Unveiling antimicrobial peptide–generating human proteases using PROTEASIX

DOI:
10.1016/j.jprot.2017.02.016

Document Version
Accepted author manuscript

ABSTRACT

Extracting information from peptidomics data is a major current challenge, as endogenous peptides can result from the activity of multiple enzymes. Proteolytic enzymes can display overlapping or complementary specificity. The activity spectrum of human endogenous peptide-generating proteases is not fully known. Hence, the indirect study of proteolytic enzymes through the analysis of its substrates is largely hampered. Antimicrobial peptides (AMPs) represent a primordial set of immune defense molecules generated by proteolytic cleavage of precursor proteins. These peptides can be modulated by host and microorganismal stimuli, which both dictate proteolytic enzymes' expression and activity. Peptidomics is an attractive approach to identify peptides with a biological role and to assess proteolytic activity. However, bioinformatics tools to deal with peptidomics data are lacking. PROTEASIX is an excellent choice for the prediction of AMPs-generating proteases based on the reconstitution of a substrate's cleavage sites and the crossing of such information with known proteases' specificity retrieved by several publicly available databases. Therefore, the focus of the present tutorial is to explore the potential of PROTEASIX when gather information concerning proteases involved in the generation of human AMPs and to teach the user how to make the most out of peptidomics results using PROTEASIX.

Keywords: Bioinformatics, Antimicrobial Peptides, Proteases, Knowledgebase, Pathophysiology

Introduction

Making sense of peptidomics results is a major current challenge, as endogenous peptides can result from the activity of multiple enzymes. Such proteolytic enzymes can display overlapping or complementary specificity. Furthermore, the activity spectra of human endogenous peptide-generating proteolytic enzymes are not fully known. Hence, the indirect study of proteolytic enzymes via alterations at their substrate level is largely hampered.

Antimicrobial peptides (AMPs) constitute a family of defense molecules extraordinarily diverse with a continuum of activity spectra, ranging from two to more than 150 amino acids, acting either intracellularly or extracellularly, directly over microorganisms or through activation of other immune mediators. Most AMPs are found extracellularly, primarily at the defense barriers where they are thought to be generated by proteolytic cleavage of precursor proteins. Accordingly,

thousands of fragmentation peptides can be found in the extracellular milieu [1–3]. Among these, a growing number of AMPs has been identified and the corresponding antimicrobial activity discerned [4–7]. For instance, the salivary short form of Thymic stromal lymphopoietin (isoform 2) results from proteolytic cleavage of Thymic stromal lymphopoietin and displays antimicrobial properties [4,8]. Similarly, eotaxin-3/CCL26 is generated by mast cell protease-mediated cleavage during allergic inflammatory responses to bacterial infections of the airways and is active against several airway pathogens, including *Streptococcus pneumoniae* and *Staphyloccous aureus* [5]. However, with few exceptions [9–11], little is known concerning how variations at the peptidome level may be associated with human diseases and mediated or triggered by pathogenic microorganisms. Therefore, the identification and prediction of proteases implicated in AMPs generation might help to understand some disease mechanisms and the role of disease-associated as well as pathogen-associated peptides as well as to pinpoint potential therapeutic targets.

With the flourishing of proteomics, many tools for protein-protein interactions' annotation and prediction (e.g. STRING [12]) as well as curated knowledgebase (e.g. UniProt [13]) have emerged. In contrast, in the field of peptidomics there is a lack of supporting bioinformatics tools, which may stem from the larger complexity of the human peptidome over its proteome counterpart and from the fact that most protein-centered studies employ enzymatic treatment (e.g. tryptic digestion), which obliterates any information provided by endogenous peptide fingerprints. In order to address this challenge, PROTEASIX [14] can be of tremendous utility, since it allows the accurate elucidation of cleavage site from peptide sequence inputs, thus uncovering the proteases potentially implicated in endogenous peptide formation. Still, its correct exploitation requires users to be fully elucidated on how to use this tool and to know what it can provide them.

With the application of PROTEASIX [15] to a set of proteolysis products, the user can find the answers to the following questions: What are the known proteases and their target cleavage sites (observed and predicted)? For a given peptide and the respective precursor protein, what are the cleavage sites that led to its production? Are these peptides the product of observed or predicted proteolysis? What are the functions and cellular locations of proteases and their substrate proteins in the species in question? Which are the specific cleavage sites for a given protease? Therefore, the focus of the present tutorial is to explore the potential of PROTEASIX in gathering information concerning proteases involved in the generation of human AMPs and to teach the user how to make the most out of peptidomics data using PROTEASIX.

