Title	Peptigram: a web-based application for peptidomics data visualization
Authors(s)	Manguy, Jean; Jehl, Peter; Dillon, Eugène T.; Davey, Norman E.; Shields, Denis C.; Holton, Thérèse A.
Publication date	2016-12-02
Publication information	Journal of Proteome Research, 16 (2): 712-719
Publisher	ACS Publications
Link to online version	https://www.ncbi.nlm.nih.gov/pubmed/?term=peptigram; https://pubs.acs.org/doi/10.1021/acs.jproteome.6b00751
Item record/more information	http://hdl.handle.net/10197/10155
Publisher's statement	Accepted Manuscript version of a Published Work that appeared in final form in the Journal of Proteome Research, copyright © 2016 American Chemical Society after peer review and technical editing by the publisher. To access the final edited and published work see http://pubs.acs.org/doi/abs/10.1021/acs.jproteome.6b00751
Publisher's version (DOI)	10.1021/acs.jproteome.6b00751

Peptigram: a web-based application for peptidomics data visualization

Jean Manguy,[†,‡,¶] Peter Jehl,[‡,¶] Eugène T. Dillon,[†,‡,§] Norman E. Davey,[‡,¶] Denis C. Shields,[*,†,‡,¶] and Thérèse A. Holton[†,‡,¶]

†*Food for Health Ireland,*

‡*UCD Conway Institute of Biomolecular and Biomedical Research,*

¶*School of Medicine,*

§*School of Biomolecular and Biomedical Science,*

University College Dublin, Belfield, Dublin 4, Ireland

E-mail: denis.shields@ucd.ie

Abstract

Tandem mass spectrometry (MS/MS) techniques, developed for protein identification, are increasingly being applied in the field of peptidomics. Using this approach, the set of protein fragments observed in a sample of interest can be determined to gain insights into important biological processes such as signaling and other bioactivities. As the peptidomics era progresses, there is a need for robust and convenient methods to inspect and analyze MS/MS derived data. Here, we present Peptigram, a novel tool dedicated to the visualization and comparison of peptides detected by MS/MS. The principal advantage of Peptigram is that it provides visualizations at both the protein and peptide level, allowing users to simultaneously visualize the peptide distributions of one or more samples of interest, mapped to their parent proteins. In this way rapid comparisons between samples can be made in terms of their peptide coverage and

abundance. Moreover, Peptigram integrates and displays key sequence features from external databases and links with peptide analysis tools to offer the user a comprehensive peptide discovery resource. Here, we illustrate the use of Peptigram on a data set of milk hydrolysates. For convenience, Peptigram is implemented as a web application, and is freely available for academic use at http://bioware.ucd.ie/peptigram.

Keywords

data visualization; peptides; peptidomics; proteomics; mass spectrometry; sample comparison; proteolysis

Introduction

Over the past number of years significant technological advances in instrumentation and methodologies have improved the identification and quantification capacity of tandem MS/MS techniques.[1,2] Accordingly, there is intense interest in these techniques for the characterization and comparison of complex biological samples.[2–4] While the use of tryptic peptides to aid identification has long been a feature of MS-based proteomics, due to advances in MS techniques, leading to higher resolution and accuracy, it is now feasible to study the complete endogenous peptide profiles of samples of interest. In line with this, a new discipline is emerging from proteomics, namely peptidomics.[5]

Peptide focused research has been undertaken for over 100 years,[6] concentrating particularly on neuropeptides[7] and peptides with various bioactivities (e.g. antimicrobial,[8] cell penetrating[9] and ACE-inhibition[10]). Until recently, the high throughput study of peptidomes was not achievable and as such endogenous proteolytic events occurring in peptidomes of interest remained largely unexplored. Peptidomics facilitates the study of native peptides with important roles in biological systems and has been successful in identifying numerous novel functional peptides.[5,6] Other important applications of peptidomics include the dis-

covery of new biomarkers[6] and the study of food from both a characterization and digestive perspective.[5]

As the advancement of proteomics,[2] and indeed peptidomics, continues it is clear that there is a need for bioinformatic solutions to integrate and manage proteomic data.[2,11] The visualization of MS-derived peptides relative to their parent protein is recognized as an important means of rapidly gaining an appreciation for experimental samples of interest.[5,11,12] Currently, there are a number of dedicated tools for the visualization of MS peptide-protein coverage (also known as peptide mapping), including iPiG,[13] PepEx[14] and EnzymePredictor.[11]

The iPiG[13] visualization tool specifically integrates peptide data with genomic data and provides the user with a graphical output that is enriched with peptide quality information. While iPiG provides a comprehensive means of combining proteomic and genomic data, the visualizations are limited to organisms for which genomic information is available, and no information about peptide abundance is incorporated. PepEx[14] is another software that processes MS-peptide data to facilitate the creation of visualizations. This tool boasts two significant features in that it accounts for quantitative information and allows for the comparison of multiple samples, however, it does not directly produce visualizations.

In previous work we described EnzymePredictor,[11] a web-based tool that maps peptides and known proteolytic cleavage sites to precursor proteins. Although not the primary function of this software (for more see Vijayakumar et al.[11]), peptide maps are generated from a single user input file and the output is presented to the user in the form of a downloadable PDF. While such features make EnzymePredictor extremely user-friendly, a major limitation of this tool is that it does not allow for the comparison of multiple samples of interest. Further to this, quantification information is not conveyed in an intuitive manner and there is little scope for implementing user specifications about the data.

