Statistical modeling of in vitro pepsin specificity

Ousmane Suwareh, David Causeur, Julien Jardin, Valérie Briard-Bion, Steven Le Feunteun, Stéphane Pezennec, Francoise Nau

▶ **To cite this version:**

Ousmane Suwareh, David Causeur, Julien Jardin, Valérie Briard-Bion, Steven Le Feunteun, et al.. Statistical modeling of in vitro pepsin specificity. Food Chemistry, 2021, 362, pp.130098. 10.1016/j.foodchem.2021.130098 . hal-03248112

HAL Id: hal-03248112
https://hal.inrae.fr/hal-03248112v1

Submitted on 3 Jun 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

Food Chemistry 362 (2021) 130098

Contents lists available at ScienceDirect

Food Chemistry

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

Statistical modeling of *in vitro* pepsin specificity

Ousmane Suwareh [a], David Causeur [b], Julien Jardin [a], Valérie Briard-Bion [a], Steven Le Feunteun [a], Stéphane Pezennec [a], Françoise Nau [a,*]

[a] STLO, INRAE, Institut Agro, 65 rue de Saint-Brieuc, 35042 Rennes, France
[b] IRMAR UMR6625, CNRS, Institut Agro, 65 rue de Saint-Brieuc, 35042 Rennes, France

ARTICLE INFO

Keywords:
Pepsinolysis
Cleavage specificity
Prediction model
Hydrophobicity
Charge
Structure

ABSTRACT

The specificity of pepsin, the major protease of gastric digestion, has been previously investigated, but only regarding the primary sequence of the protein substrates. The present study aimed to consider in addition physicochemical and structural characteristics, at the molecular and sub-molecular scales. For six different proteins submitted to *in vitro* gastric digestion, the peptide bonds cleaved were determined from the peptides released and identified by LC–MS/MS. An original statistical approach, based on propensity scores calculated for each amino acid residue on both sides of the peptide bonds, concluded that preferential cleavage occurred after Leu and Phe, and before Ile. Moreover, reliable statistical models developed for predicting peptide bond cleavage, highlighted the predominant role of the amino acid residues at the *N*-terminal side of the peptide bonds, up to the seventh position (P7 and P7′). The significant influence of hydrophobicity, charge and structural constraints around the peptide bonds was also evidenced.

1. Introduction

Digestion is a complex process that consists of multiple steps not yet completely understood, despite a significant amount of studies on this topic. For instance, a nutrient may be present in the food without being released and absorbed during the digestion process, *i.e.* becoming bioavailable. As a result, beyond the sum of its components, the nutritional quality of food greatly depends on the structure and the physicochemical properties of the matrix (Fardet, Souchon, & Dupont, 2013). Especially, the microstructure of food influences the kinetics and extent of protein hydrolysis, as well as the release of certain peptides (Nyemb et al., 2014; Nyemb-Diop et al., 2016), some of which being potentially allergenic (Franck et al., 2002). Moreover, the phenomena that drive digestive protease activity at a molecular level are not all known either.

To deepen our knowledge of the effects of food structure on the enzymatic digestion of proteins, the main factors that influence the working mode of gastro-intestinal proteases need to be identified and hierarchised. Yet, it is especially difficult to define the substrate specificity of pepsin, the major protease of the gastric phase, and chronologically the first encountered by food. In fact, pepsinolysis releases a great variety of peptides during the gastric phase, and is even considered to be poorly reproducible according to Ahn, Cao, Yu, and Engen (2013). Studies have identified amino acid residues (AAR) the presence of which close to a peptide bond would facilitate, or in the contrary disfavour its hydrolysis by pepsin (Fruton, 1970; Hamuro, Coales, Molnar, Tuske, & Morrow, 2008). The presence of such AAR in the positions flanking a peptide bond is considered as an indicator to predict its cleavage (Tonda, Grosvenor, Clerens, & Le Feunteun, 2017). However, the cleavage frequency tables established so far, based on the nature of the residues, remain to be perfected. These tables are useful, but are insufficient to predict the peptides stemmed from a given protein. It can be assumed that pepsin activity is modulated by protein structural characteristics as well, in relation to the more or less easy access of pepsin to the peptide bonds. As a result, taking into account the protein structural

Abbreviations: BLG, β-lactoglobulin; ConA, concanavalin A; Lip, soy lipoxygenase; Lys, lysozyme; MG, myoglobin; OVA, ovalbumin; AAR, amino acid residue; SASA, solvent accessible surface area; ROC, receiver operating characteristic; PTM, post translational modification; GRAVY, grand average of hydropathy; AIC, Akaike information criterion; AUC, area under the curve; LC–MS/MS, liquid chromatography coupled to tandem mass spectrometry; VMD, visual molecular dynamics.

* Corresponding author.
E-mail addresses: ousmane.suwareh@inrae.fr (O. Suwareh), david.causeur@agrocampus-ouest.fr (D. Causeur), julien.jardin@inrae.fr (J. Jardin), valerie.briard-bion@inrae.fr (V. Briard-Bion), steven.le-feunteun@inrae.fr (S. Le Feunteun), stephane.pezennec@inrae.fr (S. Pezennec), francoise.nau@agrocampus-ouest.fr (F. Nau).

https://doi.org/10.1016/j.foodchem.2021.130098
Received 4 February 2021; Received in revised form 13 April 2021; Accepted 11 May 2021
Available online 13 May 2021
0308-8146/© 2021 Elsevier Ltd. All rights reserved.

features could improve the reliability of models for pepsin activity.

The previous studies dedicated to determining which characteristics make a peptide bond more susceptible to pepsin hydrolysis than another only focused on the nature of the residues along the primary sequence of the protein substrate (Fruton, 1970; Hamuro et al., 2008). In contrast, a major challenge for the present study was to consider in addition different physicochemical factors, at the molecular and sub-molecular scales, which may legitimately be regarded as indicators of different phenomena such as steric hindrance, electrostatic, and hydrophobic interactions. In particular, the 3D-structure of the protein substrates was taken into account in order to assess the effect of the local environment of each peptide bond, beyond the sole primary sequence. This original approach was developed on a set of six proteins with varied physicochemical characteristics, which were all submitted to *in vitro* gastric digestion. Using experimental identification of the peptides released and statistical modelling approaches, new assumptions could be put forward with regard to determining factors of pepsin activity.

2. Materials and methods

2.1. Proteins

β-Lactoglobulin (BLG; Uniprot entry P02754; PDB entry: 1BSQ; Cat. No. L3908), concanavalin A (ConA; P02866; 1GKB; Cat. No. C2010), soy lipoxygenase (Lip; P08170; 1YGE; Cat. No. L7395) and myoglobin (MG; P68082; 1AZI; Cat. No. M9267) were purchased from Sigma-Aldrich. Ovalbumin (OVA; P01012; 10VA) was prepared following the procedure established by Croguennec, Nau, Pezennec, and Brule (2000), and lysozyme (LYS; P00698) was provided by Liot SA (France). The structural characteristics and size of the proteins were collected from Uniprot (UniProt Consortium, 2018) and their physicochemical characteristics were determined from the ProtParam tool of ExPaSy (Gasteiger et al., 2003).

2.2. Peptidomic data origin

The six proteins were individually submitted to digestion in gastric conditions according to the static *in vitro* digestion model described by Minekus et al. (2014), except that the gastric phase lasted 60 min only, and purified bile salts (NaGC and NaGCDC, from Sigma-Aldrich) replaced bile extract. All the digestions were performed in triplicate as described by Torcello-Gomez, Dupont, Jardin, Briard-Bion, Deglaire, Risse, Mechoulan, and Mackie (2020). Protein solutions (5 mg/mL) were diluted with simulated gastric fluid (50:50) and the pH was adjusted to 3.0 before porcine pepsin addition (2000 U/mL in the final volume; Cat. No. P7012 from Sigma-Aldrich). Samples (200 μL) were taken from each protein solution after 30 s, 2, 5, 10, 20, 30 and 60 min of digestion and pepsinolysis was immediately stopped by adding 5 μL of 0.73 mM Pepstatin A in each sample.

The samples were then analysed by LC-MS/MS for peptide identification as described by Torcello-Gómez et al. (2020). Briefly, mass spectrometry analysis was performed using a nano-LC Dionex U3000 system fitted to a Q-Exactive mass spectrometer (Thermo Scientific, San Jose, USA) equipped with a nano-electrospray ion source. Peptides were separated on a C18 PepMap RSLC column (Dionex) using an acetonitrile gradient. The Proxeon source operated in positive ion mode using the m/z range 250–2000, with a resolution of the mass analyzer set to 70,000 for MS and 17,500 for MS/MS. Peptides were identified from the MS/MS spectra using the X!Tandem pipeline software (Langella et al., 2017) against a database composed of the sequences of the proteins studied to which was added the common Repository of Adventitious Protein (http://thegpm.org/crap). Database search parameters were specified as follows: non-specific enzyme cleavage, and serine phosphorylation, methionine oxidation, and deamidation of glutamine or aspartic acid as putative modifications. A minimum score corresponding to an e-value $<$ 0.01 was required to validate peptide identification.

