recognised in UNIFI, were excluded from the analysis. The hydrolysis only yielded peptides of which the sequence occurred once in the protein sequence of α-LA.

Data processing: Analysis of large peptides and intact protein

After automated peptide identification, the chromatogram was manually checked for the presence of unidentified large peptides. One abundant high Mw peptide and intact protein were present in the data but initially not identified in the UNIFI data-processing. In case of proteolysis of α-LA by pepsin, Laureto *et al.* identified peptide 53-123 with SDS-PAGE, comprising 70 amino acid residues with a mass of 8,343 Da [32]. To annotate this peptide in a (regular) peptide analysis, the sensitivity and resolution of the mass spectrometer need to be sufficient to recognise of the 100% ^{12}C isotope peak, in order to calculate the correct mass. Both were sufficient for the mass spectrometer used in this study (**Annex 5.1**). However, the non-specific UNIFI annotation method did -by software- not annotate peptides above 40 amino acids. Therefore, the data was additionally processed with a specific processing method, in which the enzyme was specific for all amino acids in the α-LA sequence. The resulting peptides were added to the results of the regular analysis. Using this additional analysis, peptide 53-123 was identified with 17 MS/MS fragments. In some samples, even intact α-LA was annotated by the UNIFI with clear fragmentation spectrum (**Annex 5.2**). In the cases intact protein was in the chromatogram but not identified by the software, it was added manually.

Data processing: Removal of in-source fragments

The annotated and filtered peptide list exported from UNIFI contained approximately ~15 % of in-source fragments. These entries were removed using PeptQuant, an in-house written script in Matlab v2018b. Annotations were considered as in-source fragment if the peptide and potential in-source fragment eluted at the same retention time, the in-source fragment included the same sequence as the peptide and the in-source fragment had a lower MS intensity than the peptide.

Data processing: Peptide quantification

Each individual peptide was quantified based on the UV peak area at 214 nm. The UV chromatograms were integrated in Masslynx V4.2, according to **Chapter 2**. UV_{214} peak areas were included when ≥ 150 µAU x min. The molar extinction coefficient was predicted for each peptide based on Kuipers *et al*. [33]. The peptide concentrations were calculated using **equation 3**.

$$C_{peptide}\ [\mu M] = \frac{A_{214} \cdot Q}{\varepsilon_{214} \cdot l \cdot V_{inj} \cdot k_{cell}} \qquad \text{(Eq. 3)}$$

where A_{214} [µAU min] is the UV peak area at 214 nm, V_{inj} [µL] is the volume of sample injected, Q [µL min^{-1}] is the flow rate, ε_{214} [L Mol^{-1} cm^{-1}] the predicted molar extinction coefficient of the peptide and l [cm] is the path length of the UV cell, which is 1 cm according to the manufacturer. The cell constant, k_{cell} for the UV detector was 0.78 (**Chapter 2**). In this study,~44 to 85 % of the peptides shared the integrated UV peak area with at least one other peptide. In this case, the UV area was divided based on MS intensity and the predicted molar extinction coefficient of the peptides, similar to **Chapter 2**. The division of UV area based on MS intensity negatively affected the accuracy of the determined concentrations, as shown previously in **Chapter 2**. Nevertheless, over- or underestimation is expected to be low, as the cumulative concentration of each individual amino acid in the peptides is comparable with the concentration of the protein that was initially hydrolysed.

Parameters to evaluate completeness of analysis

The completeness of the peptide analysis was described by the coverage parameters previously introduced by Butré *et al*. [27]. The amino acid sequence coverage described whether all amino acids of the protein sequence were covered in at least one peptide (**equation 4**). The peptide sequence coverage considered that expected peptides are in some cases not identified, and that multiple peptides could be present that cover similar part of the protein sequence (**equation 5**). Consider **Chapter 2** for elaborate explanation.

$$Amino\ acid\ sequence\ coverage\ [\%] = \frac{\#\ unique\ annotated\ amino\ acids}{\#\ amino\ acids\ in\ protein\ sequence} \cdot 100\ \% \qquad \text{(Eq. 4)}$$

$$Peptide\ sequence\ cov.[\%] = \frac{\#\ AA\ (annotated\ peptides)}{\#\ AA\ (annotated\ peptides) + \#\ AA\ (missing\ peptides)} \cdot 100\ \% \qquad \text{(Eq. 5)}$$

The completeness of quantification was evaluated by the protein recovery. After quantification of all peptides individually, the total concentration of peptides including an amino acid in a certain position of the protein sequence was calculated. The protein recovery was the average observed amino acid concentration in perspective to the starting protein concentration (**equation 6**).

$$Protein\ recovery\ [\%] = \left(\frac{\left(\frac{\sum C_n}{\#AA_{protein}} \right)}{C_0} \right) \cdot 100\ \% \quad \text{(Eq. 6)}$$

where C_n [µM] is the concentration of each individual AA (n) in the protein sequence, and $\#AA_{protein}$ is the number of amino acids in the initial protein and C_0 [µM] is the initially injected protein concentration.

Evaluation of completeness of analysis

The amino acid sequence of α-LA was covered for 100 % by the peptides, in all time points (**Table 5.2**). The peptide sequence coverages, that considered missing peptides, ranged from 76 ±10 % (pH 5) to 92 ±6 % (pH 1) (**Table 5.2**). Typically, lower coverages were observed for early time points than for end-point digests in this data-set. For instance at pH 2, the peptide sequence coverage after 30 seconds was 79 ±2 %, whereas that of the 2 hour digest was 97 ±0 %. The observed amino acid and peptide sequence coverages were in the same range with coverages observed for α-LA hydrolysates produced by bovine trypsin and chymotrypsin (**Chapter 2 & 4**) .

Table 5.2. Sequence coverages and protein recovery for pepsin hydrolysates of α-LA. The average and standard deviation were calculated of all time points for both duplicates.

pH	Amino acid sequence coverage (%)	Peptide sequence coverage (%)	Protein recovery (%)
1.0	100 ±0	92 ±6	85 ±11
2.0	100 ±0	91 ±7	85 ±18
3.0	100 ±0	92 ±9	90 ±5
4.0	100 ±0	82 ±13	77 ±4
5.0	100 ±0	76 ±10	77 ±2

After peptide identification, each individual peptide was quantified based on UV_{214} peak area and its predicted molar extinction coefficient. In many cases, multiple peptides covered the same part of the protein sequence. Therefore, the concentration of each amino acid in the protein sequence was calculated by summation of the peptide concentrations including that amino acid (**Figure 5.1**). Protein recoveries for the hydrolysates ranged from 77 ±4 % (pH 4) to 90 ±5 % (pH 3).

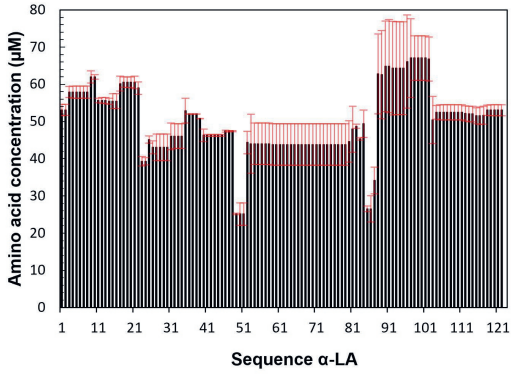


Figure 5.1. Concentration of all amino acids in the respective sequence positions of α-LA, after 2 hours hydrolysis at pH 2. Standard deviation (red) is calculated over both hydrolyses. The injected protein concentration was 56 µM.

Enzyme selectivity

The absolute peptide concentrations were used to calculate the concentration of cleavage site products $C_{t,j}$ for cleavage site j at time point t (**equation 7**).

$$C_{t,j}[\mu M] = \sum\{C_{peptide}[x-y] \mid j = x-1 \cup j = y\} \qquad \text{(Eq. 7)}$$

where $C_{t,j}$ [µM] equals the sum of all peptide concentrations ($C_{peptide}$) with sequence x-y, which are released after amino acid j or which end by amino acid j.

For each cleavage site, the product formation in time was described with first order reaction kinetics (**equation 8**) and the sequential first order kinetics (**equation 9**), similarly as previously described by Vorob'ev *et al.* [14].

$$C_{j,t}[\mu M] = 2(C_0 - C_0 \times e^{-k_{j,app} \times t}) \qquad \text{(Eq. 8)}$$

$$C_{j,t}[\mu M] = 2(C_0 \times (1 + \frac{k_1 \cdot e^{-k2 \cdot t} - k_2 \cdot e^{-k1 \cdot t}}{k_2 - k_1})) \qquad \text{(Eq. 9)}$$

where $C_{j,t}$ [µM] is the concentration cleavage site products for cleavage site j at time point t, C_0 [µM] is the (expected) protein concentration and $k_{j,app}$ [s^{-1}] is the apparent hydrolysis rate constant for cleavage site j. The C_0 concentration was 560 µM, which is the α-LA concentration during hydrolysis. For each cleavage site, four fits were created: With C_0 fixed & variable and with normal & demasking kinetics using an automated in-house script in Matlab version 2019b. Cleavage site product concentrations were fitted up to the maximum observed concentration. The fit with C_0 as variable was used as best fit when the fitted C_0 was ≤ 80% of the injected C_0. To check for demasking kinetics, the R^2 was calculated on the time points up to 20 % of the maximum observed concentration and accepted when R^2 increased by 0.05 relative to the regular fit. In case not enough data was obtained to determine k, but products were formed above 10 µM, an X was shown instead of the fitted value for k. When insufficient data were obtained and the maximum concentration was below 10 µM the cleavage sites were excluded from further analysis. A final manual check was done, where in few cases the choice of best fit was adjusted based on the results obtained at other pH values. In some cases very low hydrolysis rates lead to problems distinguishing direct from demasking kinetics. In those cases, the fit was manually adjusted to direct.

Selectivity (%) was calculated by dividing $k_{j,\,app}$ by the total hydrolysis rate, k_{tot} (**equation 10**).

$$Selectivity\ [\%] = \frac{k_i}{\sum k_i} \times 100\% \qquad \text{(Eq. 10)}$$

where k_i, [s^{-1} mg^{-1} enzyme] is the hydrolysis rate constant calculated with the first order or sequential first order fit. The total conversion rate [s^{-1} M_{pepsin}^{-1}] was calculated by dividing k_{tot} by the amount of pepsin, and was used as parameter to describe the enzyme activity.

Clustering of cleavage sites

To relate the amino acids in the binding site positions (P3-P3') to pepsin activity, the cleavage sites were divided into clusters. The cleavage sites with established hydrolysis rate constants were divided in three clusters using k-means clustering, using the common logarithm of the average selectivity over all pH conditions. The 32 cleavage sites were clustered in high selectivity sites (HSS, 12 sites, average selectivity ≥ 1 %), intermediate selectivity sites (ISS, 15 sites, average selectivity between 0.1 and 1 %) and low selectivity sites (LSS, 5 sites, average selectivity ≤ 0.1 %). Cleavage sites with significant hydrolysis, but without established k value were in the very low selectivity site cluster (VLSS, 8 sites). For the other cleavage sites no hydrolysis was observed with product formation above 10 µM (NH, 82 sites).

Results and Discussion

Secondary and tertiary protein structure

After removal of the calcium ion, the (apo) α-LA lost its globular tertiary structure (**Annex 5.3**). This was in line with previous near-UV spectra for α-LA at low ionic strength [34]. The secondary structure elements (at 37 °C) were similar for (apo) α-LA, regardless of the pH applied. On average, 29 ±3 % of the protein was present as α-helix, 19 ±4 % as β-sheet, 46 ±1 % as random coil and 7 ±2 % as β-turn. These experiments match earlier statements that apo α-LA does not have a globular structure, but has fixed secondary structural elements, typical for the molten globule state [17]. It is expected that due to the similar state of the substrate at each pH, the substrate will induce no pH effect on the peptide release kinetics.

Degree and rate of hydrolysis at different pH

The total hydrolysis rate by pepsin was similar at pH 1, 2, 3 (**Figure 5.2**), with an average rate of 2.2 ±0.4 $\cdot 10^3$ s^{-1} M^{-1}. At pH 4 and 5, the pepsin activity decreased ~100x relative to the lower pH values (average rate of 19 ±14 s^{-1} M^{-1}). This means that pH affects the (total) enzyme activity, which could be corrected for when comparing rates of individual cleavage sites at different pH conditions. The degree of hydrolysis after 2 hours incubation (DH_{max}) reached at pH 2 was 10.1 ±0.1 % (**Figure 5.2**), which was close to previously reported DH_{max} values after gastric digestion of α-LA at pH 2 (9.5 %) [35]. The DH_{max} values obtained at pH 3 and 4 were respectively 11.2 ±0.2 % and 8.7 ±5.7 %. The DH_{max} reached at pH 5 was considerably lower (1.5 ±2.1). A similar pH dependency of the DH_{max} was reported by Salelles *et al.* for casein substrates. The reproducibility between the duplicate hydrolyses with pH-stat was considerably less at pH 4 and 5 than at pH 3 and lower. This could be explained by the low value for dissociation constant α of the newly formed carboxyl and amino group: The amount of acid titrated is little and more prone to slight variations in conditions.

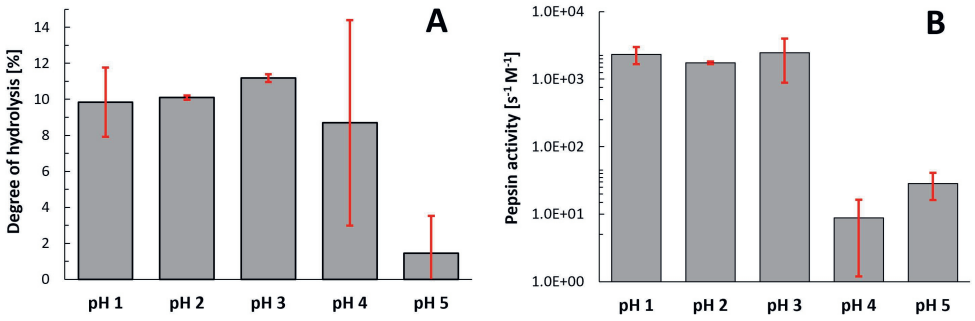


Figure 5.2. Degree of hydrolysis [%] + standard deviation after 2 hour digestion (**A**) and total hydrolysis rate [s^{-1} M^{-1}] + standard deviation (**B**) for hydrolysis of 1% α-LA by pepsin at an enzyme to substrate ratio of 1:100 at 37 °C. Total hydrolysis rate was calculated from the peptides and their concentrations in time.

The effect of pH on numeric, non-quantitative specificity

The specificity and preference of pepsin are in literature typically derived from the identified peptides after digestion (**Figure 5.3**). At pH 1-3, pepsin released peptides by hydrolysis at C-terminal site of all amino acids except proline and arginine. Most frequent, peptides were formed by hydrolysis at the C-terminal site of phenylalanine, leucine, tyrosine and aspartic acid. At pH 4-5, hydrolysis was observed after a smaller number of amino acids. For instance, no peptides were identified that were formed by hydrolysis after asparagine, serine, threonine, methionine, valine, isoleucine and cysteine. The numeric specificity indicates that pH affects the type of bonds hydrolysed by pepsin. This seems to indicate an effect of pH on specificity, although the effect is opposite as reported in the past [5]. Possibly, these 'apparent' differences in specificity could be explained by considering the kinetics and concentrations of the peptides.

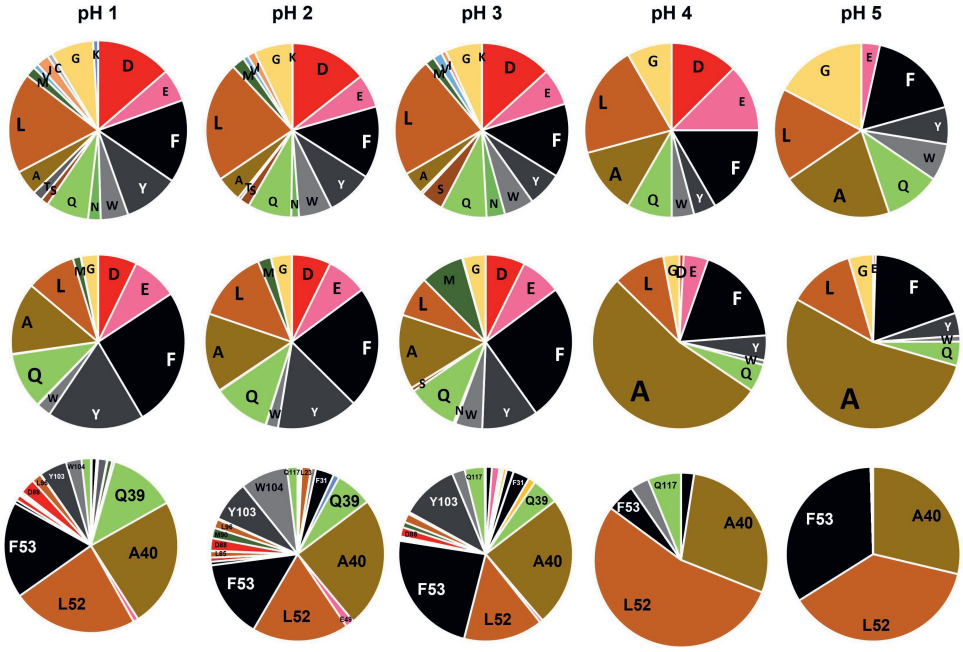


Figure. 5.3. Numeric specificity (**top**), preference for the P1 position (**middle**) and selectivity (**bottom**) of porcine pepsin upon proteolysis of α-LA at pH 1 to 5. Specificity was calculated from peptide sequences identified after 2 hours without considering peptide concentrations. Preference was calculated from peptide concentrations after 2 hours and corrected for the frequency of occurrence of each amino acid.

The effect of pH on peptide concentrations in time

Large differences were observed between peptide concentrations at pH1-3 and 4-5. The concentrations of individual peptides during hydrolysis were affected similarly by pH as the total hydrolysis rate (**Figure 5.4**). Within pH 1 to 3, peptides were released and degraded to similar concentrations in time, except that a few peptides showed slightly different concentrations at pH 3. To correct for the large differences in hydrolysis rate between data from pH 1-3 and 4-5, the time of the x-axis was normalised with the total hydrolysis rate. After correction, the data at pH4-5 followed the same curve as the peptides at pH1-3. This leads to the conclusion that peptides were released with similar relative kinetics at each pH (**Figure 5.4**). Differences in peptide composition at each pH seem mostly affected by total enzyme activity rather than changes in specificity to hydrolyse certain peptide bonds. To dive deeper in the affinity of pepsin towards certain peptide bonds, hydrolysis kinetics for individual cleavage sites were studied.

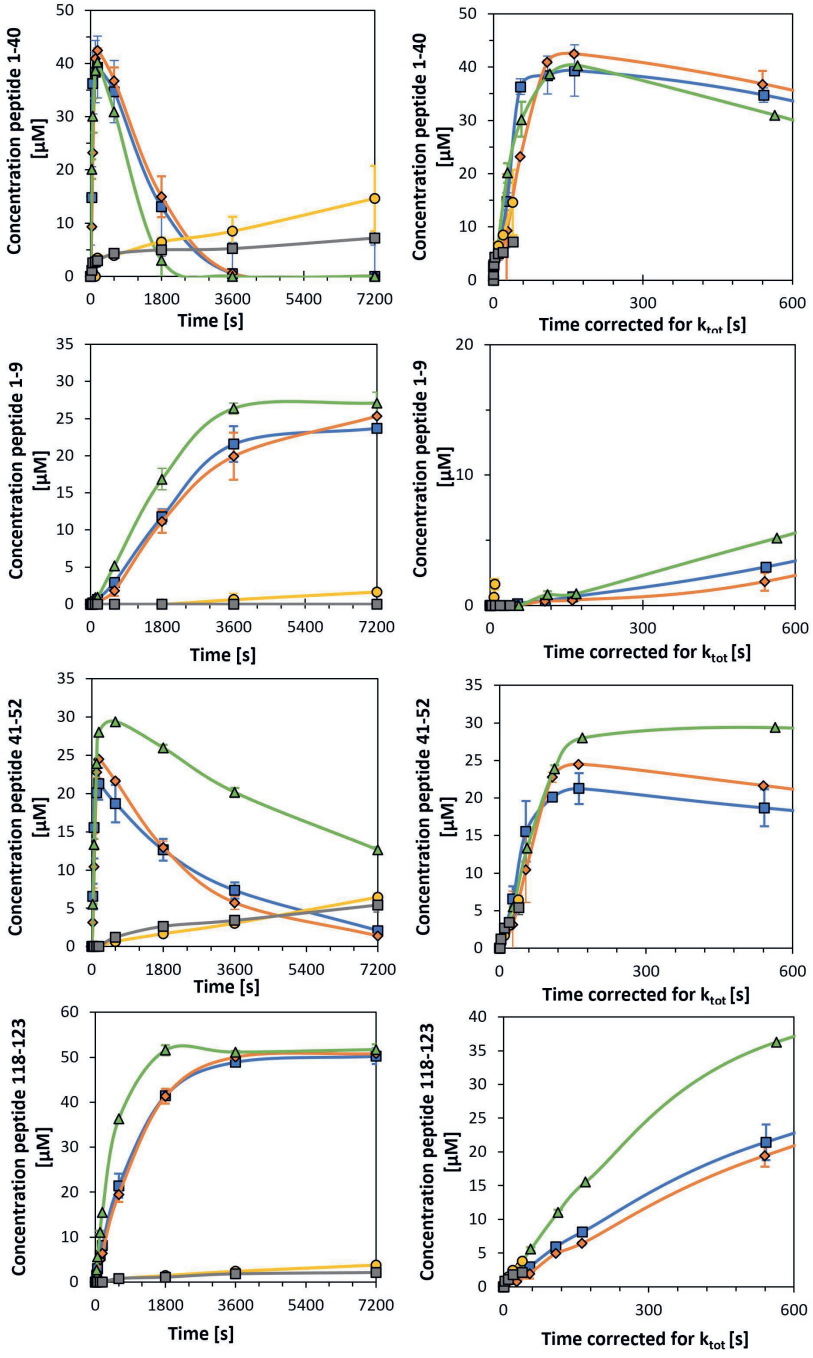


Figure 5.4. Concentrations in time for peptides 1-40, 1-9, 41-52, 118-123 during α-LA hydrolysis by pepsin, at pH 1 (■), 2 (◆), 3 (▲), 4 (●), 5 (■). Horizontal axis represents time and time corrected for total hydrolysis rate (k_{tot}).

Determining hydrolysis rate constants of individual cleavage sites

To be able to compare different pH values, the hydrolysis rate constants of individual peptide bonds were calculated (**Annex 5.5**) and (similarly) expressed relative to the total hydrolysis rate (selectivity). Hydrolysis rates for individual cleavage sites varied 4 decades (ranging from $2.1 \cdot 10^{-2}$ s^{-1} (F53, pH 3) to $3.5 \cdot 10^{-6}$ s^{-1} (Y103, pH 5)(**Annex 5.4**)). Cleavage sites had an average relative standard deviation of~34 % for k between the duplicate experiments. Approximately 35 % of the cleavage sites showed a clear delay in the onset of hydrolysis, and were fitted using demasking kinetics, as discussed later in the text. When a cleavage site showed such a delay, this was the case at each pH.

Pepsin selectivity and pH

In general, the relative hydrolysis rates were all comparable between the pH conditions 1 to 5. This means that the relative hydrolysis rate of a particular peptide bond relative to the total hydrolysis rate is more or less similar, independent of the pH applied. For instance, cleavage sites A40, L52, F53, Y103 and Q117 had major contributions to the total hydrolysis rate at all tested pH values (**Table 5.3**). Peptide bonds that were not hydrolysed by pepsin at pH 2, remained in most cases also intact at other pH conditions. A large number of cleavage sites hydrolysed at pH 1-3, did not show clear product formation at pH 4-5. This was most likely due to the ~100 x lower total hydrolysis rate at pH 4-5. Cleavage sites with intermediate or low rates did not reach detectable product concentrations within the hydrolysis time. Some minor effects of pH were noticed. For example, the selectivity of Y103 at pH 3 was ~2x higher than at pH 1 (**Annex 5.5**). It was expected that bonds with amino acid residues as glutamic acid and aspartic acid in the P1 position might become more susceptible towards pepsin hydrolysis at low pH, because the carboxyl group is in the protonated state. However, for these bonds also no significant difference was observed between the pH conditions 1-3.

What kinetics underlie the release of peptides?

The dataset showed us that there were no major changes induced by pH in selectivity. In addition, it allowed us to provide more insight into the different phenomena underlying the release of peptides. Since the selectivity was relatively similar at each pH, the kinetics measured at pH 2 were shown, which are also representative for pH 1 and 3. Three different kinetics were distinguished: (1) cleavage sites which were hydrolysed in intact protein, (2) cleavage sites that became accessible during hydrolysis and (3) cleavage sites that became less accessible during hydrolysis.

Table 5.3. Average hydrolysis rates for α-LA cleavage sites during pepsin hydrolysis at pH 1 to 5, determined with first order kinetics. An X indicates that significant product formation was observed, but k could not be established from the data. Cleavage sites are sorted from high to low average selectivity. The cleavage sites were clustered based on average selectivity [%] in high selectivity sites (HSS), intermediate selectivity sites (ISS), low selectivity sites (LSS) and very low selectivity sites (VLSS). Cleavage sites that were not hydrolysed above the detection limit at any pH were excluded. Amino acids in the cleavage site positions were shaded when acid (red), basic (green), aromatic (blue) or proline (yellow).