Here we present Peptigram, an online, user-friendly and dynamic visualization tool for peptidomics researchers. Peptigram requires a single input file, and visualizations are made

available to the user directly in their web browser. Peptigram firstly provides a global summary view of the peptide profile of samples and additionally provides a detailed peptide alignment map that dynamically integrates and links to external sequence resources according to the users requirements. For convenience, Peptigram's publication quality visualizations may be shared with other users via a hyperlink or can be downloaded. In this article we describe the development of Peptigram and illustrate its application by analyzing a panel of bovine milk hydrolysates. The Peptigram web application is freely available for academic users at http://bioware.ucd.ie/peptigram.

Material and Methods

Experimental data

To demonstrate the utility of Peptigram we use a subset of the peptide data from Holton et al.[15] of bovine milk samples digested with various endopeptidases to represent exemplar food hydrolysates. Given the complexity of milk and the increasing application of peptidomics in the analysis of food hydrolysates (see Dallas et al.[5]), this data set represents a pertinent test case. The samples described in this article include: ArgC, which was generated with the ArgC protease; LysC which was generated using the LysC protease; and ArgC-LysC which was generated using a combination of both the ArgC and LysC proteases. The experimental approach, including sample generation, mass spectrometry, database searching and relative abundance extraction, is described fully in Holton et al.[15]

Implementation

Peptigram is implemented as a web application and therefore precludes the need for the user to download and install any software. The Peptigram software interface and visualizations are built using standard web technologies; meaning this tool is accessible using any combination of operating system and web browser compliant with modern standards. Compatibility

of Peptigram has been tested on the following operating systems and web browsers: Windows: Firefox, Chrome, Opera and Edge; MacOS: Safari, Firefox and Chrome; and Ubuntu: Firefox, Chrome and Opera. Peptigram is written in a combination of PHP, Python 3 and JavaScript. The Symfony framework integrated with Bootstrap is used to create the web interface and Python scripts are used to process input data, which is stored on the server.

Peptigram creates two kinds of publication quality visualizations: peptide profiles and peptide alignment maps (see Figure 1). Peptide profiles are created using the D3 JavaScript library,[16] while the protein visualization tool ProViz[17] is used to create the peptide alignment maps. ProViz retrieves data such as alignments, post-translational modifications, motifs and sequence variations from multiple protein sequence databases. Cleavage patterns of common endopeptidases, obtained from PeptideCutter[3] and EnzymePredictor,[11] are incorporated into the peptide alignment maps. User documentation is available from the Peptigram homepage.

Input file format and data extraction

A single input file in CSV format is required containing protein IDs from the UniProt database. Although Peptigram is predominantly designed to work with the "peptides.txt" output file from the peptide search engine MaxQuant with Andromeda[18] (version 1.5.2.8), any CSV file with the required minimum fields will be accepted as input. Accordingly, suitable output files from other peptide search engines may also be used. The required input file columns are: "Peptide", "UniProt ID", "Start position", "End position" and the sample intensity columns. Peptigram is intended to compare multiple samples but can accept data from a single sample for the purposes of visualization. Equally, although Peptigram is optimized for the comparison of peptide intensity data, other measures, such as spectral counts or identification confidence scores, may be supplied. An example file is available in the Supporting Information.

A script performs data extraction and subsequent storage within the database (see Fig-

ure 1). Extracted data for each job are stored on the server database for 30 days to facilitate sharing and re-examination of results. For each protein in the input data set, the UniProt accession number is verified and information pertaining to the protein length and signal peptide (if present) are retrieved from UniProt. Some MS database search tools may disregard the signal peptide when determining the position of peptides within their parent protein. As such Peptigram checks the correspondence between the input peptide position and the true position in the UniProt protein sequence and adjusts them accordingly where required. This step represents the only alteration made by Peptigram to the user's input data.

Results

After submission of an input file (see Figure S1), the user is redirected to a waiting page, which is automatically refreshed until the data extraction process is complete. Peptigram then automatically redirects the user to a summary page. Each job summary page is assigned a unique identifier thus allowing data visualizations to be shared and revisited. On the summary page all proteins present in the data set are listed in a table and their associated peptide profiles can be viewed (see Figure S2 for the summary page for our test data). From this page, the user can access the peptide alignment map for each of the detected proteins. Here, we use the results from a milk hydrolysis experiment to illustrate the use of Peptigram.

Summary table with protein coverage visualization

Precursor proteins are summarized in a table with a visual schematic of peptide coverage for each protein (see Figure S2). For each individual protein the following information is displayed: protein name, organism, UniProt ID, number of peptides attributed to this protein (across all samples), the intensity of the most abundant peptide and a plot with the percentage coverage of peptides detected. Depending on the number of precursor proteins in the input file, this protein table may extend across multiple pages. For convenience the table

can be sorted dynamically and a search field is provided to allow the user to quickly locate proteins of interest. The default sorting order places the most abundant proteins (as defined by the maximum abundance across all samples) at the top. By selecting a precursor protein from this summary table the user can navigate to the more detailed peptide alignment map. In Figure S2, we see that our test data was found to contain 225 precursor proteins. A total of 2606 peptides were detected across all samples, with precursor proteins containing between 1 and 278 peptides. Protein α_{S2}–Casein is seen to have a peptide coverage of 77% and contains the peptide with the highest ion intensity (1.326×10^{11}) in our data set.