For five out of the six proteins digested, *i.e.* BLG, ConA, OVA, LYS and Lip, the digestion was performed during the *In vitro protein digestibility (Allergestion)* project (Torcello-Gómez et al., 2020) and peptidomic data were directly provided by the authors. The digestion of the 6th protein, *i.e.* MG, especially chosen for the current project, was performed and the peptidomic data obtained strictly following the same protocol as Torcello-Gómez et al. (2020).

2.3. Generation of the data tables

Data processing was carried out with the R software (R Core Development Team, 2020).

2.3.1. Identification of cleavage sites

The sequences of all identified peptides were collected from the peptidomic data. For each peptide identified in at least two out of the three replicates, the sequence was matched with the complete sequence of the protein. This matching made possible the identification of the two peptide bonds (N- and C-terminal ends), the hydrolysis of which has led to the release of the peptide. The hydrolysed peptide bonds (hereinafter referred to as "cleavage sites") were identified at each digestion time for each protein.

2.3.2. Construction of the data tables

A data table was constructed for each protein. Each row of each data table referred to one of the protein's peptide bonds and each column to a variable describing the peptide bonds likely to give information about the rules of pepsin activity. Fourty variables were selected for describing the peptide bonds as such (20 variables), their physicochemical (11 variables) and structural (2 variables) environment, and their cleaved/non-cleaved status at each digestion time (7 variables).

2.3.2.1. Description of the peptide bonds (20 variables). The nature of 10 AAR on each side of each peptide bond along the primary sequence were the first 20 variables. Conventionally, the position occupied by these 20 AAR from the closest to the furthest from the peptide bond are called P1, P2, P3, P4, P5, P6, P7, P8, P9 and P10 (*N*-terminal side), respectively and P1′, P2′, P3′ P4′, P5′, P6′, P7′, P8′, P9′ and P10′ (*C*-terminal side), respectively.

2.3.2.2. Physicochemical environment of the peptide bonds (11 variables). The accessibility of the peptide bonds, illustrated by the Solvent Accessible Surface Area (SASA), was calculated from the Protein Data Base (PDB) file of each protein with the "measure sasa" function of Visual Molecular Dynamics (VMD) (Humphrey, Dalke, & Schulten, 1996), using 1.4 Å as the probe radius. Firstly, the accessibility of the peptide bond within the protein which it is part of ("SASA protein") was calculated, *i.e.* considering the 3D folding of the protein. Secondly, the accessibility considering the peptide bond as isolated ("SASA isolated") was calculated. Finally, the only accessibility variable retained for analysis was the ratio of "SASA protein" to "SASA isolated", hereinafter referred to as "SASA ratio".

The next variables referred to the number of potentially influent post-translational modifications (PTM) (disulphide bridge, phosphorylation, glycosylation or acetylation) around each peptide bond within four different radiuses: 3, 6, 9 and 15 Å ("PTM 3A", "PTM 6A", "PTM 9A" and "PTM 15A", respectively). The distances between the considered peptide bond and all the residues carrying a PTM were calculated from the PDB file of the protein using the coordinates ("X", "Y" and "Z") of the alpha carbons of the AAR. The number of PTMs inside each radius around the peptide bond was then determined.

Three variables were generated for reporting the net charge of the residues surrounding each peptide bond considering three different radiuses: 3, 6 and 9 Å ("charge 3A", "charge 6A", "charge 9A", respectively). It is noteworthy that 9 Å is the radius length allowing to consider

the eight residues generally assumed to be bound by the active site of pepsin (Polverino de Laureto, Frare, Gottardo, van Dael, & Fontana, 2002). For each peptide bond, the distances to all residues of the protein were calculated as explained above, before selecting the residues located within the considered radiuses. The charge of each selected residue was calculated as the sum of the charges of all the atoms contained in it. The PQR file containing these charges was generated thanks to the PDB2PQR / PROPKA program (http://server.poissonboltzmann.org/pdb2pqr) (Dolinsky, Nielsen, McCammon, & Baker, 2004; Olsson, Søndergaard, Rostkowski, & Jensen, 2011; Søndergaard, Olsson, Rostkowski, & Jensen, 2011) using the CHARMM force-field at pH 3.0.

Three variables referred to the total hydrophobicity (GRAVY) of the residues surrounding each peptide bond, also considering radiuses of 3, 6 and 9 Å ("GRAVY 3A", "GRAVY 6A", "GRAVY 9A", respectively). Using a similar procedure as for the net charge, the GRAVY score was calculated for each radius using the "hydrophobicity" function of the package "peptides" (Osorio, Rondón-Villarreal, & Torres, 2015) and the Kyte-Doolittle scale.

2.3.2.3. Structural environment of the peptide bonds (2 variables). The "secondary structure" variable indicated the conformation of the protein segment, according to the PDB file, containing the two AAR on both sides of the considered peptide bond. In order to enable further analyses, the nomenclature adopted by the PDB was simplified to minimize the number of modalities, and so to have enough observations for each of them. To this end, the different α-helix types (H, G, I) were grouped into the single modality "helix", the different β-sheet types (E and B) into the single modality "sheet", and finally the less ordered structures (S, T and C) into the single modality "coil". This grouping was decided in accordance with Reeb and Rost (2019). As a result, six different modalities were defined for the "secondary structure" variable: HH, EE, CC, HE, HC and EC.

The distance between each peptide bond and the nearest unstructured zone (coil) along the primary sequence, and expressed in number of AAR, was another generated variable, namely "Distance from coil".

2.3.2.4. Cleaved/non-cleaved status (7 variables). For a given peptide bond, there were seven cleavage variables, one per digestion time ("cleavage T05", "cleavage T2", "cleavage T5", "cleavage T10", "cleavage T20", "cleavage T30" and "cleavage T60", respectively). The cleavage variable was set to "1" if the peptide bond was cleaved, *i.e.* if among the peptides identified there was at least one peptide generated by the cleavage of the peptide bond in question, and to "0" otherwise.

2.4. Calculation of the protein density

The density of a protein can be defined as the mean number of AAR per unit of volume. In that aim, theoretical volumes represented by spheres with 5 Å radius were equitably distributed throughout each whole protein, in three dimensions (X, Y, Z). The centres of the spheres were placed at each node of a 3D grid, the unit element of which was a 3 Å cube. All the spheres without residues inside were excluded of the calculation of the protein density as they were assumed to be located outside the protein. Finally, the protein density was calculated as the mean number of residues per sphere.

2.5. Statistical and descriptive analyses

All the analyses were performed using the R software (R Core Development Team, 2020). The data table used to construct the table of cleavage frequencies, logistic regression models and the optimal model was a merge data of all the different protein data tables (cf. 2.3.2.). Therefore, a variable referring to the protein to which the peptide bond belongs ("protein") was added to the variables.

2.5.1. Table of cleavage frequencies

The cleavage frequency represents the percentage of P1-P1′ peptide bonds which were cleaved at least once, to the total number of occurrences of the P1-P1′ combination in the six proteins digested. The table of cleavage frequencies was constructed using the "coltable" function of the SensoMineR package (Husson & Lê, 2009). The column names represented the nature of the AAR at the P1 position, while row names represented the nature of the AAR at the P1' position.

2.5.2. Peptides coverage maps

For a given protein, all the peptides identified during the course of the *in vitro* digestion were mapped on the complete protein sequence using the "segments" function of the graphics package (R Core Development Team, 2020). Each peptide was drawn on a separate horizontal line, from the peptide with the smallest starting residue number on the top, to the one with the highest residue number at the bottom.

2.5.3. Logistic regression models

Maximum-likelihood estimation of logistic regression models was implemented using the "glm" function of the R package stats (R Core Development Team, 2020). Studentized propensity scores and t-tests for the significance of adding the variable were obtained using the summary function of the same package. Analyses of deviance tables were obtained from the different variables assessed by the logistic regression models constructed, for type II tests, using the "Anova" function of the car package (Fox & Weisberg, 2019) or for type I tests, using the "anova.glm" function of the R package stats (R Core Development Team, 2020).

2.5.4. Feature selection

A stepwise search of a regression model including only a subset of explanatory variables and minimizing the Akaike Information Criterion (AIC) was implemented using the forward/backward algorithm of the function "stepwise" in the R package "RcmdrMisc" (R Core Development Team, 2020). The former recursive algorithm starts with no variable in the model and updates the current model by adding the explanatory variable that leads to the largest decrease of the AIC. The relevance of adding this variable to the model was automatically assessed by a likelihood ratio test. The area under the sensitivity/specificity curve (AUC) was chosen as the criterion for evaluating the model.

