Journal of proteome research
Article
pubs.acs.org/jpr

EnzymePredictor: A Tool for Predicting and Visualizing Enzymatic Cleavages of Digested Proteins

Vaishnavi Vijayakumar,[†] Andrés N. Guerrero,[§] Norman Davey,[∥] Carlito B. Lebrilla,[§] Denis C. Shields,[†] and Nora Khaldi*,[†,‡]

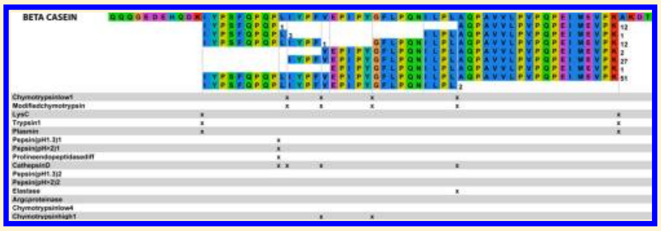
[†]UCD Conway Institute of Bio molecular and Biomedical Research, School of Medicine and Medical Sciences, and UCD Complex and Adaptive Systems Laboratory, University College Dublin, Dublin, Republic of Ireland

[‡]Department of Food Science and Technology, University of California, Davis, California 95616, United States

[§]Department of Chemistry, University of California—Davis, One Shields Avenue, Davis, California 95616, United States

[∥]Structural and Computational Biology Unit, European Molecular Biology Laboratory (EMBL), Meyerhofstrasse 1, D-69117 Heidelberg, Germany

Supporting Information



ABSTRACT: Mass spectrometric analysis of peptides contained in enzymatically digested hydrolysates of proteins is increasingly being used to characterize potentially bioactive or otherwise interesting hydrolysates. However, when preparations containing mixtures of enzymes are used, from either biological or experimental sources, it is unclear which of these enzymes have been most important in hydrolyzing the sample. We have developed a tool to rapidly evaluate the evidence for which enzymes are most likely to have cleaved the sample. EnzymePredictor, a web-based software, has been developed to (i) identify the protein sources of fragments found in the hydrolysates and map them back on it, (ii) identify enzymes that could yield such cleavages, and (iii) generate a colored visualization of the hydrolysate, the source proteins, the fragments, and the predicted enzymes. It tabulates the enzymes ranked according to their cleavage counts. The provision of odds ratio and standard error in the table permits users to evaluate how distinctively particular enzymes may be favored over other enzymes as the most likely cleavers of the samples. Finally, the method displays the cleavage not only according to peptides, but also according to proteins, permitting evaluation of whether the cleavage pattern is general across all proteins, or specific to a subset. We illustrate the application of this method using milk hydrolysates, and show how it can rapidly identify the enzymes or enzyme combinations used in generating the peptides. The approach developed here will accelerate the identification of enzymes most likely to have been used in hydrolyzing a set of mass spectrometrically identified peptides derived from proteins. This has utility not only in understanding the results of mass spectrometry experiments, but also in choosing enzymes likely to yield similar cleavage patterns. EnzymePredictor can be found at http://bioware.ucd.ie/∼enzpred/Enzpred.php

KEYWORDS: *hydrolysate, protein digestion, enzyme cleavage, peptides, fermentation, mass spectrometry, mass spectrometry visualization*

■ INTRODUCTION

Food fermentation is the use of certain bacterial organisms (Food grade ones, such as *Lactobacillus sp.*) to digest and degrade the food elements, such as cleaving the food proteins with the protease produced by many bacteria. It is one of the oldest ways of preservation that ensures required levels of quality from the initial time of manufacturing right until consumption.[1] Fermented products have a preservative effect, enabled by limiting the growth of spoilage in the food product. The utilization and harvesting of fermentation by humans is thought to date back approximately 8000 years.[1] The reality is most likely that human consumption of fermented food preceded this date, but was not cultured per se. Research has shown that fermentation can inhibit pathogenic bacteria that otherwise could cause disorders. Toxins and antinutritive factors can also be reduced. Besides, the nutritive value can be enriched as a result of fermentation. For example, the fermentation of soy not only makes the end product more digestible, it can also improve flavor and texture, appearance and aroma, synthesize vitamins, destroy undesirable flavors, reduce carbohydrates, decrease cooking time, and transform what might otherwise be agricultural wastes into tasty and nutritious human food. In this work, we focus on pure enzymatic degradation of food proteins, but our work can also extend to bacterial fermentation. The end products of both bacterial and enzymatic digestion of a food sources are referred to as “hydrolysates”.

It is only in the past few years that the technology has developed capabilities to efficiently analyze the products and

Received: August 3, 2012
Published: October 25, 2012

ACS Publications © 2012 American Chemical Society

Table 1. Enzyme Names Classified by Alphabetical Order Used in EnzymePredictor, the Known Cleavage Pattern, and Their Corresponding References

	cleavage pattern						
enzyme	P4	P3	P2	P1	P1′	P2′	ref
Arg-C proteinase	-	-	-	R	-	-	20
Asp-N endopeptidase[a]	-	-	-	-	D	-	20
Caspase 1	F, W, Y, or L	-	H, A, or T	D	not P, E, D, Q, K, or R	-	21
Caspase 2	D	V	A	D	not P, E, D, Q, K, or R	-	21
Caspase 3	D	M	Q	D	not P, E, D, Q, K, or R	-	21
Caspase 4	L	E	V	D	not P, E, D, Q, K, or R	-	21
Caspase 5	L or W	E	H	D	-	-	21
Caspase 6	V	E	H or I	D	not P, E, D, Q, K, or R	-	21
Caspase 7	D	E	V	D	not P, E, D, Q, K, or R	-	21
Caspase 8	I or L	E	T	D	not P, E, D, Q, K, or R	-	21
Caspase 9	L	E	H	D	-	-	21
Caspase 10	I	E	A	D	-	-	21
Chymotrypsin high specificity (C-term to [FYW], not before P)	-	-	-	F or Y	not P	-	20
	-	-	-	W	not M or P	-	20
Chymotrypsin low specificity (C-term to [FYWML], not before P)	-	-	-	F, L, or Y	not P	-	20
	-	-	-	W	not M or P	-	20
	-	-	-	M	not P or Y	-	20
	-	-	-	H	not D, M, P, or W	-	20
Enterokinase	D or N	D or N	D or N	K	-	-	22
Factor Xa	A, F, G, I, L, T, V, or M	D or E	G	R	-	-	23
Formic acid	-	-	-	D	-	-	24
Glutamyl endopeptidase	-	-	-	E	-	-	25
GranzymeB	I	E	P	D	-	-	21,26
Hydroxylamine	-	-	-	N	G	-	27
Iodosobenzoic acid	-	-	-	W	-	-	28
Lys-C	-	-	-	K	-	-	20
NTCB (2-nitro-5-thiocyanobenzoic acid)[a]	-	-	-	-	C	-	29
Pepsin (pH1.3)	-	not H, K, or R	not P	not R	F, L, W, or Y	not P	20
	-	not H, K, or R	not P	F, L, W, or Y	-	not P	20
Pepsin (pH > 2)	-	not H, K, or R	not P	not R	F or L	not P	20
	-	not H, K, or R	not P	F or L	-	not P	20
Proline-endopeptidase	-	-	H, K, or R	P	not P	-	20,30
Staphylococcal peptidase I	-	-	not E	E	-	-	20
Thrombin	-	-	G	R	G	-	20
	A, F, G, I, L, T, V, or M	A, F, G, I, L, T, V, W, or A	P	R	not D or E	not D or E	20
	-	-	-	K or R	not P	-	20
Trypsin	-	-	W	K	P	-	20
	-	-	M	R	P	-	20
	-	-	C or D	K	D	-	20
	-	-	C	K	H or Y	-	20
	-	-	C	R	K	-	20
	-	-	R	R	H or R	-	20
Modified chymotrypsin	-	-	-	F, W, Y, or L	not P	-	Used by CRUX[10]
Elastase	-	-	-	A, V, I, L, G, or R	G, P, A, L, or F	-	Used by CRUX[10]
Cyanogen bromide	-	-	-	M	-	-	31
Proline endopeptidase diff	-	-	-	P	-	-	Used by CRUX[10]
Plasmin	-	-	-	K or R	-	-	32