Peptide profile of precursor proteins

Sequence-level peptide coverage diagrams are usually used to show the occurrence of endogenous peptides in a sample by highlighting their position directly within the precursor protein sequence. Such representations can quickly become overcrowded and difficult to interpret when numerous peptides are mapped to the same region.[5] For example, in our test data, some proteins are found to contain more than 200 peptides, spanning up to 96% of the precursor protein, thus making a sequence-based visualization of peptide coverage across multiple samples difficult to comprehend. To overcome this, Peptigram creates a graphical summary representation of the precursor protein, capturing its entire length.[12] Here, a vertical bar is drawn at every amino acid position that is covered by at least one peptide. In cases where multiple peptides are found at an amino acid position, the height of the bar is extended to be proportional to the number of distinct peptides found in this position. A separate peptide coverage track for each sample is included. In this way, Peptigram summarizes in a legible manner the peptide coverage depth for each amino acid residue of the precursor protein across multiple samples (see Figure 2).

A further challenge for coverage visualizations is encountered when peptide intensity data is considered, as not all peptides are present in the same quantity across proteins or samples. To address this, we incorporated an additional variable of color into our profile plots to cap-

ture peptide intensities. Using a green color gradient, the vertical bars are colored according to their associated peptide intensity. Specifically, the color intensity of a given bar represents the sum of peptide abundances at this position, with light green representing a low relative abundance and dark green representing a high relative abundance. As such, the difference between regions of the protein covered by the highest and lowest peptide abundances can be visualized effectively and in an intuitive manner (see Figure 2). Peptigram can normalize peptide intensities, as the user requires, either according to the global intensity scale for the entire data set, or by a sub range of intensities specific to each protein. Essentially, each peptide profile plot efficiently represents the relative abundance and diversity of peptides for various samples at every position in a given protein. Selection buttons allow the user to choose which proteins and samples to display and which normalization scale to use. Plots can be updated according to the users specifications and are available for download as Scalable Vector Graphics (SVG) files.

In Figure 2, a peptide profile for our test data set was generated. Here, three samples (namely ArgC, LysC and ArgC–LysC) are visualized on successive tracks that span the length of the precursor protein β–lactoglobulin. Differences between the three samples are clearly evident, with neither peptide counts nor peptide intensities displaying uniformity across samples. In all samples, peptide coverage is high between amino acid positions 90-120 of β–lactoglobulin; however, variation in intensity is evident (Figure 2). In this region, a low relative intensity is observed for the ArgC sample (Figure 2c), while, conversely, for the LysC and ArgC–LysC samples a high relative intensity is displayed (Figure 2e). In terms of peptide counts, we can see a difference between the ArgC–LysC sample and the other two samples in the region around amino acid position 80. Here, no peptides are detected for the ArgC and LysC samples, as denoted by the dashed lines, while ArgC–LysC is found to contain peptides in this region (Figure 2). In general we can see from Figure 2 that the LysC and ArgC–LysC samples share a high degree of similarity in the region from approximately amino acid position 85 to the C-terminus of β–lactoglobulin. In contrast, the region towards

the beginning of β–lactoglobulin (amino acid positions 15-79 approximately) fails to display any visible similarity between the three samples.

Peptide alignment map with peptide intensities

As the peptide profiles only provide a global summary of the MS sample data, users do not get to compare the individual peptide sequences associated with an area of interest across samples. Accordingly, Peptigram also provides a detailed, interactive peptide alignment map visualization that is sequence-based. An individual peptide alignment map is generated for each precursor protein in the MS data set and the user can navigate to such visualizations by selecting their precursor protein of interest on the job summary page (see Figure S3).

Peptide alignment maps consist of the precursor protein sequence (Figure 3a) annotated with a series of information tracks (Figure 3b–e). In order to aid the comparison between samples, the peptide data for each sample is plotted in its own individual track. Therefore, for our test data set, three separate sample tracks are created: one each for the ArgC, LysC and ArgC–LysC samples (see Figure 3a). In each sample track, a green box is plotted for every peptide, in which the peptide sequence is overlaid. Each box is placed at the corresponding position of the peptide within the precursor protein. Overlapping and adjacent peptide boxes are vertically arranged in such a way that avoids confusion between distinct peptides. Similarly to the peptide profile plots, the green color intensity of each peptide box is proportional to the peptide intensity. The numerical intensity value of a peptide can be displayed on screen by hovering over the peptide box of interest or by clicking on the peptide, which will display it in an information box (see Figure S4).

