



Peptimetric: Quantifying and Visualizing Differences in Peptidomic Data

Erik Hartman^{1*}, Simon Mahdavi¹, Sven Kjellström² and Artur Schmidtchen^{1,3,4}

¹Division of Dermatology and Venereology, Department of Clinical Sciences, Lund University, Lund, Sweden, ²Division of Mass Spectrometry, Department of Clinical Sciences, Lund University, Lund, Sweden, ³Dermatology Skåne University Hospital, Lund, Sweden, ⁴Copenhagen Wound Healing Center, Bispebjerg Hospital, Department of Biomedical Sciences, University of Copenhagen, Copenhagen, Denmark

OPEN ACCESS

Edited by:

Raghu Machiraju,
The Ohio State University,
United States

Reviewed by:

William C Ray,
Nationwide Children's Hospital,
United States
Bjorn Sommer,
Royal College of Art, United Kingdom

*Correspondence:

Erik Hartman
erik.hartman@hotmail.com

Specialty section:

This article was submitted to
Data Visualization,
a section of the journal
Frontiers in Bioinformatics

Received: 08 June 2021

Accepted: 11 August 2021

Published: 25 August 2021

Citation:

Hartman E, Mahdavi S, Kjellström S
and Schmidtchen A (2021)
Peptimetric: Quantifying and
Visualizing Differences in
Peptidomic Data.
Front. Bioinform. 1:722466.
doi: 10.3389/fbinf.2021.722466

Finding new sustainable means of diagnosing and treating diseases is one of the most pressing issues of our time. In recent years, several endogenous peptides have been found to be both excellent biomarkers for many diseases and to possess important physiological roles which may be utilized in treatments. The detection of peptides has been facilitated by the rapid development of biological mass spectrometry and now the combination of fast and sensitive high resolution MS instruments and stable nano HP-LC equipment sequences thousands of peptides in one single experiment. In most research conducted with these advanced systems, proteolytically cleaved proteins are analyzed and the specific peptides are identified by software dedicated for protein quantification using different proteomics workflows. Analysis of endogenous peptides with peptidomics workflows also benefit from the novel sensitive and advanced instrumentation, however, the generated peptidomic data is vast and subsequently laborious to visualize and examine, creating a bottleneck in the analysis. Therefore, we have created Peptimetric, an application designed to allow researchers to investigate and discover differences between peptidomic samples. Peptimetric allows the user to dynamically and interactively investigate the proteins, peptides, and some general characteristics of multiple samples, and is available as a web application at <https://peptimetric.herokuapp.com>. To illustrate the utility of Peptimetric, we've applied it to a peptidomic dataset of 15 urine samples from diabetic patients and corresponding data from healthy subjects.

Keywords: peptidomics, proteomics, bioinformatics, mass spectrometry, biomarkers, visualization

1 INTRODUCTION

Although peptides have been studied since the beginning of the 20th century, the practice of analyzing the complete peptidome of large cohorts of samples has only recently been actualized by the advancement of biological mass spectrometers, tandem-mass spectrometry techniques (MS/MS) and computational algorithms [Schrader et al. (2013)]. These new techniques allow for large screening of endogenous bioactive molecules with great physiological implications which ultimately may give rise to effective and sustainable pharmaceuticals and diagnostic tools [Lai et al. (2014)]. New and more advanced mass spectrometers generate progressively more data as they're able to identify more peptides and perform searches faster [Timp and Timp (2020)]. Large

sets of peptidomic data are difficult to comprehend and visualize without advanced and dedicated programs, and there is therefore a great need for tools which enable researchers with modest experience in programming and visualization skills to analyze their samples.

The nature of peptidomic data makes it highly compatible with computational algorithms and there are already many software programs such as MsViz, PepServe and CLaSS which utilize this fact to map post translational identifications, screen for antimicrobial peptides, and map the feature space of peptidomic datasets respectively [Martin-Campos et al. (2017); Das et al. (2021); Alexandridou et al. (2011)]. In addition to these softwares there are several packages, mostly for R, with similar aims, such as seqinR and Pviz [Preston et al. (2020); Sauteraud et al. (2021)]. These tools have their purposes, however, to our knowledge, close to no programs exists which aim to visualize the peptide coverage of complete peptidomes interactively and with easy access. One useful tool for visualization of multiple peptidomic samples is Peptigram [Manguy et al. (2017)], created by Manguy et al. Peptigram allows for easy peptide visualization using an internet based framework, and connects with various resources such as Pfam [Mistry et al. (2021)] and ProViz [Jehl et al. (2016)] to create a comprehensive map of the peptidome of multiple samples. However, as with every tool, there are some limitations to Peptigram, especially when it comes to pre-processing, quantitative comparisons between sample groups and interactive sample exploration.

The process of producing MS/MS data in data dependent mode is stochastic and results in variation between samples. Additionally, variation is added by sampling techniques used prior to the liquid chromatography tandem-mass spectrometry (LC-MS/MS) analysis which leads to biases between samples. These biases are of extreme importance if one tries to find biomarkers, quantify patterns or statistically determine correlations. Therefore, peptidomic data regularly require pre-processing where normalization techniques and cutoffs are introduced, so that inter-sample biases are minimized and peptides of low abundance or certainty are removed from the dataset [Klont et al. (2019)].

Furthermore, scientific hypothesis testing generally consists of groups of samples, where a difference or correlation of an observed phenomenon is quantified and its statistical significance determined. This is especially true of medical and biological trials, as many studies compare positive samples to a negative control group. Additionally, the nature of peptidomic data is highly dependent on the connection between precursor protein and the resulting peptides. To explore these relationships effectively a high level of interactivens is required, where one can easily switch between investigating the identified specific endogenous peptides and the sequence coverage from the corresponding proteins.