Table 1. continued

	cleavage pattern						
enzyme	P4	P3	P2	P1	P1′	P2′	ref
Cathepsin D	-	-	-	A, V, L, I, P, M, F, or W	A, V, L, I, P, M, or F	-	32

[a] Enzymes represented in bold cleave at the C-terminus of residues in position P1′

benefits of the fermented foods. Mass spectrometry (MS) has helped researchers to have a deeper and more detailed investigation of these hydrolysates. For many years, MS has been widely employed for the identification and characterization of proteins.[2] There is a broad diversity in the objectives for which MS analysis has been applied to studies on food proteins.[3] These include the detection and characterization of biomolecules in complex mixtures (such as milk), structural information, or locating post-translational modifications. The sheer volume of data generated by MS warrants the need for sophisticated data handling systems prediction and visualization.[4] The objective of computationally predicting the most likely cleavage proteases is not only relevant to mass spectrometrically identified fragments from hydrolysates, but also to peptides from data sets of peptides obtained in a more targeted way from a protein preparation, such as approaches that focus on peptides with free N-termini,[5] a subset of which reflect cleavage by particular proteases.

In this work, we focus on the peptides produced in a hydrolysate. We have constructed “EnzymePredictor”, a software that allows the visualization of the hydrolysate and the prediction of sets of enzymes that have been used or can be used to generate the current hydrolysate’s identified peptides. The predictions are based on a set of known patterns of 35 characterized enzymes (Table 1). The software has a friendly visualization output that allows the user to quickly view the peptides that have been detected by MS in the hydrolysate. The user can simultaneously view the positional information of the peptides on their source protein and the possible enzymes that have been used to produce them.

■ MATERIALS AND METHODS

1. Input

1.1. Experimental Hydrolysate Generation.

1.1.1. Chemicals and Sample Set. C18 columns were purchased from Supelco, while the enzymes trypsin and α-chymotrypsin were obtained from Promega and Sigma, respectively. All reagents and solvents used were either of analytical grade or better. Human milk samples from four different mothers (obtained from an ongoing project in our research group) at the third month of lactation were combined in a pooled sample.

1.1.2. Sample Preparation. 0.1 mL of the human milk pool was mixed with 0.4 mL of ammonium bicarbonate 50 mM (pH ≈ 8.2) and centrifugated at 4 °C and 3000 rpm for 30 min. The top layer consisting of fat was carefully removed and discarded. The rest of the sample was resuspended, and three fractions of 50 μL were collected and heated up to 80 °C for 10 min. Once the fractions cooled, trypsin, chymotrypsin or both enzymes (in all cases 1 μg of total enzyme) were added. Incubation was allowed to proceed during 8 h at 37 °C with agitation and was stopped by heating at 80 °C for 15 min. The digestions were centrifugated at 15 000 rpm for 30 min in order to separate the insoluble fraction composed of undigested protein complexes and cellular residues. The peptidic content of the supernatants was purified via solid phase extraction in C18 columns. Peptides were eluted from the cartridge with a 60% ACN, 0.1% TFA solution. Finally, the samples were completely dried in the speed vac and reconstituted in 25 μL of nanopure water prior to the mass analysis.

1.1.3. Mass Spectrometry Analysis. The peptidic digestion was analyzed using an Agilent 1200 series LC system coupled to an Agilent 6520 Q-TOF mass spectrometer. The instrument has been previously described in detail.[6] The tandem mass spectra of the peptides were acquired in a data-dependent manner following LC separation on the microfluidic chip. A C-18 chip was used. Both pumps used binary solvent: A, 3.0% ACN/water (v/v) with 0.1% formic acid; B, 90% ACN/water (v/v) with 0.1% formic acid. A flow rate of 4 μL/min of solvent A was used for sample loading with 2 μL injection volume. The drying gas temperature was set at 325 °C with a flow rate of 4 L/min of filtered dry grade compressed air. MS and MS/MS spectra were acquired in the positive ionization mode with an acquisition rate of 0.63 spectra per second. MS data were acquired over a mass range of 300−3000 m/z, while MS/MS data were acquired over 100−3000 m/z mass range. Mass calibration was enabled using reference masses.

For the MS/MS analysis, peptides were subjected to collision induced disociation with nitrogen as the collision gas and using a collision energy that was dependent on the relation mass to charge of the different signals detected according to the equation $V_{\text{collision}} = m/z$ (3.8/100 Da) Volts −4.2 V.

1.1.4. Data Analysis. The tandem-MS data was extracted from the chromatogram using the MassHunter software (Agilent Technologies Inc.). Peptide identification was accomplished using the database searcher Mass Spectral-Generating Function Database (MS-GFDB)[7] against a human milk protein library. The human milk library was constructed based on a query to the UniProt database. The query returned only proteins from *Homo sapiens* and at least one of the following: “tissue specificity” keyword “milk” or “mammary”, “tissue” keyword “milk” or “mammary” or gene ontology “lactation”. This query returned a list of 1472 proteins. For the database search, masses were allowed with a 40 ppm error. No complete modifications were included, but up to four potential modifications were allowed on each peptide. Potential modifications allowed were phosphorylation of serine, threonine or tyrosine and oxidation of methionine. A nonspecific cleavage ([X]|[X]) (where ‘X’ is any amino acid) was used to search against the protein sequences. The fragmentation method selected in the search was CID, and the instrument selected was TOF. Peptides generated in the output were accepted if their p-values were less than or equal to 0.01 corresponding to confidence levels of 99%.