The peptide alignment map of α_{S2}–Casein for the ArgC, LysC and ArgC–LysC samples of our test data set is displayed in Figure 3. Using this visualization, on a global level, we can see that the peptide coverage of the LysC and ArgC–LysC samples is generally similar, with ArgC displaying a somewhat distinct profile. In this view, overall differences in peptide intensity are also clearly observed, with ArgC containing only low intensity peptides and

LysC and ArgC–LysC containing 2 and 3 high intensity peptides respectively (Figure 3). This observation makes biological sense, since the high intensity peptides detected in the LysC and ArgC-LysC samples terminate in a Lysine, consistent with the cleavage pattern of the LysC enzyme used in their generation. This visualization can additionally be used to focus on the features of particular peptides of interest across samples. For example, if we select the peptide "ALNEINQFYQK", we see that it is present in all 3 of our samples but at varying intensities (Figure 3c). As the color gradients demonstrate, the ArgC sample has a comparably low intensity of this peptide (2.79×10^8), while a much larger abundance is found in the LysC and LysC–ArgC samples (6.65×10^{10} and 5.61×10^{10} respectively).

As the potential enzymatic cleavage history of samples is of interest in peptidomic studies, the proteolysis patterns of well-known cleavage agents defined by sequence motifs have been incorporated into the Peptigram peptide alignment maps. In a dedicated track per enzyme or cleavage agent (found directly below the sample peptide tracks), predicted cleavage sites along the length of the precursor protein are represented by colored boxes (see Figure 3d). Each peptidase is assigned a unique color, which is consistent across the visualization, thus allowing patterns of hydrolysis amongst samples to be assessed. In Figure 3c, for our test data, we selected to display the possible cleavage sites of ArgC and LysC, the respective peptidases used to generate our hydrolysates. Plasmin and cathepsin D were additionally selected as these are known to be endogenous enzymes of milk (Figure 3). Cleavage sites can be tracked across the entire visualization by means of dashed red lines by clicking on the particular sites of interest. Using our test data, a cleavage site shared by plasmin, trypsin and LysC is tracked across the peptide alignment map of α_{S2}–Casein (Figure 3b). Known peptidase cleavage sites not observed in the input peptide data are represented in the visualization by a translucent box. In our test data, we see 6 missed cleavage sites for cathepsin D (Figure 3).

Peptigram enriches the peptide alignment maps with various sequence-based information for the precursor protein. Firstly, to provide an evolutionary context to the MS peptide data,

a multiple sequence alignment of the precursor protein across species is integrated into this view (Figure 3a). The protein visualization tool ProViz[17] is responsible for supplying the multiple sequence alignment data, as well as sequence feature information for the precursor proteins (when available). Supplemental protein sequence information includes: modified residues, known peptides, SNPs, predicted disordered regions and Pfam domains. This information is retrieved from various public sequence repositories (e.g. UniProt, PDB and Pfam) or is computed, as described in Jehl et al.[17] In the peptide alignment maps, the external protein sequence information is displayed in separate tracks below the sample peptide tracks, and is represented either by colored boxes or by a histogram. In Figure 3e, we can see phosphorylation sites, a casein Pfam domain and a predicted disordered region are found in this section of α_{S2}–Casein. For user convenience, peptide alignment maps are available for download as a PDF file. For further information and a description of the sequence level peptide intensity visualization, see Supporting Information.

Interactivity of the peptide alignment maps

By incorporating, and expanding upon, the interactivity features of ProViz,[17] Peptigram presents the user with a highly interactive tool that aids data exploration and understanding in a convenient manner. Given that the peptide alignment map visualization is quite information rich, the user is presented with the option to perform a more focused survey of data pertaining to a particular peptide of interest. Clicking on a peptide of interest will prompt the display of an information box (see Figure S4) summarizing the peptide intensity across samples. Here, links to external tools, such as the PepBank[19] and PeptideAtlas[20] databases, or the bioactivity prediction web-server PeptideRanker,[21] can also be found, allowing the user to obtain more in-depth knowledge of the characteristics of their peptide.

To facilitate a more targeted study of the user's data, peptide alignment maps can be easily adapted in a dynamic manner to suit the users requirements. The "Settings & filters" panel (see Figure S5) allows the user to modify the visualization in a number of ways. For

example, every data track (except the precursor protein sequence) can be hidden as required, so as to display, for instance, only samples or peptidases of interest. Moreover, peptides can be filtered by intensity to retain only peptides between user-defined thresholds. In order to define a threshold users can input the values directly, or can use a dedicated slider button. In addition, the user can select to display only a specified number of the most or least intense peptides (e.g. the top 20 most intense peptides). Finally, there is the option to highlight vertical sections of the visualization should the user wish to focus on a particular region of interest in the precursor protein (see Figure 2b and Supporting Information for more).

To illustrate the operation and utility of the interactive features of Peptigram's peptide alignment maps, we generated a peptide alignment map of a sub section of β–lactoglobulin, which corresponds to the region seen in Figure 2e. To achieve this, the default peptide alignment map of β–lactoglobulin for our test data was firstly resized to display only residues from amino acid positions 82-155, by directly inputting these values and clicking on the "Resize" button. Next, using the dedicated drop down list in the "Settings & filters" panel, we selected the ArgC and LysC samples (thus excluding the ArgC–LysC sample), along with the ArgC and LysC peptidases to display only their associated cleavage sites. We then removed the lowest intensity peptides using the dedicated slider buttons in the "Settings & filters" panel. This resulted in the visualization displayed in Figure 4. From this visualization (Figure 4), we see that the LysC sample contains noticeably more high intensity peptides than the ArgC sample, reflecting what is observed in the peptide profile (see Figure 2e). It is evident in Figure 4 that one cleavage site (between residues 107 and 108) appears to differentiate the peptide profiles of these two samples. As might be expected due to the hydrolysis history of the LysC sample, residue 107 is a Lysine. To highlight this position, the cleavage site was marked on the visualization by clicking on the peptidase box to produce a red dashed line (see Figure 4).