In previous work, computation in combination with tools such as Peptigram was utilized to explore the presence of biomarkers and endogenous antimicrobial peptides in infected wounds, further showcasing the applicability and usefulness of algorithms on peptidomic data [Hartman et al. (2021)]. In this paper, we have continued to explore and develop algorithms

TABLE 1 | Allowed column names for the input files.

Peptide sequence	Peptide, Sequence
Precursor protein ID	Accession, Protein, UniProt ID
Intensity	Area, Intensity
Spectral count	SpC, Spectral count, #Feature

allowing for massive screening and visualization of peptides in large cohorts of samples. Ultimately, we've created Peptimetric, an open-source web based application for highly interactive group comparisons and sample exploration of peptidomic data, available at <http://peptimetric.herokuapp.com>. Peptimetric allows the user to pre-process their data using different types of normalization and cutoffs to remove biases which are introduced in sample preparation and subsequent LC-MS/MS analysis. Thereafter it provides a dynamic graphical interface where the user may investigate their sample groups on a protein and peptide level to search for peptides of interest. Furthermore, the generated figures and tables are readily available to download for publications or further exploration. Peptimetric is freely available as an open GitHub repository (<https://github.com/ErikHartman/peptimetric>) under an MIT license.

2 METHODS

2.1 Implementation

Peptimetric is implemented as a web application making it accessible through a web browser without the need for package or software-installations at (<https://peptimetric.herokuapp.com/>). The application is hosted by the cloud service Heroku [Heroku (2021)]. The front end of the application is developed in Python 3.8.8 using the Dash library [version 1.2.0, Plotly (2021)] in combination with Bootstrap components. Peptimetric is therefore compatible with the following web browsers: Chrome, Opera, Microsoft Edge and Firefox and Safari. The complete repository, including the data files and requirements, are available at GitHub (<https://github.com/ErikHartman/Peptimetric>) and is freely available under an MIT license. The repository may be cloned to run Peptimetric locally.

2.2 Input Files

Peptimetric is used to visualize and dynamically explore the differences between the peptidomes of groups of samples. Therefore, at least one data-file, in a csv (comma separated values) format, per group is required to use the application. However, for statistical analysis, a minimum of three files per group is required. The files need to include four columns for the peptide sequence, intensity (or other abundance metric), spectral count (or other abundance metric) and a protein precursor id (UniProt id). To accommodate for different raw data processing softwares, such as PEAKS, the names in **table 1** (including capitalization variations) are accepted for the columns:

All input files are stored on the Heroku server during the session. A local copy of proteomes fetched from UniProt

TABLE 2 | The species currently available in the Peptimetric database, alongside the number of proteins and UniProt proteome identifier (UPID).

Species	Number of proteins	Proteome identifier
<i>Homo sapiens</i> (Human)	77,027	UP000005640
<i>Sus scrofa</i> (Pig)	49,792	UP000008227
<i>Rattus norvegicus</i> (Rat)	29,936	UP000002494
<i>Cricetulus griseus</i> (Hamster)	23,885	UP000001075
<i>Mus musculus</i> (Mouse)	55,470	UP000000589
<i>Danio rerio</i> (Zebra fish)	46,849	UP000000437
<i>Drosophila melanogaster</i>	22,114	UP000000803
<i>Caenorhabditis elegans</i>	26,625	UP000001940
<i>Candida albicans</i>	6,035	UP000000559
<i>Escherichia coli</i>	4,273	UP000000318

TrEMBL in April of 2021 is used to get the FASTA sequences and UniProt mnemonic protein identifier (e.g. HBA_HUMAN) for the submitted precursor proteins [Consortium (2020)]. The local database consists of a subset of proteomes which were chosen due to their prevalence in biological and peptidomic studies (see **table 2** to view the species currently available in Peptimetric). Additional proteomes are easily added, and we urge researchers to contact us if they wish to analyze samples from other species. The local proteome database will be updated if requested by users.

2.3 Data Processing

The first step of data manipulation performed by Peptimetric is to take the 10th logarithm (log₁₀) of all intensity values. This is commonly performed to diminish the exponential nature of MS intensities and to later provide appropriate quantification for proteins and peptides as Gaussian (normal) distribution is assumed [Dufresne et al. (2017)]. If one has already taken the logarithm of their values, or wishes to leave the intensity values unaltered, a checkbox in the normalization popup is to be checked before uploading the files.

Biases are often introduced during sample preparation LC-MS/MS analysis [Klont et al. (2019); Geyer et al. (2019)]. Some of these biases may be dealt with when processing the raw files, however, they may go undetected or one may choose to leave the samples unaltered to perform post-processing normalization. Additionally, peptidomic data may include peptides and proteins with low abundance or certainty which may not be appropriate to include in peptidomic analyses. Therefore, Peptimetric includes ways to normalize the input data, as well as to discard sets of the data based on cutoffs.

Normalization of the data from MS or MS/MS can be performed in multiple ways, however, Peptimetric accommodates for two ways of normalization: either by normalizing on the samples' global values [Griffin et al. (2010); Nusinow et al. (2020)] or by normalizing on a housekeeping protein (Karpievitch et al. (2012)). When normalizing by global values, each value in the given sample is divided by the sum of all values in the sample. This results in the centering of sample values around a common value (i.e., normalizing them), however, all values are reduced as a consequence. To reinstate the values to their original scale, all

samples are multiplied by the average sample value, see **Equation 1**.

$$\text{Global} : N_i = \frac{\sum_{j=1}^m S_j}{m} \times \frac{p_i}{\sum_{i=1}^n p_i} \quad (1)$$

where:

N normalized value.

S total sample value

m number of samples

p peptide value

n number of peptides in sample.

Normalization by a housekeeping protein is performed similarly to the global normalization, however, instead of using the total sum based on the complete sample, a subsection containing peptides from the selected precursor protein is used for normalization, see **Equation 2**.

$$\text{Housekeeping} : N_i = \frac{\sum_{j=1}^m P_j}{m} \times \frac{p_i}{\sum_{i=1}^n p_i} \quad (2)$$

where:

N normalized value.