1.2. Software Input File Format. The input file should be a tab-delimited text file with header line that contains a minimum of two columns. The first column should contain the UniProt accession number or the entry name of protein, while the second column should contain the peptide sequence (see Supporting Information Table S1 for example). Other columns may also be placed in the table, but these will not be taken into account. An initial input file is generated using a database searcher such as X!tandem, MS-GDF (as in this study), or Mascot, from the mass spectrometry reading of the hydrolysate. In some cases the UniProt accession name (or entry name) and the peptide sequence can be provided by these database searchers, but might still need to be rearranged as mentioned above to suit the software input format (into column 1 and 2, respectively). In some cases, the searcher provides a long UniProt identification for the protein source such as "sp|PO5814|CASB_BOVIN", in this case the UniProt accession name needs to be extracted and provided in the fist column as "PO5814".

2. Steps Performed in EnzymePredictor

Figure 1 represents a schematic overview of the methods section and how EnzymePredictor operates. Three main steps are performed with the input files.

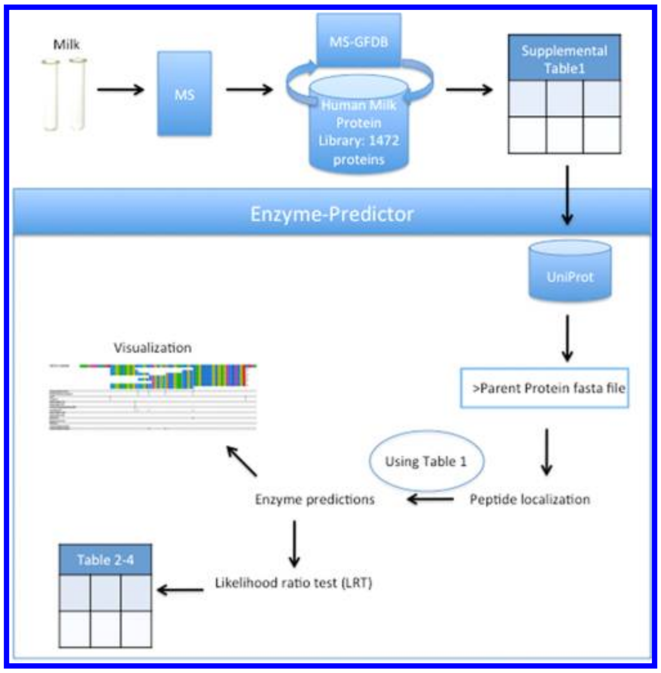


Figure 1. Schematic overview of the main steps performed by EnzymePredictor. The initial part of the figure contains the main experimental steps to generate the tables. Table numbering corresponds to the tables in the manuscript.

2.1. Data. We use the UniProt accession number or the entry name of protein provided in the first column of the input file (Supporting Information Table S1) to retrieve the corresponding FASTA sequence of the protein source (the protein containing the peptides prior fermentation/digestion). We extract the protein from the UniProt database.[8] The peptides are then mapped onto their source protein to obtain their localization and neighboring amino acids on the protein.

2.2. Enzymes Search. *2.2.1. Predicting the Enzymes Used to Generate the Hydrolysate.* We define for each peptide a pattern consisting of the amino acids that are positioned on both sides of the peptide. Four amino acids upstream and 2 downstream of the N and C terminal cleavage sites were used to build a pattern consisting of 6 amino acids. The four upstream positions correspond to the positions P4, P3, P2 and P1 (in this order reading N-to-C terminal). The positions downstream of the N-terminal are represented as P1′and P2′. The cleavage patterns we adopted for each enzyme are presented in Table 1. We currently have cleavage patterns for 35 known enzymes (Table 1), obtained from PeptideCutter[9] and Crux.[10]

2.2.2. Predicting the Enzymes That Could Cleave within the Current Peptides Found in the Hydrolysate. We distinguished between enzymes that have been used to create the hydrolysate and enzymes that could be used but most likely will not give the same overall result. To investigate this, we used the exact same approach described above to examine the number of times the predicted enzymes have been found to cleave within the sequence of the current peptides.

2.3. Displaying the Results. In order to decipher between the possible enzyme(s) that were used to obtain this hydrolysate, we tabulate the following for each enzyme: (1) the compiled total number of times each enzyme has cleaved at the termini of each peptide, (2) the total number of times the enzyme has uniquely cleaved a residue, (3) the total number of times each enzyme can possibly cleave within the identified peptides, (4) the total number of proteins that have been cleaved at peptide termini by this enzyme's pattern, and (5) the calculated odds ratio for each enzyme's tendency to cut at termini rather than the interior of peptides, along with its standard error.

In order to get a visual display of the hydrolysate, we find and collect all the information regarding the peptide's localization in the parent protein (Figure 2), the number of times each peptide was detected by MS (Figure 2), and finally the enzymes for each peptide N- and C-terminus (from the above; Figure 3).

3. Output

3.1. Tabulated Outputs. The user can choose to download a document that contains a number of tables describing the hydrolysate. The first table describes the overall picture of all the enzymes that possibly cleave the hydrolysate (Tables 2–4). All enzymes that have cleaved at least once are represented in this table (Tables 2–4). Enzymes are ranked according to their total number of peptide cleavages, from the highest to the lowest number of cleavages.

The second table that is provided by the software contains more specific information. This table provides the cleavage details of each peptide present in the hydrolysate (Supporting Information Table S2).

3.2. Visualization. EnzymePredictor generates a PDF file for each source protein detected in the hydrolysate. The source protein is displayed at the top of the PDF (Figure 3). The amino acids are shown in different colors representing their biochemical properties. Peptides that are detected by MS are represented under the source protein. The number of times a peptide has been found in the hydrolysate is displayed at their right (Figure 3). Vertical gray lines connect the peptide cleavage to the enzyme(s) that could cleave this pattern. The list of enzymes predicted to cleave the protein is placed below the peptides (Figure 3; Table 2).

A food hydrolysate may contain very small or very large numbers of peptides. The visualization is coded to take this into account and size the image according to (1) the number of peptides, (2) amino acids of the source protein, and (3) the number of predicted enzymes.