PROTEASIX: a peptide centric tool

PROTEASIX [14] is an open-source, updated, peptide-centric knowledgebase and tool for *in silico* prediction or retrieval of proteases known to be involved in native peptide generation. It can be used for both small and large-scale investigations in an automatic fashion (for the updated version, please refer to the currently available at http://proteasix.cs.manchester.ac.uk). In order to organize data from a variety of public sources, PROTEASIX employs its own ontology (PROTEASIX Ontology) [15], which re-uses parts of other Protein, Gene, Chemical and Biological Ontologies. By doing so, PROTEASIX knowledgebase supports the PROTEASIX tool, allowing the linkage of peptide fragments to their corresponding proteases as well as the uncovering of possible disease and innate response mechanisms.

Proteolytic cleavage requires the recognition of short amino acid sequences, also called motifs or cleavage sequences (CS). In turn, proteases (both endopeptidase and exopeptidase) exhibit varying binding specificity/affinity for the recognition of such motifs, spanning from being strictly restricted to one/few critical amino acids in specific positions, to being unspecific for generic amino acids or groups thereof [16,17]. In PROTEASIX, each protease is firstly associated with their corresponding CS. By aligning CS with the scissile bond (a covalent chemical bond susceptible to enzymatic cleavage such as proteolytic hydrolysis) positions on the peptide sequence provided by the user, PROTEASIX attempts to determine which enzymes generate such peptides; entries that cannot be aligned are, thus, discarded.

Since co-location of proteases and respective substrates is imperative to proteolysis, in PROTEASIX all proteases and associated substrates are annotated with the Gene Ontology Cellular Component in order to verify whether a common location can be found and thus increase the confidence of the cleavage prediction. Moreover, because this binding specificity/affinity depends on the amino acid sequence, mutations may lead to alterations in proteolytic activity and explain why some proteases/fragmentation peptides appear as deregulated across human disorders.

PROTEASIX step-by-step

1: The information required for each peptide consists of a peptide identifier (anything will do - e.g. a group identifier, an ordinal attribute, a random number) as well as the UniProt Accession Number or identifier of the parent protein from which the peptide is derived, together with its start and stop amino acid positions with respect to the parent protein's sequence (**Supplementary Table S1-**

Input Sheet). In order to retrieve proteins matching your peptides sequences, please refer to the UniProt Peptide search tool (http://www.uniprot.org/peptidesearch/) or the Protein Information Resource (PIR) tool (http://research.bioinformatics.udel.edu/peptidematch/index.htm). In order to correctly arrange these, prepare a (e.g. excel) spreadsheet consisting of 4 columns in the following order: a peptide ID (the numbers 1 to 216 in this case), the Uniprot Accession Number for the precursor protein, the Start and Stop amino acids positions mapping the peptide to its precursor protein.

2*:* Go to Proteasix website (http://proteasix.cs.manchester.ac.uk) and select "Prediction tool" (Note: by the time this tool is employed in the future, new protease/substrate combinations may have been added. For this reason and for reference purposes, the version herein used is stored in a separate link http://proteasix-v2.cs.manchester.ac.uk)

3: Copy and paste the above columns directly from an Excel or tab-delimited file to the field "Input peptide list". You can input up to 350 rows. Move to the "Next Step" (**Figure 1A**).

4: The algorithm retrieves several annotation properties for each protein class created from the UniProtKB, including its whole sequence. The input peptide sequence for any query is extracted, and the *N*- and *C*-terminus are reconstructed. The stored CS and the flanking amino acids are retrieved. The sequence of its corresponding *N*- and *C*-terminus are matched against the CS. If a successful match has been previously observed, a more detailed match is triggered. If an exact match was not previously annotated, PROTEASIX carries out a predictive approach for the N-terminus and C-terminus classes with sequences that remained unmatched by replacing or eliminating amino acids and matching the end result with known protease-substrate combinations. Further validation is achieved on the observed and predicted proteolysis, which is found by confirming if the source organism for the protease and the substrate are the same and if both the protease and the substrate are co-located.

5: When the message "The automatic reconstruction of *N*- and *C*-terminal Cleavage Sites is finished" is displayed, click "Next Step" to see the "Observed proteases" (**Figure 1B**).

6: When the message "The identification of observed Protease/Cleavage Site association is finished" is displayed, click "Next Step" to see the "Predicted proteases" (**Figure 1C**).