	Hydrolysis rate constants[1] [s^{-1}]						Amino acids in cleavage site positions					
CS	**pH 1**	**pH 2**	**pH 3**	**pH 4**	**pH 5**		**P3**	**P2**	**P1**	**P1'**	**P2'**	**P3'**
L52	1.5E-02	8.1E-03	7.3E-03	1.5E-04	2.2E-04	HSS	Y	G	L	F	Q	I
A40	1.4E-02	1.1E-02	1.3E-02	4.6E-05	2.6E-04	HSS	T	Q	A	I	V	Q
F53	1.1E-02	6.5E-03	2.1E-02	1.3E-05	2.5E-04	HSS	G	L	F	Q	I	N
Q39	7.5E-03	3.1E-03	2.6E-03			HSS	D	T	Q	A	I	V
Y103	2.9E-03	3.4E-03	8.3E-03	5.4E-06	3.5E-06	HSS	I	N	Y	W	L	A
W104	1.7E-03	4.0E-03	1.1E-03			HSS	N	Y	W	L	A	H
Q117	9.9E-04	8.8E-04	1.9E-03	8.1E-06	5.4E-06	HSS	L	D	Q	W	L	C
F31*	5.1E-04	1.5E-03	1.3E-03	3.9E-06		HSS	T	T	F	H	T	S
D88*	1.6E-03	1.0E-03	1.0E-03			HSS	T	D	D	I	M	C
L96	2.6E-03	7.4E-04	7.5E-04			HSS	K	I	L	D	K	V
L23*	9.2E-04	7.8E-04	6.1E-04	X		HSS	V	S	L	P	E	W
M90	X	1.8E-03	4.5E-04	X		HSS	D	I	M	C	V	K
E49	4.9E-04	6.8E-04	3.4E-04	X		ISS	S	T	E	Y	G	L
G35*	3.0E-04	5.0E-04	5.8E-04	X	X	ISS	T	S	G	Y	D	T
F9*	4.7E-04	1.9E-04	5.1E-04	X		ISS	E	V	F	R	E	L
D83	5.7E-04	3.5E-04	2.2E-04			ISS	L	D	D	D	L	T
L85	2.3E-04	5.1E-04	2.3E-04			ISS	D	D	L	T	D	D
F80*	3.1E-04	3.1E-04	1.8E-04			ISS	D	K	F	L	D	D
E11*	2.2E-05	3.3E-05	6.7E-04	X		ISS	F	R	E	L	K	D
D14	X		3.0E-04			ISS	L	K	D	L	K	G
G20	1.6E-05	X	3.0E-04			ISS	Y	G	G	V	S	L
W26*	1.3E-05	3.6E-04	8.7E-06			ISS	P	E	W	V	C	T
D87	2.0E-04	8.6E-05	X			ISS	L	T	D	D	I	M
D97*	6.5E-05	9.1E-05	1.9E-04	X		ISS	I	L	D	K	V	G
W118	X	X	1.7E-04			ISS	D	Q	W	L	C	E
G17*	3.0E-04	5.8E-05	2.0E-04	X		ISS	L	K	G	Y	G	G
L3	8.8E-05	9.2E-05	2.6E-05	X		ISS	E	Q	L	T	K	C
E25*	5.2E-05	3.6E-05	8.7E-06			LSS	L	P	E	W	V	C
S22	X		2.2E-05			LSS	G	V	S	L	P	E
Y18	1.5E-05		X			LSS	K	G	Y	G	G	V
D84	X	1.3E-05	1.8E-05			LSS	D	D	D	L	T	D
D46	X	X	7.2E-06			LSS	N	N	D	S	T	E
Q2			X			VLSS	-	E	Q	L	T	K
V8						VLSS	C	E	V	F	R	E
G19		X				VLSS	G	Y	G	G	V	S
Y36	X	X				VLSS	S	G	Y	D	T	Q
L81	X	X				VLSS	K	F	L	D	D	D
I95		X				VLSS	K	K	I	L	D	K
N102	X	X	X			VLSS	G	I	N	Y	W	L
L115	X		X			VLSS	E	K	L	D	Q	W

* Cleavage site product formation followed first order demasking kinetics.

High selectivity sites and zipper scenario

For 9 cleavage sites, hydrolysis started immediately after enzyme addition (**Figure 5.5A**). The hydrolysis rate constants of these cleavage sites varied widely. Cleavage sites A40 and L52 were hydrolysed fastest with rates of 1.1 ±0.4 $\cdot 10^{-2}$ s^{-1} and 8.1 ±0.6 $\cdot 10^{-3}$ s^{-1}, respectively (**Table 5.3, Annex 5.4**). Because of the fast hydrolysis of these sites, all intact protein was hydrolysed to intermediate peptides within 3 minutes, showing that pepsin hydrolysed α-LA according to a zipper scenario. It is surprising that a few sites had a much higher selectivity than other (high selectivity) sites, since all should be accessible considering that α-LA occurs as molten globule.

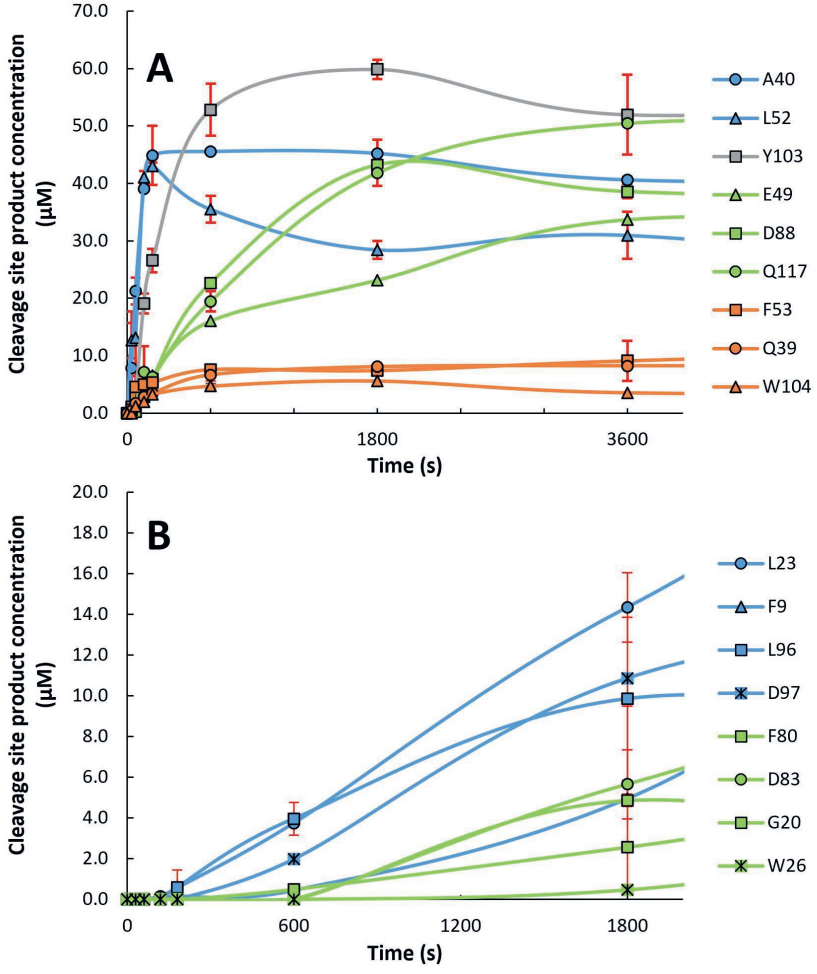


Figure 5.5. Cleavage site product concentrations in time for pepsin hydrolysis of α-LA, example of pH2. Standard deviation of the replicate hydrolysis is shown in **red**. Panel **A** shows CS for which hydrolysis starts immediately, panel **B** illustrates CS for which hydrolysis starts with a delay of 3 min (——) or 10 min (——).

Peptide bond demasking

For some cleavage sites, product formation did not start immediately after enzyme addition but seems to have a 3 min lag time (for instance F9 and L23) or even 10 min lag time (for instance F80, D83) (**Fig. 5.5B**). For these sites it seems that the initial protein (structure) has to be hydrolysed into intermediate peptides before these bonds become accessible for pepsin. This phenomenon was previously described as peptide bond 'demasking' by Vorob'ev *et al.* and was related to the denaturation of the globular protein structure [14, 36]. Our study shows that this demasking effect also occurs with hydrolysis of a molten globule structure, surprisingly. For the cleavage sites with a lag time of 10 minutes, one could debate that the loss of intact protein causes demasking, since, after 3 minutes all intact protein is already hydrolysed (**Fig. 5.5B**). Possibly, some rigid secondary structure elements as α-helices or β-sheets were still present in large peptides that limited the accessibility of pepsin. Suwareh *et al.* observed that cleavage sites in these elements had a significantly lower probability to be hydrolysed by pepsin [26]. Alternatively, cleavage sites could be inaccessible because peptides interacted with other peptides via hydrophobic interactions or disulphide bonds. Fontana *et al.* suggested that pepsin is hindered when peptides contain cysteine residues that are involved in disulphide bonds [37]. At last, the peptide bonds that show demasking kinetics might be intrinsically unstable when present in a peptide and broken non-enzymatically, potentially accelerated by the presence of protease. In a previous study using *Bacillus licheniformis* protease spontaneous cleavage of peptide bonds in presence of enzyme was also observed [38].

Masking of cleavage sites

Some cleavage sites were hydrolysed directly after enzyme addition, but reached low plateau concentrations. For instance, cleavage sites Q39, F53 and W104 had plateau concentrations of 10 to 15 % of the injected protein concentration. These three sites were next to the three cleavage sites with the highest hydrolysis rates. It seems that pepsin is able to hydrolyse these cleavage sites, given the immediate formation of product, but only when neighbouring bonds (A40-I41, L52-F53, Y103-W104) are intact. Since these neighbouring sites are hydrolysed very fast, the cleavage sites Q39, F53 and W104 become unavailable or 'masked'. For endo-protease pepsin, the binding site positions P2 and P2' need to be filled by amino acids to facilitate hydrolysis. Similar observations were done for bovine trypsin when two lysine residues were next to each other in the protein sequence [12]. The masking of cleavage sites influences the peptide composition, since the hydrolysis rate does not only depend on the amino acids in the binding site positions but also on hydrolysis rates of adjacent cleavage sites.

Does pepsin have a preference for amino acids in the P3-P3' binding site positions?

Pepsin preference for amino acids in the P1 position

The hydrolysis rate constants of the individual cleavage sites contain information to study the preference of pepsin. In previous studies, enzyme specificity, or preference for amino acids in the P1 position, has been determined by analysing the N- and C-terminal amino acids for peptides after a certain digestion time, without considering concentrations or release kinetics.

In this section, we will investigate whether considering hydrolysis rates will change our perspective on the preference of pepsin. First, the cleavage sites with established k were divided in three clusters based on their hydrolysis rate (high-, intermediate-, and low selectivity sites clusters). Phenylalanine (F) and methionine (M) were most preferred in the P1 position, with 50 % and 100 % of the cleavage sites in the high selectivity cluster (**Figure 5.6**). The preference for methionine should be considered with caution, because there is only one occurrence in the sequence of α-LA. Other high selectivity sites had in the P1 position aspartic acid residues (D), alanine (A), glutamine (Q), leucine (L) and the aromatic amino acids (Y, W). These results clearly confirm that of pepsin is tolerant to many different amino acids in the P1 position. Hydrolysis was never observed after threonine (T), proline (P), cysteine (C) and the positively charged residues (K, R, H). These results are in line with the cleavage probabilities described by Powers *et al.* and Hamuro *et al.* [7, 25], of which the cleavage probabilities were extracted and visualised (**Annexes 5.6-5.7**). For peptide bonds with amino acids that seem favourable in the P1 position, still a relatively high number were hydrolysed at lower rate or not at all. The amino acid in the P1 position seems therefore not dominant for hydrolysis and might be influenced by the amino acids in the other binding site positions.

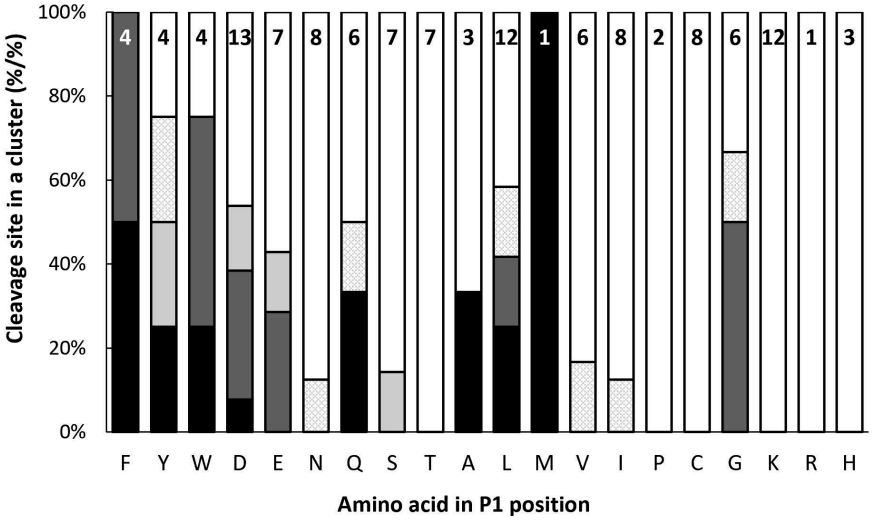


Figure 5.6. Cleavages sites hydrolysed with high selectivity (HSS ; ■), intermediate selectivity (ISS; ▩), low selectivity (LSS; □), or significantly hydrolysed but not enough data to fit k (VLSS; □) and cleavage sites never hydrolysed above detection limit (NH;□) plotted based on the amino acid in the P1 position. The numbers in the bars indicate the number of occurrences of that amino acid residue in the sequence of α-LA.

Influence of the amino acid in P1' binding site position

For the P1' position pepsin showed a preference for aromatic amino acids, which was also similar to previous studies that did not consider peptide concentrations and kinetics (**Table 5.4, Annexes 5.6-5.7**) [7]. Peptide bonds with an aromatic amino acid in the P1' position were hydrolysed 2.1x more than the average of all amino acids (**Table 5.4**). A similar high cleavage

probability was observed when P1' was occupied by leucine (1.7x average). The data derived from Hamuro *et al.* showed also a higher cleavage probability when the P1' position was occupied by an aromatic amino acid (30 %) than by other amino acids (12 %). In our data, high selectivity sites had a preferred amino acid in both the P1 and P1' positions, as for instance cleavage sites L52-F53 and Y103-W104. But for other examples, the P1' position was occupied by an aromatic residue, whereas the P1 position was occupied by a non-preferred residues as glycine (G35-Y36) or glutamic acid (E49-Y50). Still, both examples were ISS and HSS, respectively. Based on the numbers and observations, we observe that the amino acid in the P1' position had almost a similar influence in the probability of cleavage as the amino acid in the P1 position. We see two possible explanations:

1. The interaction of the amino acid in P1' with the S1' subsite is as important as the interaction of P1 with the S1 subsite. A good interaction between P1' and S1' could compensate for a less favourable amino acid in the P1 position.
2. Pepsin cleaves in some cases at the N-terminus of (certain) amino acids instead of the C-terminus. This would means that the N-terminus and C-terminus of the substrate sequence were oriented in opposite direction in the pepsin binding groove.

Table 5.4. Number of cleavage sites with type of amino acid residue in the P1 and P1' position that were hydrolysed (HSS/ISS/LSS/VLSS) or not hydrolysed (NH).

Amino acid in P1 position	Amino acid in the P1' position	Hydrolysed	Not hydrolysed	Total	Ratio observed /expected[1] Hydrolysed	Ratio observed /expected[1] Non-hydrolysed
F, Y, W	F, Y, W	1	0	1	3.1	0.0
	L	3	0	3	3.1	0.0
	Other	6	2	8	2.3	0.4
L	F, Y, W	1	0	1	3.1	0.0
	L	0	0	0	/	/
	Other	6	6	12	1.5	0.7
Other	F, Y, W	7	3	10	2.1	0.4
	L	6	5	11	1.7	0.7
	Other	10	66	76	0.4	1.3
Total	Total	40	82	122		

[1]The ratio observed/expected was calculated by dividing the number of cleavage sites hydrolysed or non-hydrolysed by the expected number of cleavage sites based on the total number hydrolysed or non-hydrolysed. For instance, other amino acids were hydrolysed 7 out of 10 times when F, Y, W was in P1' position. This was 2.1 x the expected number of hydrolysed occurrences, considering that 40 out of 122 cleavage sites were hydrolysed.

Influence of the amino acid in other binding site positions

The effect of amino acids flanking the P1 and P1' positions were also investigated (**Annex 5.8**). Previous studies described hindrance when histidine, lysine or arginine occupied the P3 position or when proline was in the P2' position [7, 25]. Out of 13 cleavage sites with a positively charged

residue in the P3 position, one cleavage site (L96) was hydrolysed with high selectivity. The effect of proline in the P2' position should be tested with a substrate with a higher proline content, considering there are only two proline residues in α-LA. Analysis of the amino acids in the P3, P2, P2' and P3' positions did not lead to dominant factors that influence the probability of hydrolysis and were not yet suggested previously [39].

Cause of pepsin selectivity

The amino acids in the P3-P3' binding site positions by themselves are not decisive for hydrolysis to occur, but do influence the probability that a peptide bond will be hydrolysed. The question what causes the differences in selectivity still remains. The influence of the substrate tertiary and secondary structure seems limited as well. Since α-LA is present in the molten globule state during the applied conditions, it is expected that all bonds are accessible for pepsin, especially after hydrolysis of the intact protein sequence. In a study of Suwareh *et al*., it was observed that hydrolysis by pepsin was significantly less likely when potential cleavage sites occurred in rigid secondary structure elements [26]. It was therefore expected that bonds with high hydrolysis rates would not occur in secondary structure elements as α-helices or β-sheets. However, the three peptide bonds with the highest selectivity in our data, are positioned in different structural elements, respectively in a β-sheet, β-turn and α-helix (**Annex 5.9**). We hypothesise that it is most likely that the binding enthalpy of the substrate sequence with the substrate binding groove of pepsin determine the probability of hydrolysis. The binding enthalpy depends on the amino acids in the binding site positions, and their interactions with the subsite of pepsin. Since amino acids in some positions could affect the 3-dimensional orientation and subsite interactions of amino acids in other positions, it is not possible to link the primary sequence to selectivity directly from analysing the amino acids in the binding site positions. Towards predicting the selectivity of pepsin, it would be promising to use a protein-protein docking approach, which considers 3-dimensional orientation of side chains and their interactions as well as the (energy) favourable binding orientation. This study showed that different cleavage sites are hydrolysed at different rates and peptide bonds could be masked or demasked for hydrolysis, which creates perspective for modelling approaches and predictions of hydrolysis by pepsin.

Conclusion

Comparing peptides after hydrolysis indicates differences in pepsin specificity as function of pH. However, by quantification of peptide concentrations at multiple time points, we showed that these observed differences were due to the total hydrolysis rate. The relative hydrolysis rates of individual peptide bonds (enzyme selectivity) were generally similar in the pH range 1-5 for α-LA. Further analysis of the peptide release kinetics unravelled a major role of peptide bond masking and -unexpectedly- demasking. Due to both phenomena, many of the peptide bonds within the preference of pepsin, were not -or initially not- hydrolysed by pepsin, despite the intrinsic ability of pepsin to hydrolyse these bonds.

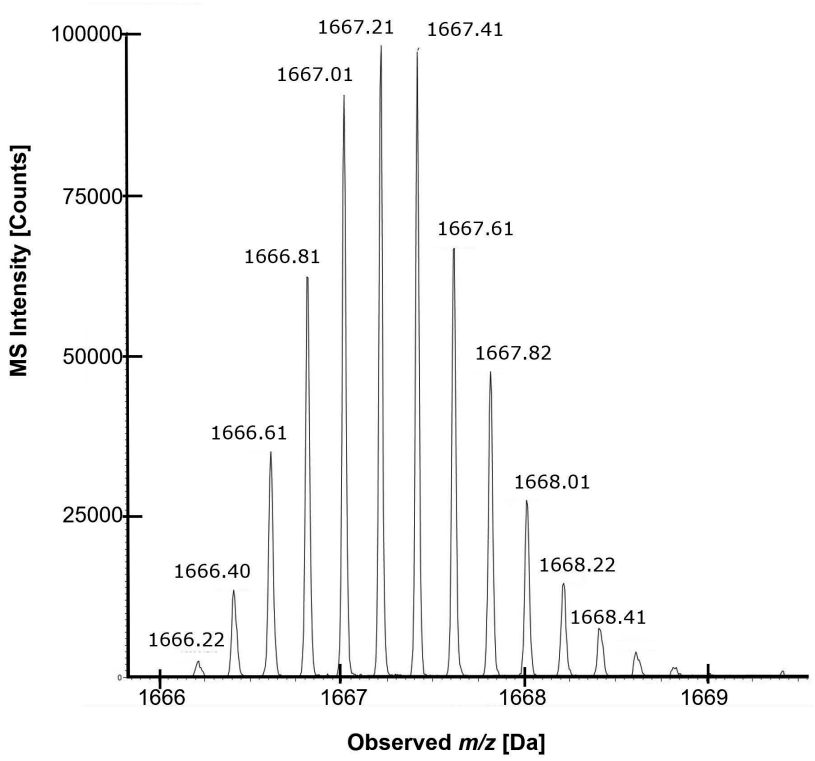
References

1. Minekus, M., Alminger, M., Alvito, P., Ballance, S., Bohn, T., Bourlieu, C., Carriere, F., Boutrou, R., Corredig, M., Dupont, D., Dufour, C., Egger, L., Golding, M., Karakaya, S., Kirkhus, B., Le Feunteun, S., Lesmes, U., Macierzanka, A., Mackie, A., Marze, S., McClements, D. J., Menard, O., Recio, I., Santos, C. N., Singh, R. P., Vegarud, G. E., Wickham, M. S. J., Weitschies, W., Brodkorb, A. (**2014**). A standardised static *in vitro* digestion method suitable for food - an international consensus. *Food & Function*, 5, 1113-1124.
2. Malagelada, J.-R., Longstreth, G. F., Summerskill, W. H. J., Go, V. L. W. (**1976**). Measurement of gastric functions during digestion of ordinary solid meals in man. *Gastroenterology*, 70, 203-210.
3. Piper, D. W., Fenton, B. H. (**1965**). pH stability and activity curves of pepsin with special reference to their clinical importance. *Gut*, 6, 506-508.
4. Kheroufi, A., Brassesco, M. E., Campos, D. A., Mouzai, A., Boughellouta, H., Pintado, M. E. (**2022**). Whey protein-derived peptides: The impact of chicken pepsin hydrolysis upon whey proteins concentrate on their biological and technological properties. *International Dairy Journal*, 134, 105442.
5. Lockridge, O., Adkins, S., La Du, B. N. (**1987**). Location of disulfide bonds within the sequence of human serum cholinesterase. *Journal of Biological Chemistry*, 262, 12945-12952.
6. Palashoff, M. H. (**2008**). Determining the specificity of pepsin for proteolytic digestion. Northeastern University.
7. Hamuro, Y., Coales, S. J., Molnar, K. S., Tuske, S. J., Morrow, J. A. (**2008**). Specificity of immobilized porcine pepsin in H/D exchange compatible conditions. *Rapid Communications in Mass Spectrometry*, 22, 1041-1046.
8. Reddy, I. M., Kella, N. K. D., Kinsella, J. E. (**1988**). Structural and conformational basis of the resistance of β-lactoglobulin to peptic and chymotryptic digestion. *Journal of Agricultural and Food Chemistry*, 36, 737-741.
9. Loveday, S. M., Peram, M. R., Singh, H., Ye, A., Jameson, G. B. (**2014**). Digestive diversity and kinetic intrigue among heated and unheated β-lactoglobulin species. *Food & Function*, 5, 2783-2791.
10. Dubois, V., Nedjar-Arroume, N., Guillochon, D. (**2005**). Influence of pH on the appearance of active peptides in the course of peptic hydrolysis of bovine haemoglobin. *Preparative Biochemistry and Biotechnology*, 35, 85-102.
11. Sanchez-Reinoso, Z., Cournoyer, A., Thibodeau, J., Said, L. B., Fliss, I., Bazinet, L., Mikhaylin, S. (**2021**). Effect of pH on the antimicrobial activity and peptide population of pepsin hydrolysates derived from bovine and porcine hemoglobins. *ACS Food Science & Technology*, 1, 1687-1701.
12. Deng, Y., van der Veer, F., Sforza, S., Gruppen, H., Wierenga, P. A. (**2018**). Towards predicting protein hydrolysis by bovine trypsin. *Process Biochemistry*, 65, 81-92.
13. Vorob'ev, M. M. (**2013**). Quantification of two-step proteolysis model with consecutive demasking and hydrolysis of peptide bonds using casein hydrolysis by chymotrypsin. *Biochemical Engineering Journal*, 74, 60-68.
14. Vorob'ev, M. M., Butré, C. I., Sforza, S., Wierenga, P. A., Gruppen, H. (**2016**). Demasking kinetics of peptide bond cleavage for whey protein isolate hydrolysed by *Bacillus licheniformis* protease. *Journal of Molecular Catalysis B: Enzymatic*, 133, S426-S431.
15. Vorob'ev, M. M. (**2022**). Modeling of Proteolysis of β-Lactoglobulin and β-Casein by Trypsin with Consideration of Secondary Masking of Intermediate Polypeptides. *International Journal of Molecular Sciences*, 23.
16. Adler-Nissen, J. (**1976**). Enzymic hydrolysis of proteins for increased solubility. *Journal of Agricultural and Food Chemistry*, 24, 1090-1093.
17. Kataoka, M., Tokunaga, F., Kuwajima, K., Goto, Y. (**1997**). Structural characterization of the molten globule of α-lactalbumin by solution X-ray scattering. *Protein Science*, 6, 422-430.