3.3. Ranking the Enzymes. We wanted to rank the enzymes on the basis of the enrichment of their cleavage sites at termini over the interior of a series of identified peptides. To do

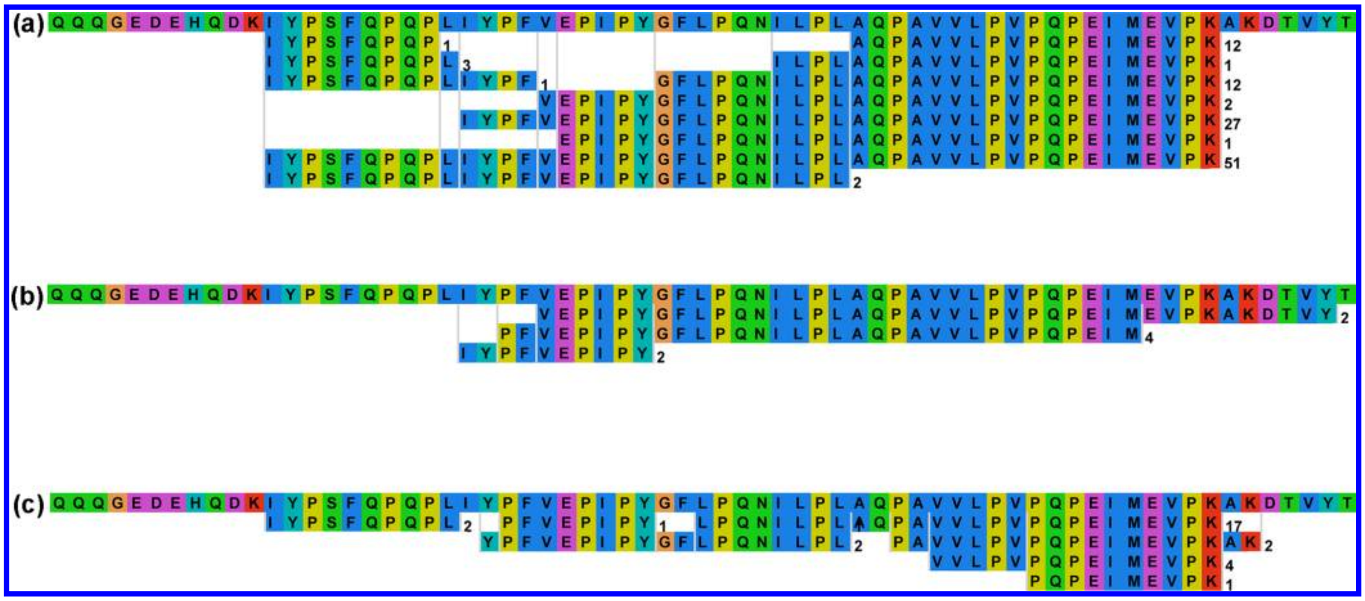

(a)
QQQGEDEHQDKIYPSFQPQPLIYPFVEPIPYGFLPQNILPLAQPAVVLPVPQPEIMEVPKAKDTVYT
12
1
12
2
27
1
51
2
(b)
QQQGEDEHQDKIYPSFQPQPLIYPFVEPIPYGFLPQNILPLAQPAVVLPVPQPEIMEVPKAKDTVYT
2
4
2
(c)
QQQGEDEHQDKIYPSFQPQPLIYPFVEPIPYGFLPQNILPLAQPAVVLPVPQPEIMEVPKAKDTVYT
2
1
17
2
2
4
1

Figure 2. Part of the visualization output for human breast milk β-casein resulting from trypsin, chymotrypsin, and the combination of both trypsin and chymotrypsin digestions. β-Casein resulting from (a) trypsin digestion; (b) chymotrypsin digestion; and (c) the combination of both trypsin and chymotrypsin. In each panel the top sequence represents part of the source protein (β-casein). The sequences under the source protein represent the peptides detected in the hydrolysate by mass spectrometry. The amino acid residues are colored depending on their biochemical properties. The colors we used are red, blue, green, cyan, yellow, orange, pink and magenta.

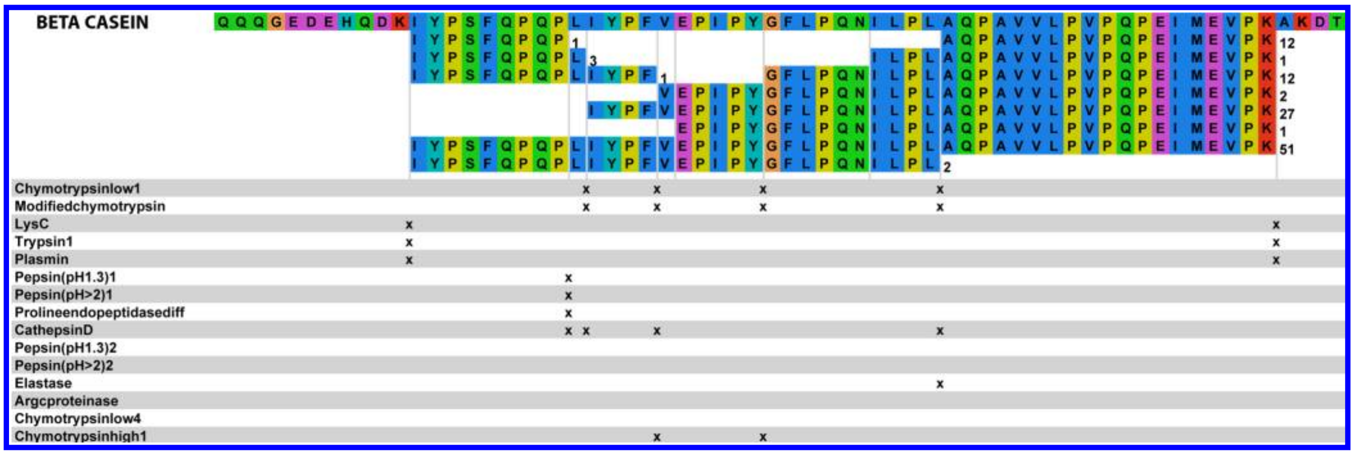


Figure 3. Part of the visualization output for human breast milk β-casein resulting from trypsin digestion and the corresponding predicted enzymes. The entry name of the protein, or its UniProt accession number, is displayed at the left side of the sequence. This allows the users to easily know which protein sequence is represented. Gray vertical lines show the enzymatic cleavage sites representing the starting and ending position of each peptide. Each layer below the peptide visualization corresponds to an enzyme. The presence of "x" indicates that the enzyme can cleave at this position. The numbers represented after each peptide indicate the number of times the peptide has been detected by MS in the hydrolysate.

this we used the information collected by EnzymePredictor. The odds ratio (OR) is calculated for each enzyme (A) as

$$\mathrm{OR} = \frac{a \times d}{b \times c}$$

where a is the total number of sites cleaved by enzyme "A" at the peptide termini (column 4 entitled "total cleavage" in Tables 2−4); b is the total number of all sites cleaved by any enzyme except "A"; c is the total number of sites enzyme "A" could have cleaved within the current peptides (column 6 entitled "number of expected cleavages within the peptide" in Tables 2−4); and finally d is the number of sites all enzymes, aside from A, could have cleaved within the current peptides. We also calculated the standard error (s.e.), as follows:

$$\text{s.e.} = \exp\left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}\right)^{\frac{1}{2}}$$

When one or more of the four groups is zero, 0.5 was added to each cell[11] (see Tables 2−4).