7: When the message "The probability of Protease/Cleavage Site association has been calculated" is displayed, click "Next Step" to see the "Detailed results" (**Figure 1D**).

8: Results can be downloaded by selecting "Copy and paste detailed results" and then "Download detailed results". Detailed results can be sorted as desired by the user prior to download and can

then be exported as table format (**Figure 2**).

9: Curate the downloaded results according to the desired confidence in the assignments and predictions. For instance, remove predictions based on cleavages observed in different substrates or positions, or those observed on the same substrates and positions but in different taxa. If high-confidence assignments or predictions were achieved, the user may completely exclude medium- and low-confidence ones (go through the different sheets in **Supplementary Table S1**).

10: Curated results can be subject of further analysis using other bioinformatics tools. In this case, we used Cytoscape's plug-in ClueGO for easy visualization of the Proteasix output and to extract biological information [18,19]. For this purpose, create a (e.g. Excel) table containing 3 columns: a column containing substrate identifiers (**Substrate** in **Figure 3A**), a column containing interactor identifiers (**CS sequence** in **Figure 3A**) and a column containing protease identifiers (**Protease** in **Figure 3A**). Open Cytoscape and create a new empty network (**Figure 3B**). Import a new network from file and choose the Excel table you have just created (**Figure 3C**). From the scroll-down menu just above each column, the user specifies what should be used as a target (red-orange dot) and as a source (green dot) node. Irrespectively of which column is assigned as target or source nodes, the same information should be achieved in the end, because our network is undirected (the same would not be possible when using directed networks as in signaling cascades). However, because our search went from peptides to proteases, use the first as source and the second as target. In this case, the interactor (edge) should always be the cleavage sequence (**Figure 3D**). Once the network is imported, the user may modify the layout and perform further analysis using the built-in tools as described elsewhere [18,19].**11:** Test protease prediction/assignment against a random dataset as a control in order to rule out random assignments. Use an unrelated but similar-size peptidome dataset and follow the exact same steps as above described. Check the distribution of the most common proteases. If the same proteolytic enzymes appear as the more frequent ones in both datasets (test and control), its representativeness should be questioned (see below). This control should become common practice by the acquainted user.

PROTEASIX-based analysis of AMPs for proteases prediction

Our group is currently working on expanding a curated in-house database of human AMPs identified throughout distinct body fluids (mostly saliva, milk, blood, sweat and urine samples) and active against forty bacterial, fungal or viral species (*Bastos P et al., Human Antimicrobial Peptides in Bodily Fluids: Current Knowledge and Therapeutic Perspectives on the Post-antibiotic Era.* 2017

Medicinal Research Reviews DOI: 10.1002/med.21435). Data contained in this repertoire served herein as an illustration of a big data ("omics") problem that can be tackled using PROTEASIX. Provided with the corresponding amino acids sequences, we employed PROTEASIX to depict for the first time which human proteases generate antimicrobial peptides in a wide array of human biofluids.

Starting from 229 unique sequences from AMPs present throughout human body fluids and encoded by 79 different genes, PROTEASIX provided 1,696 outputs (i.e. protease-CS-substrate pairs) (**Supplementary Table S1- Output Sheet**). As a valuable feature of PROTEASIX, these outputs are discriminated between taxonomic classification (even though we are referring to human peptides and proteases, few studies employed orthologue peptides of animal origin when performing *in vitro* studies as these are cheaper, but PROTEASIX accurately discriminate these). Also, PROTEASIX discriminates between proteolytic cleavages observed at the same or at different but similar substrates or positions, thus attributing weighted confidence levels (low, medium or high). As such, we removed those highly similar but observed only in slightly different substrates or positions (identified only as a last resource when using the interactive algorithm employed by PROTEASIX) as well as those observed on different taxa or predicted with medium to low confidence. Most importantly, PROTEASIX allows for almost automatic curation by the user as the outputs are duly annotated and can be sorted as required. By doing so, we reduced the initial list to 523 specific protease-CS-substrate pairs implicating 61 proteolytic enzymes (**Figure 4**, **Supplementary Table S2**), which were then visualized on Cytoscape (**Figures 5 and 6**).