Peptigram offers the user the ability to emphasize peptides of interest across the peptide alignment map. To do this, regular expressions or the actual peptide sequence can be used

as input for the "Search peptides" panel. This feature can be useful to highlight peptides that start at, end at, or overlap particular cleavage sites, or to focus on peptides that contain a specific sequence motif of interest. In Figure 4, we searched for peptides matching the following regular expression: (NENK$)|(^VLVL)|(NENKVLVL), where $ represents the end of a sequence and ^ represents the start. This regular expression selects all peptides ending with "NENK", or starting with "VLVL", or with the motif "NENKVLVL" within them. Peptides not matching this pattern are obscured in the background of Figure 4, while those matching the pattern are prominent in the foreground. In doing this, we see the peptide "IPAVFKIDALNENK" has a far greater intensity in the LysC sample than in the ArgC sample (Figure 4), which may warrant further investigation. To facilitate this, the user can click on a peptide to bring up the "Peptide Information Box" (see Figure S4). From this panel, external peptide resources can be queried with a peptide sequence of interest. For the "IPAVFKIDALNENK" peptide we find that it has an entry in PeptideAtlas, which allows the user to cross-reference information about this peptide from other studies. The "IPAVFKIDALNENK" peptide achieves a PeptideRanker[21] score of 0.28, indicating that it is unlikely to be interesting from a bioactivity point of view.

Limitations

As Peptigram functions on all data from UniProt, it has the advantage of being generally applicable across a broad diversity of organisms. Conversely, this also poses a significant disadvantage, as data from other important sequence databases or *de novo* sequencing cannot be considered in the current version. Secondly, Peptigram does not represent post-translational modifications in its present form. Peptigram, like other server-based tools is limited by the capacity of our server to handle large files. However, it is worth noting that during testing, no performance issues were encountered with a data file (2.6 MB) of 50 samples with over 3000 peptides and 500 proteins. Finally, while use of input data from MS database search engines other than MaxQuant with Andromeda (e.g. X!Tandem, SEQUEST and PEAKS)

is currently feasible, this does require some user preparation of their data in order to be consistent with Peptigram input requirements. Further to this, as Peptigram does not alter the input data, it may be necessary for the user to transform or normalize their data before upload. Despite these limitations, Peptigram represents an important progression in the visualization and interpretation of peptide data from multiple samples.

Conclusion

Peptigram is the first software tool to provide both protein and sequence level visualizations of peptide intensities for the interactive comparison of multiple samples of interest. To demonstrate the utility of Peptigram we described its application on a test data set of milk hydrolysates. Milk serves as a suitable material for testing comparative peptidomics methods, as it is complex, containing numerous endogenous peptides across a large expression range.[15] When presented with such data our software performed robustly, providing a convenient and dynamic method for peptidomic data visualization and comparison, delineating differences and similarities between samples in an intuitive manner.

The visualization of actual peptide sequence data is of great importance to gain an in depth understanding of experimental sample composition. Given that "peptide mapping" for even a single sample can often result in complex and overcrowded visualizations, Peptigram aims to offer the user an interactive sequence level approach for comparing peptide maps that are both manageable and digestible. As demonstrated through the use of our test data set, the peptide map visualizations of Peptigram are intuitive, comprehensive and most importantly, provide for user customization. In this way, users can choose to focus on the data they deem relevant and filter out that which is of limited interest to their present study. Integration with relevant sequence-based information, such as cleavage sites and the location of protein domains, and external peptide assessment tools adds useful context to the sample data.

Peptigram has a wide variety of potential applications. These include the study of peptide generation in foods, assessing batch-to-batch variation, identifying potential biomarkers, and comparison of various hydrolysis or fermentation regimes. It is also relevant to other applications of peptidomic approaches, including the analysis of intracellular extracts, extracellular media, serum and other biological fluids. Importantly, the use of Peptigram is not limited to peptidomics experiments as it can equally be applied in a proteomic workflow for the analysis of tryptic peptide coverage and splice variation.

Author contributions

Project design was by TAH, with further development by DCS and JM. Software implementation was by JM with inputs from TAH, DCS and ND. PJ provided supplementary coding and support for integration with ProViz. ETD prepared the milk hydrolysates and processed the raw MS/MS data. JM and TAH wrote the manuscript, with contributions from DCS and ND. All authors reviewed and commented on the software and on the manuscript.

Acknowledgement

The authors thank Nessa Noronha (Food for Health Ireland, University College Dublin, Belfield, Dublin 4, Ireland), Matthias Wilm and Gerard Cagney (UCD Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Belfield, Dublin 4, Ireland) for useful discussion.

This work was funded by Enterprise Ireland grant (TC2013001) to Food for Health Ireland (supporting JM, ETD and TAH); Irish Research Council Bioinformatics and Systems Biology Graduate Research Education Programme (supporting JM), Science Foundation Ireland Principal Investigator grant (08/IN1/B1864 to DCS); Science Foundation Ireland Starting Investigator Research Grant (13/SIRG/2193 to NED); Science Foundation Ireland Industry Fellowship grant (15/IFB/3601 to TAH); and Science Foundation Ireland Principal

Investigator grant (11/PI/1034 supporting PJ).