P total protein value

m number of samples

p peptide value

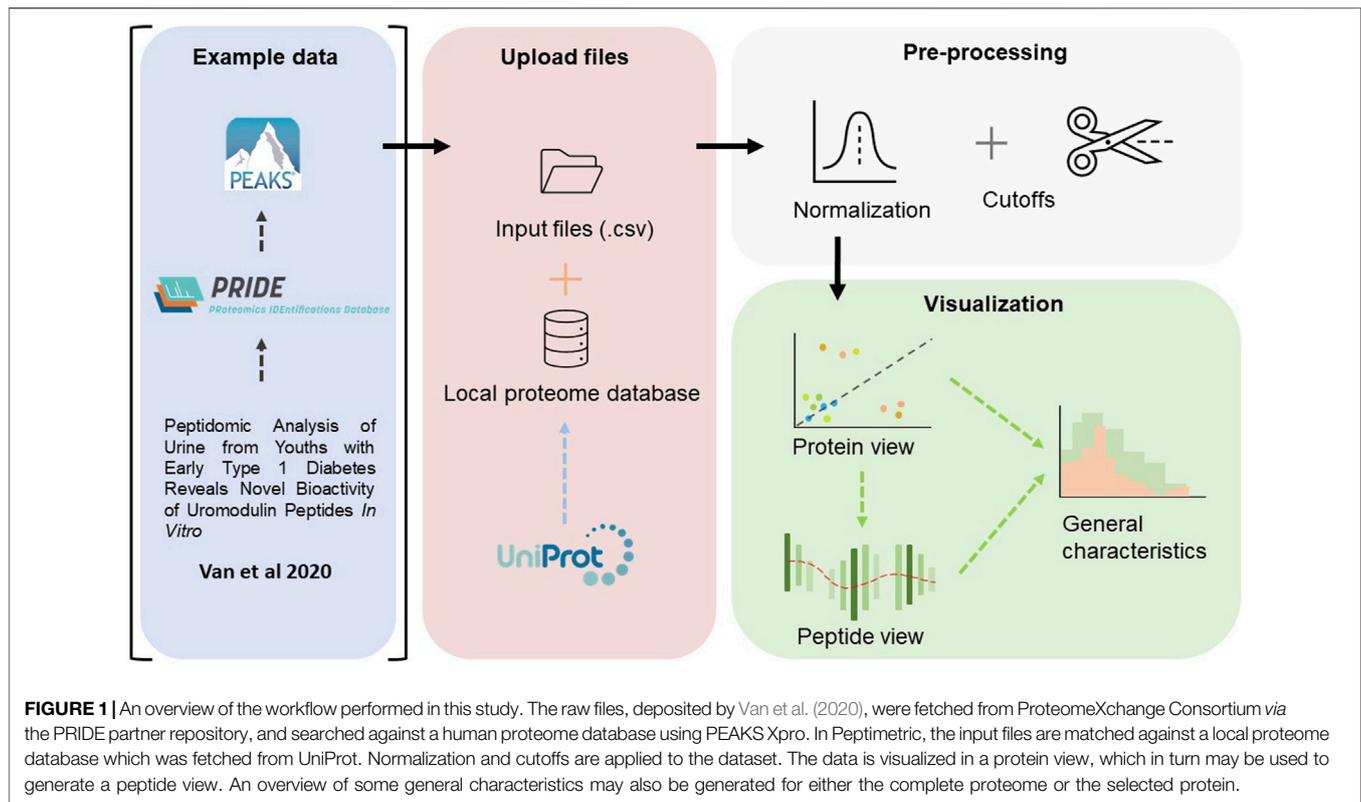
n number of peptides in protein.

The housekeeping protein may be defined by the user. If the housekeeping protein is not present in a given sample, the intensity and spectral count values in the sample are left unaltered.

Cutoffs are used to remove proteins and/or peptides below given thresholds and are applied to peptides and proteins separately. Peptides may be discarded based on intensity, spectral count and for being retention time (RT) and/or collision cross section (CCS) outliers. Outliers are defined to be situated three standard deviations from the mean value as is common practice when assuming a Gaussian distribution (the procedure is often denoted as the three sigma rule or 68-95-99.7-rule). Protein cutoffs may thereafter be applied to remove proteins with cumulative intensities or spectral counts below a given threshold. Additionally, a cutoff may be applied to remove proteins with few peptides. Note that protein cutoffs are applied after peptide cutoffs. Therefore, proteins which contain many peptides of low abundance will be removed if the peptide cutoffs are applied appropriately.

2.4 Experimental Data

To illustrate the utility and typical usage of Peptimetric, a dataset generated from a study by Van et al. (2020), describing peptidomic analysis of urine, was fetched from an online repository. The study profiles the urinary peptides from 15 patients with type-1-diabetes (D) and uses corresponding data from non-diabetic (ND) subjects for comparison. This study is well suited for analysis using Peptimetric as it contains groups of samples and uses a discovery peptidomics approach. Additionally, the study collected data from a relatively large population and conducted thorough qualitative and



quantitative analyses. Details about study population, the sample generation and preparation are described further in Van et al. (2020).

The raw files were retrieved from ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD012210 [Protomecentral (2021)]. Van et al. (2020) used MaxQuant (version 1.5.3.28) to analyze the raw files whereas PEAKS Xpro was used in this study. Similar settings were applied for the search, using a maximum of two modifications per peptide, allowing the same modifications (methionine and proline oxidation, N-terminal acetylation). After the database search the identified peptides were filtered with a cut off using 1% FDR and a minimum of two peptides per protein. In total, 6,559 and 9,024 unique peptides were identified in the samples from non-diabetics and type-1-diabetics respectively. Comparatively, the search conducted by Van et al. (2020) resulted in fewer peptides, finding 5,011 and 5,708 unique peptides for the respective groups. The complete files are uploaded to the GitHub repository, as well as to the Peptimetric server, and the data may be investigated by any user by clicking “Load Example Files”.