3. Results and discussion

3.1. Main features of the six proteins digested

The panel of the six proteins investigated in the present study offers a wide range of physicochemical properties (Suppl. 1). In that respect, the molecular weight of the proteins varies by a factor of eight between the largest protein (lipoxygenase (Lip), 94.369 kDa) and the smallest one (lysozyme (LYS), 14.313 kDa). Their isoelectric points (pI) range from acidic values (pI = 4.83 for BLG) to basic ones (pI = 9.32 for LYS). The proteins also differ in terms of secondary structure: myoglobin (MG) consists exclusively of α-helices (79.7%) whereas β-lactoglobulin (BLG) and concanavalin-A (ConA) mainly consist of β-sheets (54.3% and 48.3%, respectively). Some of these proteins carry PTMs as disulfide bridges (LYS, BLG and ovalbumin (OVA)), phosphorylations (OVA and MG), or glycosylations (OVA), in different proportions. Similarly, their stability index, aliphatic index, and GRAVY scores vary by factors of 4, 1.6, and 78, respectively. Therefore, it can be assumed that this panel of proteins provides the opportunity to assess how the physicochemical and structural characteristics of proteins can affect the pepsin activity.

3.2. Pepsin cleavage sites are not uniformly distributed along the protein sequences

The six proteins presented above have been individually submitted to *in vitro* gastric digestion during which samples were collected at seven

digestion times. The peptides were unambiguously identified by LC-MS/MS analysis. Throughout the 60-min digestion, the total number of different peptides identified varied from only six for LYS, up to 345 for Lip (Suppl. 2). Most of the peptides (50 to 93%) were identified as early as 30 s of digestion for all the proteins, and only a few additional peptides were generated at a later stage of the digestion, meaning pepsin was very active from the very first moments of the digestion. It is noteworthy that in the present study, pepsin was added in protein solutions at pH 3.0, at which pepsin activity is almost maximum (Luo, Chen, Boom, & Janssen, 2018). *In vivo,* however, the gastric pH quickly increases while food enters into the stomach (except with highly acidic foods) before decreasing very progressively all over the stomach (Nau et al., 2019). In real conditions, proteolysis by pepsin is therefore probably not as quick as observed in the present *in vitro* study.

The identified peptides covered a high proportion of the protein sequences they stemmed from for all the proteins (from 68.3% up to 100% protein coverage), except LYS (25.6%) (Fig. 1). The peptide coverage maps also highlight how the structural particularities of each protein can affect the identification and/or the release of peptides during digestion. Most of the protein segments not covered by the identified peptides despite the 60-min digestion contained structural particularities (Fig. 1). Especially, most of these segments contained Cys residues involved in disulfide bridges (Fig. 1B, D, F). It can be assumed that these fragments did actually exist in the digested samples, but were not identified by LC-MS/MS because they did not correspond to theoretical linear sequences that were used for bioinformatics identification. Surprisingly, we

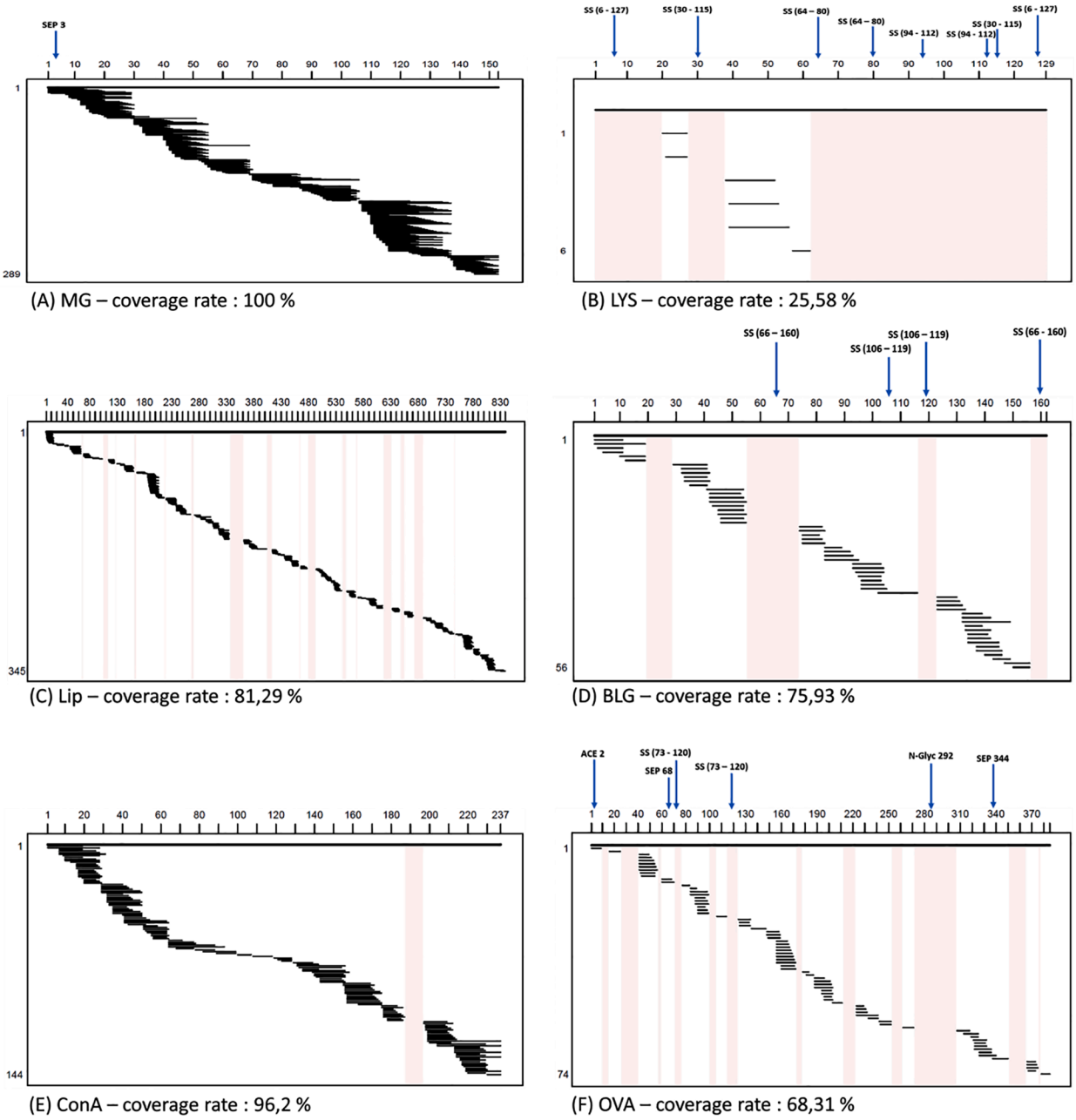

Fig. 1. Peptide coverage maps of (A) myoglobin (MG); (B) lysozyme (LYS); (C) lipoxygenase (Lip); (D) ß-lactoglobulin (BLG); (E) concanavalin-A (ConA); (F) ovalbumin (OVA). The size of each identified peptide is relative to the length of the complete protein sequence represented at the top of each graph. Blue arrows indicate cysteine residues involved in a disulfide bridge (SS), a glycosylation site (*N*-Glyc), phosphoserine residues (SEP), or an acetylated residue (ACE). The coloured areas of the maps represent the segments of protein sequences not recovered in the peptide populations (missing segments). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

identified in BLG digesta a peptide including a Cys residue (Cys_{106}) involved in a disulfide bridge (Fig. 1D). This peptide probably stemmed from denatured BLG molecules in which the disulfide bridge Cys_{106}-Cys_{119} was cleaved. Such a denaturation could result from the process applied to produce the BLG powder and/or from the storage conditions of the protein powder (Norwood et al., 2016).

Peptides were also missing in the OVA sequence around the glycation site (Fig. 1F). The length of the missing segment (35 AAR) suggests that the proteolysis could have been hindered due to the steric hindrance of the glycosylated group. The decrease of pepsin accessibility to its cleavage sites due to glycosylation has been previously reported (Niu et al., 2016). In contrast, neither acetylation (Gly_1, Fig. 1F) nor phosphorylation (Ser_3, Fig. 1A; Ser_{344} Fig. 1F) seemed to disturb pepsinolysis.

There were also missing segments that did not present any structural specificity. Some of them, of five AAR or less (10 zones in Lip and three in OVA), could be too small to be identified by LC-MS/MS because of technical limits (5–6 AAR). These small fragments are inherently indicative of protein zones actively and/or preferentially hydrolysed by pepsin. However, there were also longer missing segments (nine, five, one and one in Lip, OVA, BLG and ConA, respectively) without PTM or other apparent constraint that could explain why these zones were not identified in the peptide populations (Fig. 1C, D, E, F).