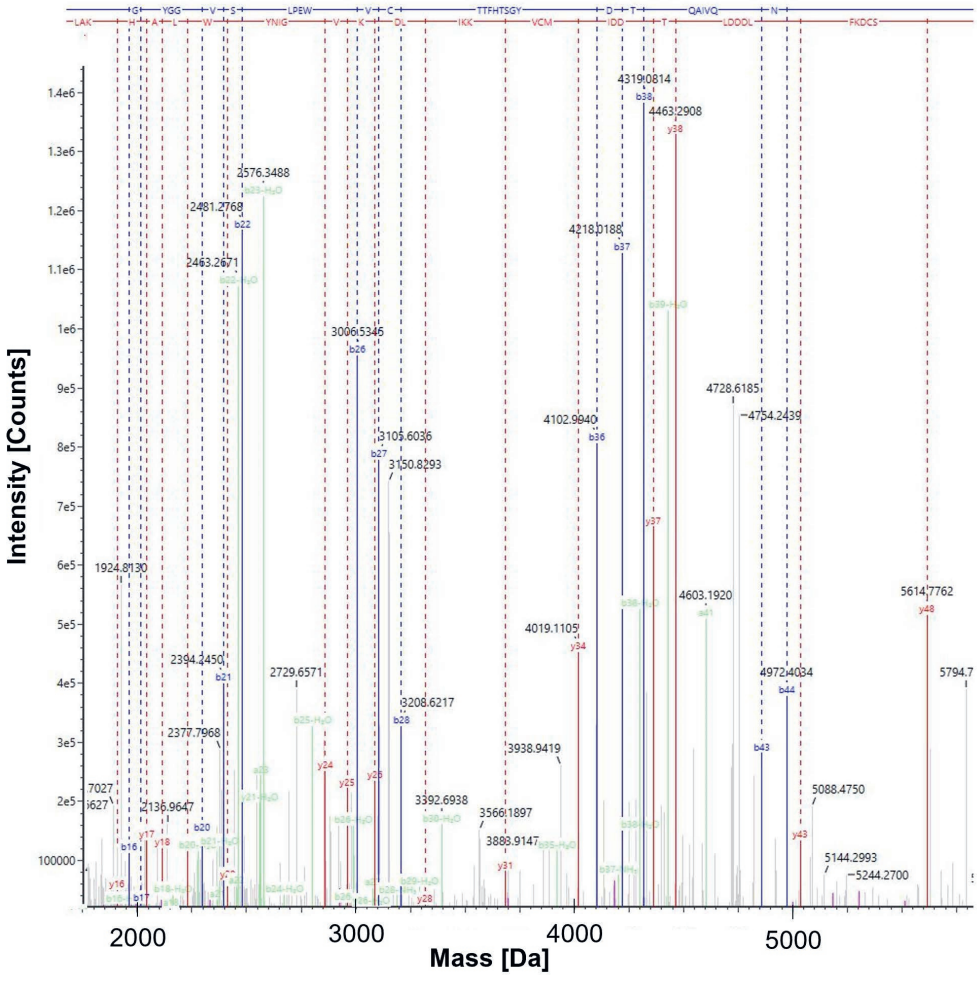
18. Fernández, A., Menéndez, V., Riera, F. A. (**2012**). α-Lactalbumin solubilisation from a precipitated whey protein concentrates fraction: pH and calcium concentration effects. *International Journal of Food Science & Technology*, 47, 467-474.
19. Giussani, L., Fois, E., Gianotti, E., Tabacchi, G., Gamba, A., Coluccia, S. (**2008**). The determination of pepsin dimensions at different pH values: A simulation study. *Nuovo Cim della Soc Ital di Fis B*, 123, 1477-1483.
20. Salelles, L., Floury, J., Le Feunteun, S. (**2021**). Pepsin activity as a function of pH and digestion time on caseins and egg white proteins under static *in vitro* conditions. *Food & Function*, 12, 12468-12478.
21. Mat, D. J. L., Cattenoz, T., Souchon, I., Michon, C., Le Feunteun, S. (**2018**). Monitoring protein hydrolysis by pepsin using pH-stat: *In vitro* gastric digestions in static and dynamic pH conditions. *Food Chemistry*, 239, 268-275.
22. Miralles, B., Del Barrio, R., Cueva, C., Recio, I., Amigo, L. (**2018**). Dynamic gastric digestion of a commercial whey protein concentrate. *Journal of the Science of Food and Agriculture*, 98, 1873-1879.
23. Andreeva, N. S., Rumsh, L. D. (**2001**). Analysis of crystal structures of aspartic proteinases: On the role of amino acid residues adjacent to the catalytic site of pepsin ‐ like enzymes. *Protein Science*, 10, 2439-2450.
24. Hofer, F., Kraml, J., Kahler, U., Kamenik, A. S., Liedl, K. R. (**2020**). Catalytic site pKa values of aspartic, cysteine, and serine proteases: Constant pH MD simulations. *Journal of Chemical Information and Modeling*, 60, 3030-3042.
25. Powers, J. C., Harley, A. D., Myers, D. V. (**1977**). Subsite specificity of porcine pepsin. In: Tang J, editor. Acid Proteases: Structure, Function, and Biology. New York, NY: Springer US, p. 141-157.
26. Suwareh, O., Causeur, D., Jardin, J., Briard-Bion, V., Le Feunteun, S., Pezennec, S., Nau, F. (**2021**). Statistical modeling of *in vitro* pepsin specificity. *Food Chemistry*, 362, 130098.
27. Butré, C. I., Sforza, S., Gruppen, H., Wierenga, P. A. (**2014**). Introducing enzyme selectivity: A quantitative parameter to describe enzymatic protein hydrolysis. *Analytical and Bioanalytical Chemistry*, 406, 5827-5841.
28. Kim, J.-S., Monroe, M. E., Camp, D. G., Smith, R. D., Qian, W.-J. (**2013**). In-Source fragmentation and the sources of partially tryptic peptides in shotgun proteomics. *Journal of Proteome Research*, 12, 910-916.
29. Deng, Y., Butré, C. I., Wierenga, P. A. (**2018**). Influence of substrate concentration on the extent of protein enzymatic hydrolysis. *International Dairy Journal*, 86, 39-48.
30. Margot, A., Flaschel, E., Renken, A. (**1994**). Continuous monitoring of enzymatic whey protein hydrolysis. Correlation of base consumption with soluble nitrogen content. *Process Biochemistry*, 29, 257-262.
31. Butré, C. I., Wierenga, P. A., Gruppen, H. (**2014**). Influence of water availability on the enzymatic hydrolysis of proteins. *Process Biochemistry*, 49, 1903-1912.
32. Polverino de Laureto, P., De Filippis, V., Di Bello, M., Zambonin, M., Fontana, A. (**1995**). Probing the molten globule state of alpha-lactalbumin by limited proteolysis. *Biochemistry*, 34, 12596-12604.
33. Kuipers, B. J. H., Gruppen, H. (**2007**). Prediction of molar extinction coefficients of proteins and peptides using UV absorption of the constituent amino acids at 214 nm to enable quantitative reverse phase high-performance liquid chromatography-mass spectrometry analysis. *Journal of Agricultural and Food Chemistry*, 55, 5445-5451.
34. Heijnis, W. H., Wierenga, P. A., Van Berkel, W. J. H., Gruppen, H. (**2010**). Directing the oligomer size distribution of peroxidase-mediated cross-linked bovine α-lactalbumin. *Journal of Agricultural and Food Chemistry*, 58, 5692-5697.
35. Xie, D., Du, L., Lin, H., Su, E., Shen, Y., Xie, J., Wei, D. (**2022**). In vitro-in silico screening strategy and mechanism of angiotensin I-converting enzyme inhibitory peptides from α-lactalbumin. *LWT*, 156, 112984.
36. Vorob'ev, M. M., Vogel, V., Mäntele, W. (**2013**). Demasking rate constants for tryptic hydrolysis of β-casein. *International Dairy Journal*, 30, 33-38.

37. Fontana, A., Polverino de Laureto, P., De Filippis, V., Scaramella, E., Zambonin, M. (**1997**). Probing the partly folded states of proteins by limited proteolysis. *Folding and Design*, 2, R17-R26.
38. Butré, C. I., Buhler, S., Sforza, S., Gruppen, H., Wierenga, P. A. (**2015**). Spontaneous, non-enzymatic breakdown of peptides during enzymatic protein hydrolysis. *Biochimica et Biophysica Acta - Proteins and Proteomics*, 1854, 987-994.
39. Tonda, A., Grosvenor, A., Clerens, S., Le Feunteun, S. (**2017**). *In silico* modeling of protein hydrolysis by endoproteases: a case study on pepsin digestion of bovine lactoferrin. *Food & Function*, 8, 4404-4413.

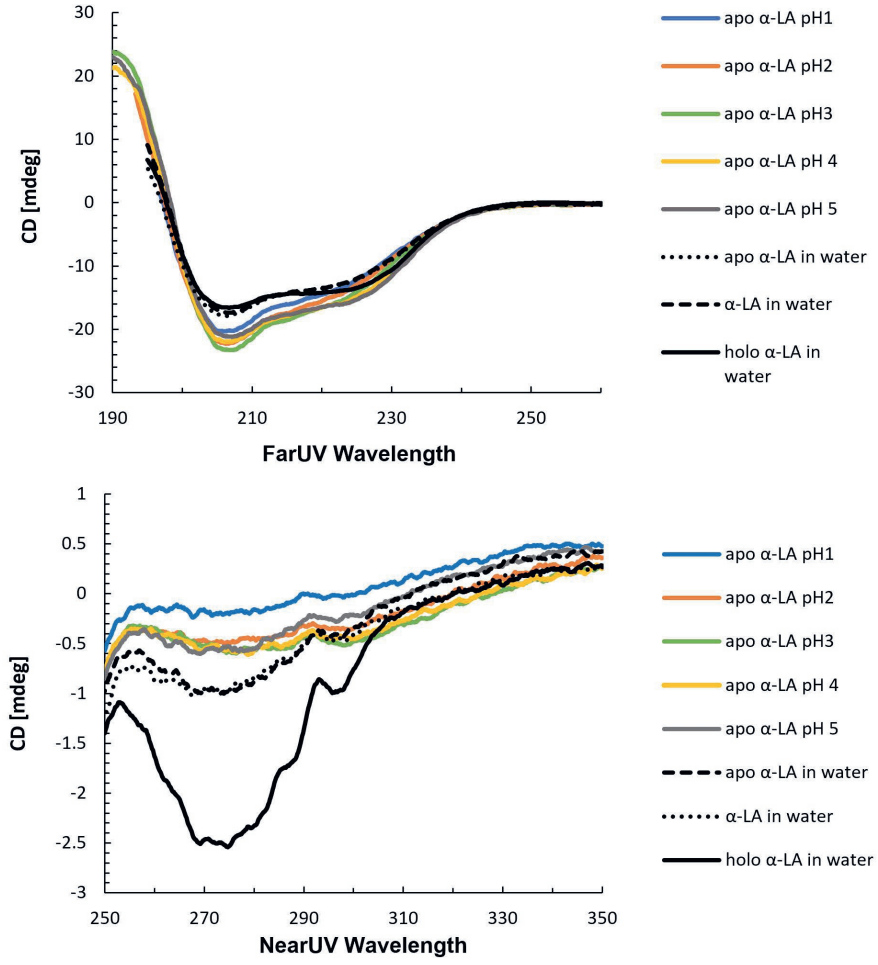
Annexes chapter 5



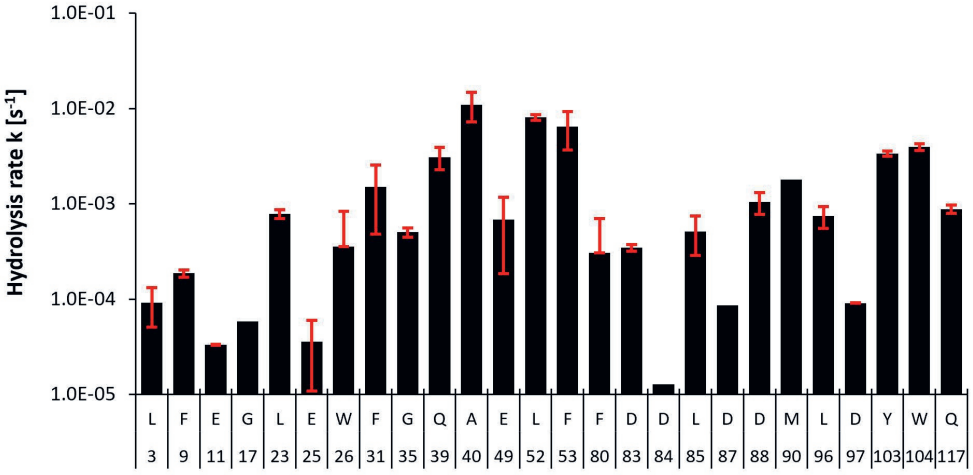
Annex 5.1. Peptide α-LA 53-123 [M+5H] identified with UNIFI, after 1 minute of pepsin hydrolysis at pH 2. The separation of the isotope peaks allowed automated identification of this peptide in semi-specific peptide analysis.



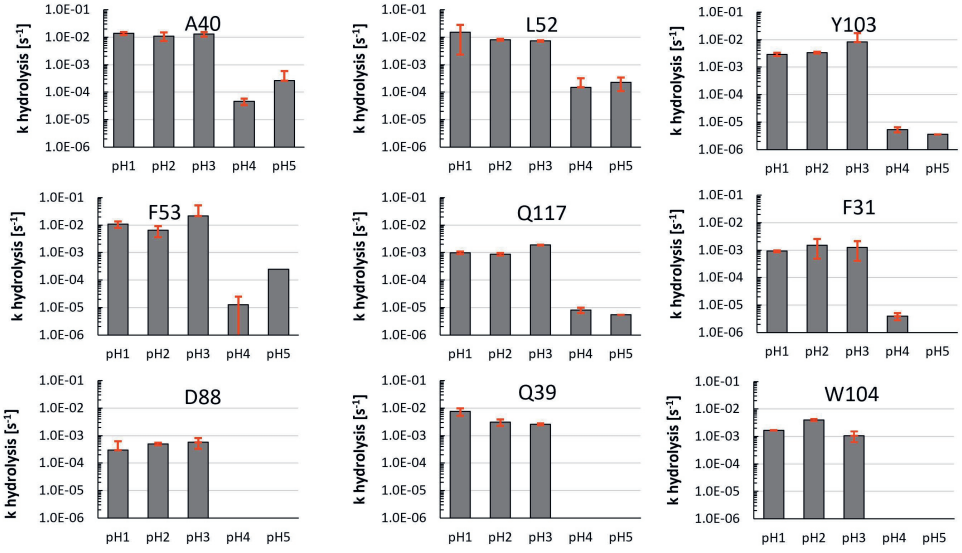
Annex 5.2. Part of the MS/MS spectrum of intact α-LA at pH 2 after 30 s, annotated in UNIFI.



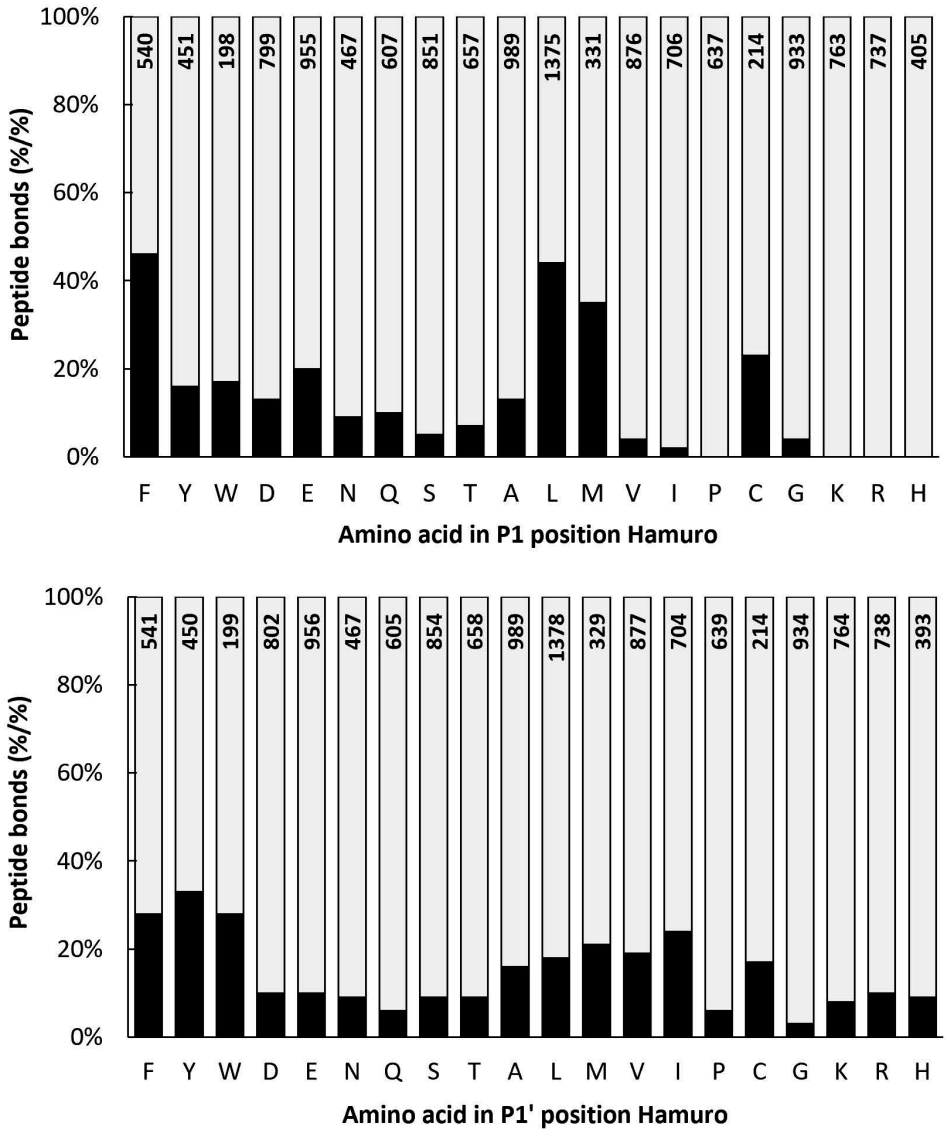
Annex 5.3. Far-UV spectra (top) and Near-UV spectra (bottom) for α-LA at 37 °C.



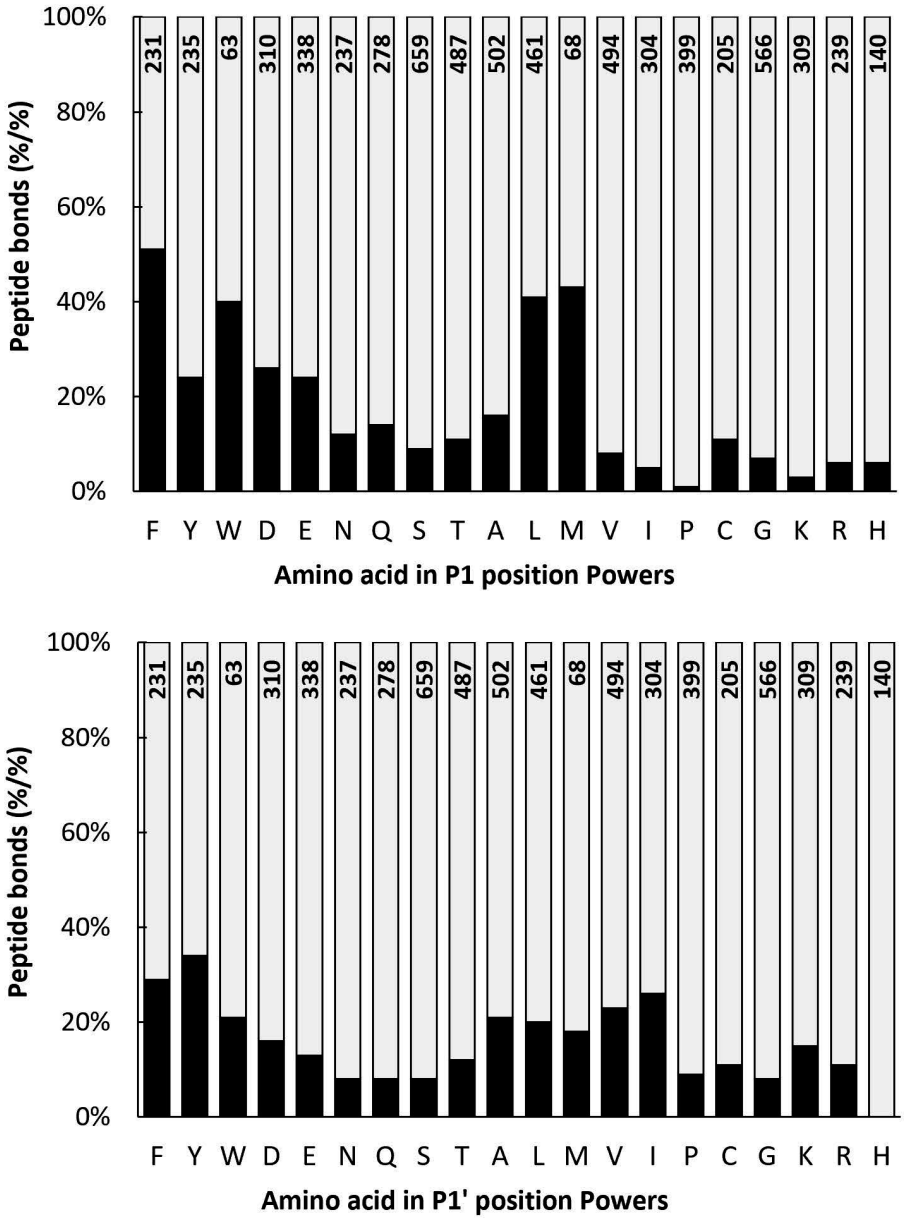
Annex 5.4. Hydrolysis rates of pepsin cleavage sites α-LA at pH 2. Standard deviation (**red**) is calculated over two biological replicates. Selectivity is plotted on a logarithmic scale.



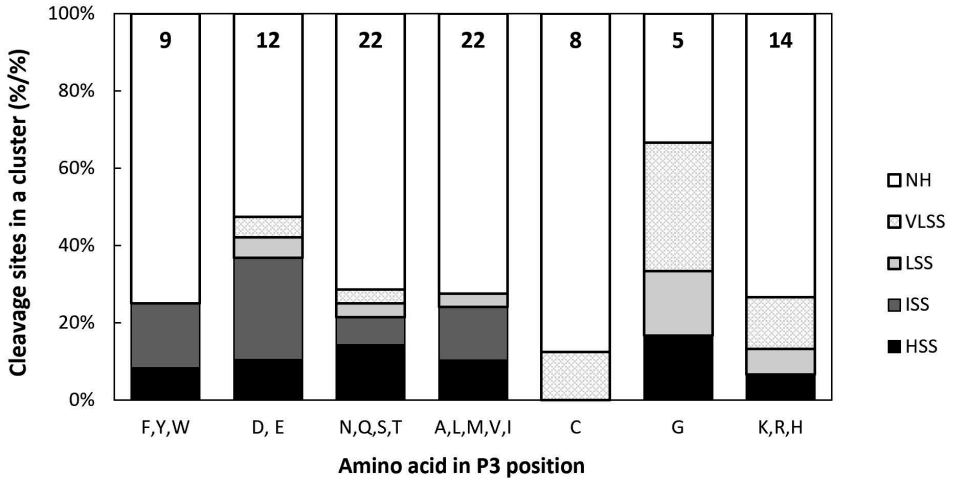
Annex 5.5. Hydrolysis rate constants [s^{-1}] of several cleavage sites in α-LA at pH 1-5 during pepsin hydrolysis. Standard deviation (**red**) is calculated over the duplicate hydrolyses.

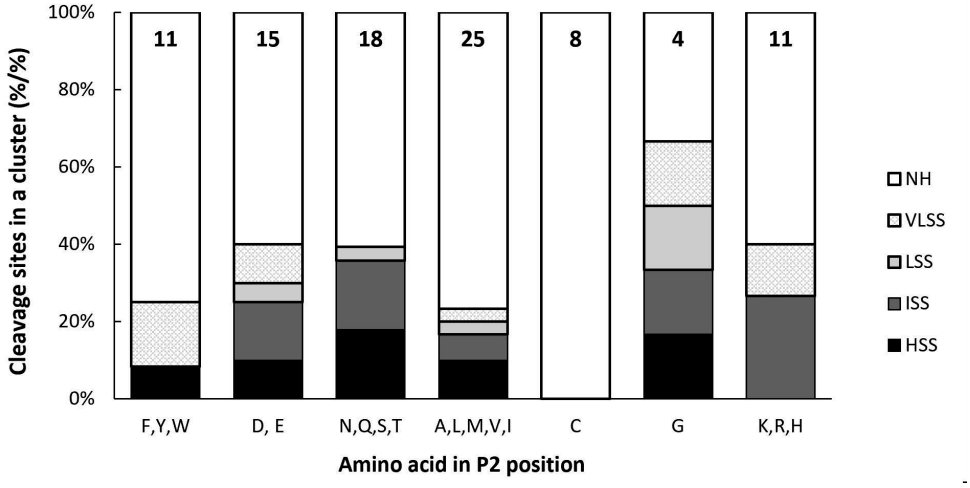


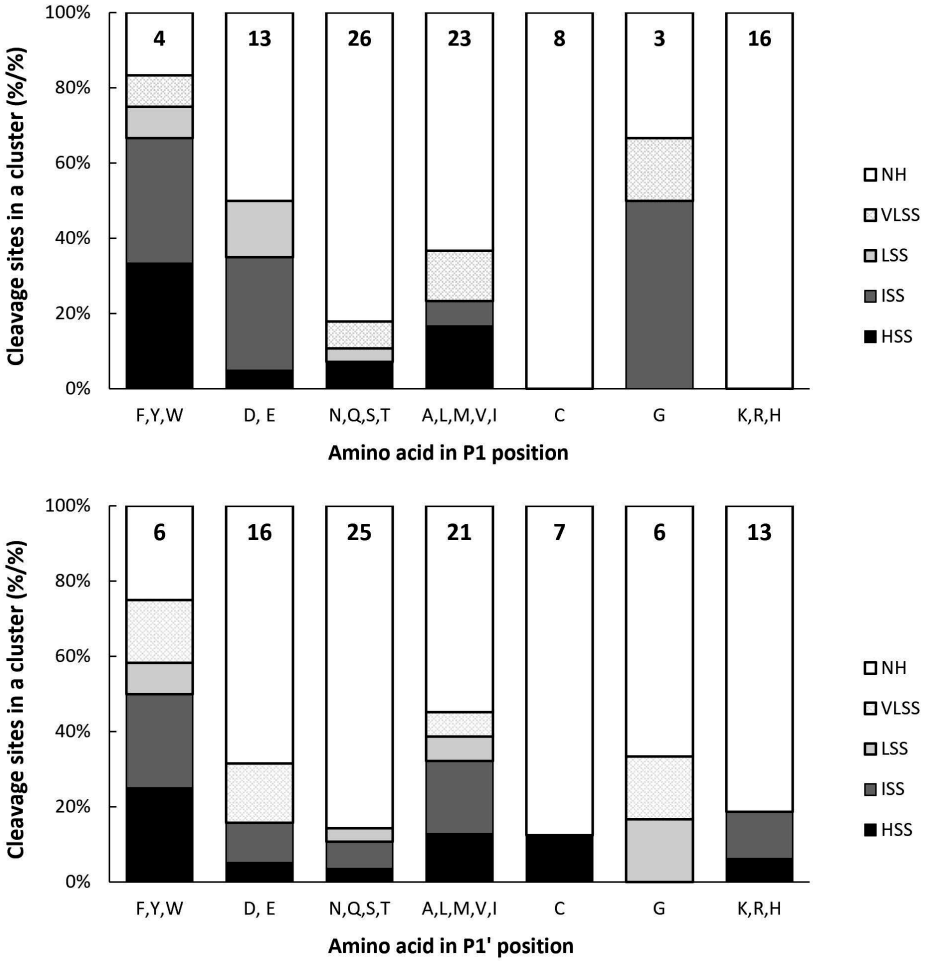
Annex 5.6. Cleavage site probabilities extracted from data of Hamuro *et al.* Percentage of hydrolysed (■) and intact (□) observations was categorised based on the amino acid in the P1 position (top) and P1' position (bottom). The number in the bar indicates the total number of observations.