As observed in **Figure 4**, the contribution of the different proteases for the human antimicrobial peptidome is uneven. Interestingly, protease frequency distribution was not clustered by families of proteases. For instance, one can find both prominent and poor contributors among the matrix metalloproteinase (MMP) (e.g. 26 cleavage sequences for MMP9 versus 4 for MMP1), kallikrein (KLK) (e.g. 21 cleavage sequences for KLK4 versus 2 for KLK14) and cathepsin (CTS) (e.g. 44 cleavage sequences for CTSD versus 2 for CTSK) family of proteases (**Figure 4**). This suggests that i) the mechanism of action of human proteases does not dictate and cannot predict if its products will have antimicrobial activity and ii) each protease is unique concerning its role in AMPs' generation and cannot be replaced even by similar enzymes. On the other hand, the differential contribution of very similar enzymes (belonging to the same family) to the human antimicrobial peptidome might be the result of expression in different tissue locations and body fluids, which is accompanied by different substrate availability.

Some proteases appear to be involved in the generation of a more diverse set of AMPs (**Supplementary Table S2** and **Figure 5**). Such enzymes include granzyme A (GZMA), cathepsin B (CTSB), cathepsin D (CTSD), cathepsin S (CTSS), cathepsin L1 (CTSL), neutrophil elastase (ELANE), kallikrein-2 (KLK2), kallikrein-4 (KLK4), kallikrein-6 (KLK6), meprin A subunit alpha (MEP1A), meprin A subunit beta (MEP1B), 72 kDa type IV collagenase 2 (MMP2), matrilysin (MMP7), Matrix metalloproteinase-9 (MMP9), macrophage metalloelastase 12 (MMP12), mollagenase 3 (MMP13), matrix metalloproteinase-14 (MMP14), signal peptidase complex catalytic subunit SEC11C (SEC11C). However, diversity does not strictly dictate how crucial a given enzyme may be, as it may actually turn out to be rather unspecific.

Prominent role and broader specificity and activity spectra of some of these proteases (CTSB, CTSD, ELANE, KLK4, MMP9, MMP12) can be depicted by larger outer nodes in **Figure 6.** In addition, some cleavage peptides such as cathelicidin antimicrobial peptide (CAMP), dermcidin (DCD), eotaxin (CCL11), basic salivary proline-rich protein 2 (PRB2), stromal cell-derived factor 1 (CXCL12), histone H4 (HIST1H4A), beta-casein (CSN2), protachykinin-1 (TAC1) were also much more interconnected than others, reflecting that these are targeted by multiple proteases as depicted by larger inner nodes on **Figure 6**. Therefore, some enzymes appear to generate multiple antimicrobial peptides by cleaving several different substrates and some substrates may also be amenable to proteolytic cleavage by several different proteases, showing considerable redundancy among antimicrobial peptide-generating human proteases.

Most of the predicted proteases were endopeptidases. Because endopeptidases are significantly more specific than exopeptidases, AMP-generating proteolytic cleavage in humans must be a considerably controlled and targeted process. Among those, metalloendopeptidases, serine hydrolases and cysteine-type endopeptidases are the most prevalent (**Figure 7**). Metalloendopeptidases are metalloproteases with endopeptidase activity whose catalytic functions require divalent metal cations as cofactors (e.g. Zinc and Calcium ions). These enzymes constitute the most diverse group of known proteases, containing active sites which can vary considerably. Despite being classically known for their extracellular/tissue remodeling involvement, metalloendopeptidases are also known to play a role in a host of cellular processes (including cellular signaling, cancer metastasis and developmental processes, and are activated by inflammatory processes, regulating immune cell function and development and being involved in disease progression in a substrate-specific way [20–22].

Cysteine-type endopeptidases include intracellularly-acting proteases such as calpain-1 and calpain-2 catalytic subunits, which act upon substrates involved in cytoskeletal remodeling and signal transduction, and caspase-1, which is involved in several processes including apoptosis and cell responses to the environment [23]. Other significant contributors for the human extracellular antimicrobial peptidome are aspartic-type endopeptidases such as pepsin A-3 and gastricsin (both involved in the food digestion process), cathepsin E (recognized for its role in the immune response primarily during MHC class II-mediated antigen presentation, activation-induced lymphocyte depletion in the thymus and in glial cell activation in the brain [24]), presenilin-1, which may regulate intracellular signaling and gene expression and cleaves E-cadherin, promoting the disassembly of the E-cadherin/catenin complex and increasing the pool of cytoplasmic beta-catenin, thus negatively regulating Wnt signaling [25], and cathepsin D, an intracellularly-acting protease also involved in MHC class II-mediated antigen presentation as well as the autophagy process. Together, these observations allow one to see that AMP-generating human proteases display a very diverse set of mechanisms of action and are involved in a wide range of biological functions not necessarily implicated in defense mechanisms.