By contrast, some zones of the protein sequences were much more covered than others, *i.e.* they were included in a high number of different peptides (Fig. 1A, C, D, E, F). This suggests these highly-covered zones could be relatively resistant ones, located between two highly susceptible zones in which many different peptide bonds could be hydrolysed by pepsin.

3.3. Sensitivity to pepsinolysis depends on protein density and PTMs

Proteolysis degree is commonly expressed as either the percentage of hydrolysed peptide bonds over the total number of peptide bonds, or as the concentration ratio of the hydrolysed proteins to the initial proteins (w/w). According to the latter definition, a high protein sensitivity means a high proportion of hydrolysed protein molecules, even if the number of peptide bonds hydrolysed in each protein molecule is low.

In the present study, peptidomic data could only be used to identify released peptides but not to measure their absolute abundance. Therefore, the usual proteolysis degree mentioned above could not be deduced. However, the proportion of peptide bonds cleaved at least once during the 60-min digestion (hereinafter referred to as cleavage rate) could be considered as an indicator of the protein sensitivity to pepsinolysis, although slightly different to the parameters commonly used (Deng, van der Veer, Sforza, Gruppen, & Wierenga, 2018). Hence, a high cleavage rate indicates that the protein sequence is susceptible to pepsinolysis at many places. This criterion of protein sensitivity to pepsin is the one used in the sections below.

3.3.1. High molecular density increases resistance to pepsinolysis

Since pepsin activity requires a physical access to peptide bonds, we assumed that the sensitivity to pepsin hydrolysis could depend on protein density. Fig. 2 shows the relationship between the mean protein density (estimated as explained in Section 2.4.) and the cleavage rate, which highlighted huge differences between the six proteins studied. As expected, the protein with the highest density, namely LYS, has the lowest cleavage rate, *i.e.* 7.0% (Fig. 2). In contrast, MG has a much lower density and the highest cleavage rate, *i.e.* 67.8%. In all, considering LYS, BLG, ConA and MG, a negative relationship between protein density and protein cleavage rate seemed to emerge (Fig. 2).

However, Lip and OVA displayed dissimilar patterns. Lip (839 AAR) is approximately six times as big as the other proteins, except OVA (385 AAR). Then, regardless of the protein density, the percentage of peptide bonds easily accessible for pepsin is statistically lower in a large protein in comparison to a smaller one (Hamuro et al., 2008). This could explain

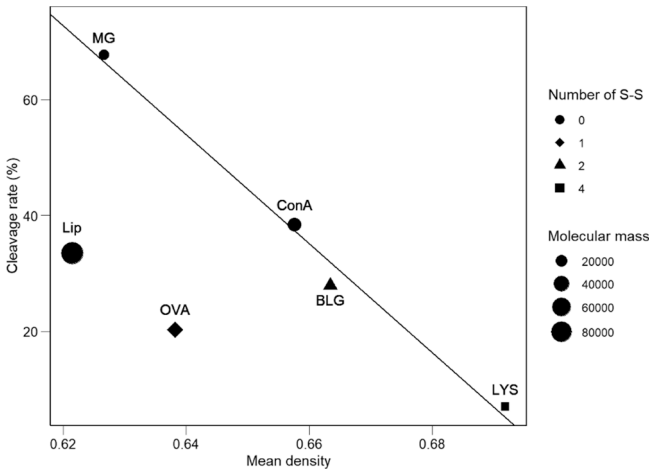

Fig. 2. Relationship between cleavage rate (%) and mean density of proteins. MG, LYS, Lip, BLG, ConA, and OVA are respectively the abbreviations for myoglobin, lysozyme, lipoxygenase, beta-lactoglobulin, concanavalin-A, and ovalbumin. The mean density is the mean number of residues in a sphere 5 Å in radius. The linear regression line is based on LYS, BLG, ConA, and MG only.

why, despite a mean density slightly lower than that of MG, the cleavage rate was lower for Lip (33.5%) than for MG. Concerning OVA, since glycans are not described in the 3D structure established from X-ray crystal diffraction, they could not be taken into account in the mean density estimation, thus possibly leading to underestimation.

3.3.2. Disulfide bridges and glycosylation increase resistance to pepsinolysis

As shown in Figs. 1 and 2, LYS seems to be the most resistant protein to pepsin hydrolysis considering that it has the lowest cleavage rate (7.0%). This is consistent with the low number of peptides generated from LYS (Suppl. 2), and with the literature in which LYS is described as particularly resistant throughout the gastric phase, probably due to its very compact and rigid structure stabilised by four disulfide bridges in a sequence of 129 AAR only (Jiménez-Saiz, Martos, Carrillo, López-Fandiño, & Molina, 2011). Indeed, beyond the compactness of a protein, reflected by the mean protein density score presented above, protein flexibility is another key parameter. Protein flexibility can increase susceptibility to proteolysis as it is assumed that the active site of proteases needs access to an unfolded stretch of up to 12 AAR (Fontana, Polverino de Laureto, De Filippis, Scaramella, & Zambonin, 1997) and binding eight to 10 AAR of them (Polverino de Laureto et al., 2002). It is therefore noteworthy that the flexibility of LYS increases at very low pH (1.2 – 2.0) but remains low at pH 3.0 (Polverino de Laureto et al., 2002), *i.e.* the pH of the *in vitro* digestion. Similarly, BLG in its native state is generally considered as highly resistant to pepsin hydrolysis due to a low flexibility, partly explained by the presence of two disulfide bridges (Gekko, Kimoto, & Kamiyama, 2003).

OVA seems to be also characteristically resistant to gastric digestion considering a cleavage rate of 20.3% (Fig. 2). The relative resistance of OVA to pepsinolysis has been explained as a consequence of the protein glycosylation (Dupont et al., 2010). The presence of *N*-glycosylated groups would help to stabilise the tertiary structure of OVA, as it was reported for other proteins due to interactions between the oligosaccharide core and the AAR side chains (Opdenakker, Rudd, Ponting, & Dwek, 1993). This stabilising effect should be higher in the acidic conditions of *in vitro* gastric digestion (Niu et al., 2016). In addition, the steric hindrance due to the *N*-linked glycan (Niu et al., 2016; Opdenakker et al., 1993) may contribute to the resistance of OVA to pepsinolysis.

The three proteins that had the highest cleavage rates (Lip, ConA, and MG, Fig. 2) share one common feature which is the absence of disulfide bridges, or glycosylation. This could result in a higher

accessibility of pepsin to the peptide bonds in these protein sequences, because not hindered by structure constraints. Moreover, high values for both the cleavage rate (Fig. 2) and coverage rate (Fig. 1) calculated for these proteins could be indicative of an easier peptide identification in the absence of PTMs.

Consistently with the present study, the high susceptibility of Lip to hydrolysis by pepsin has been previously reported (Astwood, Leach, & Fuchs, 1996). Other results of the present study may appear less consistent with the available literature. MG was reported to be resistant during *in vitro* gastric digestion (Li et al., 2020), whereas the highest cleavage rate was measured for this protein in the present study. ConA was reported to be as susceptible as OVA to pepsinolysis (Takagi, Teshima, Okunuki, & Sawada, 2003), contrary to the present results. BLG seemed to be more susceptible to pepsin than reported in literature which mentions BLG as much more resistant than OVA (Dupont et al., 2010), or of equivalent susceptibility (Astwood et al., 1996; Takagi et al., 2003). In contrast, cleavage rate measurement indicated OVA as more resistant to pepsin hydrolysis than generally reported in literature, considering that no intact proteins remain after 60 min of gastric digestion (Takagi et al., 2003; Torcello-Gómez et al., 2020). As mentioned above, these discrepancies most likely arise from the different experimental approaches (SDS-PAGE, free NH_2 release, etc) and the related calculations used to assess the protein sensitivity to pespinolysis. Moreover, the variability in the experimental conditions (pH, pepsin and substrate concentrations, etc.) used for the digestion experiments may also contribute to these discrepancies.

3.4. The nature of the AAR at P1 position is more critical for pepsin activity than that at P1′ position

Pepsin specificity is usually determined by identifying the AAR the presence of which near a given peptide bond would favour its pepsinolysis (Fruton, 1970; Hamuro et al., 2008), introducing the concept of "preferential" cleavage sites. Hamuro et al. (2008) reported the residues at the P1 and P1′ positions as having the greatest influence on pepsin hydrolysis. In order to make the present study comparable to that of Hamuro et al. (2008) regarding the specificity of pepsin, only the peptide bond cleavages observed after 30 s of digestion were considered for analysis in the following sections.