Comparison of two enzymes is facilitated by comparing their OR values. If the OR values differ by more than the sum of their two standard errors, this provides an indication that this difference is more likely to be significant. If the two OR values for two enzymes are not clearly distinguished (standard error ranges overlap) then this is a good sign that the number of peptides sampled in the analysis is too small to distinguish between the patterns inferred to be associated with the two enzymes under comparison. Bear in mind that these comparisons make assumptions regarding the true hydrolysis

Table 2. Predicted Enzymes and Counts for Trypsin Digestion of Human Breast Milk

enzymes	N-terminus cleavage count	C-terminus cleavage count	total cleavage	unique cleavage	number of expected cleavages within the peptide	number of proteins cleaved	odds ratio	std error
Plasmin	132	153	285	0	15	28	67.857143	1.3074944
Trypsin1	130	150	280	0	5	28	199.41645	1.5731303
Lys-C	73	86	159	0	10	26	48.699429	1.3886571
Arg-C proteinase	59	67	126	0	5	25	74.517621	1.5802984
Elastase	20	22	42	0	254	20	0.4060516	1.1862273
Pepsin (pH 1.3) 1	19	13	32	0	290	11	0.264299	1.2092908
Pepsin (pH > 2) 1	16	11	27	0	205	10	0.3250176	1.2313509
Asp-N endopeptidase	9	7	16	0	68	11	0.6060326	1.3235303
Modified chymotrypsin	9	3	12	0	321	9	0.0866544	1.3449995
Chymotrypsin low1	9	3	12	0	292	9	0.0964266	1.3456895
Cathepsin D	7	1	8	0	378	5	0.0476912	1.4322391
NTCB	0	6	6	0	0	5	34.007292	4.340791
Chymotrypsin low4	2	3	5	0	25	4	0.5179786	1.634392
Pepsin (pH > 2) 2	4	1	5	0	189	5	0.0642993	1.5755456
Pepsin (pH 1.3) 2	4	1	5	0	277	5	0.0423284	1.5726549
Chymotrypsin high1	3	1	4	0	154	4	0.0639516	1.6616281
Cyanogenbromide	3	0	3	0	23	2	0.3374099	1.8495536
Proline endopeptidase diff	2	0	2	0	161	2	0.030442	2.0389672
Chymotrypsin low3	2	0	2	0	18	1	0.287683	2.1090899
Thrombin2	0	1	1	1	0	1	7.8098694	5.1212737
FactorXa	1	0	1	1	0	1	7.8098694	5.1212737
Thrombin1	1	0	1	1	0	1	7.8098694	5.1212737

patterns of the enzymes, which are not always true, and may be particularly dependent on the lysis conditions as well as peculiarities of the proteins being digested. Thus, the comparison is valid, so long as the user bears in mind the various assumptions that are made.

The user is required to take two factors into account when interpreting the output: first, which enzymes account for more potential terminal cleavages, and to what extent do those enzymes show a strong enrichment for terminal over internal cleavages. The enzymes were ranked according to their total number of cleavages, since an enzyme that is key in the hydrolysis of a sample would present a high number of cleavages. The OR is then used to distinguish between the enzymes with the highest cleavage totals that most likely have been important in hydrolyzing the sample. An enzyme with a combined high total of cleavages and a high OR is to be considered as a leading candidate in the creation of the hydrolysate.

■ RESULTS AND DISCUSSION

1. Enzyme Prediction

To assess the software's performance, we experimentally generated a hydrolysate. We used human breast milk as the raw material, and performed three independent digestions (see Materials and Methods). We chose to use human milk as opposed to a "clean" protein digestion because of its complex nature, which represents the reality of a food hydrolysate. We carried out a digestion with trypsin, one with chymotrypsin, and a digestion using the combination of both trypsin and chymotrypsin (see Materials and Methods). The latter digestion is to examine if the software is sensitive to the usage of a combination of enzymes. The resulting hydrolysates were then passed through mass spectrometry and the MassHunter software (Agilent Technologies Inc.) to yield the list of peptides (see Materials and Methods).

These results are represented in Table 2 for trypsin digestion. Table 3 represents the results for chymotrypsin, and Table 4 for an experiment in which there was digestion of the samples with a combination of both trypsin and chymotrypsin. As explained in methods, the most likely enzymes to have contributed to the shaping of the hydrolysate are the ones with a combined highest number of cleavage sites and highest OR.

2. Digestion of Human Milk Using One Enzyme

2.1. Trypsin. EnzymePredictor correctly predicts trypsin and places this enzyme at the top of all 22 predicted enzymes. Indeed, trypsin has the highest number of cleavages, and the highest OR (Table 2).

The OR values indicate that certain enzymes are over-represented and others under-represented. Given the similarities of many cleavage patterns for many enzymes, it is not surprising that the digestion patterns for multiple enzymes are enriched at peptide termini. However, the software clearly distinguishes the first enzyme from all others with this particular sample. Interpretation of results should take into account the possibility of hydrolysis of the sample by enzymes other than those added experimentally. Milk is a complex mixture; it contains many elements including enzymes.[12] Certain enzymes are transferred from blood into milk, such as plasmin.[13] The presence of plasmin as a likely contributor to the creation of this hydrolysate is not surprising, given that this enzyme is a blood enzyme that is known to make its way into milk. Our software however also shows other predicted enzymes with a similar likelihood (based on OR and total number of cleavages), and these are lysC and Arg-c (Table 2). It is important to note that both enzymes overlap with plasmin specificity (Table 1). The milk enzyme plasmin cleaves at lysine

Table 3. Predicted Enzymes and Counts for Chymotrypsin Digestion of Human Breast Milk

enzymes	N-terminus cleavage count	C-terminus cleavage count	total cleavage	unique cleavage	number of expected cleavages within the peptide	number of proteins cleaved	odds ratio	std error
Modified chymotrypsin	31	37	68	0	51	22	3.2091097	1.2174368
Chymotrypsin low1	31	30	61	0	50	22	2.8775684	1.2235983
Pepsin (pH 1.3) 2	17	30	47	0	41	18	2.6235512	1.2501768
Chymotrypsin high1	17	21	38	0	14	17	6.2613636	1.3758716
Cathepsin D	17	14	31	0	176	19	0.3189099	1.2282394
Pepsin (pH > 2) 2	10	11	21	0	35	15	1.2861789	1.3272827
Elastase	11	7	18	0	113	14	0.3053097	1.2991589
Plasmin	11	4	15	0	32	12	0.9925	1.3759276
Trypsin1	11	4	15	0	31	12	1.0258065	1.3780937
Lys-C	7	2	9	0	19	7	1.0033154	1.5059106
Pepsin (pH 1.3) 1	5	4	9	0	47	9	0.3915229	1.4465523
Chymotrypsin low3	3	5	8	0	10	7	1.7089005	1.6134515
Cyanogen bromide	3	5	8	0	11	7	1.5516421	1.598042
Chymotrypsin high2	0	7	7	0	1	6	15.078329	2.9178048
Chymotrypsin low2	0	7	7	0	1	6	15.078329	2.9178048
Iodosobenzoic acid	0	7	7	0	1	6	15.078329	2.9178048
Pepsin (pH > 2) 1	5	2	7	0	39	7	0.3688157	1.5146531
Arg-C proteinase	4	2	6	0	13	6	0.9771635	1.6444685
Asp-N endopeptidase	2	3	5	0	41	3	0.2486538	1.6124701
Proline endopeptidase diff	2	0	2	0	97	2	0.0387395	2.048521
NTCB	0	1	1	1	3	1	0.7052271	3.1782752

or arginine at position P1, while lys-C cleaves at lysine at position P1, and Arg-c cleaves arginine at position P1. This overlap is likely to be the reason all three enzymes are highly ranked. But as explained above, it is only plasmin that is most likely the enzyme that has participated in the creation of this hydrolysate.