The complete workflow is described in **Figure 1** below. In summary, the workflow follows the common paradigm in visual analytics: analyse first - show the important - zoom, filter and analyse further - details on demand [Schneiderman (1996) Keim et al. (2008)]. Firstly, an overview is rendered, where all proteins are visualized. One may zoom and filter this overview, and thereafter select a protein to subsequently present a view of the peptide coverage within that protein. Lastly, one may

generate figures to showcase some general characteristics of the peptides within the selected protein or for the complete peptidome. The iterative back-and-forth between the protein view and detailed peptide view and general characteristics, in combination with filtering and normalization, mimics the sense-making loop which is paramount in understanding and visualization [Wijk (2005)].

3 RESULT

After submitting input files into each group, the files are concatenated and matched against the local database to retrieve the FASTA sequences and UniProt mnemonic identifiers. If a submitted UniProt id is not found in the database, the peptides belonging to the precursor protein will be discarded from the dataset.

3.1 Protein View

After uploading the files and applying potential normalization and cutoffs, it is possible to generate a protein graph which showcases all the precursor proteins present in the samples in the format of a scatter plot, see **Figure 2A**. The protein graph allows the user to visualize the abundance of the proteins based on one of the following metrics: sum of intensities, mean of intensities, sum of the spectral counts, and mean of the spectral counts. The means are calculated by taking the mean of the abundance metric of all the peptides in each protein, whereas the sum is calculated by taking the sum of the abundance metric of all the peptides in

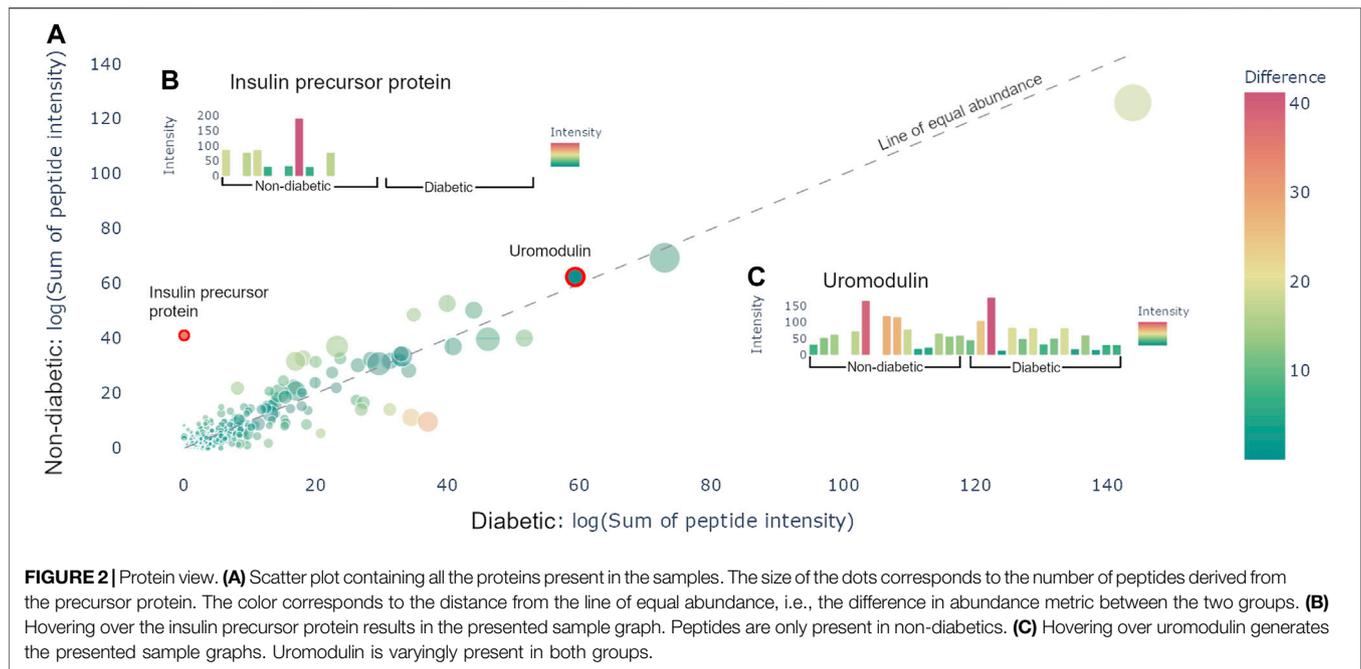


TABLE 3 | Protein table. The table contains the top five most abundant proteins in both groups.

Protein	UniProt id	Peptides ND	Peptides D	Intensity ND \pm SD	Intensity D \pm SD
ALBU_HUMAN	P02768	3,134	2,051	131.46 \pm 249.49	168.08 \pm 231.32
HBA_HUMAN	P69905	2,690	752	68.78 \pm 197.96	134.78 \pm 511.65
K2C6A_HUMAN	P02538	1,110	1,605	60.40 \pm 187.98	50.20 \pm 119.98
SPRR3_HUMAN	Q9UBC9	695	1,302	59.83 \pm 161.34	44.62 \pm 112.55
UROM_HUMAN	P07911	489	263	50.09 \pm 50.61	57.62 \pm 48.66

each protein, for each individual sample. Thereafter, the resulting values are averaged across the samples. This results in a value for each group, which is plotted as X- and Y-coordinates in the scatter plot. Alongside with the abundance metric chosen by the user the protein graph also provides another two dimensions of visualizing differences between the sample groups, namely: the number of peptides derived from each precursor protein, which correlates to the dot size, and the absolute value of the difference between the chosen abundance metric between the groups, which is proportional to the color. The color scale used is a standard diverging scale provided by Plotly (RdYlGn). Since the protein abundance for each group is plotted along the X- and Y-axis, the diagonal represents the point of equal abundance between the groups. Furthermore, the standard deviation of the abundance metric is off by default but can be displayed in the protein graph.