For the six proteins of the present study, and the 400 (20 × 20) possible peptide bonds depending on the nature of the residues at the P1 and P1′ positions, the cleavage frequencies observed are indicated in Table 1. The highest cleavage frequencies (>40%) concern 115 P1-P1′ combinations (28.75% of the total, red cells). The medium cleavage frequencies (20 to 40%) concern 86 combinations (21.5% of the total, orange cells). The lowest cleavage frequencies (<20%) concern 47 combinations (11.75% of the total, yellow cells). Some P1-P1′ combinations (128, 32% of the total) were never hydrolysed (white cells), and 24 P1-P1′ combinations (NaN, grey cells) did not exist in the data set.

Consistently with Hamuro et al. (2008), the present study highlights that pepsin cleaves preferentially certain peptide bonds, whereas some others are never or very rarely hydrolysed (Table 1). This makes it clear that the nature of the residues at both P1 and P1′ positions is a strong determinant for the activity of pepsin. Thus, the highest average cleavage frequencies were observed after Phe (F), Leu (L) and Glu (E) residues (*vs* after Phe, Leu and Met according to Hamuro et al. (2008)), and before Phe, Leu and Tyr (Y) residues (*vs* before Tyr, Phe and Trp according to Hamuro et al. (2008)). The present results thus confirm the affinity of pepsin for hydrophobic AAR. However, some P1-P1′ combinations not including these AAR were also highly hydrolysed by pepsin: 100% of the Gln-Asn (Q-N), Gln-Gln (Q-Q), Gln-Arg (Q-R), Val-Trp (V-W) and His-Thr (H-T) combinations were hydrolysed (Table 1). Moreover, unlike Hamuro et al. (2008), some cleavages after His, Lys, Arg and Pro (P) were observed. Despite the frequencies of these cleavages were low, it demonstrates that these four AAR at P1 position do not definitely prohibit cleavage by pepsin. This might be due to different experimental conditions as compared to the ones used by Hamuro et al. (2008) who, in particular, used immobilised pepsin.

Table 1
Cleavage frequencies of all possible peptide bonds after 30 s of digestion. Cleavage frequency means percentage of cleavages of a given P1-P1′ combination to the total number of this combination in the six proteins of the dataset. The column names give the nature of the AAR (according to the international codification) at the P1 position. The row names give that of the residue at the P1′ position. The higher (]40%-100%]), medium (]20%-40%]) and lower (]0%-20%]) frequencies are respectively coloured in red, orange and yellow. The P1-P1′ combinations never hydrolysed appear in white. NaN means that the combination did not exist in the data set. The "Ave." column (respectively "Ave." row) shows the average cleavage frequencies for all the peptide bonds for which the corresponding AAR is located at the P1 position (respectively P1′ position).

	F	L	E	M	D	W	Y	A	Q	T	N	V	R	S	I	G	K	H	P	C	Ave.
F	50	70	42.9	50	100	NaN	NaN	75	0	33.3	25	28.6	50	20	0	50	66.7	0	25	0	44
L	42.9	54.5	75	50	50	50	42.9	37.5	50	42.9	14.3	43.8	40	11.8	18.2	35.7	27.3	0	40	0	38.5
Y	75	87.5	0	100	20	100	100	50	0	33.3	50	37.5	0	25	25	0	20	0	20	0	38.1
I	100	66.7	50	50	66.7	0	0	62.5	50	42.9	40	16.7	0	33.3	22.2	33.3	10	28.6	0	0	37.8
V	100	71.4	37.5	100	22.2	28.6	16.7	45.5	50	25	40	62.5	40	30.8	0	20	11.1	0	16.7	0	33.6
E	40	72.2	23.1	20	14.3	33.3	50	50	66.7	75	0	11.1	0	14.3	14.3	0	33.3	33.3	0	0	30.2
K	80	58.8	46.2	20	16.7	0	33.3	12.5	0	28.6	33.3	0	0	30	25	40	0	0	0	0	30
M	66.7	66.7	75	0	NaN	0	0	37.5	0	0	50	0	0	0	50	0	0	NaN	0	0	30
Q	75	50	25	100	0	100	0	40	100	20	50	0	0	0	0	25	0	0	0	0	29.2
W	NaN	100	25	NaN	NaN	33.3	NaN	0	NaN	0	50	100	0	66.7	25	NaN	50	0	0	NaN	29
H	100	50	NaN	0	0	NaN	50	33.3	50	75	0	0	50	33.3	50	16.7	20	0	16.7	0	28.8
D	60	50	20	0	33.3	0	75	36.4	0	9.1	14.3	0	40	14.3	66.7	12.5	25	0	0	NaN	26.2
P	25	47.1	33.3	66.7	100	100	25	33.3	0	12.5	0	0	0	0	0	50	0	0	0	0	25.6
R	40	60	0	NaN	100	0	60	20	100	33.3	25	12.5	25	20	16.7	11.1	0	0	0	0	24.3
A	33.3	11.1	63.6	33.3	44.4	0	40	21.4	28.6	37.5	0	14.3	0	22.7	22.2	16.7	22.2	14.3	0	0	24
N	75	66.7	40	33.3	0	100	50	20	100	20	33.3	0	14.3	14.3	9.1	14.3	0	NaN	0	0	22.8
S	100	38.5	44.4	0	50	0	25	14.3	16.7	0	25	18.2	0	13.3	20	0	9.1	25	0	0	20.9
T	50	42.9	40	50	20	50	0	33.3	20	50	12.5	0	14.3	9.1	0	7.7	0	100	16.7	NaN	20.6
G	33.3	38.5	20	0	16.7	100	0	0	14.3	16.7	42.9	20	22.2	0	20	0	12.5	16.7	0	0	18.8
C	33.3	0	0	NaN	NaN	0	0	0	0	NaN	0	0	0	0	NaN	0	NaN	NaN	0	NaN	4.3
Ave.	57.3	54.9	42.1	36.6	35.9	35.5	34.9	32.5	29.2	27.5	23.1	20.5	20.3	18.3	17.9	17.8	17.4	13.5	7.9	0	28.6

Moreover, as reported by Hamuro et al. (2008), the nature of the AAR at the P1 position seems to have a stronger influence on pepsin cleavage than the AAR at the P1′ position. Indeed, the distribution of the average cleavage frequencies for the different AAR at the P1 position has a high standard deviation (14.1), with three AAR (Phe, Leu and Glu) after which a cleavage was very frequently observed (more than 42%), and one AAR (Cys, C) after which cleavage was never observed. This absence of cleavage after the Cys residues may be partly due to their involvement in disulfide bridges (14 out of the 23 Cys of the data set), making these AAR less accessible for pepsin. In contrast, a low standard deviation (8.5) was observed for the distribution of the average cleavage frequencies depending on the AAR at the P1′ position. Interestingly, moderate cleavage frequencies (between 20.6 and 38.5% on average) were observed after 17 out of the 20 different AAR at the P1′ position.

However, some P1-P1′ combinations are much more frequent than others in the sequences of the six proteins studied, whereas some are even missing (Suppl. 3). This should prompt us to consider with caution the table of cleavage frequencies. For example, 100% cleavage was calculated for the combination Leu-Trp (L-W) (Table 1), but this combination occurred only once in the data set. Moreover, the table of cleavage frequencies was constructed by pooling the results from the six proteins, without considering that the location of each P1-P1′ combination probably differs from a protein to another, regarding structure features in particular, with potential consequences on its accessibility. This means that conclusions on pepsin preferences, based on observing the cleavage frequencies as calculated in Table 1 only, sounds risky. To address this issue, a statistical approach is proposed thereafter, based on propensity scores, and in which all the limitations exposed above have been considered.

Pepsin specificity can be viewed as the propensity of pepsin to cleave a sequence of AAR not randomly along this sequence. Hereafter, a logistic regression model of the probability of a cleavage at a given peptide bond has been introduced to identify which AAR located at P1 and P1′ positions favour cleavage. In order to account for the aforementioned highly different cleavage rates across proteins, the protein to which the peptide bond belongs was also considered. As a result, the model introduced three sets of propensity parameters: a first set of six propensity parameters (one for each protein) and two sets of 20 propensity parameters (one for each AAR at P1 and P1′ positions, respectively). For convenience, each set of propensity scores sum to zero, meaning that a peptide bond with all P1 and P1′ propensity scores equalling zero would have the same probability of a cleavage than all other peptide bonds within the same protein. A largely positive propensity parameter, for a given AAR at a given position, would indicate that this AAR favours cleavage, whereas a largely negative value would indicate that this AAR disfavours cleavage. The Cys (C) residue was excluded from this step as poorly represented in the dataset (Suppl. 3). Moreover, it is the only one after which no cleavage was observed. Consequently, its strongly negative propensity score (−13.2) hinders the analysis of the propensity scores of the other AAR at P1 position. The overall goodness-of-fit of the model was first assessed by the squared Pearson correlation of fitted probabilities of a cleavage for all pairs of AAR at P1 and P1′ positions against observed probabilities, and clearly indicates a non-random fitting (Suppl. 4).