Other enzymes are also predicted by EnzymePredictor but are ranked very low, given the combined OR and total number of cleavages; in other words, they have most likely not contributed to the creation of the hydrolysate. The prediction of these enzymes by EnzymePredictor may reflect three scenarios. First, many of the enzymes in our database are nonspecific, compared to the more specific ones like trypsin. In other words, they share common patterns to other enzymes, such as the example above where Lys-C shares a similar pattern to that of plasmin. Second, because our database contains only 35 enzymes (Table 1) some of the enzymes in our data set may overlap in specificity with ones that we do not have so far. As our database will grow, this will be less of an issue. Finally, the cleavages may reflect off-target or nonenzymatic cleavage of the proteins.

2.2. Chymotrypsin. Similarly to trypsin, our software correctly predicts chymotrypsin as the main enzyme that has been used. Table 3 shows that three versions of chymotrypsin are highly ranked. The highest being chymotrypsin high affinity with the highest combined OR and total number of cleavages (Table 3). Other enzymes such as cathepsin D also have high number of cleavages but a low OR, which indicates a low likelihood that these enzymes were the main drivers behind the creation of the hydrolysate. Cathepsin D detection by EnzymePredictor comes as a result of this enzyme being a naturally occurring milk enzyme. Although it seems that this enzyme participates highly in the cleavage of many of the peptides, EnzymePredictor results show that it is not a key enzyme in the creation of this hydrolysate.

Interpretation can be made on the basis of various attributes, depending on the criteria the user wants to take into consideration, to sort the enzymes. A sample interpretation was performed for the hydrolysate containing the peptide fragments digested by chymotrypsin. A scatter plot graph was plotted against the total number of sites cleaved by an enzyme at the termini and its odds ratio (Figure 4). This highlights four enzymes that have very strong enrichment of terminal cleavage (and high OR). Of these, modified chymotrypsin accounts for many more of the total set of terminal cleavages, and is the leading candidate enzyme for cleaving this hydrolysate. This graph is relatively easy to interpret, and the highlighted enzyme is indeed the enzyme that was used experimentally (modified chymotrypsin in this case). However, in other cases the user needs to weigh up the evidence favoring one enzyme that has a high OR but accounts for relatively little terminal cleavage versus an enzyme with a lower OR but accounting for a greater proportion of terminal cleavage. Visualization such as that in Figure 4 can greatly assist in such interpretation.

Table 4. Predicted Enzymes and Counts for Trypsin and Chymotrypsin Digestion of Human Breast Milk

enzymes	N-terminus cleavage count	C-terminus cleavage count	total cleavage	unique cleavage	number of expected cleavages within the peptide	number of proteins cleaved	odds ratio	std error
Modified chymotrypsin	14	30	44	0	84	20	1.6968919	1.2169433
Chymotrypsin low1	13	27	40	0	81	19	1.5852144	1.2251235
Plasmin	24	12	36	0	24	19	5.0087719	1.3107984
Trypsin1	24	11	35	0	22	18	5.3061224	1.3217471
Pepsin (pH 1.3) 2	9	23	32	0	68	18	1.4933696	1.2498985
Cathepsin D	12	17	29	0	182	19	0.4492585	1.232969
Chymotrypsin high1	6	21	27	0	24	15	3.6602564	1.332562
Lys-C	17	6	23	0	10	13	7.4895775	1.4675798
Elastase	10	12	22	0	127	16	0.5055737	1.2699228
Arg-C proteinase	7	6	13	0	14	10	2.9307241	1.4767146
Pepsin (pH > 2) 2	3	9	12	0	60	10	0.6043716	1.3797828
Pepsin (pH 1.3) 1	5	6	11	0	72	10	0.4554193	1.3899913
Pepsin (pH > 2) 1	5	5	10	0	63	9	0.4757591	1.4129552
Chymotrypsin low3	4	4	8	0	9	8	2.7795796	1.6316134
Cyanogen bromide	4	4	8	0	10	8	2.4994595	1.6129868
Asp-N endopeptidase	3	3	6	0	50	4	0.36	1.5467835
Chymotrypsin high2	1	3	4	0	3	4	4.1461676	2.1513003
Chymotrypsin low2	1	3	4	0	3	4	4.1461676	2.1513003
Iodosobenzoic acid	1	3	4	0	3	4	4.1461676	2.1513003
Formic acid	2	1	3	0	51	3	0.174902	1.8168062
Chymotrypsin low4	2	0	2	0	10	2	0.6148936	2.1746516
NTCB	1	1	2	0	3	2	2.0620567	2.4962674
Glutamylendopeptidase	1	0	1	1	52	1	0.0568251	2.749245
StaphylococcalpeptidaseI	1	0	1	1	46	1	0.0645831	2.7526479
Proline endopeptidase diff	1	0	1	1	95	1	0.0299037	2.7374629

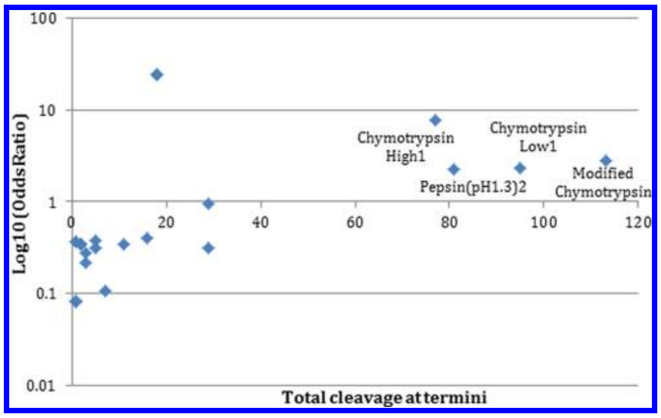


Figure 4. A logarithmic scatter plot graph plotted using odds ratio and the total sites cleaved by the enzyme at termini. The plot represents the odds ratio. The total number of sites cleaved at the termini by an enzyme is plotted in the *X*-axis with the odds ratio plotted in the *Y*-axis.