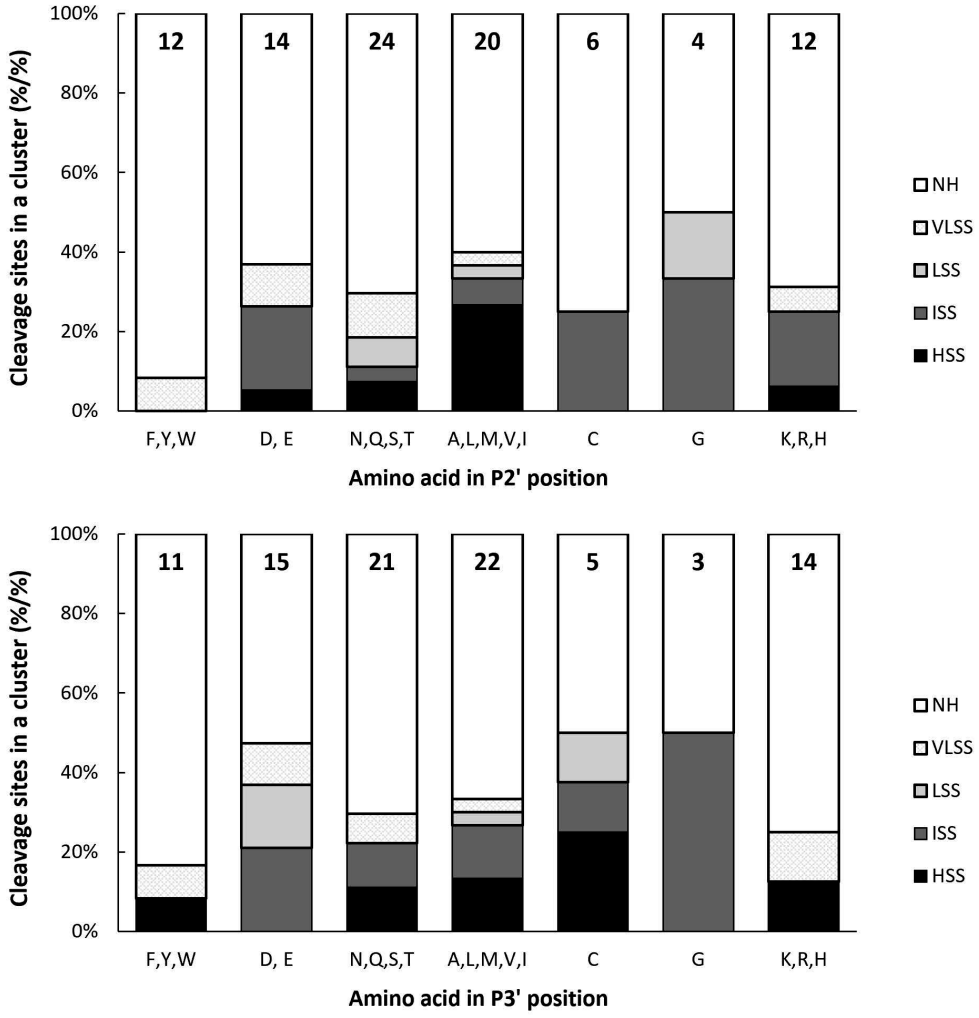


Annex 5.7. Cleavage site probabilities extracted from data of Powers *et al.* Percentage of hydrolysed (■) and intact (□) observations was categorised based on the amino acid in the P1 position (top) and P1' position (bottom). The number in the bar indicates the total number of observations.

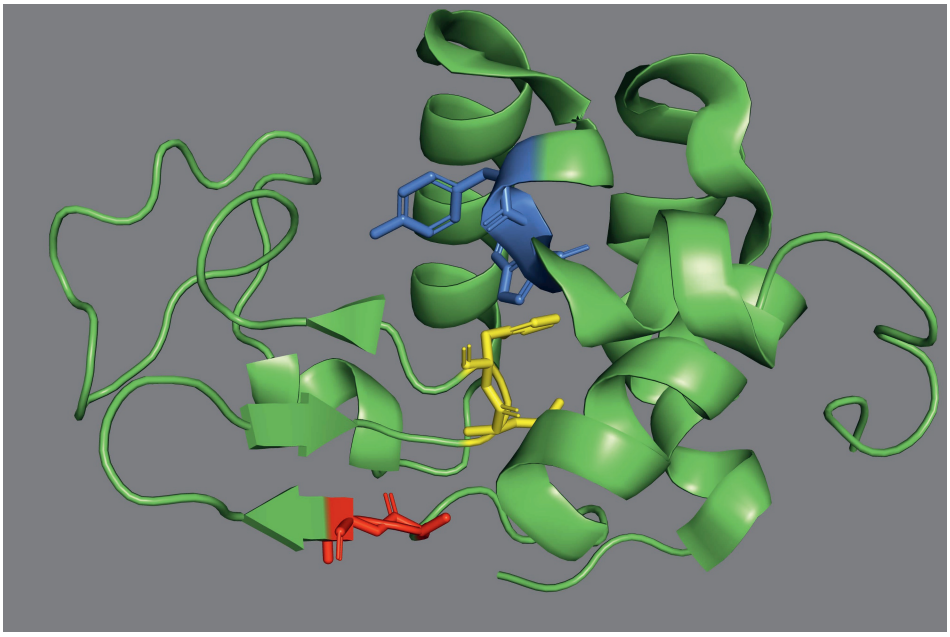

100%
80%
60%
40%
20%
0%
Cleavage sites in a cluster (%/%)
9
12
22
22
8
5
14
F,Y,W
D, E
N,Q,S,T
A,L,M,V,I
C
G
K,R,H
Amino acid in P3 position
NH
VLSS
LSS
ISS
HSS


100%
80%
60%
40%
20%
0%
Cleavage sites in a cluster (%/%)
11
15
18
25
8
4
11
F,Y,W
D, E
N,Q,S,T
A,L,M,V,I
C
G
K,R,H
Amino acid in P2 position
NH
VLSS
LSS
ISS
HSS


100%
80%
60%
40%
20%
0%
Cleavage sites in a cluster (%/%)
4
13
26
23
8
3
16
F,Y,W
D, E
N,Q,S,T
A,L,M,V,I
C
G
K,R,H
Amino acid in P1 position
NH
VLSS
LSS
ISS
HSS
6
16
25
21
7
6
13
Amino acid in P1' position



Annex 5.8. The division of cleavage sites in selectivity clusters HSS (■), ISS (■), LSS (□) VLSS (□) and NH (□) based on the type of amino acid occupying the positions P3-P3'.



Annex 5.9. apo α-LA visualised using PyMOL (crystal structure 1F6R). The coloured peptide bonds are A40-I41 (**red**) in a β-sheet, L52-F53 (**yellow**) in a β-turn, Y103-W104 (**blue**) in a α-helix.

CHAPTER 6

General discussion

MAIN OUTCOME OF THE THESIS

In this thesis, a method was introduced for automated identification and quantification of peptides (**Chapter 2**). Previous studies in our laboratory manually assigned peptide sequences to LC-MS spectra, which was extremely time-consuming: Analysis of one chromatogram took approximately a day of work. To limit the number of annotation options, the choice for substrates and enzymes was limited to relatively small proteins and specific proteases as trypsin and *Bacillus licheniformis* protease [1, 2]. By automation of the annotation process with UNIFI, the time required per sample was reduced to approximately one minute, for comparable protease-substrate combinations. The automation allowed us to apply the method on more complex hydrolysates, created from plant protein extracts or by proteases with a less defined specificity. This development, validated with a manual reference analysis, allowed us to analyse these hydrolysates while maintaining the quality and confidence in the list of peptides as obtained with manual annotation.

The method automated in this thesis (**Figure 6.1**) had many advantages over traditional proteomics processing for the analysis of food hydrolysates. The most important advantages were the large range of peptides (from tri-peptides up to intact α-LA), the 97 % repeatability for peptides in a tryptic mixture of α-LA, β-LG, β-cas and the successful removal of in-source fragments. Moreover, the method benefits from the possibility to -absolute and label-free- quantify all individual peptides and to evaluate the completeness of the analysis. Many of the calculations and routines performed in the experimental chapters, were automated with scripts in Matlab. With these, for instance, in-source fragments were removed, the annotated peptides were correctly matched to UV peak areas and hydrolysis rates were derived from the product formation of individual cleavage sites (**Figure 6.1**). The scripts were developed in such a way that data of multiple injections can be processed at once and results are exported to a single (Microsoft Excel) file with several overview tables of the results.

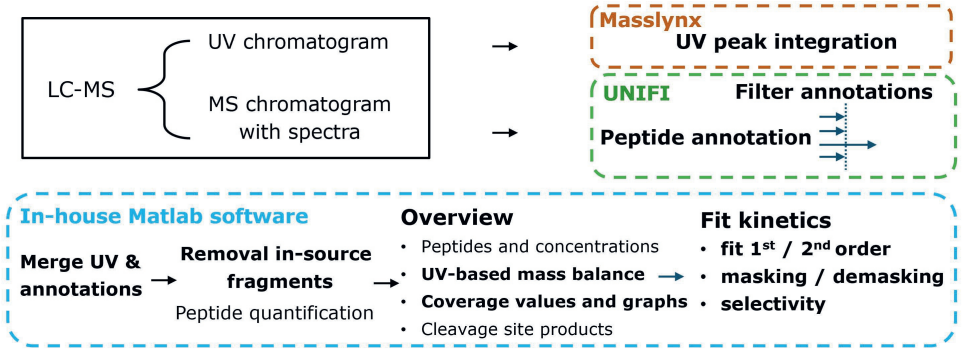


Figure 6.1. Steps in processing of LC-MS data and required software. The steps in bold were developed or improved during this PhD project.

The manual data-processing helped to define processing parameters in such a way that the automated analysis yielded peptide lists of similar quality and repeatability as one would achieve with manual processing. The automated analysis method yielded amino acid sequence coverages of 99-100 % (**Chapter 2**), coverages that are typically not obtained with other automated approaches from proteomics. To study peptide release by proteases during hydrolysis, every individual peptide is important since it explains part of the protease selectivity. In contrast, information on individual peptides is less important in proteomics analyses, since there the identification of a protein depends on multiple unique peptides combined. A high repeatability in replicate analysis was an important requirement for the method. The repeatability depended on detection of the parent ion and on the number of MS/MS fragment ions, which were both correlated mostly to MS ion intensities. In **Chapter 2**, a parameter was introduced, the limit of annotation. Peptides above this limit, explaining ~99% of the total annotated MS intensity, were 100 % repeatably annotated in replicate analyses of simple hydrolysates. Such repeatability on peptide-level is much higher than reported for proteomics approaches, which have typically a repeatability on peptide-level of 35 - 60 % [3]. Another requirement of the method was that it could annotate a wide range of peptide lengths, since these vary from long to short during the hydrolysis process. Typical proteomics approaches target medium-size peptides of 7 to 40 amino acids, which is the default length in MaxQuant. Smaller peptides are not considered since these are not unique for one protein in a database. The method introduced was able to annotate peptides consisting of 3 amino acids but also peptides of 7,000 Da. By automation of the method it appeared that in-source fragmentation of peptides and their formed *m/z* signals should be dealt with. It is not possible to distinguish a low abundance peptide from an in-source fragment, by mass, MS/MS fragments or intensity. In this thesis, a routine was developed to recognize and remove in-source fragments, using retention time and information of the parent peptide. This worked successfully for the applications described in this thesis. The automated analysis made it possible to analyse hydrolysates of higher complexity. In **Chapter 3**, complexity was increased by using a legume seed protein extract containing multiple proteins occurring in multiple isoforms. In **Chapter 4 & 5**, complexity was increased by using broadly or a-specific proteases. The increase in complexity increased the number of signals per chromatogram and had an even more dominant effect on the number of annotation options for the software. This was reflected in the processing time in UNIFI. For example, a chromatogram of hydrolysates obtained after a-specific hydrolysis, or from many and long protein sequences, or with multiple post-translational modifications considered, is processed in UNIFI generally within 10 minutes. When complexity relates to all three aspects simultaneously, the software requires up to 1 hour per sample or freezes when analysing multiple chromatograms in one analysis. In order to analyse these complex samples, the minimum signal intensity was increased in MS processing such that precursor ions without sufficient intensity were excluded prior to the identification stage in UNIFI instead of during filtering in UNIFI. After optimisation, the UNIFI software was able to process these data. For instance, ten chromatograms could be processed of a-specific hydrolysates of collagen A1 (1056

amino acids) and A2 (1038 amino acids) in ~8 hours in one go, with up to two hydroxyprolines considered.

The automated peptide identification method was combined with absolute label-free quantification. In the field of proteomics, accurate and easy quantification is a continuing struggle. Absolute quantification methods require isotopically labelled standards and therefore protein concentrations are based on only a few peptides. Relative quantification techniques can be used to describe differences between peptide quantities in different samples, but are not suitable to determine molar concentrations of individual peptides. Previous studies showed that UV absorbance gives a linear response with molar peptide concentrations and is a successful approach to determine individual peptide concentrations in hydrolysates [2]. The automated method presented in this thesis was also equipped with UV-based quantification, which allowed us to quantify peptides in complex hydrolysates. A few software settings were optimised for automated, fast and reproducible UV-peak integration in Masslynx (**Chapter 2**). For peptides with an intensity above the LOA, the relative standard deviation for individual peptide concentrations in replicate injections was only 4 % (**Chapter 2**) and thereby as accurate as absolute quantification techniques that involve labelling [4]. Peptide quantification based on UV_{214} absorbance would be a good option for other researchers that want to quantify identified peptides in food hydrolysates.

After identification or quantification of all the peptides present in a hydrolysate, one can evaluate the completeness of the peptide analysis by using the suggested coverage parameters e.g. amino acid sequence coverage, peptide sequence coverage, protein recovery and molar sequence coverage. By calculating the amino acid sequence coverage, one can describe how much of the protein sequence was covered by the identified peptides. This is also often done in proteomics. However, since peptides can be quantified with our method, one can now make a mass-balance and determine how much of the initial protein was included in the results, and, determine which specific regions of the protein sequence were quantified less than one would expect based on the injected protein concentration. Both are very useful when characterising hydrolysates with a low solubility. For instance, the protein recovery was used to describe that on average 14 % of the protein material in the yellow pea seeds was included in the results of the analysis (**Chapter 3**). Although introduced by Butré *et al.* in 2014, the coverage parameters are currently not used much outside our laboratory. The outcome of these coverage parameters would be very useful to objectively compare different hydrolysates or for instance the effect of equipment, quantification technique and sample preparation on completeness of the analysis.

In literature, typically only the endpoint of the hydrolysis process is analysed for the peptide composition to determine protease (secondary) specificity and preference. In **Chapter 4** and **Chapter 5**, it was shown that by sampling during hydrolysis, one can follow the formation and degradation of peptides during the process. With the peptide concentrations at various timepoints, the hydrolysis rates and kinetics (normal / demasking) of individual cleavage sites

can be determined. This resulted in a very detailed description of the hydrolysis process and thereby novel insights in protease activity. The automation of the peptide annotation made it possible to do this for proteases with a less defined specificity.

As described above, the method successfully fulfilled the need for an automated and robust method to identify and quantify the peptides in food hydrolysates. The second part of the discussion will evaluate remaining challenges, their solutions and possible method extensions, divided in the topics peptide identification and peptide quantification, for simple and complex systems. The third part of the discussion will evaluate the insights in chymotrypsin and pepsin and thereby the predictability of peptides formed during protein hydrolysis of a-specific proteases.

THE METHOD TO IDENTIFY AND QUANTIFY PEPTIDES IN FOOD HYDROLYSATES

Acquisition of the LC-MS data

Choice of the mass spectrometer and its influence on the data obtained

The final list of identified peptides is strongly influenced by the choices in data processing approach (**Chapter 2**), but also by the LC-MS hardware used to acquire the data [5]. Highly sensitive and accurate mass spectrometers will in some cases be able to detect more peptides, but are also expensive. Halfway this PhD project, the Synapt G2Si, used in **Chapter 2**, was replaced by the Select Series Cyclic IMS. Both were used in this case to evaluate the choice for hardware. The choice of hardware can influence (1) the accuracy in determination of the *m/z* and (2) the recognition of the precursor ion.

1. To determine the peptide mass, it is important that the *m/z* is determined as accurately as possible, since it allows a narrow mass error window in automated processing and limits the number of tentative sequences that match the observed mass. In **Chapter 2**, the use of two lock mass components highly enhanced the mass accuracy of data acquired with the Synapt G2Si, especially for large molecular weight peptides. The mass spectrometer used in the other chapters (Select Series Cyclic IMS) had a comparable mass accuracy as the Synapt G2Si, despite that the software allowed only one lock mass. The Select Series IMS did not show a mass error dependency with increasing mass as was observed when measuring with a single lockmass on the Synapt G2Si (**Chapter 2**). Since both mass spectrometers had comparable mass accuracy, the replacement did not allow us to further narrow down thresholds on mass error and did not affect the number of identified peptides. When mass accuracy would increase much more (< 0.5 ppm), it will not have a large effect on the number of tentative annotation options per *m/z*, since the differences in elemental composition result in peptide mass differences larger than 1 ppm or are isobaric (exactly same mass) [6, 7].
2. To recognize the precursor ion in the processing software, the peak shape and peak resolution (full width at half maximum) are important. These rely on the signal intensity and the type of TOF detector as well as its tuning of the ion beam, respectively. For instance, the peak resolution of LeuEnk was ~2 $\cdot 10^4$ in Synapt G2Si and improved to ~4-5

$\cdot 10^4$ in Select Series Cyclic IMS. A high MS peak resolution allows the processing software to detect the peak shape of lower intensity peaks. The replacement of the Synapt G2Si by the Select Series Cyclic IMS led to a 10x decrease in limit of detection. As a result, low intensity peaks could be identified that were not (reliably) identified before. For a tryptic α-LA hydrolysate, the number of identified α-LA peptides increased from 50 to 60 by change from Synapt G2Si to Select Series Cyclic IMS. The change in mass spectrometer led to a larger increase in number of identified signals for complex hydrolysates. For an a-specific collagen digest, the number of identified MS signals increased from 10,000 to 66,000; the number of MS/MS signals from 15,000 to 123,690 and the number of peptides from 200 to 1,000. Although the differences in number of recognised peptides between Synapt and Cyclic seem striking, one should keep in mind that the differences only manifest in the low abundance signals. The effect of machine peak resolution on the number of identified signals and peptides can also be estimated by an *in silico* simulation as for instance done by Geromanos *et al.* [8].

A high quality mass spectrometer contributes to the number of (low abundant) peptides one can reliably identify and should be considered when comparing results obtained with different LC-MS approaches. In some cases, the lowly abundant peptides contain the most important information. This is the case for identification of proteins which come in many different variants. The lowly abundant peptides are the ones that code uniquely for a particular genetic variant, whereas the abundant peptides originate from sequences generic for all variants. An example of such a protein with high sequence identity between the genetic variants is patatin from *Solanum tuberosum* [9], for which over 40 sequences are reported in the Uniprot database.

Peptide fragmentation techniques

The reliability of annotated peptides depends on the quality of the fragmentation spectra. In this thesis, MS^E was used [10]. With MS^E, the ions that elute from the LC at a certain retention time are exposed to a ramp of collision energies. The benefit of this technique is that no pre-selection of ions is required, and therefore all precursor ions are fragmented. In peak processing, the fragment ions are matched to the precursor ions based on their chromatographic peak. For the data obtained in this thesis, no issues were observed in matching precursor ions with MS/MS fragments. A collision energy ramp was used since the energy required for fragmentation varies with peptide size and the type of amino acids in the peptide sequence [11, 12]. Using this ramp, all peptides were successfully fragmented as long as the MS intensity was sufficient. Some studies indicated that peptides without positively charged residues formed incomplete series of b- and y-ions [13] and peptides with serine phosphorylation showed reduced backbone fragmentation [14]. In our study, no issues were encountered in fragmentation and identification of both types of peptides. For the peptides with high MS intensities, approximately 50 to 100 % of the b and y fragments were formed and annotated, which leaves no doubt about the peptide sequence identified and thereby no room for improvement. For low intensity MS signals, the quality of the fragmentation data decides whether the peptide can be included in

the result, or does not match the minimum fragment criteria for reliable annotation. Little improvement in the fragmentation information can have a relatively large effect on the number of peptides identified. A possible improvement would be to consider also other type of fragment ions as a-ions and b- or y- ions with a water or ammonium loss, which are now already identified in UNIFI, but not used in the annotation criteria for the MS/MS fragments.

For very complex hydrolysates, in which all peaks co-elute, issues might occur in assigning the observed MS/MS fragments to the MS data. For these, it can be useful to remove redundancy in the fragmentation data. At first, fragmentation data could -in theory- be improved by using the Cyclic ion-mobility cell [15, 16]. This dimension allows separation of co-eluting peptides based on slight differences in ion-mobility. Fragmentation can be done in the transfer-cell, after a single pass through the cyclic ion-mobility cell. By doing so, the travel time of the ions can be used in UNIFI processing to see which fragments came from which precursor ion [17]. In practice, saturation of the ion-mobility cell led to limited ion intensities in the TOF-detector. The ion-mobility cell was saturated at MS intensities of $1 \cdot 10^5$ counts, whereas the TOF-analyser was not yet saturated at intensities of $2 \cdot 10^7$ counts. As a result, the amount of ions available for fragmentation was similarly lower in the ion-mobility analysis. For a tryptic digest of α-LA, addition of the IMS dimension resulted in 34 % less identified peptides and 6 % lower fragment recovery for shared annotations. So, to make the ion-mobility cell useful for our application, the ion load capacity should be increased. Cook *et al.* described the same issue when combining the ion-mobility in analysis of volatile compounds [18].

Alternatively, redundancy in the fragmentation spectra can be removed by switching to a different ion-selection strategy, in which only the *m/z* of the peptide of interest is fragmented rather than all ions of co-eluting peptides, as done with a trapping MS. Fragmentation of the most abundant ions in each spectrum would not be useful (Data Dependent Acquisition), since in case of co-elution of peptides, the high intensity ions might all come from one peptide. Alternatively, one can predefine *m/z* windows (of for instance 25 Da) that are selected for fragmentation (Data Independent Acquisition) [19]. With this strategy, no pre-selection of ions is needed, but co-eluting peptides have individual fragmentation spectra. This looks promising, but the total MS cycle time available needs to be divided over all the *m/z* windows [19]. A consensus needs to be made between the number of *m/z* windows and the (time for different) collision energies used which influence the acquisition time per window and thereby fragment intensities.

Choice of LC-MS and mobile phase additive on peptide quantification

UV-based quantification is most accurate when peptides are baseline separated in the LC. In case of co-elution, UV-area has to be divided over peptides using their MS-intensities, which decreases the accuracy (**Chapter 2**). Here again, the choice of LC-setup has a major influence on the quality of the acquired data (**Figure 6.2**).