The protein graph may be manipulated in multiple ways. All graphs created in Peptimetric are generated with Plotly's graphing tools, which provides the user with a built-in modebar containing tools for e.g., downloading the plots. Hovering over a protein in the protein graph displays the metrics for that specific protein. Simultaneously, hovering over a protein will produce a sample graph (see **Figures 2B,C**),

containing the abundance metric for each individual sample for the protein.

All proteins present in the samples are summarized in a protein table (see **table 3**). The protein table contains the UniProt mnemonic protein identifier, UniProt id, number of peptides, abundance metric value, the standard deviation and the *p*-value (although the *p*-value is only shown if the given protein is present in three or more samples in each group) for the given abundance metric between the two groups. The table may be filtered with regular expression syntax, e.g.: >, \leq , \geq , = (for more see <https://dash.plotly.com/datatable/filtering>).

The user is also able to search for all proteins present in the samples, that have not been removed by any potential cutoffs. Searching for, clicking on, or selecting a protein in the table highlights it in the protein graph, making it possible to create a peptide graph for the highlighted protein.

3.2 Peptide View

Once a protein is selected in the protein view, a peptide view may be generated to view the peptides of the selected protein. Doing so results in a peptide graph alongside a peptide table. The peptide graph follows a visualization convention where the peptides are

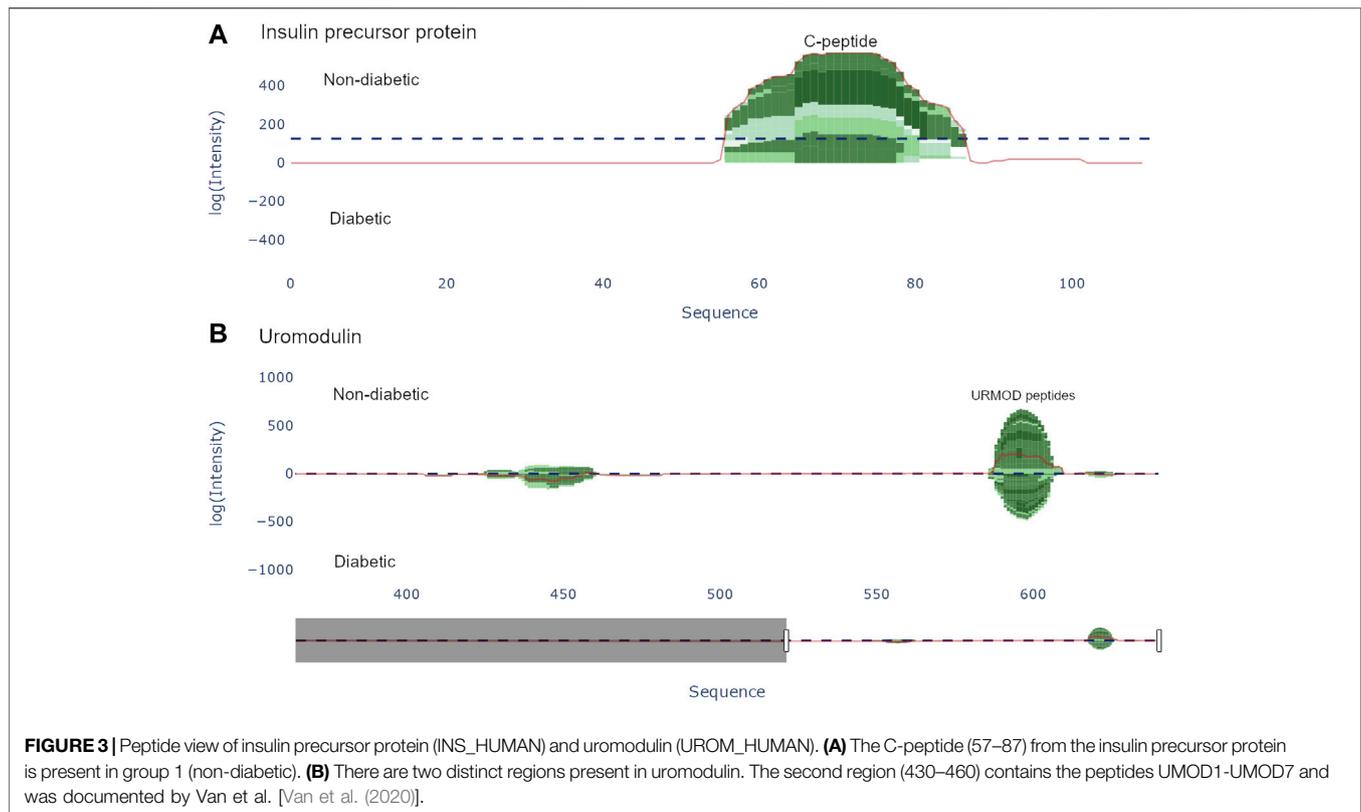


TABLE 4 | Peptide table for uromodulin (UROM_HUMAN). The table contains the five most abundant peptides in both groups.

Peptide	Start	End	Intensity ND \pm SD	Intensity D \pm SD
SGSVIDQSRVLNLGPITRKGVQ	588	610	4.68 \pm 14.16	0.59 \pm 2.30
VGGTGMFTVRM	449	460	2.39 \pm 5.31	2.11 \pm 5.70
SGSVIDQSRVLNLGPITRK	588	607	2.24 \pm 6.00	0.18 \pm 0.71
VGGTGMFTVRMA	449	461	1.95 \pm 5.17	2.18 \pm 5.09
LQPMVSALNIR	438	449	1.62 \pm 4.42	2.14 \pm 5.70

mapped onto the complete protein sequence as barcharts. To allow for easy visualization and quantification of peptide abundance and coverage, the color of the bars is proportional to the number of overlapping peptides at each amino acid position, whereas the height of the bars is proportional to the used abundance metric, see **Figure 3**. The discrete color scale applied to the bars is a subset derived from the “algae” color scale provided by Plotly. Similarly to the protein view, spectral count and intensity are available abundance metrics, and the user may easily switch between the two.