3. Digestion of Human Milk Using a Combination of Two Enzymes, Trypsin, and Chymotrypsin

EnzymePredictor correctly identifies both enzymes trypsin and chymotrypsin (Table 4). Indeed both these enzymes have the highest combined total number of cleavages and OR (Table 4). Although three specificities of chymotrypsin are represented as likely enzymes, the OR output highlights chymotrypsin-high-affinity as the main contributor. The prediction of modified chymotrypsin and chymotrypsin-low-affinity is most likely a result of them sharing a very similar cleavage pattern (see Table 1). The next most likely enzyme is pepsin, known to have broad cleavage specificity. The prediction of pepsin is most likely due to the overlap of specificity of this enzyme with both trypsin and chymotrypsin's specificity (Table 1).

The results above show that the software predicts the correct enzyme(s) for all three digestions (Tables 2–4). The software also identifies other enzymes for each digestion (Tables 2–4). Some such as plasmin, a blood enzyme, make it through into milk and act on cleaving the milk proteins. Relatively few instances of unique cutting sites for enzymes such as proline endopeptidase and thrombin were observed, which may reflect in part enzymatic activities present in the milk itself. Finally a reasonable explanation to their presence in milk is that the cleavages they are responsible for were mistakenly assigned to these enzymes by our software. Indeed our database of enzyme is currently limited to 35 enzymes, and some enzymes may share very similar patterns, or are very unspecific to a degree whereby they can share many cleavage patterns.

4. Visualization

Figure 3 represents the software's visualization for β-casein, a major milk protein. Many of the peptides are present in one or more copies. Some regions are densely represented, such as the region running from position 55 to 102 (the position count includes the signal peptide). These densely represented regions maybe important as they may carry a bioactivity.[14] The coloring of the different amino acids is a great way to visually quickly assess pockets of similar amino acids (similar biochemical properties), for example, pockets of charged residues.

Both Figures 2 and 3 show that mass spectrometry failed to identify the fragments in many regions of the proteins. For example we have no information for region 44–53 in β-casein (Figure 3). The poor coverage can be a problem, especially when researchers want to identify what exactly is occurring in these regions, or identify possible bioactive peptides.[4] Our software helps shed light on the possible cleaved peptides in this area. Indeed, knowing the enzymes that have been used to cleave the proteins will help the researchers to assess what peptides may occur in these obscure areas of the protein.[4]

Our database is currently restricted to 35 enzymes, and despite this small number, the software can assign many enzymes to each cleavage. This is interesting because it allows the researcher to replace an enzyme by another one that cleaves part of the protein that they choose, increasing or decreasing specificity as required. It is also interesting in the scenario where a bacterium was used to generate a hydrolysate. And although this hydrolysate might have some excellent aspects, the bacteria may have some negative effects (e.g., not be approved for food use). Replacing the bacterium by some enzymes (such as food grade enzymes) that show similar patterns to those of the bacteria is a great solution to generate a similar hydrolysate with no side effects, such as bacteria toxicity.

CONCLUSIONS

Our software successfully identifies the enzymes that have been used to generate the hydrolysate. This was the case for both single enzyme use (trypsin, chymotrypsin), or a when a combination of enzymes was used (Figures 2, 3, Tables 2–4).

Besides identifying the enzymes, our software "EnzymePredictor" provides a clear visualization of the MS output. The visualization is realized for each protein independently and shows the areas that MS identifies with the highest coverage. It helps the researchers rapidly assess pockets of peptides with similar biochemical properties, peptides densely represented, and gray areas that mass spectrometry fails to identify.[4]

Figure 3 shows that these gray areas represent an important part of the proteins. This maybe for a variety of reasons, one of which maybe post-translational modification such as phosphorylation or glycosylation.[15] But regardless of the reasons underlining the absence of information in those regions, knowing the enzymes that were used to generate the hydrolysate can shed light on what exactly is happening in these regions. By using our predicted enzyme(s), researchers can get the fragments (using PeptideCutter tool[9] or similar software) that cannot be seen by MS, hence filling the gaps especially when it comes to searching for bioactive peptides.[4]

The growing interest in generating and examining a hydrolysate comes as a result of their effects in an array of human health benefits. These human health benefits scan a diverse array of diseases such as diabetes,[16] inflammation,[17] and even cancer.[18] The effects of the hydrolysate come as a result of the food proteins releasing peptides via digestive enzymes.[4] It has been discussed that some of these regions that contain bioactive peptides may be under positive selection.[4,19] These peptides happen to carry a bioactivity with a beneficial functionality. However, identifying these active peptides in a hydrolysate has been a challenge. Even though the hydrolysate shows activity in a given assay, the large number of peptides released from enzymatic digestion in the mixture makes it hard to identify the bioactive(s) that is performing that function of interest.[4] By combining the possible enzymes that have cleaved the hydrolysate, and the densely represented regions of the proteins, this software may allow the user to pinpoint bioactive regions and peptides.

ASSOCIATED CONTENT

Supporting Information

Supporting Tables S1, S2. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: nora.khaldi@ucd.ie. Phone: +353-1-716-5335. Fax: +353-1-716-5396.

Author Contributions

VV coded the software with input from NK; NK designed the idea, wrote the specification with some input from DS, supervised, and wrote the paper with VV and DS. VV and NK carried out the analyses. AG and CL carried out the enzymatic digestion of human breast milk, the MS and the data analysis of MS through MS-GFDB. ND and VV coded the visualization.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank Niall Haslam and Fergal Duffy for helping with the bioware server. This work was funded by an Enterprise Ireland grant to Food Health Ireland and the Irish Research Council for Science, Engineering and Technology, cofunded by Marie Curie Actions under FP7, and SFI grant 08/IN.1/B1864.

REFERENCES

(1) Paul Ross, R.; Morgan, S.; Hill, C. Preservation and fermentation: past, present and future. *Int. J. Food Microbiol.* **2002**, *79* (1–2), 3–16.

(2) Cristoni, S.; Bernardi, L. R. Bioinformatics in mass spectrometry data analysis for proteomics studies. *Expert Rev. Proteomics* **2004**, *1* (4), 469–83.

(3) Leonil, J.; Gagnaire, V.; Molle, D.; Pezennec, S.; Bouhallab, S. Application of chromatography and mass spectrometry to the characterization of food proteins and derived peptides. *J. Chromatogr., A* **2000**, *881* (1–2), 1–21.

(4) Khaldi, N. Bioinformatics approaches for identifying new therapeutic bioactive peptides in food. *Funct. Foods Health Dis.* **2012**, *2* (10), 1–17.

(5) Prudova, A.; auf dem Keller, U.; Butler, G. S.; Overall, C. M. Multiplex N-terminome analysis of MMP-2 and MMP-9 substrate degradomes by iTRAQ-TAILS quantitative proteomics. *Mol. Cell. Proteomics* **2010**, *9* (5), 894–911.