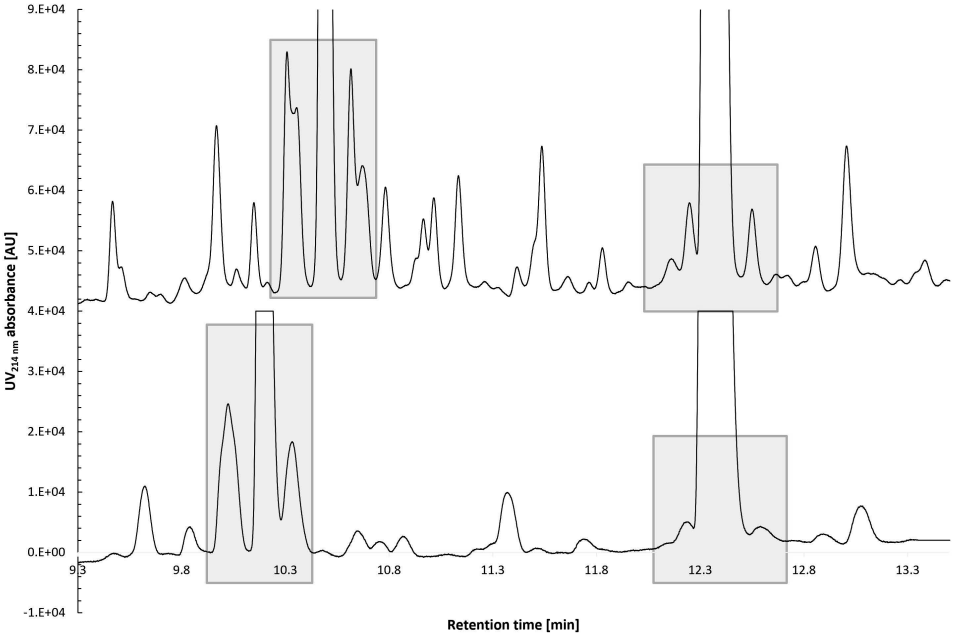


Figure 6.2. Chromatogram of a tryptic digest of α-LA with the H-class Acquity LC (bottom) and the Premier LC (top) from Waters both equipped with C18 peptide column. The UV absorbance on the Premier LC was given an off-set of $4 \cdot 10^4$ AU. The shadings highlight the differences in peak separation.

Some peptides that co-eluted on the H-class Acquity LC were separated using the Premier LC (**Figure 6.2**). Besides the LC, also the mass spectrometer influences the quantification. It was discussed in the previous paragraph that a more sensitive mass spectrometer contributes to the number of peptides that can be detected. Despite having a larger dynamic range for signal intensities in the MS, a sensitive mass spectrometer seems also more sensitive for fouling of the MS, in the long run. To avoid that, the suppliers suggested to avoid MS intensities above $1 \cdot 10^6$ counts, where possible. This implied for analysis in the Select Series Cyclic IMS that hydrolysates had to be diluted an additional 2-3x times, relative to the Synapt G2Si. Obviously, dilution of the sample led to lower peptide concentrations, smaller UV_{214} peak areas and thereby less peptides that exceeded the limit of quantification relative to the undiluted hydrolysates. Therefore, the higher number of identified peptides with the Select Series Cyclic IMS relative to the Synapt G2Si did not lead to a higher number of quantified peptides.

Many other LC-MS methods make use of formic acid (FA) instead of trifluoroacetic acid (TFA) in the solvents. FA is often favoured over TFA since it does not show ion suppression [20] and the formation of TFA clusters in the MS chromatogram [21]. TFA has been described to yield a better peak separation in the LC than FA, but, the differences we observed (with the Premier LC) were negligible. For trypsin specific peptides in our data, signals were suppressed by 60-85% with TFA, relative to FA. Therefore, when FA is used instead of TFA, one should inject

considerably lower peptide concentrations to prevent pollution of the MS-detector. This will negatively affect the number of peptides that can be quantified with UV.

The number of peptides that exceed the limit of quantification can be increased by making the peptide concentration in the LC higher than the amount entering the mass spectrometer. Potentially, one can use a splitter between LC and MS that let only a part of the solution pass from LC to MS. Similarly, a dynamic range enhancement lens (DRE-lens) can be used to restrict the amount of ions going to the MS during data acquisition (not possible in the current Cyclic software). Thirdly, increasing the path length in the UV-detector would help a bit. At last, a sample can be injected at high concentration for quantification, (which goes afterwards to the waste bin) and a sample can be injected at low concentration for identification.

The completeness of quantification

Analysis of the mass balance and visualisation of the unique amino acid concentrations allowed us to describe how much of the expected protein material was recovered by the peptide concentrations. For yellow pea, 14 % of the protein in the seeds was described by the peptides in the hydrolysates (**Chapter 3**). Here, the quantification is incomplete due to peptide losses before injection. These issues were related most likely to the complexity of the sample. e.g. peptide interactions with the substantial non-protein part (comprising ~40% w/w) or intrinsic low solubility. However, also for hydrolysates of milk proteins (protein content >90 % w/w), the unique amino acid concentration in the peptides were in some cases not in line with the expected concentration. Remarkably, the analyses with (apo) α-LA were more complete than the analyses with β-LG and especially β-cas, regardless of the protease (trypsin, chymotrypsin or pepsin) used. It seems therefore that the protease used is not (directly) the cause for the low molar coverage.

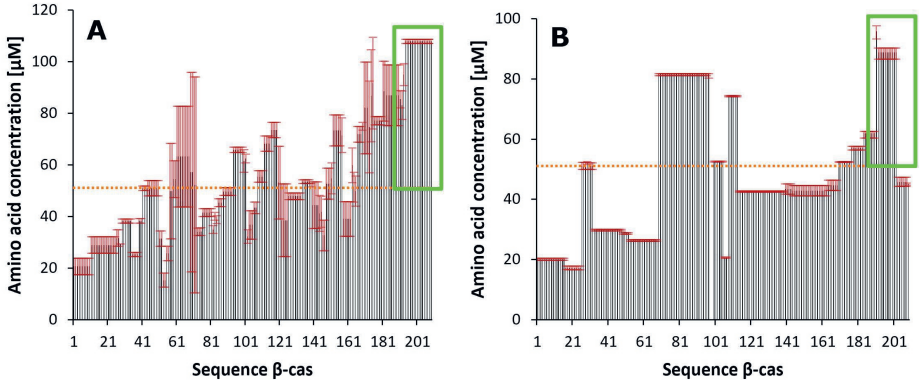


Figure 6.3. Molar sequence coverage plot of β-cas hydrolysed by chymotrypsin for 1 hour (A) and by bovine trypsin for 2 hours (B). The injected protein concentration is 50 µM, indicated by the orange dotted line. Standard deviation for chymotrypsin is from two replicate digestions, standard deviation for trypsin is from four replicate injections. Amino acid region 194-209 is highlighted by the green box.

Possibly, the molar extinction coefficients of peptides are in some cases not predicted correctly. For instance, in β-cas hydrolysates by chymotrypsin, peptide 194-209 (QEPVLGPVRGPFPIIV) was quantified 2x the injected protein concentration (**Figure 6.3**, **Chapter 4**). For this example, there were no clues that the peak area was integrated wrongly, or co-elution could have led to inaccurate concentrations. This peptide had potentially a ~2x higher molar extinction coefficient than predicted maybe due to the 4 proline residues. The same region of the sequence was also over-estimated in hydrolysates of bovine trypsin (**Chapter 2**). In the study of Kuipers *et al.* all (synthetic) peptides that contained proline (GPRP; RPPGFSP and RPPGFSPFR) also showed a 16 % higher experimentally measured extinction than predicted [22]. In the analysis of Deng *et al.* for hydrolysates with bovine, porcine and human trypsin, this peptide region of β-cas was not clearly over-estimated and depended on the degree of hydrolysis [23]. It would be valuable to re-evaluate the role of proline, since the current UV-quantification is fully dependent on the predicted molar extinction coefficients. Preferably, synthetic peptides should be tested with different position and number of proline residues.

Increasing complexity in data-processing and the comparison with proteomics

Repeatability in a-specific data processing for raw and processed data

Up to now, information on the repeatability and reliability of non-specific data-processing approaches are rather limited. Traditionally in proteomics, proteases with a clearly defined specificity as trypsin are used for hydrolysis and the database search. Last years, also semi-specific and non-specific data-processing became available despite that that gives a computational challenge due to the numerous tentative annotation options [24]. For instance, Guo *et al.* performed proteomics analyses with seven proteases in different combinations, among which chymotrypsin and elastase, with the aim to improve amino acid sequence coverages [25]. Theoretically, annotation of peptides that are fully specific with the given enzyme specificity, should be similarly annotated regardless of the type of processing analysis e.g. fully-specific, semi-specific or a-specific. We observed that the type of processing analysis chosen matters for the results of the peptide identification in UNIFI. For instance, in-source fragments are often recognised as such in the fully- or semi-specific analysis, but are annotated as unique peptide in the a-specific analysis. These remaining in-source fragments are removed in later processing based on sequence and retention time. However, when the in-source fragment is annotated to a different part of the sequence or different protein than the parent peptide, the developed routine is not able to recognise these in-source fragments. This issue did not occur during a-specific analysis of single protein hydrolysates but might occur for analysis with multiple substrates. To investigate the effect of the processing type on annotated peptides, in-source fragments and repeatability, a mixture of tryptic digests of α-LA, β-LG and β-cas was injected five times and processed semi-specifically and a-specifically. Both analyses of the same data led to large differences in the raw list of annotated peptides. The data before filtering consisted of ~1,500 annotated peptides after a-specific processing and~150 peptides after semi-specific processing. A-specific processing showed that only ~10 % of the entries in the unfiltered data were annotated in all five replicates. The low repeatability was mostly in the peptides with

relatively low intensities. Of the 100 most abundant peptides, 87 were annotated similarly in all five replicates. For the filtered a-specific output, 53 % of the entries were identified in all 5 replicates and for the peptides with an intensity above the LOA (> $2.1 \cdot 10^5$ Counts), 82 % of the peptides were repeatably identified in all five replicates. This repeatability is considerably lower than obtained with the semi-specific analysis of the mix (97 %), but still higher than repeatability reported on peptide-level with other automated approaches. The difference in repeatability between the data with and without filtering stresses once more the importance of setting criteria to MS/MS fragmentation and intensity to ensure a reliable and repeatable peptide list.

The correctness of substrate sequences used in data-processing

The methodology described in this thesis requires (expected) substrate sequences as input. In other untargeted approaches, available databases are used as for instance "*Bos taurus*" or *"Homo sapiens"* for supply of protein sequences. Both proteomics approaches and our approach rely on the availability and correctness of the protein sequences and will not identify peptides that do not match the protein sequences outside the database. For the analysis of the pea extracts, a few peptides were clearly present but do not seem to come from the protein they were annotated to (**Chapter 3**). For instance, the region 103-107 ((E)-KEEED), assigned to legumin B, had -unexpectedly- the highest molar coverage in the pea vicilin fraction. Most likely these peptides originate from vicilin or provicilin. However, there were no (semi-specific) annotation options in these proteins. Here, we question therefore the correctness of the protein sequence used. For some other legumes, for instance lentil, the sequences of the main storage proteins are currently not reported and reviewed in Uniprot. At last, mutations can occur in the protein sequence [26], which are not known or not included in the database [27]. For instance, recently 206 mutant lines of *Sorghum bicolor* showed differences in amino acid composition attributed to sequence variation of storage proteins [28]. To investigate suspicious annotations as in **Chapter 3**, it would be useful to have an alternative approach that does not depend on sequences in a database. There are already some data-processing approaches that can annotate peptides that do not necessarily match a database. Some recent bioinformatics studies combined genomics and proteomics to identify peptides with unexpected mutations in protein sequences [29]. Other approaches build up the amino acid sequence of the peptide from the MS/MS spectra, also called *de novo* sequencing. In *de novo* annotation, the mass fragment ions and the absolute differences in their masses are used to determine the amino acid sequence of the peptides [30]. Difficulties in this approach are that some amino acids and combinations of amino acids have exactly the same mass. For instance iso-leucine and leucine (113.08 Da), A+N and G+Q (185.08 Da), E+S and T+D (216.07 Da). Moreover, MS/MS data has to be of high quality. Most methods that use *de novo* sequencing combine different fragmentation techniques and use information on which fragment masses are typically formed or not formed (in that type of fragmentation) [31]. Recently also machine learning was incorporated to improve *de novo* peptide identification [32, 33]. Since most highly abundant peptides are identified with >50 % of the theoretical b/y fragments, it could also be interesting to use *de novo* sequencing. The *de*

novo approach would also be useful to annotate *m/z* signals coming with a large UV peak area, but currently not identified.

Analysis of the reference samples with a typical approach of proteomics

In this thesis often the comparison has been made between the developed method and proteomics approaches, and the results that would be obtained for peptide mapping in food hydrolysates. Despite using comparable hardware, often reported coverages and repeatability are lower in proteomics analysis. To make a fair comparison, the tryptic mixture of α-LA, β-LG and β-cas and the individual α-LA hydrolysate, as used for development of the method in **Chapter 2** were analysed with a setup typical for proteomics by dr. S. Gregersen in Aalborg, in respectively three and one injection(s). The hydrolysates were reduced and alkylated before injection on the EASY-nLC system coupled to a Q Exactive HF mass spectrometer, both from Thermo Scientific, similar to [34, 35]. A data-dependent acquisition was used in which (up to) 20 most intense MS1 precursors were selected for HCD fragmentation, which were annotated using MaxQuant v1.6.10.43 [36] with the bovine proteome, a tryptic specific analysis and a 1% false discovery rate. Quantification was done using IBAQ [37].

For the proteomics analysis of α-LA, 114 peptides were identified and in the mixture 132 ±4 peptides. The number of identified peptides for the mixture was similar to the number of peptides identified from α-LA, β-LG and β-cas above the LOD (133) (**Chapter 2**). However, these 132 ±4 peptides in the proteomics analysis originated from ~30 proteins (**Figure 6.4**). This number is remarkably higher than the 3 proteins that are dominant in the mixture. Only 58 ±1 peptides were matched to α-LA, β-LG or β-cas, which means that the proteomics analysis did not identify more peptides than with our methodology in these relatively simple hydrolysates. The proteomics analysis did identify some remarkable proteins as for instance pancreatic ribonuclease 4 (P61823) and *O*-glucosyltransferase 2 (A0A3Q1MR70). Surprisingly, 29 peptides from α, β and κ-casein were identified in the individual α-LA hydrolysate, despite that it should not contain caseins. These cannot have been carried over from the previous injection, since the α-LA hydrolysate was analysed before the mixture. We strongly doubt the correctness of the casein annotations in the α-LA protein isolate done with the proteomics analysis.

For none of the main proteins, a 100% amino acid sequence coverage was obtained with the proteomics analysis. For β-LG, peptide (K)-IIAEK (71-75) was not identified. For α-LA, sequence 14-58 was not covered by peptides. For β-cas, the coverage was only 44 %. The missing sequences in the proteomics analysis seem to be covered by peptides that are either too small or too large to be annotated in the proteomics analysis. This stresses the strength of the method in this thesis to identify tri-peptides up to intact α-LA.

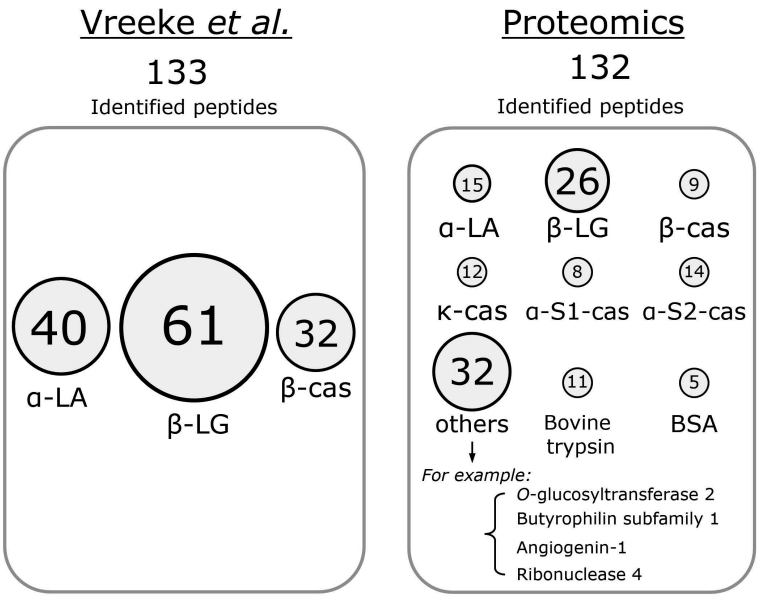


Figure 6.4. The number of peptides identified in tryptic hydrolysates of α-LA, β-LG and β-cas protein isolates (**Chapter 2**) and the number of peptides identified with a proteomics analysis.

The effect of sample complexity on quantification

Typically the amount of injected protein in LC-MS is reported in weight or weight/volume. In reality, this means that with increasing complexity, the molar concentrations of individual protein will simultaneously decrease when the amount injected is kept constant. For the analysis of hydrolysates of milk protein isolates (**Chapter 2**) (with a high protein content, few substrates and relatively short protein sequence), injected concentrations were 50-100 µM per protein, similar to a loading of 2-4 nmole on column). This resulted in high coverage values and accurate estimates of protein concentrations (**Chapter 2**). For hydrolysates of pea protein extracts, the total amount of protein was lower and many, relatively long, protein sequences were used. In consequence, the effective concentration of each protein variant in the cultivar extracts were ≤10 µM for the yellow pea cultivar extracts (similar to a loading of ≤40 pmole per protein on column). This yielded low protein concentrations and amino acid sequence coverages, even for the most abundant proteins in the pea extracts. One could argue how accurate protein quantification is in these samples, since part of the peptides falls below detection- and quantification thresholds. In proteomics, 5,000-11,000 proteins were reported in a single LC-MS run [38, 39]. In case the amount of injected protein is similar, individual protein concentrations will be even lower. For each individual protein in these samples, the molar amount loaded on column can be estimated, assuming that the injected amount is 4 µg (similar to [39]), each protein is equally present and on average 40 kDa. The molar amount loaded per protein would be 0.01 pmole for a sample containing 10,000 hydrolysed proteins. In reality, the proteins will have different molar concentrations and part of the (still detected) proteins will have amounts lower than 0.01 pmole. The individual protein and peptide concentrations will affect the number

of peptides that exceed the detection limit and thereby also the sequence coverages and protein quantification.

It would be interesting to see how the method in this thesis performs at various (individual) protein concentrations, for samples with known protein concentrations. Therefore, an additional study was conducted to quantify known amounts of protein, at different concentrations. To exclude the effect of sample losses during sample preparation, complexity was induced by mixing 9 hydrolysates of different substrates. The hydrolysates were individually analysed at 1 mg/mL and as mixture at three different protein concentrations (0.495 mg/mL, 0.1 mg/mL and 0.05 mg/mL). The protein concentrations were calculated using the average unique amino acid concentration (calculation I from **Chapter 3**). The amino acid sequence coverages ranged from 65 % (α-S2-casein) to 100 % (α-LA) for the individually analysed hydrolysates (**Table 6.1**). Mixing the 9 hydrolysates decreased the average amino acid coverage by 18 %. Injecting the mixture at lower concentration had a major impact on sequence coverage, due to a decrease in number peptides that had intensities sufficient for detection. At 0.495 mg/mL, 186 peptides were identified in the mixture. At 0.1 mg/mg, this value lowered to 75 identified peptides and at 0.05 mg/mL only 33 peptides were identified. No peptides of α-S2-casein were identified in the 0.05 mg/mL mixture, although these were present at higher injected concentrations.

Table 6.1. Amino acid sequence coverages [%] for proteins in mixing experiment.

Protein	Individual 1 mg/mL	Mixture 0.495 mg/mL	Mixture 0.1 mg/mL	Mixture 0.05 mg/mL
α-LA	100	72 ± 20	29 ± 5	13 ± 4
β-LG	94	78 ± 0	69 ± 0	55 ± 1
β-cas	87	36 ± 8	26 ± 2	17 ± 2
BSA	94	79 ± 1	21 ± 0	4 ± 1
α-S1-cas	96	78 ± 2	42 ± 2	8 ± 0
α-S2-cas	65	56 ± 4	7 ± 0	0 ± 0
Lysozyme	98	97 ± 2	57 ± 0	36 ± 3
Hemoglobulin subunit A	95	75 ± 0	51 ± 0	6 ± 0
Hemoglobulin subunit B	86	85 ± 0	37 ± 4	14 ± 5

[1]The hydrolysates of ovalbumin and pea albumin contained mostly intact protein and were not considered in further processing.

[2]The hydrolysate of α-casein contained peptides from both the S1 and S2 isoform. Similarly, the hemoglobulin occurred as mixture of subunit A and B.

The absolute protein concentrations in the 0.5 mg/mL mixture had on average a 27 % relative standard deviation from the expected absolute protein concentrations. This shows that we were able to accurately determine protein concentrations for samples of 10 proteins regardless of a ~10x increase in peptide numbers and thereby the number of co-eluting peptides and amount of UV area that had to be divided with MS-intensity.

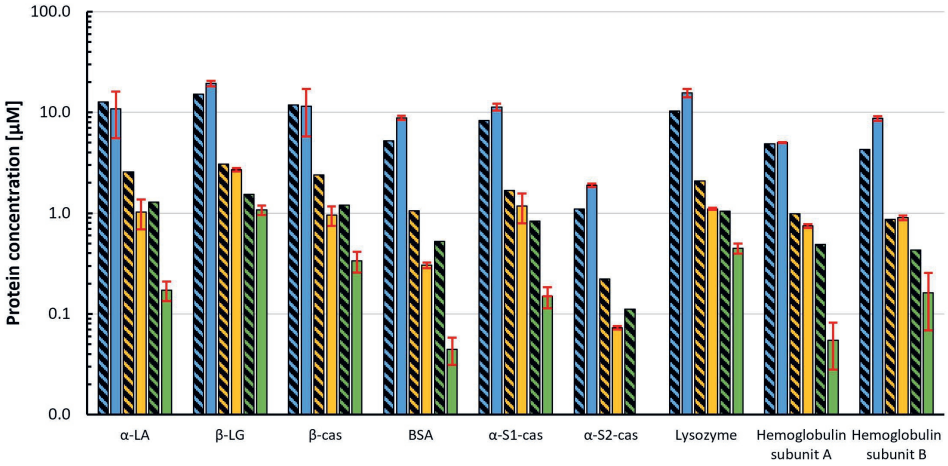


Figure 6.5. Absolute concentrations [µM] with standard deviation of proteins in the mixture injected at 0.5 mg/mL ( ), 0.1 mg/mL ( ) and 0.05 mg/mL ( ) determined from three replicate injections. The expected protein concentration is indicated by the diagonally striped bar. Concentrations are plotted on logarithmic scale.

Quantification was not accurate for mixtures injected at 0.1 mg/mL and 0.05 mg/mL, which represented peptide concentrations as one would analyse hydrolysates from 50 or 100 proteins with similar amount of injected protein as the non-diluted mixture. All absolute protein concentrations were lower than expected. For instance, the absolute concentration of α-LA was 0.2 µM, where 1.2 µM was expected (**Figure 6.5**). It seems that with our current experimental setup, protein (variant) concentrations could be quantified only when (injected) concentrations are approximately above 2 µM. At lower concentrations, part of the peptides are not identified (reflected by the low amino acid sequence coverage), and estimates for concentrations cannot be determined by averaging the unique amino acid concentrations. This is important to consider for applications in the future, as the analysis of digests of food ingredients or full meals. For instance, a traditional Dutch pea soup, containing pea, potato, several other vegetables and meat will certainly contain over 50 proteins, of which most at concentrations below 2 µM.

PROTEOLYSIS BY A-SPECIFIC PROTEASES

Describing proteolysis for digestive proteases without clearly defined specificity

Using the automated method discussed above, hydrolyses of different digestive proteases were studied. In literature, protease specificity and preference are often determined by analysis of the peptides after hydrolysis of protein substrates [40, 41]. Both parameters are determined by counting the type of amino acids on the termini of the peptides. The majority of studies focusing on proteases did not quantify the peptides present and did not analyse multiple timepoints, needed to describe quantitative peptide release kinetics. The automated method for peptide identification and quantification made it possible to do both and investigate proteolysis for the

digestive proteases in much more detail than ever before (**Chapter 4 and 5**). The UV-based quantification was used to calculate concentrations of all peptides, and thereby providing a quantitative description of protease preference (**Chapter 4**).

In this thesis, hydrolyses of simple substrates with different digestive proteases were studied. The method was developed using trypsin, which was highly specific to hydrolyse peptide bonds after lysine and arginine. In previous research in our laboratory, bovine trypsin hydrolysed only 41 % of the potential cleavage sites with high selectivity [1]. For the cleavage sites that were hydrolysed slowly or not at all, 74 % was explained by a secondary specificity [1]. The second protease, chymotrypsin, hydrolysed after all amino acids (except glycine and arginine), which directly indicates the contrast with trypsin in complexity. Chymotrypsin showed a clear preference for aromatic residues, methionine and leucine. For the residues that were preferred, ~73 % of the cleavage sites were hydrolysed with high- or intermediate selectivity and for the other cleavage sites, 49 % were hindered by a proline in the positions P3, P1' or P2' as secondary specificity (**Chapter 4**). The last digestive protease studied was pepsin, for which the preference was again less clearly defined. Also for amino acids within the preference, only ≤ 50 % of the occurrences were hydrolysed and these were not explainable with a secondary specificity (**Chapter 5**). This was in line with cleavage probabilities determined by others [41, 42].

The specificity, preference and secondary specificity described in this way were found to be only part of the parameters needed to describe the kinetics of formation and subsequent hydrolysis of peptides during hydrolysis. For instance, some cleavage sites were hydrolysed with a delay in onset (demasking) or became in-accessible due to the hydrolysis of adjacent bonds (masking). The question rises whether the concepts used to characterise bovine trypsin also apply for the a-specific digestive proteases and how the new insights fit in the existing concepts. The question rises whether peptide release kinetics by the digestive proteases can be predicted and whether new concepts are needed to describe the protein hydrolysis process. The concepts that influence the peptides and their abundance at a certain moment of hydrolysis are: Enzyme (secondary) specificity and preference (as discussed above), the hydrolysis scenario (zipper vs one-by-one), enzyme activity and product inhibition, hydrolysis kinetics (demasking and masking) and hydrolysis rates (selectivity) (**Figure 6.6**). All these seem protease and or substrate dependent and can depend on the hydrolysis conditions. In this part of the discussion, all these concepts for chymotrypsin and pepsin will be discussed.