In the logistic regression framework, under the hypothesis that the pepsin activity would not depend on the nature of the AAR at P1 position (respectively P1′), all P1 (respectively P1′) propensity scores would be zero, which is addressed by a likelihood-ratio test. Results clearly show that pepsin activity strongly depends on the nature of the residue both at P1 (p-value $< 2.2e{-}16$) and at P1′ (p-value $= 2.52e{-}08$) positions. The lower significance level at P1′ position confirms our previous hypothesis of a greater influence on pepsin activity of the residue at P1 position than that at P1′.

3.5. Huge differences of cleavage propensity scores between AAR at P1 position, evidence of positive or negative effect on pepsin activity

Fig. 3 displays the estimated P1 and P1′ studentized propensity scores (z-scores) for a cleavage. Studentization offers standard rules to identify AAR with significant propensity scores (if absolute value exceeds 1.96, then the propensity score is significant at level 0.05). Non-significant AAR (grey square on Fig. 3) will not be discussed below.

Leu (L) and Phe (F) residues have strongly positive studentized propensity scores at both P1 and P1′ positions, meaning that their presence on both sides of a peptide bond would favour in all cases its hydrolysis by pepsin. Opposite conclusions may be drawn for Gly (G) residues, associated with strongly negative studentized propensity scores at both P1 and P1′ positions. Therefore, Gly would disfavour pepsin activity in all cases. Glu (E) and Met (M) residues have positive studentized propensity scores, and Pro (P), Lys (K), His (H) and Ser (S) residues have negative scores when located at P1 position, whereas their studentized propensity scores are not significant when located at P1′ position. Therefore, these AAR influence pepsin activity only when located at P1 position: Glu and Met would favour pepsinolysis, whereas Pro, Lys, His and Ser would disfavour pepsin activity. In contrast, Thr (T) would disfavour pepsin activity when located at P1′ position, but would have no effect when located at P1 position. At last, Ile (I) is the only AAR which significantly disfavours pepsinolysis when located at P1 position, but favours pepsinolysis when located at P1′ position. In summary, only 10 AAR out of the 20 would have a significant influence on pepsin activity when located at P1 position, and only five at P1′ position (Fig. 3).

It is noticeable that the results of the present study were overall consistent with the conclusions of Hamuro et al. (2008) who tested 39 proteins, compared to only six proteins in the present dataset. One can assume that the general statistical framework chosen here avoids the potential bias due to a smaller dataset in which all the P1-P1′ combinations are not equally present. In contrast, some conclusions drawn from simple counting of P1-P1′ actually cleaved (Table 1) were not confirmed by the studentized propensity score methodology. Thus, Tyr (Y) and Trp (W) residues have no significant propensity scores at both P1 and P1′ positions. Therefore, although Hamuro et al. (2008) identified the positive impact of Trp at P1′ position on pepsin activity, it is not confirmed by the present study. It might be due to the critically low number (31) of Trp residues in the present dataset.

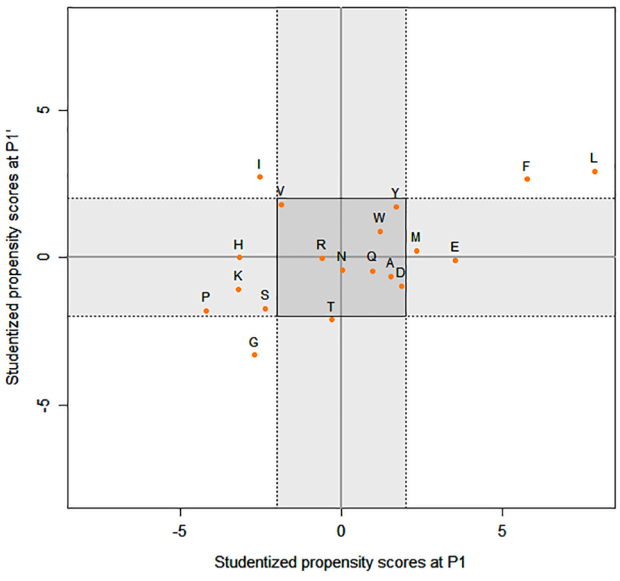

Fig. 3. Studentized propensity scores of residues at the P1 and P1′ positions for all proteins after 30 s of digestion. Studentized propensity scores were obtained using a logistic regression model.

3.6. Extended characterization of pepsin specificity beyond P1 and P1′ with a greater influence of the N-terminal side

The former logistic regression model of the probability of a cleavage is now extended in order to investigate the impact of AAR at all positions beyond P1 and P1′ from P10 at the *N*-terminal side, to P10′ at the *C*-terminal side. The analysis of deviance table of the model is provided in Table 2. Note that the relative locations Pi' at the *C*-terminal side and Pi at the *N*-terminal side of AAR have been added successively in the model according to their distance to the peptide bond (Type-I analysis of deviance). An alternative testing approach was also conducted where locations at Pi are introduced before locations at Pi', with similar conclusions (data not shown).

As expected, the residue at P1 position had a greater influence than the one at P1′ position (p-value $<$ 2.2e-16 and p-value $=$ 2.3e-04, respectively). This is generally true up to the eighth position around the peptide bond. Thus, unlike what Hamuro et al. (2008) reported, the nature of AAR in the environment of a peptide bond affects pepsin activity far beyond the third AAR on both sides. This makes sense, knowing that pepsin activity requires an interaction of the protease with a stretch of eight to 10 AAR (Polverino de Laureto et al., 2002). However, this questions previous studies concluding that pepsin would interact with seven residues at most (Powers, Harley, & Myers, 1977).

In the present study, a significant effect was observed even for the ninth and tenth AAR, at least at the *N*-terminal side (Table 2). However, in that case, it is likely that it reflected a conformational effect. Indeed, some AAR, away from a peptide bond along the protein sequence, may eventually be near because of the protein 3D structure. In such a case, these AAR could influence pepsin activity, because of their effect on the physicochemical environment of the peptide bond. Therefore, it seemed relevant to investigate the effect of such physicochemical and structural parameters.

3.7. Hydrophobicity, 3D-structure and charge around the peptide bonds also influence pepsin activity

In order to assess how much the physicochemical environment and structural features do affect the cleavage of a given peptide bond, compared to the nature of AAR from P10 to P10′ positions, the statistical model described above has been completed with covariates providing a description of the close environment of peptide bonds. All the explanatory variables described in Section 2.3.2. were candidate covariates to enter the model. A stepwise feature selection algorithm was implemented to identify a minimal subset of covariates leading to an optimal fit.

The resulting model kept 21 variables (Table 3). The quality of the model was assessed through the Area Under the ROC Curve (AUC) which measures the ability of the model to predict the occurrence of a cleavage. A poor prediction performance is measured by a low AUC value, close to 0.5, whereas a large value, close to 1, indicates a good prediction performance (Fawcett, 2006). The high AUC value obtained here, 0.9338, therefore validates the selection of features to explain the cleavage mechanism. Table 3 presents the type-II analysis of deviance of the selected model.

As expected, the nature of the AAR flanking the peptide bonds, up to the seventh residues on both sides (p-value $<$ 1.3e-04), had the most significant influence on pepsin action. The nature of the tenth residue on the *N*-terminal side was also significant (p-value $=$ 1.3e-05), but likely due to conformational effects, as mentioned above.

The nature of the protein ("protein" variable) digested also provides highly significant supplementary information (p-value $=$ 7.4e-14). This is consistent with literature that already established that some proteins are digestible by pepsin much more than others, probably due to different physicochemical and structural properties. In fact, the "protein" variable likely sums up influencing variables not taken into account in the model. For instance, protein density and rigidity did not enter the model, which focused on peptide bond characteristics only (peptide bond scale). Yet, the very compact and rigid structure of LYS has also been mentioned to explain high resistance to pepsinolysis (Jiménez-Saiz et al., 2011) . On the contrary, MG, the less dense protein in the present dataset, was found to be more susceptible to pepsinolysis (Fig. 1). If available, enthalpy of denaturation or protein compressibility might be relevant parameters to consider for further studies, at the protein scale.

The "secondary structure" variable also had a significant influence on pepsin activity (p-value $=$ 8.1e- 04). This is consistent with the thermodynamically disadvantageous conditions for proteolysis of rigid elements of secondary structures, and especially of helices (Fontana et al., 1997), despite some exceptions apply, for instance with caspase-3 protease and glutamyl endopeptidase which cleave in helices nearly as frequently as in unstructured loops (Timmer et al., 2009). It is

Table 2
Analysis of deviance Table (type I) successively including significance testing of positions of AAR around a peptide bond. Positions were included according to their distance to the peptide bond.