Supporting Information Available

- Supplementary Text: Details about the peptide alignment interactivity features. Supplementary Figures S1–S6: screenshots of the interface and details of the interface.
- Supplementary File: Example Peptigram input file.

This material is available free of charge via the Internet at `http://pubs.acs.org/`.

References

(1) Michalski, A.; Damoc, E.; Hauschild, J.-P.; Lange, O.; Wieghaus, A.; Makarov, A.; Nagaraj, N.; Cox, J.; Mann, M.; Horning, S. Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. *Mol. Cell Proteomics* **2011**, *10*, M111.011015.

(2) Altelaar, A. M.; Munoz, J.; Heck, A. J. Next-generation proteomics: towards an integrative view of proteome dynamics. *Nature Reviews Genetics* **2013**, *14*, 35–48.

(3) Kinter, M.; Sherman, N. E. *Protein sequencing and identification using tandem mass spectrometry*; John Wiley & Sons, 2005; Vol. 9.

(4) Wilm, M. Quantitative proteomics in biological research. *Proteomics* **2009**, *9*, 4590–4605.

(5) Dallas, D. C.; Guerrero, A.; Parker, E. A.; Robinson, R. C.; Gan, J.; German, J. B.; Barile, D.; Lebrilla, C. B. Current peptidomics: applications, purification, identification, quantification, and functional analysis. *Proteomics* **2015**, *15*, 1026–1038.

(6) Schrader, M.; Schulz-Knappe, P.; Fricker, L. D. Historical perspective of peptidomics. *EuPA Open Proteomics* **2014**, *3*, 171–182.

(7) Hökfelt, T.; Bartfai, T.; Bloom, F. Neuropeptides: opportunities for drug discovery. *The Lancet Neurology* **2003**, *2*, 463–472.

(8) Phoenix, D. A.; Dennison, S. R.; Harris, F. Antimicrobial peptides: their history, evolution, and functional promiscuity. *Antimicrobial peptides* **2013**, 1–37.

(9) Heitz, F.; Morris, M. C.; Divita, G. Twenty years of cell-penetrating peptides: from molecular mechanisms to therapeutics. *British Journal of Pharmacology* **2009**, *157*, 195–206.

(10) Hartmann, R.; Meisel, H. Food-derived peptides with biological activity: from research to food applications. *Current Opinion in Biotechnology* **2007**, *18*, 163–169.

(11) Vijayakumar, V.; Guerrero, A. N.; Davey, N.; Lebrilla, C. B.; Shields, D. C.; Khaldi, N. EnzymePredictor: A Tool for Predicting and Visualizing Enzymatic Cleavages of Digested Proteins. *J. Proteome Res.* **2012**, *11*, 6056–6065.

(12) Lambers, T. T.; Gloerich, J.; van Hoffen, E.; Alkema, W.; Hondmann, D. H.; van Tol, E. A. Clustering analyses in peptidomics revealed that peptide profiles of infant formulae are descriptive. *Food Sci Nutr* **2015**, *3*, 81–90.

(13) Kuhring, M.; Renard, B. Y. iPiG: Integrating Peptide Spectrum Matches into Genome Browser Visualizations. *PLoS ONE* **2012**, *7*, e50246.

(14) Guerrero, A.; Dallas, D. C.; Contreras, S.; Chee, S.; Parker, E. A.; Sun, X.; Dimapasoc, L.; Barile, D.; German, J. B.; Lebrilla, C. B. Mechanistic peptidomics: factors that dictate specificity in the formation of endogenous peptides in human milk. *Mol. Cell Proteomics* **2014**, *13*, 3343–3351.

(15) Holton, T. A.; Dillon, E. T.; Robinson, A.; Wynne, K.; Cagney, G.; Shields, D. C. Optimal computational comparison of mass spectrometric peptide profiles of alternative hydrolysates from the same starting material. *LWT - Food Science and Technology* **2016**, *73*, 296–302.

(16) Bostock, M.; Ogievetsky, V.; Heer, J. D3: Data-Driven Documents. *IEEE Trans Vis Comput Graph* **2011**, *17*, 2301–2309.

(17) Jehl, P.; Manguy, J.; Shields, D. C.; Higgins, D. G.; Davey, N. E. ProViz - a web-based visualization tool to investigate the functional and evolutionary features of protein sequences. *Nucl. Acids Res.* **2016**, *44*, W11–W15.

(18) Cox, J.; Neuhauser, N.; Michalski, A.; Scheltema, R. A.; Olsen, J. V.; Mann, M. Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. *J. Proteome Res.* **2011**, *10*, 1794–1805.

(19) Shtatland, T.; Guettler, D.; Kossodo, M.; Pivovarov, M.; Weissleder, R. PepBank - a database of peptides based on sequence text mining and public peptide data sources. *BMC Bioinformatics* **2007**, *8*, 280.

(20) Desiere, F.; Deutsch, E. W.; King, N. L.; Nesvizhskii, A. I.; Mallick, P.; Eng, J.; Chen, S.; Eddes, J.; Loevenich, S. N.; Aebersold, R. The PeptideAtlas project. *Nucl. Acids Res.* **2006**, *34*, D655–D658.