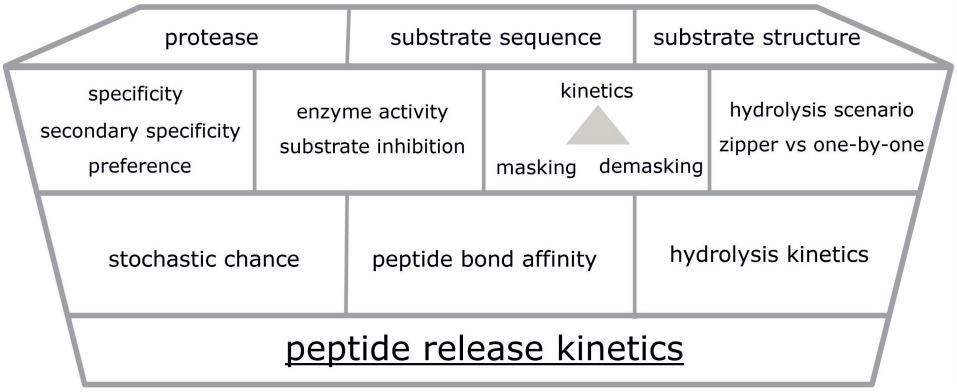


Figure 6.6. Factors that influence peptide release kinetics during hydrolysis.

The hydrolysis scenario of a-specific proteases

The affinity to hydrolyse intact protein or intermediate peptides will influence the type of peptides present during intermediate stages of hydrolysis. According to the Linderstrøm-Lang theory, the protease affinity to intact protein depends on the denaturation state of the substrate [43]. In literature, the hydrolysis scenario is determined in a rather limited number of cases [44-46], despite its importance in digestion. For chymotrypsin, the affinity to hydrolyse intact protein (pH 8.0, 37 °C) was much higher for apo α-LA (molten globule) and β-cas (random coil), than for β-LG (globular), which matches the Linderstrøm-Lang theory (**Chapter 4**).

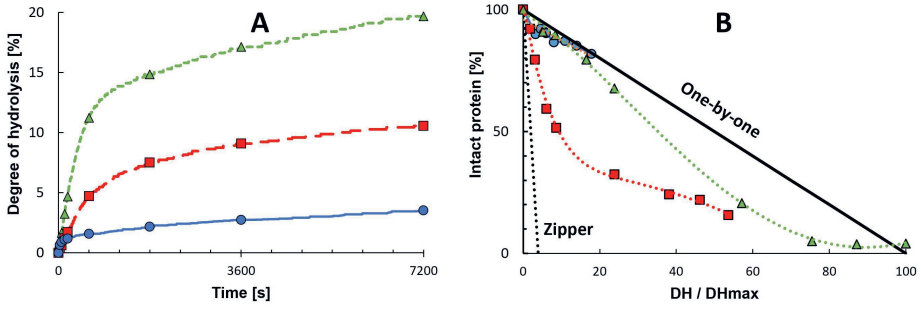


Figure 6.7. Degree of hydrolysis measured on-line with pH-stat (**A**) and protein concentration measured by off-line analysis of LC-MS samples (**B**) plotted against DH/DH_{max} for pepsin hydrolysis of native β-LG(—●), heated β-LG for 15 min at 90 °C (—■) and β-LG with chemically reduced disulphide bonds (- -▲). Hydrolyses were performed at E:S ratio of 1:100 at 37 °C. The theoretical lines for the zipper and one-by-one scenario are indicated.

For pepsin, the observations also matched the Linderstrøm-Lang theory (pH 3.0, 37 °C): apo α-LA (molten globule) was hydrolysed following the zipper scenario and β-LG (globular) followed the one-by-one scenario. Heating the β-LG to its unfolded state [47, 48] (15 min, at 90 °C) caused a shift from one-by-one to a zipper scenario (**Figure 6.7**), similar to the result of Reddy *et al.* [49].

Reduction of the di-sulphide bonds [50] prior to digestion yielded an opened but globular folding state [51]. This led to a 2x higher degree of hydrolysis than with the heated β-LG, exactly similar as described before [49]. The reduced β-LG did not show the clear shift from one-by-one to zipper scenario as observed with the completely unfolded β-LG.

All these observations seem to be in line with the theory. However, other examples from the past were not in line with the theory and suggest that the hydrolysis scenario does not only depend on the substrate. For instance, the affinity to intact protein depends on the conditions of the hydrolysis. Deng *et al.* reported for BLP a shift from the zipper scenario to the one-by-one scenario by increasing the substrate concentration and the reverse for bovine trypsin [52]. Similarly contradicting to the theory, Kosters *et al*. showed that hydrolysis of the same substrate (β-LG) under similar conditions by two proteases with similar specificity (BLP and V8), followed the zipper scenario for BLP and the one-by-one scenario for V8 [53]. This would not be expected based on the Linderstrøm-Lang theory. Therefore, it seems that there is currently no clear understanding of what determines the hydrolysis scenario. The affinity to hydrolyse intact protein (or intermediate peptides) depends on the substrate (state), the protease and conditions, of which the latter two are not considered in the Linderstrøm-Lang theory. Predicting the hydrolysis scenario was found more difficult than the current theory suggests. The automated methodology allowed peptide identification from small peptides to intact protein, enabling to follow peptide release kinetics for both scenario's.

The effect of enzyme and substrate concentration on hydrolysis

For chymotrypsin, the DH_{max} reached after hydrolysis of α-LA, β-LG and β-cas was 2.2 ± 0.3 lower at E:S ratio of 1:100 than at an E:S ratio of 1:25 at similar substrate concentration (**Chapter 4**). Similarly, bovine trypsin poorly hydrolysed 0.1 % apo α-LA at E:S ratio of 1:100 (DH_{max} of 1.5 %), but considerably better at E:S ratio of 1:25 (DH_{max} of 4.0 %) [52]. In the research of Deng *et al*., the hydrolysis of 0.1 % α-LA with bovine trypsin followed the one-by-one scenario, which yielded small peptides at (relatively) low DH [52]. Deng *et al.* hypothesised that these small peptides bound to the protease and thereby inhibited further hydrolysis. At higher substrate concentrations, the hydrolysis followed the zipper scenario and the small peptides were not formed at relatively low DH and no inhibition was observed. Inhibition of proteases by peptides is also a common phenomenon in ACE-inhibition [54, 55] and pathogenic organisms [56, 57]. For pepsin, it is generally accepted that β-LG is resistant to hydrolysis [49, 58, 59], however, we observed that pepsin was able to hydrolyse β-LG and release peptides. Hydrolysis at an E:S ratio of 1:100 yielded a degree of hydrolysis of 3.2 ±0.3 % after 2 hours and 12.3 % after 24 hour incubation. In literature, a few others also reported the hydrolysis of native β-LG and the release of peptides by pepsin [60-63]. The degree of hydrolysis reached for pepsin depended also on the E:S ratio (**Figure 6.8**). The hydrolysis followed strongly a one-by-one scenario, which gave us the hypothesis that inhibiting peptides might be formed rather than that the globular structure caused the resistance of β-LG to pepsin hydrolysis. Loveday *et al.* also suggested the formation of inhibiting peptides for β-LG and pepsin [64].

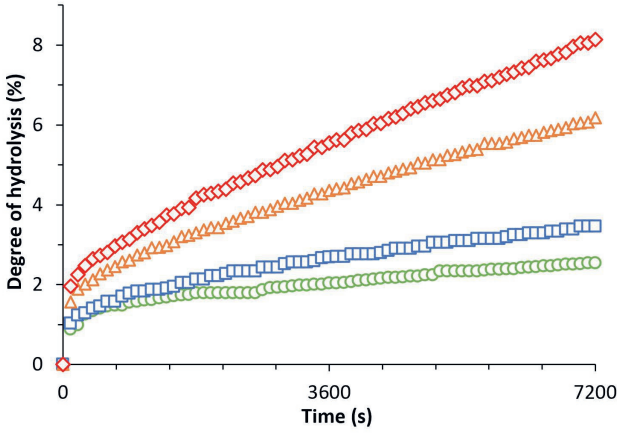


Figure 6.8. Degree of hydrolysis versus time for 1% β-LG hydrolysed by porcine pepsin at enzyme to substrate ratios of 1:200 (○), 1:100 (□), 1:25 (△) and 1:10 (◇).

To test the hypothesis of inhibitory peptides, two experiments were performed. At first, a pepsin β-LG endpoint hydrolysate (1 mL), containing small peptides, was added to an (ongoing) pepsin α-LA hydrolysis in the pH-stat (10 mL). If inhibiting peptides would be formed in the β-LG hydrolysis, these would stop or slow down the α-LA hydrolysis. The β-LG hydrolysate was prior to addition equilibrated to the same pH and temperature, to exclude any pH effect of the addition. In contrast to our hypothesis, the hydrolysis of α-LA was similar with and without the addition of the β-LG hydrolysate. In the second experiment, a 1% α-LA solution was added to a β-LG pepsin hydrolysis after two hours. The hydrolysis of the added α-LA reached a similar degree of hydrolysis as when a second dose of α-LA was added to a pepsin α-LA hydrolysate. Both experiments reject the hypothesis that inhibiting peptides are formed during β-LG hydrolysis. Still, the observation remains that β-LG was hydrolysed by pepsin to some extent. A possible alternative hypothesis would be that β-LG is present in an equilibrium between the folded and unfolded, or non-native, state [47]. Pepsin hydrolyses the unfolded fraction of β-LG, which leads to the increase in degree of hydrolysis and peptide release. The remaining native β-LG slowly unfolds and becomes available for hydrolysis. This would explain why hydrolysis is observed to some extent, and the amount of protease affects simply the kinetics of hydrolysis. The hypothesised inhibitory effect of α-LA peptides on bovine trypsin seems to have a different cause as the low susceptibility of β-LG for pepsin.

Peptide bond demasking for a-specific proteases

Both chymotrypsin and pepsin hydrolysed some peptide bonds with a delay at the onset of hydrolysis (**Chapter 4 and 5**). To describe these data, product formation was described using demasking kinetics. This phenomenon was initially named “demasking” by Vorob’ev *et al*. and attributed to cleavage sites that were initially inaccessible for the protease, due to the folding state of the substrate [65]. It was hypothesised that hydrolysis of the intact protein structure made cleavage sites buried inside the protein accessible. Surprisingly, in **Chapter 5**, demasking

was observed also for substrates with a flexible tertiary structure, as (apo) α-LA. The cleavage sites that showed demasking did generally not have distinctive amino acids in binding site positions relative to cleavage sites hydrolysed directly. Since the demasking could not be explained by change in substrate structure, the initial hypothesis seems unlikely and three alternative causes were hypothesised and tested.

Possibly, the demasking sites might be a result of non-enzymatic hydrolysis. Peptide bonds have a high stability in intact protein, but maybe lower stability when present in a peptide. This "spontaneous cleavage" of peptide bonds was observed previously for synthesised peptides, as well as peptides formed by enzymatic hydrolysis at pH 8 [66]. To test peptide stability under pepsin conditions, a chymotrypsin hydrolysate of α-LA was incubated at pH 3 for 0, 1, 2 and 24 hours. The pH-stat did not show any acid consumption and the peptide composition with LC-MS was similar for all samples. This seems to indicate that there is no spontaneous hydrolysis effect due to the low pH. In a similar way, a hydrolysate of α-LA with pepsin was incubated at pH 8, to test bond stability under incubation conditions of trypsin and chymotrypsin. Also for these peptides, no spontaneous hydrolysis was observed.

A second hypothesis for demasking can be that hydrolysis of these cleavage sites might be a result of activity of minor protease isoforms present. For porcine gastric juice, authors reported that the majority of pepsin occurred as isoform A, with minor amounts of gastricsin, chymosin and isoform B present [67]. Possibly, one of these minor isoforms is present in the enzyme preparation and responsible for hydrolysis of the cleavage sites that show demasking. To test the isoforms in which pepsin was present, a pepsin solution (1 %) was hydrolysed with BLP at an E:S ratio of 1:25. The released peptides were identified in UNIFI using the sequences of pepsin A (P00791), pepsin B (F1S636), cathepsin D (P00795), gastricsin (P30879) and chymosin (F1S626) with a semi-specific analysis. The sequence of pepsin A was covered for 90 % by the peptides, but, not a single peptide was identified coming from the other isoforms. Therefore, the other pepsin isoforms seem to be absent in our enzyme preparation and does not explain the demasking.

A third explanation for the peptide bond demasking can be that peptide stretches interact with each other via hydrophobic interactions in the intact protein [68, 69] and become accessible after partial hydrolysis of the protein. The same line of reasoning was used to explain demasking of peptide bonds during β-casein hydrolysis [70]. In our data, the demasking sites did not have more hydrophobic amino acids than cleavage sites hydrolysed directly, which would argue against this explanation. Taken together, the phenomenon that causes the delay in hydrolysis onset remains therefore unidentified for unfolded proteins.

The incubation time will determine whether "demasking" cleavage sites will be intact or hydrolysed. Although, the demasking cannot always be explained, it has to be considered for peptide release kinetics and its influence on peptide concentrations in time. Hamuro *et al.* and Suwareh *et al.* determined pepsin cleavage frequencies after 30 s^{-1} of hydrolysis [40, 42]. Cleavage sites might be classified as intact, although these can be demasked and hydrolysed in later stages of hydrolysis. Thereby, the activity of pepsin might be underestimated and preference could be different than when determined after extensive hydrolysis. The automated

annotation method allowed us to analyse many time points during the hydrolysis process and describe which cleavage sites showed demasking kinetics. For the future, the kinetic data can be used to study the mechanism behind demasking and see whether it can be predicted from the primary, secondary or tertiary protein structure.

Protease synergy: Specificity, selectivity and its masking effect

In the previous chapters, hydrolysis was performed with a single protease but *in vivo*, multiple proteases act together in protein digestion. Since pepsin is the first protease in the digestive tract, one would expect that it has the role of disassembling the initial protein structure, to make it accessible for the pancreatic proteases, similarly as the acidic conditions in the stomach loosen the protein structures. In reality, surprisingly, pepsin does not seem efficient in doing so for multiple reasons. (i) Pepsin is hindered by insoluble food matrices as gel-like structures [44]. (ii) Pepsin has the highest activity up to pH 3, but, the food bolus has a pH ≥4 at the start of the digestion [71]. (iii) Globular proteins as β-LG are not or slowly hydrolysed [49, 58]. (iv) Pepsin has a preference for generally hydrophobic amino acid residues, which are typically inaccessible because of their position inside the protein. (v) Proteins that are accessible and soluble tend to have a short transit time in the gastric phase, relative to structured foods [72]. Counterintuitively, trypsin and chymotrypsin can hydrolyse folded substrates as β-LG (**Chapter 2 & 4**). The trypsin has even a specificity that would be very suitable for a (first) protease, since the positively charged residues tend to be on the outside of the protein. Therefore, the evolutionary role of pepsin for protein digestion in adults seems minor. Possibly, pepsin is (more) important for protein digestion of newborns [73]. For these, pepsin activity is developed relatively early to the intestinal proteases [74, 75].

In this thesis, trypsin and chymotrypsin were studied individually with intact proteins as substrates. *In vivo*, both proteases act simultaneously on the substrate left after the gastric digestion by pepsin. Depending on all factors described before as hydrolysis scenario, pepsin activity and hydrolysis time, the resulting digest will contain (a combination of) small peptides, large peptides and intact protein. The questions arise how the gastric phase affects the activities of trypsin and chymotrypsin and how the trypsin (selectivity) influences the chymotrypsin (selectivity) and vice versa. Maybe, the selectivity of different proteases is not additive, due to similarities in specificity and preference. Pepsin and chymotrypsin are both a-specific proteases and bonds with amino acids preferred by chymotrypsin (F, Y, W, M, L) will in some cases be hydrolysed already by pepsin. For instance, α-LA bond 31-32 and bond 53-54 were hydrolysed with high selectivity by pepsin as well as chymotrypsin (**Figure 6.9**). Although trypsin and chymotrypsin have a different preference for amino acids, both can still influence the hydrolysis of one another by making cleavage sites unavailable (masking) by hydrolysis of neighbouring cleavage sites. For pepsin, peptide bond 39-40 and 53-54 in α-LA became masked due to the fast hydrolysis of adjacent bonds 40-41 and 52-53 and released only product in the first minutes of hydrolysis. Pepsin, trypsin and chymotrypsin are all endo-proteases, which require the P2-P2' binding site positions to be filled by amino acids for cleavage to occur. Most likely, the digestive proteases will mask (theoretical) cleavage sites of the other endo-proteases during the

digestion. To illustrate the complexity of the *in vivo* situation, cleavage sites of pepsin and chymotrypsin were visualised for part of the α-LA sequence (**Figure 6.9**).

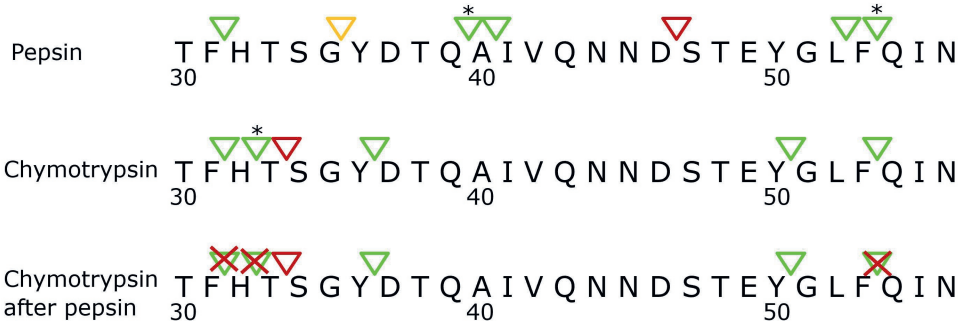


Figure 6.9. Part of the protein sequence of α-LA with experimentally determined selectivity of porcine pepsin and bovine chymotrypsin, and the theoretical selectivity of chymotrypsin after a pepsin hydrolysis. The colour matches the selectivity: High (■), Intermediate (■) or low (■). * indicate pepsin and chymotrypsin cleavage sites that showed masking. The red cross indicates chymotrypsin cleavage sites that are not available because of masking (H32 and F53) or because these were already hydrolysed by pepsin (F31).

Peptide bond masking is important for the peptide release kinetics in the gastric and intestinal phase during *in vivo* digestion. However, during *in vivo* digestion, exo-proteases in the small intestine and brush-border cells will hydrolyse cleavage sites that are masked for endo-proteases. A previous study indicated that exo-proteases in a simulated brush-border phase increased the DH from 32 % to 90 % for a digest of apo α-LA [76]. Picariello reported an increase in DH from 36 % to 76 % by brush-border enzymes for sodium caseinate [77]. Both results indicate that when protein digestion is simulated *in vitro* with the aim to estimate nutrient availability, the brush border membrane needs to be incorporated in the model. The three considerations discussed in this chapter, (i) the unclear role of pepsin, (ii) the combined action of multiple proteases and (iii) the exo-protease activity, indicate the complexity behind the peptide release kinetics though the digestive tract.

Peptide identification and quantification after intestinal digestion

In vitro digestion was performed to compare the actual peptide composition with the peptides that were expected to be released from the selectivity of the endo-proteases individually. A whey protein isolate was digested with porcine pepsin and subsequently with porcine pancreatin. In addition to trypsin and chymotrypsin, pancreatin also contains elastase and carboxypeptidases. The automated method identified 576 ±1 peptides after intestinal digestion, of which 85 % shared between both replicates. The automated method seems therefore to be able to identify and quantify peptides in these digests despite the complexity of having multiple protein sequences (α-LA, β-LG, BSA) and multiple proteases. The relative standard deviation of peptide concentrations between the duplicates was 12 % for peptides in the intestinal digest.

Table 6.2. Peptides identified and quantified after digestion of whey protein isolate with porcine pepsin (1 h, pH 3) and porcine pancreatin (2 h, pH 8) that covered amino acid sequence 30-60 of α-LA. Peptides are divided based on absolute UV-based concentrations in highly abundant (>50 % of the C_{inj}), medium abundant (20-50 % of the C_{inj}) and low abundant (10-20 % of the C_{inj}). Peptides unexpected according to the selectivity of individual proteases are in bold (**Figure 6.9**).

	Start	**End**	**Peptide after gastric phase**	**Start**	**End**	**Peptide after gastric + intestinal digestion**
High abundance	32	40	(F)-HTSGYDTQA	32	40	(F)-HTSGYDTQA
	41	49	(A)-IVQNNDSTE	37	40	(Y)-DTQA
	41	52	(A)-IVQNNDSTEYGL	**41**	**44**	**(A)-IVQN**
				41	49	(A)-IVQNNDSTE
				53	**57**	**(L)-FQINN**
				54	**57**	**(F)-QINN**
				54	58	(F)-QINNK
				59	79	(K)-IW…
Medium abundance	32	39	(F)-HTSGYDTQ	32	36	(F)-HTSGY
	36	40	(G)-YDTQA	**32**	**38**	**(F)-HTSGYDT**
	50	53	(E)-YGLF	32	39	(F)-HTSGYDTQ
	53	80	(L)-FQINNKIW..	36	40	(G)-YDTQA
	54	80	(F)-QINNKIW..	**41**	**51**	**(A)-IVQNNDSTEYG**
				50	52	(E)-YGL
				53	58	(L)-FQINNK
Low abundance	32	36	(F)-HTSGY	40	49	(Q)-AIVQNNDSTE
	36	39	(G)-YDTQ	41	50	(A)-IVQNNDSTEY
	40	49	(Q)-AIVQNNDSTE	**47**	**51**	**(D)-STEYG**
	53	83	(L)-FQINNKIW..			
	53	85	(L)-FQINNKIW..			
	54	83	(F)-QINNKIW..			

From the remaining peptides after intestinal digestion (**Table 6.2**), it seems that pepsin cleavage sites as F31 and L52 were not present as intact bonds in the peptides anymore and therefore efficiently hydrolysed. Similarly, tryptic cleavage site K58, high selectivity according to Deng *et al.* [1], was fully hydrolysed after the intestinal phase. High selectivity cleavage sites for chymotrypsin as H32 and F53 were still intact, possibly due to the masking by pepsin hydrolysis, as hypothesised. Alternatively, the chymotrypsin in the porcine pancreatin could have a different selectivity as the bovine chymotrypsin, similarly as selectivity differed between bovine and porcine trypsin for similar substrate [78]. Some cleavages did neither match the selectivity of pepsin, trypsin nor chymotrypsin, as for instance hydrolysis of cleavage sites N44 and N57 was observed. Possibly, hydrolysis of these bonds is the result of the other endo-proteases in pancreatin or exo-protease activity. It would be very interesting to follow peptide release kinetics during *in vitro* (intestinal) digestion and thereby determine the selectivity of cleavage sites. This will broaden our view on how experiments with individual proteases relate to the *in vivo* situation. The concepts used to describe protein hydrolysis form a good basis to mechanistically study peptide kinetics during digestion.

Towards predicting protein hydrolysis

As described before, peptide release kinetics of a-specific digestive proteases depend on many factors which are hard to predict. Here, two potential next steps towards predicting selectivity are suggested.

First, we should confirm that selectivity is determined by the affinity proteases have for cleavage sites and the stochastic chance of hydrolysis (based on cleavage site distribution). This has been suggested by Butré *et al.* but has not been proven yet. The assumption that selectivity is caused by these two factors underlies an *in silico* model for peptide release kinetics [79]. To confirm that selectivity is determined by protease affinity for cleavage sites and the stochastic chance of hydrolysis, a few peptide sequences can be synthesized that contain the amino acids of high selectivity cleavage sites, with the amino acids relevant for surrounding binding site positions. The difference in hydrolysis rate of the synthesised peptide and the same cleavage site in the intact protein, should be caused by the stochastic chance of hydrolysis.

Secondly, it would be interesting to understand why some cleavage sites are hydrolysed much faster than others, especially for proteins hydrolysed according to the zipper scenario. For pepsin hydrolysis of α-LA and chymotrypsin hydrolysis of α-LA and β-cas, a few cleavage sites were hydrolysed at much higher rates than all other bonds. For these substrates, the influence of protein secondary and tertiary structure seems limited. Therefore, we consider that the interaction between the amino acids in the binding site positions and the subsite underlie the affinity to be hydrolysed, despite that this cannot easily be derived from the amino acid sequence directly. The total energy required for protease-substrate binding, the catalytic reaction and release of the peptide should in theory match the affinity of proteases to a certain cleavage site. Therefore, it would be interesting to use a peptide-protease docking approach as in [80-84] to see the favourable conformation and amino acid interactions during hydrolysis of these cleavage sites. Possibly, (3-dimensional) interactions come to light that can explain differences in affinity. For instance, a certain combination of amino acids might fit well in the subsite, due to spatial distribution of charges or distribution of hydrophobic interactions. The interactions will strongly depend on the energy-favourable binding orientation and flexibility of the protease-substrate complex. Docking approaches generally study pepsin-inhibitor interactions [85-87] but examples to predict hydrolysis are (still) scarce. An example is the prediction of early cleavage events during pepsin hydrolysis of insulin by Koliński *et al.* [88]. The most favourable positions that result from the 3-dimensional docking can be used to calculate various (3-dimensional) descriptors. These can eventually be used in a quantitative structure-activity relationship model as in [89, 90] to predict hydrolysis rates for cleavage sites. The hydrolysis rates measured in **Chapter 4 and 5** form an excellent dataset to build such models and work towards predicting hydrolysis by a-specific proteases.

Concluding remarks

The automated method to identify and quantify peptides allowed us to analyse food hydrolysates, with similar confidence as obtained with manual annotation. The elaborate evaluation of completeness helped to monitor the data quality for complex hydrolysates. Using the developed method, peptide release kinetics by a-specific proteases can be studied, which were before mostly characterised by (secondary) specificity or preference. Although some of the observations were not fully understood, the data made it possible to study concepts as demasking and masking. The results of this thesis showed that these concepts should be considered for peptide release kinetics of a-specific proteases. The concepts selectivity, masking and demasking should be measured and considered when studying proteases in simple systems but most likely also for *in vivo* digestion. We invite other researchers to also quantify peptides with UV_{214nm} and analyse multiple time points, to be able to analyse selectivity and continue to unravel the mysteries behind protein hydrolysis.