Variable	P-value	Significance
Protein	$<$ 2.2e-16	***
P1′	2.3e-04	***
P1	$<$ 2.2e-16	***
P2′	7.7e-04	***
P2	9.0e-09	***
P3′	2.5e-04	***
P3	1.7e-11	***
P4′	4.0e-05	***
P4	4.1e-06	***
P5′	1.3e-05	***
P5	3.8e-07	***
P6′	2.0e-03	**
P6	6.0e-04	***
P7′	1.9e-03	**
P7	5.9e-06	***
P8′	6.8e-02	
P8	1.6e-01	
P9′	9.5e-01	
P9	8.3e-03	**
P10′	8.3e-02	
P10	4.0e-05	***

Table 3
Analysis of deviance Table (type II) for the logistic regression model of the probability of a cleavage introducing physicochemical effects within the environment of a peptide bond. The p-value for each effect reflects its importance with respect to all other effects.

Variable	P-value	significance
Protein	7.4e-14	***
P1	$<$ 2.2e-16	***
P1′	2.9e-07	***
P2	4.4e-09	***
P2′	2.8e-10	***
P3	6.9e-15	***
P3′	1.8e-06	***
P4	9.5e-12	***
P4′	9.7e-06	***
P5	3.1e-14	***
P5′	6.5e-10	***
P6	5.2e-05	***
P6′	1.3e-06	***
P7	8.3e-10	***
P7′	1.3e-04	***
P10	1.3e-05	***
charge 3A	8.6e-03	**
charge 9A	1.7e-02	*
GRAVY 9A	1.1e-03	**
Secondary structure	8.1e-04	***
Distance from coil	8.2e-04	***

noteworthy that the variable "distance from a coil" was similarly significant (p-value = 8.2e-04). The negative coefficient associated to the variable "distance from a coil" (Suppl. 5) indicates that pepsin would preferentially hydrolyse peptide bonds in or nearby coils, in accordance with Fontana et al. (1997) who stated that pepsin needs access to an unfolded stretch of up to 12 AAR to be able to cleave a peptide bond.

Hydrophobicity (GRAVY score) around the peptide bond in a 9 Å radius (p-value = 1.1e-03) is another significant variable thus providing supplementary information. Hence, the negative coefficient associated to this variable (Suppl. 5) indicates that the higher the hydrophobicity, the lower the cleavage probability. This is consistent with the location of hydrophobic areas that are mainly located at the core of globular proteins like those here studied. Therefore, we can assume that such areas may be less accessible to pepsin than hydrophilic areas, mainly exposed to the surrounding water.

Finally, the net charge in the environment of the peptide bonds turns out to have also a significant impact on pepsin activity (8.6e-03 and 1.7e-02 for the charge around the peptide bond in 3 Å and 9 Å radiuses, respectively). The negative coefficient associated to "charge 9A" (Suppl. 5) indicates that the presence of positive charges around the peptide bond inhibits cleavage by pepsin. According to the charge calculation with the PDB2PQR / PROPKA program, at pH 3.0, pepsin surface is mainly positively charged (data not shown) and could be preferentially attracted by negatively charged areas of the protein. On the contrary, there is a positive coefficient associated to the "charge 3A" variable, indicating that positive charge very close to the peptide bond favours pepsinolysis. This would be consistent with the presence of two Asp residues playing a key role in the active site of pepsin (Sepulveda, Marciniszyn, Liu, & Tang, 1975).

4. Conclusions

The present study overall confirmed, but also completed the results previously published regarding pepsin specificity. Using an original statistical approach, based on propensity scores calculated for each position on both sides of a peptide bond, the inherent bias in the content of the protein dataset (occurrence of each P1-P1′ combination) could be eliminated, and reliable statistical models could be proposed for predicting peptide bond cleavage. This methodology enabled to conclude that neither Trp, nor Tyr had a significant effect on pepsinolysis when located at P1 or P1′ position, unlike previously reported. Moreover, the nature of AAR on both sides of a peptide bond proved to be critical for pepsin activity up to the seventh level (P7 and P7′) that is on a much longer stretch than previously reported; more specifically, the AAR at the *N*-terminal side of the peptide bonds have proven to be more significant. However, despite the nature of AAR is confirmed to be a key determinant of pepsin activity, it is not the only one. The nature of the protein appears as a major factor, likely due to differences in terms of accessibility of its different peptide bonds. Especially, protein density would influence the sensitivity to pepsin, as well as PTMs, which appear generally unfavourable. Moreover, physicochemical parameters and structural features of the environment of the peptide bonds must be considered as well. Thus, pepsin would have more affinity for negatively charged and non-hydrophobic areas, but would preferentially cleave peptide bonds the close environment of which is positively charged, while secondary structure elements might limit pepsin activity. However, it should be noted that the structural and physicochemical parameters introduced in the statistical models have been determined based on crystallographic structures of the proteins, which may be different from the real structures in the acidic conditions (pH 3.0) applied. Unfortunately, protein structural changes depending on pH are not yet determined for the six proteins studied here, making the physicochemical parameters and structural features impossible to adjust to the effective conditions of *in vitro* digestion. Another limitation of the study lies in that protein dynamics could not be considered due to a lack of available data, whereas it is a major parameter for protein properties and interactions with enzymes (Li, Pan, Yang, Rao, & Chen, 2021; Timmer et al., 2009). The "static" PDB structures used in the present study, that are "mean" structures, does not enable to reflect this dynamic dimension of protein features. To overcome these limitations related to pH effect and dynamics of the protein systems, new experimental strategies still need to be invented. Lastly, some protein characteristics, not investigated in the present study, such as protein flexibility, would be also interesting to consider in further studies.

Funding

This work was supported by INRAE and Conseil Régional de Bretagne.

CRediT authorship contribution statement

Ousmane Suwareh: Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing - original draft. **David Causeur:** Conceptualization, Formal analysis, Funding acquisition, Methodology, Project administration, Software, Supervision, Validation, Writing - review & editing. **Julien Jardin:** Data curation, Investigation. **Valérie Briard-Bion:** Data curation, Investigation. **Steven Le Feunteun:** Writing - review & editing. **Stéphane Pezennec:** Software, Writing - review & editing. **Françoise Nau:** Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Validation, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank EFSA for allowing the use of LC-MS/MS results produced under the EFSA contract OC/EFSA/GMO/2017/01 " *In vitro protein digestibility*".

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.foodchem.2021.130098.

References

Ahn, J., Cao, M.-J., Yu, Y. Q., & Engen, J. R. (2013). Accessing the reproducibility and specificity of pepsin and other aspartic proteases. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics, 1834*(6), 1222–1229. https://doi.org/10.1016/j.bbapap.2012.10.003.

Astwood, J. D., Leach, J. N., & Fuchs, R. L. (1996). Stability of food allergens to digestion in vitro. *Nature Biotechnology, 14*(10), 1269–1273. https://doi.org/10.1038/nbt1096-1269.

Croguennec, T., Nau, F., Pezennec, S., & Brule, G. (2000). Simple rapid procedure for preparation of large quantities of ovalbumin. *Journal of Agriculture and Food Chemistry, 48*(10), 4883–4889. https://doi.org/10.1021/jf991198d.

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. https://doi.org/10.1016/j.procbio.2017.11.006.

Dolinsky, T. J., Nielsen, J. E., McCammon, J. A., & Baker, N. A. (2004). PDB2PQR: An automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Research, 32*(Web Server), W665–W667. https://doi.org/10.1093/nar/gkh381.

Dupont, D., Mandalari, G., Molle, D., Jardin, J., Léonil, J., Faulks, R. M., … Mackie, A. R. (2010). Comparative resistance of food proteins to adult and infant *in vitro* digestion models. *Molecular Nutrition & Food Research, 54*(6), 767–780. https://doi.org/10.1002/mnfr.200900142.

Fardet, A., Souchon, I., Dupont, D., 2013. Structure des aliments et effets nutritionnels.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, ROC Anal. Pattern Recognition, 27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010.

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*(2), R17–R26. https://doi.org/10.1016/S1359-0278(97)00010-2.

Fox, J., Weisberg, S., 2019. Fox J, Weisberg S (2019). An R Companion to Applied Regression, Third edition. Sage, Thousand Oaks CA. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Franck, P., Moneret Vautrin, D. A., Dousset, B., Kanny, G., Nabet, P., Guénard-Bilbaut, L., & Parisot, L. (2002). The allergenicity of soybean-based products is modified by food technologies. *International Archives of Allergy and Immunology, 128*, 212–219. https://doi.org/10.1159/000064254.

Fruton, J.S., 1970. The Specificity and Mechanism of Pepsin Action, in: Nord, F.F. (Ed.), Advances in Enzymology - and Related Areas of Molecular Biology. John Wiley & Sons, Inc., Hoboken, NJ, USA, pp. 401–443. DOI:10.1002/9780470122785.ch9.

Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. D., & Bairoch, A. (2003). ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Research, 31*, 3784–3788. https://doi.org/10.1093/nar/gkg563.