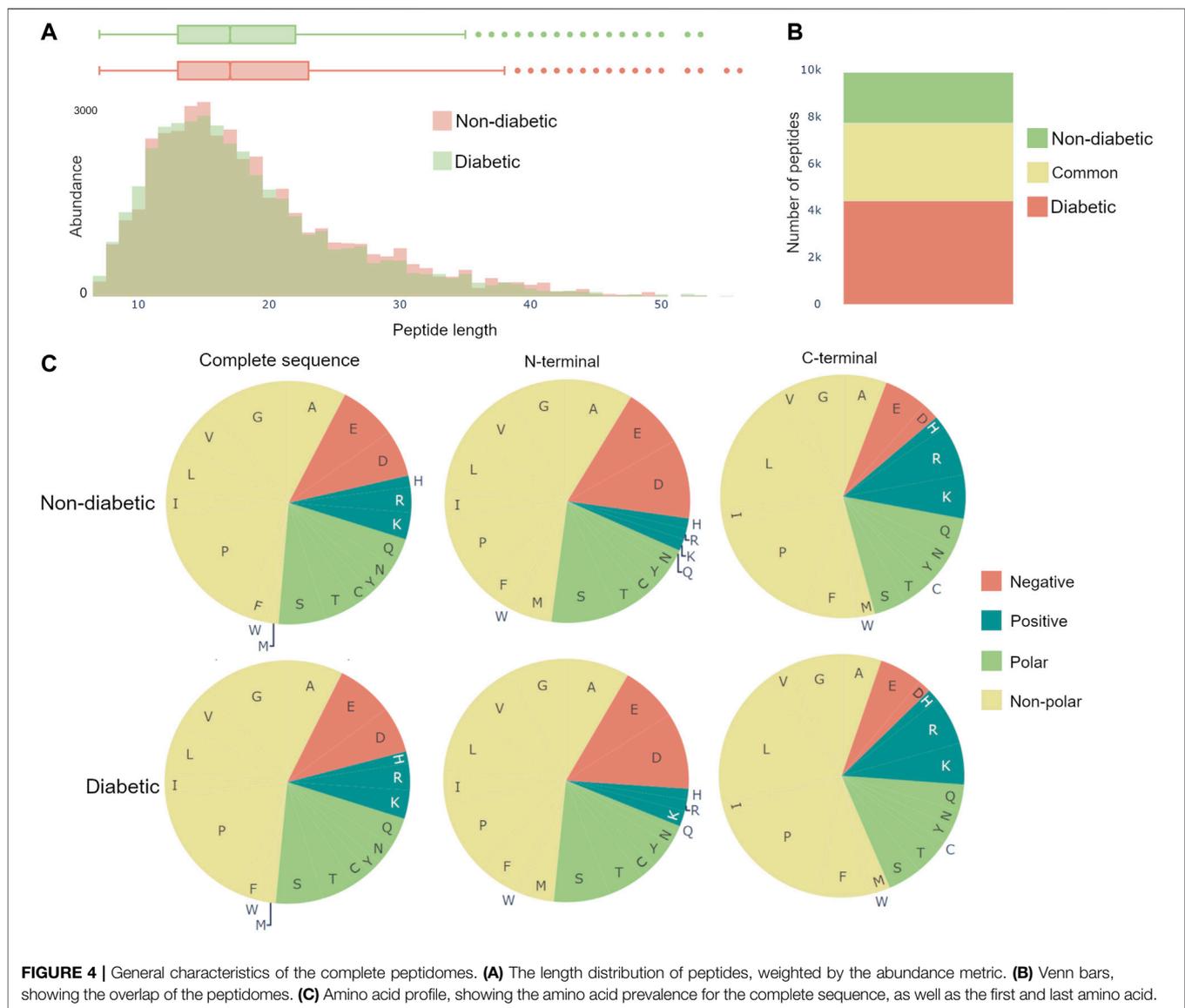
There are two options for graphically representing the peptidome of the input files: either as individual samples, or by averaging each group. If viewing each individual sample, the samples are stacked, whereas when viewing the group average, the intensity at each position is averaged. To guide the user to interesting regions of difference, a line corresponding to the difference of the height at each position is plotted. Additionally, a horizontal line representing the weight, i.e. the

sum of the chosen abundance metric. All the components in the graph are displayed as individual traces, and may be hidden/shown.

A table showing the peptides in the samples is situated adjacent to the peptide graph. The table shows the peptide amino acid sequence, the start and end position of the sequence, as well as the group mean, standard deviation and p -value of the chosen abundance metric (see **table 4**). If a peptide isn’t present in three or more samples in both groups, the p -value will not be calculated and -1 is presented. The table is filterable and sortable in a similar manner as the protein table. Each sequence in the table is selectable and doing so highlights the peptide region in the peptide graph.

3.3 General Characteristics

The dynamics of the peptidome are largely due to enzymatic activity, resulting in variations in peptide length and amino acid profile of the peptides N- and C-terminals. Therefore, Peptimetric



includes three relevant general characteristics, which may be applied to either the peptides in the entire proteome or in the chosen protein. These include: a histogram over peptide length, a bar chart showing peptide overlap between the groups, and pie charts displaying the amino acid profile, see **Figure 4**.

The length distribution is created by weighting each peptide with the value of the specific abundance metric, and thereafter counting the number of peptides with each length. Therefore, if spectral count is chosen as the abundance metric, a peptide with the spectral count value of 10 will contribute twice as much to the length distribution as a peptide with a spectral count value of 5.