References

1. Deng, Y., van der Veer, F., Sforza, S., Gruppen, H., Wierenga, P. A. (**2018**). Towards predicting protein hydrolysis by bovine trypsin. *Process Biochemistry*, 65, 81-92.
2. Butré, C. I., Sforza, S., Gruppen, H., Wierenga, P. A. (**2014**). Introducing enzyme selectivity: A quantitative parameter to describe enzymatic protein hydrolysis. *Analytical and Bioanalytical Chemistry*, 406, 5827-5841.
3. Tabb, D. L., Vega-Montoto, L., Rudnick, P. A., Variyath, A. M., Ham, A.-J. L., Bunk, D. M., Kilpatrick, L. E., Billheimer, D. D., Blackman, R. K., Cardasis, H. L. (**2010**). Repeatability and reproducibility in proteomic identifications by liquid chromatography- tandem mass spectrometry. *Journal of Proteome Research*, 9, 761-776.
4. Schulze, W. X., Usadel, B. (**2010**). Quantitation in mass-spectrometry-based proteomics. *Annual review of plant biology*, 61, 491-516.
5. Elias, J. E., Haas, W., Faherty, B. K., Gygi, S. P. (**2005**). Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nature Methods*, 2, 667-675.
6. Clauser, K. R., Baker, P., Burlingame, A. L. (**1999**). Role of accurate mass measurement (±10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Analytical Chemistry*, 71, 2871-2882.
7. Zubarev, R. A., Håkansson, P., Sundqvist, B. (**1996**). Accuracy requirements for peptide characterization by monoisotopic molecular mass measurements. *Analytical Chemistry*, 68, 4060-4063.
8. Geromanos, S. J., Hughes, C., Golick, D., Ciavarini, S., Gorenstein, M. V., Richardson, K., Hoyes, J. B., Vissers, J. P., Langridge, J. I. (**2011**). Simulating and validating proteomics data and search results. *Proteomics*, 11, 1189-1211.
9. Pots, A. M., Gruppen, H., Hessing, M., van Boekel, M. A. J. S., Voragen, A. G. J. (**1999**). Isolation and characterization of patatin isoforms. *Journal of Agricultural and Food Chemistry*, 47, 4587-4592.
10. Plumb, R. S., Johnson, K. A., Rainville, P., Smith, B. W., Wilson, I. D., Castro-Perez, J. M., Nicholson, J. K. (**2006**). UPLC/MSE; a new approach for generating molecular fragment information for biomarker structure elucidation. *Rapid Communications in Mass Spectrometry*, 20, 1989-1994.
11. Dongré, A. R., Jones, J. L., Somogyi, Á., Wysocki, V. H. (**1996**). Influence of peptide composition, gas-phase basicity, and chemical modification on fragmentation efficiency: Evidence for the mobile proton model. *Journal of the American Chemical Society*, 118, 8365-8374.
12. Paizs, B., Suhai, S. (**2005**). Fragmentation pathways of protonated peptides. *Mass Spectrometry Reviews*, 24, 508-548.
13. Bensadek, D., Monigatti, F., Steen, J. A. J., Steen, H. (**2007**). Why b, y's? Sodiation-induced tryptic peptide-like fragmentation of non-tryptic peptides. *International Journal of Mass Spectrometry*, 268, 181-189.
14. Hoffert, J. D., Knepper, M. A. (**2008**). Taking aim at shotgun phosphoproteomics. *Analytical Biochemistry*, 375, 1-10.
15. Tomczyk, N., Giles, K., Richardson, K., Ujma, J., Palmer, M., Nielsen, P. K., Haselmann, K. F. (**2021**). Mapping isomeric peptides derived from biopharmaceuticals using high-resolution ion mobility mass spectrometry. *Analytical Chemistry*, 93, 16379-16384.
16. de Bruin, C. R., Hennebelle, M., Vincken, J. P., de Bruijn, W. J. C. (**2023**). Separation of flavonoid isomers by cyclic ion mobility mass spectrometry. *Analytica Chimica Acta*, 1244, 340774.
17. Baker, E. S., Livesay, E. A., Orton, D. J., Moore, R. J., Danielson, W. F., III, Prior, D. C., Ibrahim, Y. M., LaMarche, B. L., Mayampurath, A. M., Schepmoes, A. A., Hopkins, D. F., Tang, K., Smith, R. D., Belov, M. E. (**2010**). An LC-IMS-MS platform providing increased dynamic range for high-throughput proteomic studies. *Journal of Proteome Research*, 9, 997-1006.
18. Cook, G. W., LaPuma, P. T., Hook, G. L., Eckenrode, B. A. (**2010**). Using gas chromatography with ion mobility spectrometry to resolve explosive compounds in the presence of interferents. *Journal of Forensic Sciences*, 55, 1582-1591.

19. Ludwig, C., Gillet, L., Rosenberger, G., Amon, S., Collins, B. C., Aebersold, R. (**2018**). Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. *Mol Syst Biol*, 14, e8126.
20. García, M. C., Hogenboom, A. C., Zappey, H., Irth, H. (**2002**). Effect of the mobile phase composition on the separation and detection of intact proteins by reversed-phase liquid chromatography–electrospray mass spectrometry. *Journal of Chromatography A*, 957, 187-199.
21. Stout, S. J., daCunha, A. R. (**1989**). Tuning and calibration in thermospray liquid chromatography/mass spectrometry using trifluoroacetic acid cluster ions. *Analytical Chemistry*, 61, 2126-2128.
22. Kuipers, B. J. H., Gruppen, H. (**2007**). Prediction of molar extinction coefficients of proteins and peptides using UV absorption of the constituent amino acids at 214 nm to enable quantitative reverse phase high-performance liquid chromatography-mass spectrometry analysis. *Journal of Agricultural and Food Chemistry*, 55, 5445-5451.
23. Deng, Y., Gruppen, H., Wierenga, P. A. (**2018**). Comparison of protein hydrolysis catalysed by bovine, porcine or human trypsins. *Journal of Agricultural and Food Chemistry*, 66, 4219-4232.
24. Rolfs, Z., Millikin, R. J., Smith, L. M. (**2020**). An algorithm to improve the speed of semi and non-specific enzyme searches in proteomics. *Current Bioinformatics*, 15, 1065-1074.
25. Guo, X., Trudgian, D. C., Lemoff, A., Yadavalli, S., Mirzaei, H. (**2014**). Confetti: A Multiprotease Map of the HeLa Proteome for Comprehensive Proteomics *Molecular & Cellular Proteomics*, 13, 1573-1584.
26. Alonso-Blanco, C., Aarts, M. G., Bentsink, L., Keurentjes, J. J., Reymond, M., Vreugdenhil, D., Koornneef, M. (**2009**). What has natural variation taught us about plant development, physiology, and adaptation? *The Plant Cell*, 21, 1877-1896.
27. Alfaro, J. A., Ignatchenko, A., Ignatchenko, V., Sinha, A., Boutros, P. C., Kislinger, T. (**2017**). Detecting protein variants by mass spectrometry: a comprehensive study in cancer cell-lines. *Genome Medicine*, 9, 62.
28. Khan, A., Khan, N. A., Bean, S. R., Chen, J., Xin, Z., Jiao, Y. (**2023**). Variations in total protein and amino acids in the sequenced sorghum mutant library. *Plants*, 12.
29. Choi, S., Paek, E. (**2020**). MutCombinator: identification of mutated peptides allowing combinatorial mutations using nucleotide-based graph search. *Bioinformatics*, 36, i203-i209.
30. Olson, M. T., Epstein, J. A., Yergey, A. L. (**2006**). *De Novo* peptide sequencing using exhaustive enumeration of peptide composition. *Journal of the American Society for Mass Spectrometry*, 17, 1041-1049.
31. Medzihradszky, K. F., Chalkley, R. J. (**2015**). Lessons in *de novo* peptide sequencing by tandem mass spectrometry. *Mass Spectrometry Reviews*, 34, 43-63.
32. McDonnell, K., Howley, E., Abram, F. (**2023**). Critical evaluation of the use of artificial data for machine learning based de novo peptide identification. *Computational and Structural Biotechnology Journal*, 21, 2732-2743.
33. Tran, N. H., Qiao, R., Xin, L., Chen, X., Liu, C., Zhang, X., Shan, B., Ghodsi, A., Li, M. (**2019**). Deep learning enables *de novo* peptide sequencing from data-independent-acquisition mass spectrometry. *Nature Methods*, 16, 63-66.
34. Gregersen, S., Kongsted, A.-S. H., Nielsen, R. B., Hansen, S. S., Lau, F. A., Rasmussen, J. B., Holdt, S. L., Jacobsen, C. (**2021**). Enzymatic extraction improves intracellular protein recovery from the industrial carrageenan seaweed *Eucheuma denticulatum* revealed by quantitative, subcellular protein profiling: A high potential source of functional food ingredients. *Food Chemistry: X*, 12, 100137.
35. Jafarpour, A., Gomes, R. M., Gregersen, S., Sloth, J. J., Jacobsen, C., Sørensen, A.-D. M. (**2020**). Characterization of cod (*Gadus morhua*) frame composition and its valorization by enzymatic hydrolysis. *Journal of Food Composition and Analysis*, 89, 103469.
36. Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., Mann, M. (**2011**). Andromeda: a peptide search engine integrated into the MaxQuant environment. *Journal of Proteome Research*, 10, 1794-1805.
37. Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., Selbach, M. (**2011**). Global quantification of mammalian gene expression control. *Nature*, 473, 337-342.

38. Graumann, J., Hubner, N. C., Kim, J. B., Ko, K., Moser, M., Kumar, C., Cox, J., Schöler, H., Mann, M. (**2008**). Stable isotope labeling by amino acids in cell culture (SILAC) and proteome quantitation of mouse embryonic stem cells to a depth of 5,111 proteins. *Molecular & Cellular Proteomics*, 7, 672-683.
39. Muntel, J., Gandhi, T., Verbeke, L., Bernhardt, O. M., Treiber, T., Bruderer, R., Reiter, L. (**2019**). Surpassing 10 000 identified and quantified proteins in a single run by optimizing current LC-MS instrumentation and data analysis strategy. *Molecular Omics*, 15, 348-360.
40. Suwareh, O., Causeur, D., Jardin, J., Briard-Bion, V., Le Feunteun, S., Pezennec, S., Nau, F. (**2021**). Statistical modeling of *in vitro* pepsin specificity. *Food Chemistry*, 362, 130098.
41. Powers, J. C., Harley, A. D., Myers, D. V. (**1977**). Subsite specificity of porcine pepsin. In: Tang J, editor. Acid Proteases: Structure, Function, and Biology. New York, NY: Springer US, p. 141-157.
42. Hamuro, Y., Coales, S. J., Molnar, K. S., Tuske, S. J., Morrow, J. A. (**2008**). Specificity of immobilized porcine pepsin in H/D exchange compatible conditions. *Rapid Communications in Mass Spectrometry*, 22, 1041-1046.
43. Adler-Nissen, J. (**1976**). Enzymic hydrolysis of proteins for increased solubility. *Journal of Agricultural and Food Chemistry*, 24, 1090-1093.
44. Luo, Q., Boom, R. M., Janssen, A. E. M. (**2015**). Digestion of protein and protein gels in simulated gastric environment. *LWT - Food Science and Technology*, 63, 161-168.
45. Abrahamse, E., Thomassen, G. G. M., Renes, I. B., Wierenga, P. A., Hettinga, K. A. (**2022**). Assessment of milk protein digestion kinetics: effects of denaturation by heat and protein type used. *Food & Function*, 13, 5715-5729.
46. Beaubier, S., Framboisier, X., Fournier, F., Galet, O., Kapel, R. (**2021**). A new approach for modelling and optimizing batch enzymatic proteolysis. *Chemical Engineering Journal*, 405, 126871.
47. Delahaije, R. J. B. M., Gruppen, H., van Eijk–van Boxtel, E. L., Cornacchia, L., Wierenga, P. A. (**2016**). Controlling the Ratio between Native-Like, Non-Native-Like, and Aggregated β-Lactoglobulin after Heat Treatment. *Journal of Agricultural and Food Chemistry*, 64, 4362-4370.
48. Apenten, R. K. O., Khokhar, S., Galani, D. (**2002**). Stability parameters for β-lactoglobulin thermal dissociation and unfolding in phosphate buffer at pH 7.0. *Food Hydrocolloids*, 16, 95-103.
49. Reddy, I. M., Kella, N. K. D., Kinsella, J. E. (**1988**). Structural and conformational basis of the resistance of β-lactoglobulin to peptic and chymotryptic digestion. *Journal of Agricultural and Food Chemistry*, 36, 737-741.
50. Hoppenreijs, L. J. G., Brune, S. E., Biedendieck, R., Krull, R., Boom, R. M., Keppler, J. K. (**2023**). Fibrillation of β-lactoglobulin at pH 2.0: Impact of cysteine substitution and disulfide bond reduction. *Food Hydrocolloids*, 141, 108727.
51. Brune, S. E., Hoppenreijs, L. J. G., Kühl, T., Lautenbach, V., Walter, J., Peukert, W., Schwarz, K., Imhof, D., Boom, R. M., Krull, R., Keppler, J. K., Biedendieck, R. (**2023**). Precision fermentation as a route to modify β-lactoglobulin structure through substitution of specific cysteine residues. *International Dairy Journal*, 105772.
52. Deng, Y., Butré, C. I., Wierenga, P. A. (**2018**). Influence of substrate concentration on the extent of protein enzymatic hydrolysis. *International Dairy Journal*, 86, 39-48.
53. Kosters, H. A. (**2012**). Controlling the aggregation and gelation of β-lactoglobulin by the addition of its peptides. PhD thesis dissertation. Wageningen University.
54. Xu, Z., Wu, C., Sun-Waterhouse, D., Zhao, T., Waterhouse, G. I. N., Zhao, M., Su, G. (**2021**). Identification of post-digestion angiotensin-I converting enzyme (ACE) inhibitory peptides from soybean protein Isolate: Their production conditions and in silico molecular docking with ACE. *Food Chemistry*, 345, 128855.
55. Li, X., Feng, C., Hong, H., Zhang, Y., Luo, Z., Wang, Q., Luo, Y., Tan, Y. (**2022**). Novel ACE inhibitory peptides derived from whey protein hydrolysates: Identification and molecular docking analysis. *Food Bioscience*, 48, 101737.

56. Gambacorta, N., Caputo, L., Quintieri, L., Monaci, L., Ciriaco, F., Nicolotti, O. (**2022**). Rational discovery of antiviral whey protein-derived small peptides targeting the SARS-CoV-2 main protease. *Biomedicines*, 10, 1067.
57. Agbowuro, A. A., Huston, W. M., Gamble, A. B., Tyndall, J. D. A. (**2018**). Proteases and protease inhibitors in infectious diseases. *Medicinal Research Reviews*, 38, 1295-1331.
58. Guo, M. R., Fox, P. F., Flynn, A., Kindstedt, P. S. (**1995**). Susceptibility of β-lactoglobulin and sodium caseinate to proteolysis by pepsin and trypsin. *Journal of Dairy Science*, 78, 2336-2344.
59. Deng, Y., Govers, C., Tomassen, M., Hettinga, K., Wichers, H. J. (**2020**). Heat treatment of β-lactoglobulin affects its digestion and translocation in the upper digestive tract. *Food Chemistry*, 330, 127184.
60. Chicón, R., López-Fandiño, R., Alonso, E., Belloque, J. (**2008**). Proteolytic pattern, antigenicity, and serum immunoglobulin E binding of β-lactoglobulin hydrolysates obtained by pepsin and high-pressure treatments. *Journal of Dairy Science*, 91, 928-938.
61. Egger, L., Ménard, O., Baumann, C., Duerr, D., Schlegel, P., Stoll, P., Vergères, G., Dupont, D., Portmann, R. (**2019**). Digestion of milk proteins: Comparing static and dynamic *in vitro* digestion systems with *in vivo* data. *Food Research International*, 118, 32-39.
62. Miralles, B., Del Barrio, R., Cueva, C., Recio, I., Amigo, L. (**2018**). Dynamic gastric digestion of a commercial whey protein concentrate. *Journal of the Science of Food and Agriculture*, 98, 1873-1879.
63. Lajterer, C., Shani Levi, C., Lesmes, U. (**2022**). An *in vitro* digestion model accounting for sex differences in gastro-intestinal functions and its application to study differential protein digestibility. *Food Hydrocolloids*, 132, 107850.
64. Loveday, S. M., Peram, M. R., Singh, H., Ye, A., Jameson, G. B. (**2014**). Digestive diversity and kinetic intrigue among heated and unheated β-lactoglobulin species. *Food & Function*, 5, 2783-2791.
65. Vorob'ev, M., Levicheva, I. Y., Belikov, V. (**1996**). Kinetics of the initial stage of milk protein hydrolysis by chymotrypsin. *Applied Biochemistry and Microbiology*, 32, 219-222.
66. Butré, C. I., Buhler, S., Sforza, S., Gruppen, H., Wierenga, P. A. (**2015**). Spontaneous, non-enzymatic breakdown of peptides during enzymatic protein hydrolysis. *Biochimica et Biophysica Acta - Proteins and Proteomics*, 1854, 987-994.
67. Nielsen, P. K., Foltmann, B. (**1995**). Purification and characterization of porcine pepsinogen B and pepsin B. *Archives of Biochemistry and Biophysics*, 322, 417-422.
68. Groleau, P. E., Morin, P., Gauthier, S. F., Pouliot, Y. (**2003**). Effect of physicochemical conditions on peptide−peptide interactions in a tryptic hydrolysate of β-Lactoglobulin and identification of aggregating peptides. *Journal of Agricultural and Food Chemistry*, 51, 4370-4375.
69. Yoo, S., Xantheas, S. S. (**2011**). The role of hydrophobic surfaces in altering water-mediated peptide-peptide interactions in an aqueous environment. *Journal of Physical Chemistry A*, 115, 6088-6092.
70. Vorob'ev, M. M., Dalgalarrondo, M., Chobert, J. M., Haertlé, T. (**2000**). Kinetics of β-casein hydrolysis by wild-type and engineered trypsin. *Biopolymers*, 54, 355-364.
71. Malagelada, J.-R., Longstreth, G. F., Summerskill, W. H. J., Go, V. L. W. (**1976**). Measurement of gastric functions during digestion of ordinary solid meals in man. *Gastroenterology*, 70, 203-210.
72. Deng, R., Mars, M., Janssen, A. E. M., Smeets, P. A. M. (**2023**). Gastric digestion of whey protein gels: A randomized cross-over trial with the use of MRI. *Food Hydrocolloids*, 141.
73. Dupont, D., Mandalari, G., Molle, D., Jardin, J., Léonil, J., Faulks, R. M., Wickham, M. S. J., Mills, E. N. C., Mackie, A. R. (**2010**). Comparative resistance of food proteins to adult and infant *in vitro* digestion models. *Molecular Nutrition and Food Research*, 54, 767-780.
74. McClean, P., Weaver, L. T. (**1993**). Ontogeny of human pancreatic exocrine function. *Arch Dis Child*, 68, 62-65.
75. DiPalma, J., Kirk, C. L., Hamosh, M., Colon, A. R., Benjamin, S. B., Hamosh, P. (**1991**). Lipase and pepsin activity in the gastric mucosa of infants, children, and adults. *Gastroenterology*, 101, 116-121.
76. Deng, Y. (**2018**). Towards predicting enzymatic protein hydrolysis. PhD thesis dissertation. Wageningen University.

77. Picariello, G., Miralles, B., Mamone, G., Sánchez‐Rivera, L., Recio, I., Addeo, F., Ferranti, P. (**2015**). Role of intestinal brush border peptidases in the simulated digestion of milk proteins. *Molecular Nutrition & Food Research*, 59, 948-956.
78. Deng, Y., Gruppen, H., Wierenga, P. A. (**2018**). Comparison of protein hydrolysis catalyzed by bovine, porcine, and human trypsins. *Journal of Agricultural and Food Chemistry*, 66, 4219-4232.
79. Butré, C. I. (**2014**). Introducing enzyme selectivity as a quantitative parameter to describe the effects of substrate concentration on protein hydrolysis. PhD thesis dissertation Wageningen University & Research.
80. Plewczynski, D., Łażniewski, M., Grotthuss, M. V., Rychlewski, L., Ginalski, K. (**2011**). VoteDock: Consensus docking method for prediction of protein–ligand interactions. *Journal of Computational Chemistry*, 32, 568-581.
81. Badaczewska-Dawid, A. E., Kmiecik, S., Koliński, M. (**2021**). Docking of peptides to GPCRs using a combination of CABS-dock with FlexPepDock refinement. *Briefings in Bioinformatics*, 22, bbaa109.
82. Kurcinski, M., Badaczewska‐Dawid, A., Kolinski, M., Kolinski, A., Kmiecik, S. (**2020**). Flexible docking of peptides to proteins using CABS-dock. *Protein Science*, 29, 211-222.
83. Ciemny, M., Kurcinski, M., Kamel, K., Kolinski, A., Alam, N., Schueler-Furman, O., Kmiecik, S. (**2018**). Protein–peptide docking: opportunities and challenges. *Drug Discovery Today*, 23, 1530-1537.
84. Chaudhury, S., Gray, J. J. (**2009**). Identification of Structural Mechanisms of HIV-1 Protease Specificity Using Computational Peptide Docking: Implications for Drug Resistance. *Structure*, 17, 1636-1648.
85. Yue, Y., Zhao, S., Liu, J., Yan, X., Sun, Y. (**2017**). Probing the binding properties of dicyandiamide with pepsin by spectroscopy and docking methods. *Chemosphere*, 185, 1056-1062.
86. Bursavich, M. G., Rich, D. H. (**2002**). Designing non-peptide peptidomimetics in the 21st Century: Inhibitors targeting conformational ensembles. *Journal of Medicinal Chemistry*, 45, 541-558.
87. Headey, S. J., MacAskill, U. K., Wright, M. A., Claridge, J. K., Edwards, P. J. B., Farley, P. C., Christeller, J. T., Laing, W. A., Pascal, S. M. (**2010**). Solution Structure of the Squash Aspartic Acid Proteinase Inhibitor (SQAPI) and Mutational Analysis of Pepsin Inhibition. *Journal of Biological Chemistry*, 285, 27019-27025.
88. Koliński, M., Kmiecik, S., Dec, R., Piejko, M., Mak, P., Dzwolak, W. (**2020**). Docking interactions determine early cleavage events in insulin proteolysis by pepsin: Experiment and simulation. *International Journal of Biological Macromolecules*, 149, 1151-1160.
89. Andini, S., Araya-Cloutier, C., Lay, B., Vreeke, G., Hageman, J., Vincken, J.-P. (**2021**). QSAR-based physicochemical properties of isothiocyanate antimicrobials against gram-negative and gram-positive bacteria. *LWT*, 144, 111222.
90. Kalli, S., Araya-Cloutier, C., Hageman, J., Vincken, J. P. (**2021**). Insights into the molecular properties underlying antibacterial activity of prenylated (iso)flavonoids against MRSA. *Scientific Reports*, 11.

Summary

Enzymatic hydrolysis is essential to produce food products as infant formula and to digest dietary proteins in humans and animals upon passage through the gastro-intestinal tract. The breakdown of proteins into peptides is facilitated by proteases. To better understand their actions, the peptide sequences and their concentrations during the hydrolysis process need to be characterized, which can be done with LC-MS. Existing data-processing approaches, as described in introductory **Chapter 1**, are typically developed to identify proteins (proteomics) and have several limitations when used for peptide mapping. For instance, peptide lists often differ between replicate analyses of the same sample and only a limited number of peptides can be accurately quantified. In this PhD project, an automated method was developed to identify and quantify peptides in complex protein hydrolysates, with focus on reliability and completeness in the annotated peptides. The method was subsequently used for three applications, with a step-by-step increase in complexity of substrate and protease specificity.

In **Chapter 2**, a guideline is given to determine criteria that yield highly reproducible peptide annotations. The guideline was used to automate our in-house UPLC-PDA-MS method for untargeted identification of peptides. Parameters important for peptide mapping were optimised, as mass accuracy, MS/MS fragment criteria and removal of in-source fragments. These parameters were set with simple tryptic hydrolysates of α-lactalbumin, β-lactoglobulin and β-casein, analysed manually and with UNIFI software. Peptides in the individual hydrolysates with an MS intensity above the limit of annotation represented 99 % of total MS intensity and were 100 % consistently annotated between four replicates. Afterwards, the three hydrolysates were mixed to evaluate the robustness in analysis of mixed hydrolysates. Peptides above the LOA were still reproducibly annotated (97 %). All individual peptides were quantified absolutely and label-free based on UV-absorbance. Concentrations of co-eluting peptides deviated 37 ±21 % from their expected concentration. The robust automated approach successfully replaced the manual method with minimal loss of quality.

The methodology was used in **Chapter 3** to investigate the applicability to characterise protein extracts of yellow pea (*Pisum sativum*). In these analyses, the protease had still a clearly defined specificity, but, more and longer substrate sequences were considered relative to **Chapter 2**. The main proteins in the pea extracts, legumin, vicilin and albumin, occur in genetic variants, at unknown quantities. Here again, purified fractions of the main proteins were first hydrolysed and subsequently analysed. The main challenge was how to convert measured peptide concentrations to protein concentrations. Analysis of the protein mass balance showed significant losses of proteins in extraction (37 %) and of peptides in further sample preparation (69 %). Different calculations were evaluated to determine individual protein concentrations while dealing with these insoluble peptides. The quantification approach using average amino acid concentrations in each position of the sequence showed most reproducible results and allowed comparison of the genetic protein composition of 8 different cultivars. For these, the extractable composition was remarkably similar.