(6) Nwosu, C. C.; Aldredge, D. L.; Lee, H.; Lerno, L. A.; Zivkovic, A. M.; German, J. B.; Lebrilla, C. B. Comparison of the human and bovine milk N-glycome via high-performance microfluidic chip liquid chromatography and tandem mass spectrometry. *J. Proteome Res.* **2012**, *11* (5), 2912–24.

(7) Kim, S.; Mischerikow, N.; Bandeira, N.; Navarro, J. D.; Wich, L.; Mohammed, S.; Heck, A. J. R.; Pevzner, P. A. The generating function of CID, ETD, and CID/ETD pairs of tandem mass spectra: applications to database search. *Mol. Cell. Proteomics* **2010**, *9* (12), 2840–52.

(8) Bairoch, A.; Apweiler, R.; Wu, C. H.; Barker, W. C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; Martin, M. J.; Natale, D. A.; O'Donovan, C.; Redaschi, N.; Yeh, L. S. The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **2005**, *33* (Database issue), D154–9.

(9) Gasteiger, E.; Hoogland, C.; Gattiker, A.; Duvaud, S.; Wilkins, M. R.; Appel, R. D.; Bairoch, A. Protein identification and analysis tools on the ExPASy server. *Methods Mol. Biol.* **1995**, *112*, 531–52.

(10) Park, C. Y.; Klammer, A. A.; Kall, L.; MacCoss, M. J.; Noble, W. S. Rapid and accurate peptide identification from tandem mass spectra. *J. Proteome Res.* **2008**, *7* (7), 3022−7.

(11) Liu, I. M.; Agresti, A. Mantel−Haenszel-type inference for cumulative odds ratios with a stratified ordinal response. *Biometrics* **1996**, *52* (4), 1223−34.

(12) Shahani, K. M.; Kwan, A. J.; Friend, B. A. Role and significance of enzymes in human milk. *Am. J. Clin. Nutr.* **1980**, *33*, 1861−8.

(13) Eigel, W. N.; Hofmann, C. J.; Chibber, B. A.; Tomich, J. M.; Keenan, T. W.; Mertz, E. T. Plasmin-mediated proteolysis of casein in bovine milk. *Proc. Natl. Acad. Sci. U. S. A.* **1979**, *76*, 2244.

(14) Korhonen, H.; Pihlanto, A. Bioactive peptides: Production and functionality. *Int. Dairy J.* **2006**, *16* (9), 945−60.

(15) Seitz, O. Glycopeptide synthesis and the effects of glycosylation on protein structure and activity. *ChemBioChem* **2000**, *1* (4), 214−46.

(16) Manders, R. J.; Wagenmakers, A. J.; Koopman, R.; Zorenc, A. H.; Menheere, P. P.; Schaper, N. C.; Saris, W. H.; van Loon, L. J. Co-ingestion of a protein hydrolysate and amino acid mixture with carbohydrate improves plasma glucose disposal in patients with type 2 diabetes. *Am. J. Clin. Nutr.* **2005**, *82* (1), 76−83.

(17) Martinez-Villaluenga, C.; Dia, V. P.; Berhow, M.; Bringe, N. A.; Gonzalez de Mejia, E. Protein hydrolysates from beta-conglycinin enriched soybean genotypes inhibit lipid accumulation and inflammation in vitro. *Mol. Nutr. Food Res.* **2009**, *53* (8), 1007−18.

(18) Mrochek, J. E.; Dinsmore, S. R.; Tormey, D. C.; Waalkes, T. P. Protein-bound carbohydrates in breast cancer. Liquid-chromatographic analysis for mannose, galactose, fucose, and sialic acid in serum. *Clin. Chem.* **1976**, *22* (9), 1516−21.

(19) Khaldi, N.; Shields, D. C. Shift in the isoelectric-point of milk proteins as a consequence of adaptive divergence between the milks of mammalian species. *Biol. Direct* **2011**.

(20) Keil, B. *Specificity of Proteolysis*; Springer Verlag: Berlin, 1992.

(21) Earnshaw, W. C.; Martins, L. M.; Kaufmann, S. H. Mammalian caspases: structure, activation, substrates, and functions during apoptosis. *Annu. Rev. Biochem.* **1999**, *68*, 383−484.

(22) Roche *Product Description Version* 3; October 1999.

(23) Fujikawa, K.; Titani, K.; Davie, E. W. Activation of bovine factor X (Stuart factor): conversion of factor Xaalpha to factor Xabeta. *Proc. Natl. Acad. Sci. U. S. A.* **1975**, *72*, 3359.

(24) Li, A.; Sowder, R.; Henderson, L. E.; Moore, S. P.; Garfinkel, D. J.; Fisher, R. J. Chemical cleavage at aspartyl residues for protein identification. *Anal. Chem.* **2001**, *73*, 5395−402.

(25) Houmard, J.; Drapeau, G. R. Staphylococcal protease: a proteolytic enzyme specific for glutamoyl bonds. *Proc. Natl. Acad. Sci. U. S. A.* **1972**, *69*, 3506.

(26) Thornberry, N. A.; Rano, T. A.; Peterson, E. P.; Rasper, D. M.; Timkey, T.; Garcia-Calvo, M.; Houtzager, V. M.; Nordstrom, P. A.; Roy, S.; Vaillancourt, J. P. A combinatorial approach defines specificities of members of the caspase family and granzyme B. *J. Biol. Chem.* **1997**, *272*, 17907−11.

(27) Bornstein, P.; Balian, G. Cleavage at AsnGly bonds with hydroxylamine. *Methods Enzymol.* **1977**, *47*, 132−45.

(28) Han, K. K.; Richard, C.; Biserte, G. Current developments in chemical cleavage of proteins. *Int. J. Biochem.* **1983**, *15*, 875−84.

(29) Degani, Y.; Patchornik, A. Cyanylation of sulfhydryl groups by 2-nitro-5-thiocyanobenzoic acid. High-yield modification and cleavage of peptides at cysteine residues. *Biochemistry* **1974**, *13*, 1−11.

(30) Fulop, V.; Böcskei, Z.; Polgar, L. Prolyl oligopeptidase: an unusual beta-propeller domain regulates proteolysis. *Cell* **1998**, *94*, 161−70.

(31) Schroeder, W.; Shelton, J. B.; Shelton, J. R. An examination of conditions for the cleavage of polypeptide chains with cyanogen bromide: application to catalase. *Arch. Biochem. Biophys.* **1969**, *130*, 551.

(32) Christensen, B.; Schack, L.; Klaning, E.; Sorensen, E. S. Osteopontin is cleaved at multiple sites close to its integrin-binding motifs in milk and is a novel substrate for plasmin and cathepsin D. *J. Biol. Chem.* **2010**, *285* (11), 7929−37.