(21) Mooney, C.; Haslam, N. J.; Pollastri, G.; Shields, D. C. Towards the improved discovery and design of functional peptides: common features of diverse classes permit generalized prediction of bioactivity. *PLoS ONE* **2012**, *7*, e45012.

Figure 1: Diagram of the Peptigram workflow. The CSV input file is stored in a database for future retrieval and sharing. The input data is used to generate peptide profiles and peptide alignment maps. For the latter ProViz integrates external data from protein sequence databases.

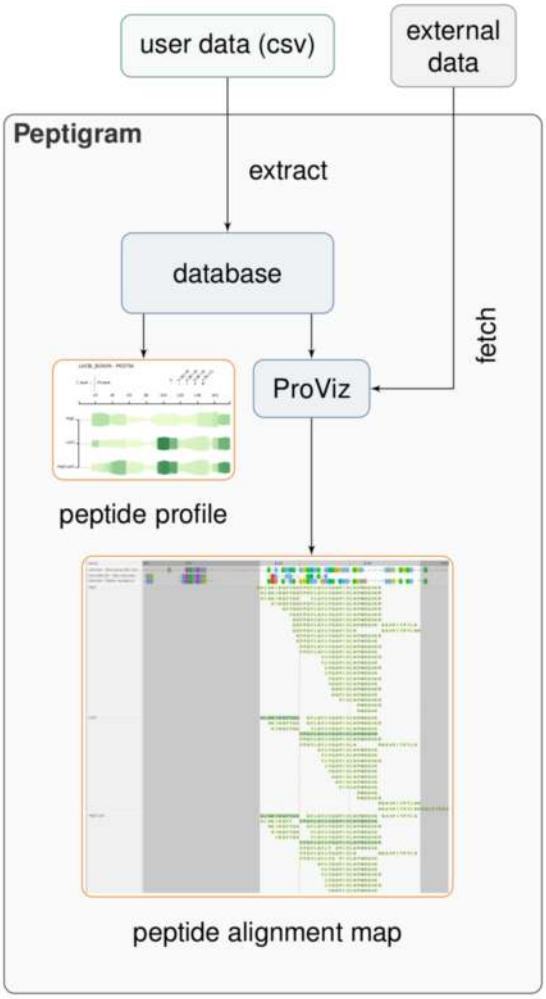


Figure 2: Peptide profile of β–lactoglobulin (P02754) from milk samples digested by ArgC, LysC or an ArgC–LysC combination. (a) At each amino acid residue along the protein, the height of the green bars is proportional to the count of peptides overlapping this position; (b) the intensity of the color (green) is proportional to the sum of the peptide of intensities overlapping this position. Therefore in region (c) of the ArgC sample there is a high count of overlapping peptides with a low sum of intensities, while region (d) of the LysC sample has a low count of peptides with a low sum of intensities. Finally, in region (e) of the LysC and ArgC–LysC samples we see that there is a high count of peptides with a high sum of intensities.

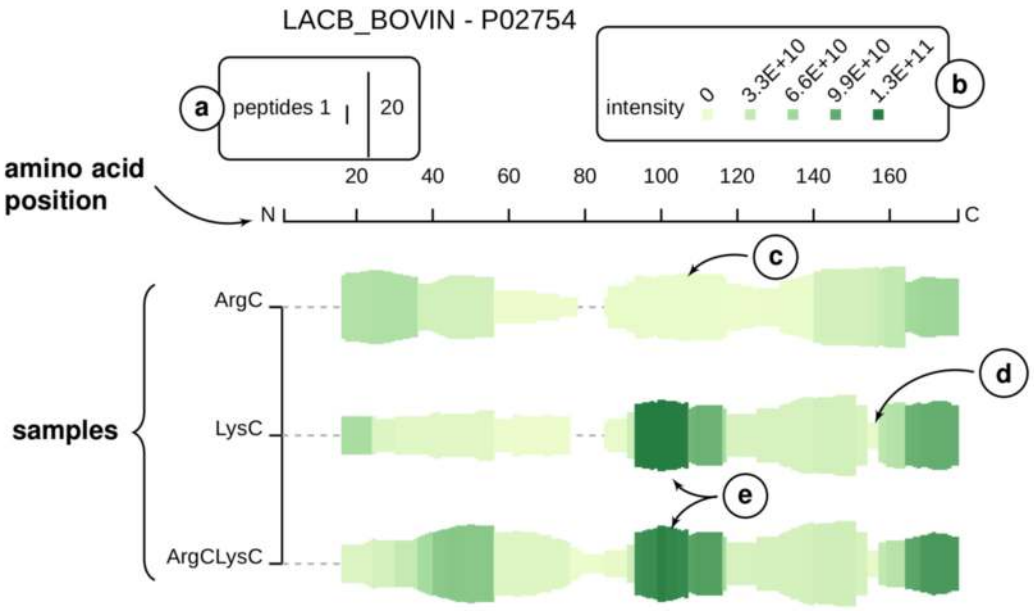


Figure 3: Peptigram peptide alignment map. Here we see α_{S2}–S2-Casein (P02663) from milk samples digested by ArgC, LysC or an ArgC–LysC combination. A species alignment (a) from GeneTree and additional information about the precursor protein (b) are retrieved by ProViz. Peptides from the input data are displayed in green boxes, where the color intensity of the peptide box is proportional to the peptide intensity. Peptides from different samples (c) are displayed on different tracks. Cleavage sites of common endopeptidases (d) can be selected for display in additional tracks. The user can choose to highlight cleavage sites of interest by selecting them; a red line will then be traced along the entire visualization at the cleavage site positions. The user can also select a region of interest (e) within the precursor protein and highlight it across all samples.