The amino acid profiles contain pie charts showing the distribution of amino acids for the complete amino acid sequence, the first amino acid (N-terminal) and the last amino acid (C-terminal), as can be seen in **Figure 4C**. The abundance of an amino acid is calculated in the same ways as in the length distribution, as described above. The color scheme for the pie

charts is created based on the polarity and acidity of the amino acids and was chosen to accommodate for color blindness [Alberts et al. (2015)].

3.4 Exploring the Dataset

Investigating the data yielded an overview of the most abundant proteins in the samples. After normalizing on global values (not applying any cutoffs) and using the sum of intensities as the abundance metric, the most abundant proteins were albumin, hemoglobin alpha, uromodulin, small proline-rich protein 3 (SPRR3) and keratin type II cytoskeletal 6A, which is in line with the results from Van et al. [Van et al. (2020)]. Additionally, apolipoprotein A1 and alpha-1-antitrypsin (A1AT) were noticeably more abundant in the diabetic samples, whereas insulin precursor protein is only present in non-diabetic samples (see **Figure 2A**). It is worth to note that the samples vary a lot, and no difference was statistically significant on the

protein level. Similar conclusions were drawn with all different abundance metrics.

Van et al. (2020) used a combination of discovery and targeted peptidomics workflows, alongside specific selection strategies to finally distinguish 7 C-terminal peptides (UMOD1-UMOD7, most significantly SGSVIDQSRVNLGPITRK, 588–606) from uromodulin as biomarkers for type-1-diabetes. When investigating the dataset, many peptides derived from this region were indeed identified (alongside the region 430–460), and differential expressions between the two groups were observed, see **Figure 3**. As mentioned above insulin precursor protein was not found in the samples taken from the patients with type-1 diabetes. Investigating insulin precursor protein in the peptide view showcases that all peptides are derived from the region 57–87, which corresponds to the well documented C-peptide [Wahren et al. (2000)]. Although there was only modest overlap between the peptidomes of the groups (~30%), there were no apparent differences regarding peptide length or amino acid profile between the groups (see **Figure 4**).

3.5 Limitations

As mentioned, Peptimetric uses a local database fetched from UniProt to process the input data to reduce loading times. This results in the disadvantage of limiting the number of species accessible for analysis as well as not having the latest version of the UniProt database.

Herokus servers only allow for a limited computation time to each process (30 s), and processes which surpass the limit are aborted. This may occur if the user tries to input several large files with a slow internet connection. If this process is aborted, the user will not be able to analyze the data, and no partly uploaded data is stored on the server. If this occurs, concatenating some of the files manually is recommended, since this will reduce processing time.

Despite these limitations, Peptimetric serves as an interactive platform for fast and user-friendly exploration of MS/MS data and for detailed investigations and visualizations of complex data sets.

4 CONCLUSION

Peptimetric allows researchers to effectively and dynamically process, investigate and explore their peptidomic data. These traits are highly sought after in a landscape where the throughput of data is ever growing and where computational algorithms play an increasingly important role. By implementing an interface where the user easily gets an overview of both the proteome and peptidome of a sample, we have created a tool with high applicability to various peptidomic projects. To illustrate the

utility of Peptimetric, we applied it to a dataset generated by Van et al. (Van et al. (2020)), where the urinary peptidomes of type-1-diabetics were studied.

We envision Peptimetric being used to complement other softwares, such as Peptigram, in the early stages of peptidomic data analysis, where overview, exploration and quantification play important roles in the identification of precursor proteins and peptides of interest. Thereafter, one may export the data retrieved from Peptimetric for further analysis of e.g., post translational modifications, enzyme cut sites, screenings for bioactive peptides or for customized statistical analyses. This methodology applies to a variety of possible studies as peptidomics has relevance in basic research and clinical studies.

DATA AVAILABILITY STATEMENT

The datasets analyzed for this study can be found in the ProteomeXchange Consortium *via* the 264 PRIDE partner repository with the dataset identifier PXD012210 [Protomecentral (2021), <http://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PX012210>]. The complete repository, including the data files and requirements, are available at GitHub <https://github.com/ErikHartman/Peptimetric> and is freely available under an MIT license.

AUTHOR CONTRIBUTIONS

EH and SM wrote the complete manuscript, which was revised by SK and AS. EH and SM wrote the code. EH, SM, SK and AS developed the ideas behind Peptimetric.

FUNDING

This work was supported by grants from Alfred Österlund Foundation, Edvard Welanders Stiftelse and Finsentistelsen (Hudfonden), the LEO Foundation, the Swedish Government Funds for Clinical Research (ALF) and the Swedish Research Council (project 2017-02341, 2020-02016).

ACKNOWLEDGMENTS

We would like to thank Arvid Larsson, who supported and advised us through the entirety of this project and did not demand nor want anything in return.

REFERENCES

- Alberts, B., Johnson, A., Lewis, J., Roberts, M. R. K., and Walter, P. (2015). *Molecular Biology of the Cell*. New York: Garland Science.
- Alexandridou, A., Dovrolis, N., Tsangaris, G. T., Nikita, K., and Spyrou, G. (2011). Pepsolve: a Web Server for Peptide Analysis, Clustering and Visualization. *Nucleic Acids Res.* 39, W381–W384. doi:10.1093/nar/gkr318

- Consortium, T. U. (2020). Uniprot: the Universal Protein Knowledgebase in 2021. *Nucleic Acids Res.* 49 (D1), D480–D489. doi:10.1093/nar/gkaa1100
- Das, P., Sercu, T., Wadhawan, K., Padhi, I., Gehrmann, S., Cipcigan, F., et al. (2021). Accelerated Antimicrobial Discovery *via* Deep Generative Models and Molecular Dynamics Simulations. *Nat. Biomed. Eng.* 5, 613–623. doi:10.1038/s41551-021-00689-x
- Dufresne, J., Florentinus-Mefailoski, A., Ajambo, J., Ferwa, A., Bowden, P., and Marshall, J. (2017). Random and Independent Sampling of Endogenous Tryptic