Gekko, Kunihiko, Kimoto, Akinobu, & Kamiyama, Tadashi (2003). Effects of disulfide bonds on compactness of protein molecules revealed by volume, compressibility, and expansibility changes during reduction. *Biochemistry, 42*(46), 13746–13753. https://doi.org/10.1021/bi030115q.

Hamuro, Yoshitomo, Coales, Stephen J., Molnar, Kathleen S., Tuske, Steven J., & Morrow, Jeffrey A. (2008). Specificity of immobilized porcine pepsin in H/D exchange compatible conditions. *Rapid Commun. Mass Spectrom. RCM, 22*(7), 1041–1046. https://doi.org/10.1002/(ISSN)1097-023110.1002/rcm.v22:710.1002/rcm.3467.

Humphrey, William, Dalke, Andrew, & Schulten, Klaus (1996). VMD: Visual molecular dynamics. *Journal of Molecular Graphics, 14*(1), 33–38. https://doi.org/10.1016/0263-7855(96)00018-5.

Husson, F., Lê, S., 2009. Francois Husson and Sébastien Lê (2009). SensoMineR: Sensory data analysis with R. R package version 1.10. http://CRAN.R-project.org/package=SensoMineR.

Jiménez-Saiz, R., Martos, G., Carrillo, W., López-Fandiño, R., & Molina, E. (2011). Susceptibility of lysozyme to in-vitro digestion and immunoreactivity of its digests. *Food Chemistry, 127*(4), 1719–1726. https://doi.org/10.1016/j.foodchem.2011.02.047.

Langella, Olivier, Valot, Benoît, Balliau, Thierry, Blein-Nicolas, Mélisande, Bonhomme, Ludovic, & Zivy, Michel (2017). X!TandemPipeline: A Tool to Manage Sequence Redundancy for Protein Inference and Phosphosite Identification. *Journal of Proteome Research, 16*(2), 494–503. https://doi.org/10.1021/acs.jproteome.6b00632.

Li, H., Pan, Y., Yang, Z., Rao, J., & Chen, B. (2021). Emerging applications of site-directed spin labeling electron paramagnetic resonance (SDSL-EPR) to study food protein structure, dynamics, and interaction. *Trends in Food Science & Technology, 109*, 37–50. https://doi.org/10.1016/j.tifs.2021.01.022.

Li, Qian, Zhao, Di, Liu, Hui, Zhang, Miao, Jiang, Shuai, Xu, Xinglian, … Li, Chunbao (2020). "Rigid" structure is a key determinant for the low digestibility of myoglobin. *Food Chem. X, 7*, 100094. https://doi.org/10.1016/j.fochx.2020.100094.

Luo, Q., Chen, D., Boom, R. M., & Janssen, A. E. M. (2018). Revisiting the enzymatic kinetics of pepsin using isothermal titration calorimetry. *Food Chemistry, 268*, 94–100. https://doi.org/10.1016/j.foodchem.2018.06.042.

Minekus, M., Alminger, M., Alvito, P., Ballance, S., Bohn, T., Bourlieu, C., … Brodkorb, A. (2014). A standardised static in vitro digestion method suitable for food – an international consensus. *Food & Function, 5*(6), 1113–1124. https://doi.org/10.1039/C3FO60702J.

Nau, Françoise, Nyemb-Diop, Kéra, Lechevalier, Valérie, Floury, Juliane, Serrière, Chloé, Stroebinger, Natascha, … Rutherfurd, Shane M. (2019). Spatial-temporal changes in pH, structure and rheology of the gastric chyme in pigs as influenced by egg white gel properties. *Food Chemistry, 280*, 210–220. https://doi.org/10.1016/j.foodchem.2018.12.042.

Niu, Canfang, Luo, Huiying, Shi, Pengjun, Huang, Huoqing, Wang, Yaru, Yang, Peilong, … Kelly, R. M. (2016). N-Glycosylation Improves the Pepsin Resistance of Histidine Acid Phosphatase Phytases by Enhancing Their Stability at Acidic pHs and Reducing Pepsin's Accessibility to Its Cleavage Sites. *Applied and Environment Microbiology, 82*(4), 1004–1014. https://doi.org/10.1128/AEM.02881-15.

Norwood, E.-A., Le Floch-Fouéré, C., Briard-Bion, V., Schuck, P., Croguennec, T., & Jeantet, R. (2016). Structural markers of the evolution of whey protein isolate powder during aging and effects on foaming properties. *Journal of Dairy Science, 99*(7), 5265–5272. https://doi.org/10.3168/jds.2015-10788.

Nyemb, K., Jardin, J., Causeur, D., Guérin-Dubiard, C., Dupont, D., Rutherfurd, S. M., & Nau, F. (2014). Investigating the impact of ovalbumin aggregate morphology on in vitro ovalbumin digestion using label-free quantitative peptidomics and multivariate data analysis. *Food Research International, 63*, 192–202. https://doi.org/10.1016/j.foodres.2014.03.041.

Nyemb-Diop, Kéra, Causeur, David, Jardin, Julien, Briard-Bion, Valérie, Guérin-Dubiard, Catherine, Rutherfurd, Shane M., … Nau, Françoise (2016). Investigating the impact of egg white gel structure on peptide kinetics profile during in vitro digestion. *Food Research International, 88*, 302–309. https://doi.org/10.1016/j.foodres.2016.01.004.

Olsson, Mats H. M., Søndergaard, Chresten R., Rostkowski, Michal, & Jensen, Jan H. (2011). PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *Journal of Chemical Theory and Computation, 7*(2), 525–537. https://doi.org/10.1021/ct100578z.

Opdenakker, Ghislain, Rudd, Pauline M., Ponting, Christopher P., & Dwek, Raymond A. (1993). Concepts and principles of glycobiology. *FASEB Journal, 7*(14), 1330–1337. https://doi.org/10.1096/fsb2.v7.1410.1096/fasebj.7.14.8224606.

Osorio, D., Rondón-Villarreal, P., Torres, R., 2015. Peptides: A Package for Data Mining of Antimicrobial Peptides. R J. 7, 4. DOI;10.32614/RJ-2015-001.

Polverino de Laureto, Patrizia, Frare, Erica, Gottardo, Rossella, van Dael, Herman, & Fontana, Angelo (2002). Partly folded states of members of the lysozyme/lactalbumin superfamily: A comparative study by circular dichroism spectroscopy and limited proteolysis. *Protein Science, 11*(12), 2932–2946.

Powers, J. C., Harley, A. D., & Myers, D. V. (1977). In *Subsite Specificity of Porcine Pepsin* (pp. 141–157). Boston, MA: Springer.

R Core Development Team, 2020. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Reeb, Jonas, & Rost, Burkhard (2019). Secondary Structure Prediction. In *Encyclopedia of Bioinformatics and Computational Biology* (pp. 488–496). Elsevier. https://doi.org/10.1016/B978-0-12-809633-8.20267-7.

Sepulveda, P., Marciniszyn, J., Liu, D., & Tang, J. (1975). Primary structure of porcine pepsin. III. Amino acid sequence of a cyanogen bromide fragment, CB2A, and the complete structure of porcine pepsin. *Journal of Biological Chemistry, 250*(13), 5082–5088.

Søndergaard, Chresten R., Olsson, Mats H. M., Rostkowski, Michał, & Jensen, Jan H. (2011). Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pKa Values. *Journal of Chemical Theory and Computation, 7*(7), 2284–2295. https://doi.org/10.1021/ct200133y.

Takagi, Kayoko, Teshima, Reiko, Okunuki, Haruyo, & Sawada, Jun-ichi (2003). Comparative Study of *in Vitro* Digestibility of Food Proteins and Effect of Preheating on the Digestion. *Biological &/and Pharmaceutical Bulletin, 26*(7), 969–973. https://doi.org/10.1248/bpb.26.969.

Timmer, John C, Zhu, Wenhong, Pop, Cristina, Regan, Tim, Snipas, Scott J, Eroshkin, Alexey M, … Salvesen, Guy S (2009). Structural and kinetic determinants of protease substrates. *Nature Structural & Molecular Biology, 16*(10), 1101–1108. https://doi.org/10.1038/nsmb.1668.

Tonda, Alberto, Grosvenor, Anita, Clerens, Stefan, & Le Feunteun, Steven (2017). In silico modeling of protein hydrolysis by endoproteases: A case study on pepsin digestion of bovine lactoferrin. *Food & Function, 8*(12), 4404–4413. https://doi.org/10.1039/C7FO00830A.

Torcello-Gomez, A., Dupont, D., Jardin, J., Briard-Bion, V., Deglaire, A., Risse, K., Mechoulan, E., Mackie, A., 2020. Human gastrointestinal conditions affect in vitro digestibility of peanut and bread proteins. Food Funct. 10.1039.D0FO01451F. DOI: 10.1039/D0FO01451F.

UniProt Consortium, T., 2018. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 46, 2699. https://doi.org/10.1093/nar/gky092.