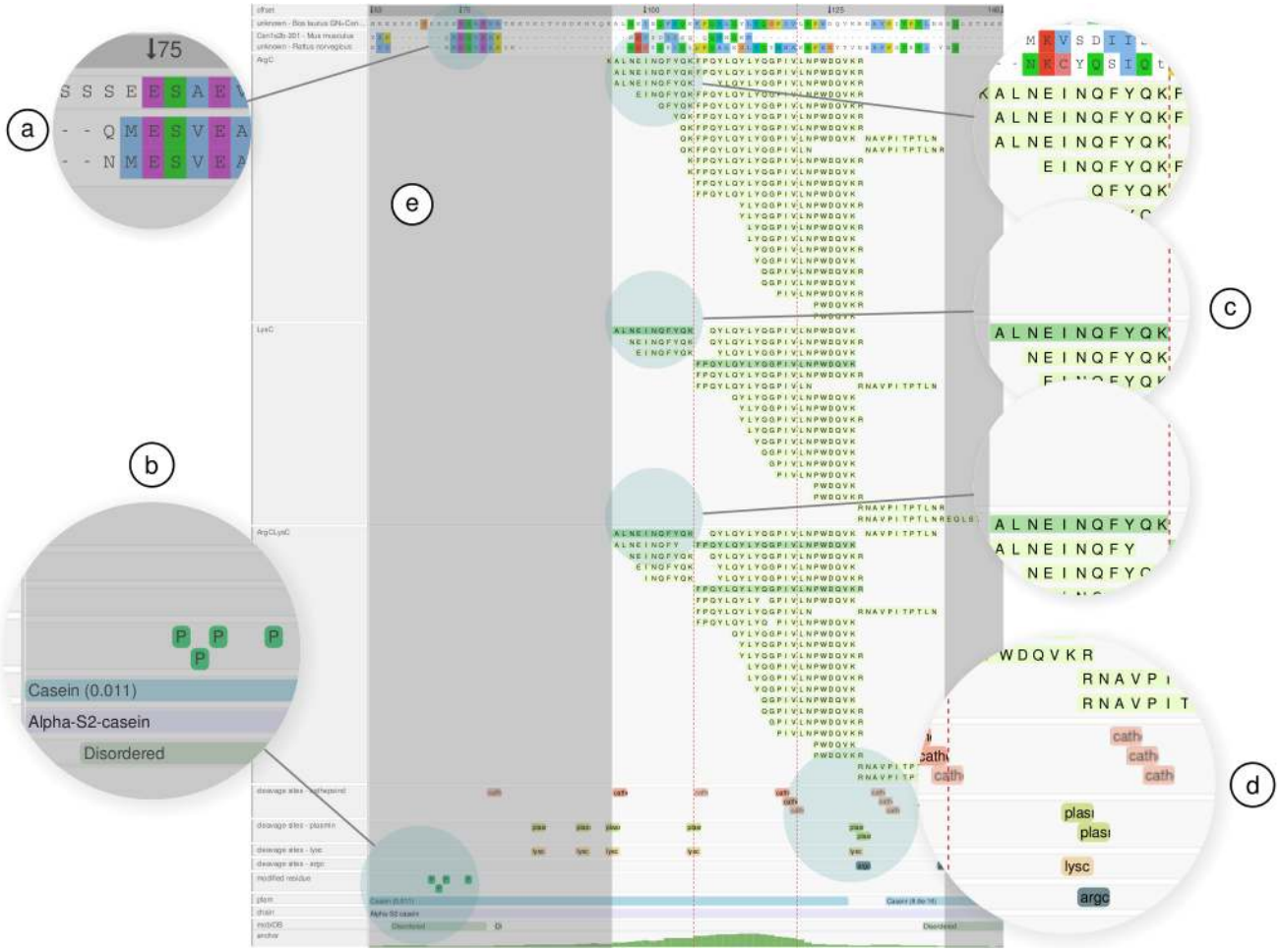
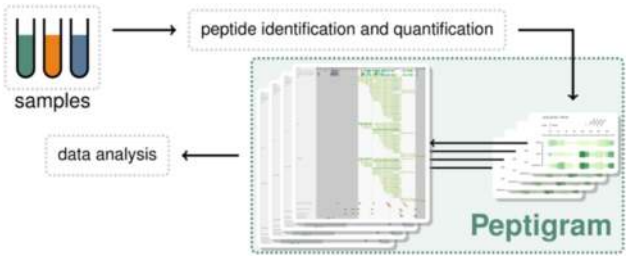


Figure 4: View of a customized Peptigram peptide alignment map. This example is generated from the default peptide alignment map of β-lactoglobulin (P02754) for our sample milk data set (see text for details of the options used to generate this customized view).

For TOC Only


peptide identification and quantification
samples
data analysis
Peptigram