In **Chapter 4**, the method was used to study the digestive protease bovine chymotrypsin, with a specificity less clearly defined than trypsin and *Bacillus licheniformis* protease. Observations about the type of bonds that are hydrolysed (specificity and preference) were in the past derived from the peptide composition after digestion or hydrolysis rates of synthetic peptides. In our study, we used the automated method to describe the path of hydrolysis, *i.e* formation and degradation of peptides for three milk proteins. The hydrolysis rates of individual cleavage sites allowed us to evaluate statements from literature on (secondary) specificity and preference. Chymotrypsin showed a preference towards aromatic amino acids, methionine and leucine, but was also tolerant to other amino acids. For the cleavage sites within this preference, ~73 % of the cleavage sites were hydrolysed with high or intermediate selectivity. For the missed cleavages within the preference, 45 % was explained by hindrance of proline, which affected hydrolysis when in positions P3, P1' or P2'. A few cleavage sites were hydrolysed extremely efficient in α-lactalbumin and β-casein. The results showed the potential to use the method to study proteases with a less defined specificity.

In **Chapter 5**, the method is used to study peptide release kinetics by porcine pepsin, a fully a-specific protease. During *in vivo* digestion, the pH in the stomach increases with food intake and decreases gradually during digestion. Our aim was to investigate whether pH affects individual hydrolysis rates of peptide bonds. α-Lactalbumin was hydrolysed by porcine pepsin systematically at pH 1 to 5 in duplicate and peptides were identified and quantified at eight time points. Apparent pH-based differences in specificity were caused by differences in total hydrolysis rate, but the relative hydrolysis rates of cleavage sites were generally independent of pH. The previously reported preference of pepsin for amino acids in the P3-P3' positions withstands when considering the hydrolysis rates of cleavage sites and was independent of pH. Despite the a-specificity of pepsin, many bonds were not or slowly hydrolysed, some cleavage sites became more accessible during hydrolysis (demasking) and some became less accessible (masking).

The main advantages of the peptide analysis method and considerations are discussed in **Chapter 6**, supplemented with additional experiments. The quality of the hardware is discussed, which affected the number of less abundant peptides that can be reliably (and repeatably) identified. In addition, several options are discussed to further improve the methodology. At last, 9 hydrolysates were analysed individually and as mixture, to show the robustness of protein quantification in such complex hydrolysates and to show how protein concentrations affect the accuracy of quantification. In the second part of **Chapter 6**, the action of the digestive proteases is discussed. The concepts used to describe hydrolysis for specific proteases are evaluated for a-specific proteases, as well as whether these concepts can be predicted.

Acknowledgements

Als eerste wil ik mijn promotor en co-promotor bedanken voor hun hulp en begeleiding. **Jean-Paul**, het was altijd fijn hoe jij mij hielp om ons enthousiasme en onze positieve houding ook in de artikelen naar voren te laten komen. Verder hield jij altijd het grotere plaatje in je achterhoofd. Onze besprekingen waren waardevol, en daarom ook nooit binnen de voorgeschreven tijd klaar. Ik waardeer het erg dat je ook zoveel tijd voor mij vrij hebt kunnen maken tijdens de afwezigheid van Peter, ondanks al je andere bezigheden zoals het besturen van onze leerstoelgroep. Ik waardeerde het ook erg als je vertelde over je reizen, eigen werkervaringen of over suikersplitsers. **Peter**, je hebt een enorme passie voor je werk, en ik ben blij dat ik daar onderdeel van geworden ben. Bij artikelen en presentaties wist je altijd goed werk naar een nog hoger niveau te tillen, door middel van de duidelijke logica in je hoofd. Met je Matlab vaardigheden ben je ontzettend belangrijk geweest voor de resultaten van mijn project, wanneer de UNIFI software of Waters mensen tekort schoten. Door je enorme betrokkenheid, kritische maar rechtvaardige blik en enthousiasme voor de wetenschap heb ik ontzettend veel van je kunnen leren.

Mijn PhD tijd was niet hetzelfde geweest zonder Jolanda. **Jolanda**, bedankt voor alle praatjes en welkome afleiding. Het was altijd gezellig met je! Vaak leverden die gezellige bezoekjes ook wat op: Geplande besprekingen, een zakje Haribo, verhuisdagen of tips over kleur van de kaft. Je bent onmisbaar voor onze leerstoelgroep.

I would like to thank all my colleagues at FCH. I really enjoyed the fun during the activities, dinners, breaks and PhD-trips. Everyone is open, friendly and willing to help, which makes we have a great working atmosphere.

I would also like to thank my lab-mates of Lab X0222. **René**, due to your supervision, we won each year the non-existing award for the tidiest laboratory. **Maud**, **Madelon**, **Carolina**, **Judith, Cas, Abel, Thore, Adrian,** I really enjoyed to work together in the lab. We had a lot of serious and non-serious talks during my pH-stat experiments. It was always nice to listen to all the good and bad music on the radio.

Loes, ik wil jou ook graag bedanken. Los van dat we goed bevriend zijn, heb ik ook veel aan je gehad tijdens mijn PhD. Je stond altijd klaar om mee te brainstormen over mijn experimenten, artikelen of andere PhD zaken. Ook heb je me laten zien hoe het er aan toe gaat op de verdiepingen boven de begane grond. Als je iets of iemand relevants voor mij tegenkwam, stuurde je dat naar me door. Je was tevens een goede BLG leverancier, in alle kleuren en smaken. Al onze gezellige brainstorm sessies, hebben uiteindelijk ook nog tot wetenschappelijke resultaten geleid.

Maud, ik wil jou bedanken voor onze samenwerking. Je was altijd mijn eerste aanspreekpunt voor alle werk gerelateerde vragen. Je hebt me geleerd hoe je goed Advanced Food Chemistry doorkomt, hoe je perfecte SDS-PAGE plaatjes maakt en alles omtrent plantaardige eiwitten. Ik vind het super tof dat onze samenwerking heeft geleid tot een prachtig artikel.

Furthermore, I would like to thank my office mates **Yuxi**, **Maud**, **Mohèb**, **Dimitris**, **Dazhi**, **Lorenz**, **Romy** and **Cas**. I appreciated our talks in between the serious work, and you were always willing to help. Also, you kept my vitamin concentration at its optimum by warning when new fruits arrived.

Ook de analisten wil ik graag bedanken. **Wouter L**, bedankt voor je hulp met de Synapt en optimalisatie van de lockmass. **Edwin**, als de kolom weer lekte stond je altijd klaar om deze even subtiel aan te draaien. Ook had je altijd nuttige input over massa spectrometrie. **Mark**, bedankt voor de jaarlijkse organisatie van AFC. **Giovanni**, bedankt voor al het pH-stat onderhoud en het uitvoeren van projecten van bedrijven. Last but not least, **René** bedankt voor alle hulp en gesprekken op het lab.

I would like to thank my BSc and MSc students **Aylin**, **Nina**, **Renee**, **Femke**, **Kevin**, **Bryan**, **Axel**, **Eddy**, and **Yoshinta**. I really enjoyed to work with you and teach you as much as possible. Some of you deserve some special words. **Aylin**, you had a great challenge to work with pepsin, with a not optimized methodology. You collected a large amount of data and gave me a lot of insight. **Femke**, sometimes we forgot that you were (still) a BSc student. How you managed the QSAR-modelling with trypsin was excellent. **Axel** and **Bryan**, I really enjoyed working with both of you. It was really nice to see how you worked together. At last, **Kevin**, it was interesting to help you in your thesis with Maud. Your initial experiments showed us the potential for making an article together.

For the most important break of the day, I would like to thank **Annemiek**, **Coen**, **Laurens**, **Loes**, **Sten**, **Patrick** and the others for our weekly CP adventures. When I forgot by accident -or intentionally- to bring my lunch, you were always happy to join me.

Goede werk efficiëntie krijg je niet voor elkaar zonder ontspanning. Daarom wil ik mijn jaarclubgenoten **Ballie**, **Chris**, **Daan**, **Martijn**, **Mauro**, **Mikey**, **Jorn**, **Quu**, **Rik**, **Stan**, en **Tino** bedanken voor alle borrels, etentjes, mountainbike avonturen, brouwdagen, golf sessies, weekendjes weg en vakanties. Jullie staan altijd voor me klaar, maken me blij en zorgen altijd voor vermaak, gekkigheid en goede gesprekken. Ongetwijfeld blijven we daar de komende jaren mee doorgaan.

Ik wil **papa**, **mama** en **Annelien** bedanken voor hun steun. Jullie ontvingen me altijd met open armen wanneer ik toe was aan wat Zeeuwse zonnestralen en bordspelletjes wilde spelen. Ondanks dat Annelien altijd wint, is het elke keer weer fijn en gezellig.

A small shout-out to Pacman, Inky, Blinky, Pinky and Clyde. Presentations and posters about protein degradation would not be the same without you.

Als laatste, maar zeker de belangrijkste wil ik **Anne** bedanken. Tijdens mijn PhD heb je me altijd gesteund, ondanks dat ik niet fatsoenlijk uit kon leggen wat ik nou precies aan het doen was. Je kalmeerde me als ik me gestrest voelde. Je liet me inzien dat ik er mag zijn, voor mezelf op mag komen en uitdagingen kan aangaan. Verder hielp je me om dingen in perspectief te plaatsen. Daarnaast leer je me fatsoenlijk Nederlands appen en begin ik eindelijk te begrijpen hoe je bepaalde woorden schrijft. Met woorden als sourdust en chickenskin hebben we ondertussen ons vocabulaire vergroot. Ik kan meestal met je lachen en ik vind het fijn om samen een team te zijn <3.

About the author

About the author

Gijs Jan Cornelis Vreeke was born on the 6th of May 1995 in Vlissingen, the Netherlands. After graduation from high school (CSW van de Perre in Middelburg), he studied the Bachelor Food Technology at Wageningen University & Research (2013-2016). As part of this Bachelor, he did the minor Supply Chain Management and performed a thesis at the laboratory of Food Chemistry on hydrolysate fractionation. Afterwards he continued with the Master Food Technology at Wageningen University & Research (2016-2019), with specialization Ingredient functionality. As part of the Master, he did a thesis in the laboratory of Food Chemistry on the quantitative structure-activity relationship of isothiocyanates against *E. coli*. Moreover, he did an internship on dairy powders at Jacobs Douwe Egberts. From February 2019 until September 2023, Gijs was employed as PhD researcher in the laboratory of Food Chemistry under supervision of dr. Peter Wierenga and Prof. dr. Jean-Paul Vincken. The results obtained in this period are presented in this thesis. During that contract, he also contributed to additional education and was involved in two projects on peptide analysis for the food industry. Following his PhD period, Gijs is employed as postdoctoral researcher at the laboratory of Food Chemistry to continue his work on the chemistry of proteins and peptides.

Contact: gijsvreeke@hotmail.com

LinkedIn: www.linkedin.com/in/gijs-vreeke-7b11a412a/

Publications

Andini, S., Araya-Cloutier, C., Lay, B., **Vreeke, G**., Hageman, J., & Vincken, J.-P. (**2021**). QSAR-based physicochemical properties of isothiocyanate antimicrobials against gram-negative and gram-positive bacteria. *LWT*, *144*, 111222. https://doi.org/10.1016/j.lwt.2021.111222

Vreeke, G. J. C., Lubbers, W., Vincken, J.-P., & Wierenga, P. A. (**2022**). A method to identify and quantify the complete peptide composition in protein hydrolysates. *Analytica Chimica Acta*, *1201*, 339616. https://doi.org/10.1016/j.aca.2022.339616

Vreeke, G. J. C., Meijers, M. G. J., Vincken, J.-P., & Wierenga, P. A. (**2023**). Towards absolute quantification of protein genetic variants in *Pisum sativum* extracts. *Analytical Biochemistry*, *665*, 115048. https://doi.org/10.1016/j.ab.2023.115048

Vreeke, G. J. C., Vincken, J.-P., & Wierenga, P. A. (**2023**). The path of proteolysis by bovine chymotrypsin. *Food Research International*, *165*, 112485. https://doi.org/10.1016/j.foodres.2023.112485

Vreeke, G. J. C., Vincken, J.-P., & Wierenga, P. A. (**2023**). Quantitative peptide release kinetics to describe the effect of pH on pepsin preference. *Process Biochemistry, 134,* 351-362. https://doi.org/10.1016/j.procbio.2023.10.021

Hoppenreijs, L. J. G., Annibal, A., **Vreeke, G. J. C.,** Boom, R., M., & Keppler, J. K. (**2024**). Food proteins from yeast-based precision fermentation: simple purification of recombinant β-lactoglobulin using polyphosphate. *Food Research International, Accepted for publication.*

Overview of completed training activities

Discipline specific activities	Organizer, location and year
Courses	
Computational Design & Discovery course[a]	Radboud University, Nijmegen, 2019
Big Data Analysis in the Life Sciences	VLAG, Wageningen, 2019
Food Proteins: Significance, Reactions and Modifications	University of Copenhagen, Online, 2020
Reaction Kinetics in Food Science	VLAG, Wageningen, 2021
Application training Cyclic IMS	Waters, Wageningen, 2021
Conferences	
33rd EFFoST International Conference	WUR, Rotterdam, 2019
Virtual International Conference on Food Digestion[a]	COST-Infogest, Online, 2021
Virtual Conference on Recent Advances in Food Analysis Conference	UCT Prague, Online, 2021
7th International Conference on Food Digestion[b]	Teagasc, Cork, 2022
General courses	
VLAG PhD Week	VLAG, Baarlo, 2019
Scientific Writing	WGS, Online, 2020
Familiarisation training Cyclic IMS	Waters, Wageningen, 2021
Matlab Onramp and Fundamentals	Mathworks, Online, 2022
Career Perspectives	WGS, Wageningen, 2022
Optional courses and activities	
Preparation of PhD research proposal	FCH, Wageningen, 2019
Food Chemistry PhD trip[a,b,c]	FCH, the Netherlands, 2021
Food Chemistry PhD trip[a]	FCH, Spain and Portugal, 2023
BSc and MSc thesis student supervision, presentations and colloquia	FCH, Wageningen, 2019-2023
PhD lunch presentations[c]	FCH, Wageningen, 2019-2023
Protein-Lipid meetings	FCH, Wageningen, 2019-2023

[a] Oral presentation, [b] Poster presentation, [c] Organising committee

The work described in this thesis was performed at the Laboratory of Food Chemistry, Wageningen University & Research, the Netherlands.

Financial support from Wageningen University for printing this thesis is gratefully acknowledged.

Cover design by Gijs J.C. Vreeke
Printed by ProefschriftMaken, the Netherlands
Edition: 150 copies

EQLTKCEVFRELKDLKGYGGVSLPEWVCTTFHTSGYDTQAIVQNNDSTEYGLFQINNKIWCKDDQNPHSSNICNISCDKFLDDDLTDDI
MCVKKILDKVGINYWLAHKALCSEKLDQWLCEKL  LIVTQTMKGLDIQKVAGTWYSLAMAASDISLLDAQSAPLRVYVEELKPTPEG
DLEILLQKWENGECAQKKIIAEKTKIPAVFKIDALNENKVLVLDTDYKKYLLFCMENSAEPEQSLACQCLVRTPEVDDEALEKFDKALKAL
PMHIRLSFNPTQLEEQCHI RELEELNVPGEIVESLSSSEESITRINKKIEKFQSEEQQQTEDELQDKIHPFAQTQSLVYPFPGPIPNSLP
QNIPPLTQTPVVVPPFLQPEVMGVSKVKEAMAPKHKEMPFPKYPVEPFTESQSLTLTDVENLHLPLPLLQSWMHQPHQPLPPTVMFP
PQSVLSLSQSKVLPVPQKAVPYPQRDMPIQAFLLYQEPVLGPVRGPFPIIV LREQPQQNECQLERLDALEPDNRIESEGGLIETWNP
NNKQFRCAGVALSRATLQRNALRRPYYSNAPQEIFIQQGNGYFGMVFPGCPETFEEPQESEQGEGRRYRDRHQKVNRFREGDIIAVP
TGIVFWMYNDQDTPVIAVSLTDIRSSNNQLDQMPRRFYLAGNHEQEFLQYQHQQGGKQEQENEGNNIFSGFKRDYLEDAFNVNRHI
VDRLQGRNEDEEKGAIVKVKGGLSIISPPEKQARHQRGSRQEEDEDEEKQPRHQRGSRQEEEEDEDEERQPRHQRRRGEEEEEDKKE
RGGSQKGKSRRQGDNGLEETVCTAKLRLNIGPSSSPDIYNPEAGRIKTVTSLDLPVLRWLKLSAEHGSLHKNAMFVPHYNLNANSIIYAL
KGRARLQVVNCNGNTVFDGELEAGRALTVPQNYAVAAKSLSDRFSYVAFKTNDRAGIARLAGTSSVINNLPLDVVAATFNLQRNEAR
QLKSNNPFKFLVPARESENRASA RSDPQNPFIFKSNKFQTLFENENGHIRLLQKFDQRSKIFENLQNYRLLEYKSKPHTIFLPQHTDA
DYILVVLSGKAILTVLKPDDRNSFNLERGDTIKLPAGTIAYLVNRDDNEELRVLDLAIPVNRPGQLQSFLLSGNQNQQNYLSGFSKNILEA
SFNTDYEEIEKVLLEEHEKETQHRRSLKDKRQQSQEENVIVKLSRGQIEELSKNAKSTSKKSVSSESEPFNLRSRGPIYSNEFGKFFEITPEK
NPQLQDLDIFVNSVEIKEGSLLLPHYNSRAIVIVTVNEGKGDFELVGQRNENQQEQRKEDDEEEEQGEEEINKQVQNYKAKLSSGDVFV
IPAGHPVAVKASSNLDLLGFGINAENNQRNFLAGDEDNVISQIQRPVKELAFPGSAQEVDRILENQKQSHFADAQPQQRERGSRETRD
RLSSV ASCNGVCSPFEMPPCGTSACRCIPVGLVVGYCRNPSGVFLRTNDEHPNLCESDADCRKKGSGNFCGHYPNPDIEYGWCFA
SKSEAEDFFSKITPKDLLKSVSTA EQLTKCEVFRELKDLKGYGGVSLPEWVCTTFHTSGYDTQAIVQNNDSTEYGLFQINNKIWCKD
DQNPHSSNICNISCDKFLDDDLTDDIMCVKKILDKVGINYWLAHKALCSEKLDQWLCEKL LIVTQTMKGLDIQKVAGTWYSLAMA
ASDISLLDAQSAPLRVYVEELKPTPEGDLEILLQKWENGECAQKKIIAEKTKIPAVFKIDALNENKVLVLDTDYKKYLLFCMENSAEPEQSL
ACQCLVRTPEVDDEALEKFDKALKALPMHIRLSFNPTQLEEQCHI RELEELNVPGEIVESLSSSEESITRINKKIEKFQSEEQQQTEDEL
QDKIHPFAQTQSLVYPFPGPIPNSLPQNIPPLTQTPVVVPPFLQPEVMGVSKVKEAMAPKHKEMPFPKYPVEPFTESQSLTLTDVENLH
LPLPLLQSWMHQPHQPLPPTVMFPPQSVLSLSQSKVLPVPQKAVPYPQRDMPIQAFLLYQEPVLGPVRGPFPIIV LREQPQQNEC
QLERLDALEPDNRIESEGGLIETWNPNNKQFRCAGVALSRATLQRNALRRPYYSNAPQEIFIQQGNGYFGMVFPGCPETFEEPQESEQ
GEGRRYRDRHQKVNRFREGDIIAVPTGIVFWMYNDQDTPVIAVSLTDIRSSNNQLDQMPRRFYLAGNHEQEFLQYQHQQGGKQEQ
ENEGNNIFSGFKRDYLEDAFNVNRHIVDRLQGRNEDEEKGAIVKVKGGLSIISPPEKQARHQRGSRQEEDEDEEKQPRHQRGSRQEEE
EDEDEERQPRHQRRRGEEEEEDKKERGGSQKGKSRRQGDNGLEETVCTAKLRLNIGPSSSPDIYNPEAGRIKTVTSLDLPVLRWLKLSA
EHGSLHKNAMFVPHYNLNANSIIYALKGRARLQVVNCNGNTVFDGELEAGRALTVPQNYAVAAKSLSDRFSYVAFKTNDRAGIARLAG
TSSVINNLPLDVVAATFNLQRNEARQLKSNNPFKFLVPARESENRASA RSDPQNPFIFKSNKFQTLFENENGHIRLLQKFDQRSKIFE
NLQNYRLLEYKSKPHTIFLPQHTDADYILVVLSGKAILTVLKPDDRNSFNLERGDTIKLPAGTIAYLVNRDDNEELRVLDLAIPVNRPGQLQ
SFLLSGNQNQQNYLSGFSKNILEASFNTDYEEIEKVLLEEHEKETQHRRSLKDKRQQSQEENVIVKLSRGQIEELSKNAKSTSKKSVSSESE
PFNLRSRGPIYSNEFGKFFEITPEKNPQLQDLDIFVNSVEIKEGSLLLPHYNSRAIVIVTVNEGKGDFELVGQRNENQQEQRKEDDEEEEQ
GEEEINKQVQNYKAKLSSGDVFVIPAGHPVAVKASSNLDLLGFGINAENNQRNFLAGDEDNVISQIQRPVKELAFPGSAQEVDRILENQ
KQSHFADAQPQQRERGSRETRDRLSSV ASCNGVCSPFEMPPCGTSACRCIPVGLVVGYCRNPSGVFLRTNDEHPNLCESDADCR
KKGSGNFCGHYPNPDIEYGWCFASKSEAEDFFSKITPKDLLKSVSTA EQLTKCEVFRELKDLKGYGGVSLPEWVCTTFHTSGYDTQA
IVQNNDSTEYGLFQINNKIWCKDDQNPHSSNICNISCDKFLDDDLTDDIMCVKKILDKVGINYWLAHKALCSEKLDQWLCEKL LIVT
QTMKGLDIQKVAGTWYSLAMAASDISLLDAQSAPLRVYVEELKPTPEGDLEILLQKWENGECAQKKIIAEKTKIPAVFKIDALNENKVLV
LDTDYKKYLLFCMENSAEPEQSLACQCLVRTPEVDDEALEKFDKALKALPMHIRLSFNPTQLEEQCHI RELEELNVPGEIVESLSSSEE
SITRINKKIEKFQSEEQQQTEDELQDKIHPFAQTQSLVYPFPGPIPNSLPQNIPPLTQTPVVVPPFLQPEVMGVSKVKEAMAPKHKEMP
FPKYPVEPFTESQSLTLTDVENLHLPLPLLQSWMHQPHQPLPPTVMFPPQSVLSLSQSKVLPVPQKAVPYPQRDMPIQAFLLYQEPVL
GPVRGPFPIIV LREQPQQNECQLERLDALEPDNRIESEGGLIETWNPNNKQFRCAGVALSRATLQRNALRRPYYSNAPQEIFIQQG
NGYFGMVFPGCPETFEEPQESEQGEGRRYRDRHQKVNRFREGDIIAVPTGIVFWMYNDQDTPVIAVSLTDIRSSNNQLDQMPRRFYL
AGNHEQEFLQYQHQQGGKQEQENEGNNIFSGFKRDYLEDAFNVNRHIVDRLQGRNEDEEKGAIVKVKGGLSIISPPEKQARHQRGSR
QEEDEDEEKQPRHQRGSRQEEEEDEDEERQPRHQRRRGEEEEEDKKERGGSQKGKSRRQGDNGLEETVCTAKLRLNIGPSSSPDIYNP
EAGRIKTVTSLDLPVLRWLKLSAEHGSLHKNAMFVPHYNLNANSIIYALKGRARLQVVNCNGNTVFDGELEAGRALTVPQNYAVAAKS
LSDRFSYVAFKTNDRAGIARLAGTSSVINNLPLDVVAATFNLQRNEARQLKSNNPFKFLVPARESENRASA RSDPQNPFIFKSNKFQ
TLFENENGHIRLLQKFDQRSKIFENLQNYRLLEYKSKPHTIFLPQHTDADYILVVLSGKAILTVLKPDDRNSFNLERGDTIKLPAGTIAYLVN
RDDNEELRVLDLAIPVNRPGQLQSFLLSGNQNQQNYLSGFSKNILEASFNTDYEEIEKVLLEEHEKETQHRRSLKDKRQQSQEENVIVKL
SRGQIEELSKNAKSTSKKSVSSESEPFNLRSRGPIYSANNEFGKFFEITPEKNPQLQDLDIFVNSVEIKEGSLLLPHYNSRAIVIVTVNEGKG
DFELVGQRNENQQEQRKEDDEEEEQGEEEINKQVQNYKAKLSSGDVFVIPAGHPVAVKASSNLDLLGFGINAENNQRNFLAGDEDNV
ISQIQRPVKELAFPGSAQEVDRILENQKQSHFADAQPQQRERGSRETRDRLSSV ASCNGVCSPFEMPPCGTSACRCIPVGLVVGYC
RNPSGVFLRTNDEHPNLCESDADCRKKGSGNFCGHYPNPDIEYGWCFASKSEAEDFFSKITPKDLLKSVSTA EQLTKCEVFRELKDL
KGYGGVSLPEWVCTTFHTSGYDTQAIVQNNDSTEYGLFQINNKIWCKDDQNPHSSNICNISCDKFLDDDLTDDIMCVKKILDKVGINY
WLAHKALCSEKLDQWLCEKL LIVTQTMKGLDIQKVAGTWYSLAMAASDISLLDAQSAPLRVYVEELKPTPEGDLEILLQKWENGEC
AQKKIIAEKTKIPAVFKIDALNENKVLVLDTDYKKYLLFCMENSAEPEQSLACQCLVRTPEVDDEALEKFDKALKALPMHIRLSFNPTQLE
EQCHI RELEELNVPGEIVESLSSSEESITRINKKIEKFQSEEQQQTEDELQDKIHPFAQTQSLVYPFPGPIPNSLPQNIPPLTQTPVVVP
PFLQPEVMGVSKVKEAMAPKHKEMPFPKYPVEPFTESQSLTLTDVENLHLPLPLLQSWMHQPHQPLPPTVF DMPIQAFLLYQEP