- Peptides from normal Human Edta Plasma by Liquid Chromatography Micro Electrospray Ionization and Tandem Mass Spectrometry. *Clin. Proteom* 14, 41. doi:10.1186/s12014-017-9176-7
- Geyer, P. E., Voytik, E., Treit, P. V., Doll, S., Kleinhempel, A., Niu, L., et al. (2019). Plasma Proteome Profiling to Detect and Avoid Sample-Related Biases in Biomarker Studies. *EMBO Mol. Med.* 11, e10427. doi:10.15252/emmm.201910427
- Griffin, N. M., Yu, J., Long, F., Oh, P., Shore, S., Li, Y., et al. (2010). Label-free, Normalized Quantification of Complex Mass Spectrometry Data for Proteomic Analysis. *Nat. Biotechnol.* 28, 83–89. doi:10.1038/nbt.1592
- Hartman, E., Wallblom, K., van der Plas, M. J. A., Petrlova, J., Cai, J., Saleh, K., et al. (2021). Bioinformatic Analysis of the Wound Peptidome Reveals Potential Biomarkers and Antimicrobial Peptides. *Front. Immunol.* 11, 620707. doi:10.3389/fimmu.2020.620707
- Heroku (2021). *Heroku*.
- Jehl, P., Manguy, J., Shields, D. C., Higgins, D. G., and Davey, N. E. (2016). ProViz—a Web-Based Visualization Tool to Investigate the Functional and Evolutionary Features of Protein Sequences. *Nucleic Acids Res.* 44, W11–W15. doi:10.1093/nar/gkw265
- Karpievitch, Y. V., Dabney, A. R., and Smith, R. D. (2012). Normalization and Missing Value Imputation for Label-free LC-MS Analysis. *BMC Bioinformatics* 13. doi:10.1186/1471-2105-13-S16-S5
- Keim, D., Mansmann, F., Schneidewind, J., Thomas, J., and Ziegler, H. (2008). *Visual Analytics: Scope and Challenges*. Springer. doi:10.1007/978-3-540-71080-6_6
- Klont, F., Bras, L., Wolters, J. C., Ongay, S., Bischoff, R., Halmos, G. B., et al. (2019). Assessment of Sample Preparation Bias in Mass Spectrometry-Based Proteomics. *Anal. Chem.* 90, 5405–5413. doi:10.1021/acs.analchem.8b00600
- Lai, Z., Petrera, A., and Schilling, O. (2014). The Emerging Role of the Peptidome in Biomarker Discovery and Degradome Profiling. *De Gruyter* 396 (3), 185–192. doi:10.1515/hsz-2014-020
- Manguy, J., Jehl, P., Dillon, E. T., Davey, N. E., Shields, D. C., and Holton, T. A. (2017). Peptigram: A Web-Based Application for Peptidomics Data Visualization. *J. Proteome Res.* 16, 712–719. doi:10.1021/acs.jproteome.6b00751
- Martin-Campos, T., Mylonas, R., Masselot, A., Waridel, P., Petricevic, T., Xenarios, I., et al. (2017). Msviz: A Graphical Software Tool for In-Depth Manual Validation and Quantitation of post-translational Modifications. *J. Proteome Res.* 16, 3092–3101. doi:10.1021/acs.jproteome.7b00194
- Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., et al. (2021). Pfam: The Protein Families Database in 2021. *Nucleic Acids Res.* 49, D412–D419. doi:10.1093/nar/gkaa913
- Nusinow, D. P., Szpyt, J., Ghandi, M., Rose, C. M., McDonald, E. R., 3rd, Kalocsay, M., et al. (2020). Quantitative Proteomics of the Cancer Cell Line Encyclopedia. *Cell.* 180 (2), 387–402.e16. doi:10.1016/j.cell.2019.12.023
- Plotly (2021). *Plotly Dash Library*.
- Preston, G. W., Yang, L., Phillips, D. H., and Maier, C. S. (2020). Visualisation Tools for Dependent Peptide Searches to Support the Exploration of *In Vitro* Protein Modifications. *PLOS ONE* 15, e0235263. doi:10.1371/journal.pone.0235263
- Protomecentral (2021). *Proteomexchange*.
- Sauteraud, R., Jiang, M., and Gottardo, R. (2021). *Pviz: Peptide Annotation and Data Visualization Using Gviz*. R package version 1.26.0.
- Schneiderman, B. (1996). *The Eyes Have it: A Task by Data Type Taxonomy for Information Visualizations*. IEEE Computer Society Press. doi:10.1109/VL.1996.545307
- Schrader, M., Schulz-Knappe, P., and Fricker, L. (2013). Historical Perspective of Peptidomics. *EuPA Open Proteomics* 3, 171–182. doi:10.1016/j.euprot.2014.02.014
- Timp, W., and Timp, G. (2020). Beyond Mass Spectrometry, the Next Step in Proteomics. *Sci. Adv.* 6, eaax8978. doi:10.1126/sciadv.aax8978
- Van, J. A. D., Clotet-Freixas, S., Zhou, J., Batruch, I., Sun, C., Glogauer, M., et al. (2020). Peptidomic Analysis of Urine from Youths with Early Type 1 Diabetes Reveals Novel Bioactivity of Uromodulin Peptides *In Vitro*. *Mol. Cell Proteomics* 19, 501–517. doi:10.1074/mcp.RA119.001858
- Wahren, J., Ekberg, K., Johansson, J., Henriksson, M., Pramanik, A., and Johansson, B.-L. (2000). Role of C-Peptide in Human Physiology. *Am. J. Physiology-Endocrinology Metab.* 278 (5), E759–E768. doi:10.1152/ajpendo.2000.278.5.e759
- Wijk, J. V. (2005). “The Value of Visualization,” in *IEEE Visualization* (Minneapolis, MN, USA: IEEE). doi:10.1109/VISUAL.2005.1532781

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Hartman, Mahdavi, Kjellström and Schmidtchen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.