In contrast to these overrepresented proteases, few seem to display beta-amyloid binding properties (**Figure 7**, e.g. beta-secretase 1, beta-secretase 2, insulin-degrading enzyme). However, such enzymes can exhibit unexpected expression patterns [26] and are also broad spectrum enzymes. For instance, insulin-degrading enzyme participates in the breakdown of insulin, islet amyloid polypeptide, glucagon, bradykinin, kallidin and other polypeptides, thus also playing a role in intercellular peptide signaling [27], but its membrane-associated isoform also acts as an entry receptor for varicella-zoster virus [28].

In contrast to endopeptidases, the human repertoire of AMP-generating exopeptidases was rather small (**Figure 7**, e.g. neprilysin, prolyl endopeptidase, macrophage metalloelastase, hepsin and beta-secretase 1). It is possible that the resulting cleavage fragments consisting only of one or two (terminal or penultimate peptide bond cleavage) amino acids together with an almost complete precursor protein would rather display poor antimicrobial activity, which could explain such underrepresentation. However, the biological implications of this observation are not clear and this smaller part of human immunity may well be proven critical as even underrepresented, exopeptidases might still have a significant role (e.g. the large MMP12 node on **Figure 7**).

As shown in **Figure 8**, most proteases participate primarily in extracellular matrix remodeling, protein processing as well as cellular responses to the environment, autophagy and other

intracellular signaling processes. However, a large portion of the observed or predicted antimicrobial peptide-generating human proteases display significant antimicrobial peptide production/secretion and neutrophil-mediated defenses functions. Moreover, while antimicrobial peptides are a part of the innate immune system, an important enrichment in adaptive immune response functions (dendrite cell maturation as well as B-cell differentiation) could be seen for their corresponding generating enzymes. Once again, this observation suggests that the generation of antimicrobial peptides in humans is a well-controlled, targeted and purposeful process.

Testing your data

We next decided to assess the robustness of PROTEASIX predictions and how confident we could be on these in order to rule out randomness in the results. Therefore, we compared the results obtained using the AMPs as input (**Supplementary Table S1- Input Sheet**) with the results obtained when the prediction was made using a different dataset. This control dataset was based on the same number of endogenous human peptides but strongly different in terms of involved biological processes and cell location as those were urinary peptides deregulated in chronic kidney disease [1].

We have thus applied the same analysis criteria, eliminating predictions using different taxa, low confidence predictions and observations in different substrates and locations. When the top 20 chronic kidney disease-associated versus AMP-generating proteases were compared, considerable differences were found (**Figure 9**). Of particular note, some enzymes such as KLK4 and MMP2 showed very high prevalence in the AMPs dataset but were found almost absent in the chronic kidney disease dataset. In contrast, others such as MMP12 and CTSB seemed to contribute to the formation of endogenous peptides associated with kidney disease much more than they do to the formation of AMPs. This result assured that the results provided by PROTEASIX were not random but specific to each inputted dataset.

FINAL REMARKS

The application of PROTEASIX to peptidomic datasets allowed the verification that most AMP-generating proteases display endopeptidase activity over serine, cysteine and aspartic acid residues and are most frequently represented by cathepsins and metalloproteases. Still, in order to fully elucidate the role of proteases in AMPs' generation, one should not disregard redundancy. Luckily, PROTEASIX informs the user of such redundancy and in this case suggested an also prominent role

of kallikreins, calpains, elastases/collagenases, caspases and beta-secretases as AMP-generating defense proteases.

PROTEASIX allows to accurately and insightfully unveil which human proteolytic enzymes are responsible for the generation of human endogenous AMPs present throughout human body fluids. PROTEASIX also depicts how these enzymes act synergistically and complementary with each other. With data retrieved from PROTEASIX, biological processes associated with such proteolytic activity activation not otherwise amenable to study at the substrate level can be depicted, which may help explaining pathophysiological phenomena and pinpointing promising candidate therapeutic targets. Altogether, PROTEASIX is a valuable tool for the exploration of interactions among proteases/cleavage peptides in a complex network of proteolytic phenomena, thus allowing not only the identification of promising targets (proteases and peptides) but also the downstream prediction of how these targets interact with and influence others. By doing so, the acquainted user may fully exploit PROTEASIX in order to unveil a whole proteolytic system starting with a set of endogenous peptides collected from specific biofluids.

ACKNOWLEDGEMENTS

Authors would like to acknowledge the initial funding partner, ProteasiX FP7-PEOPLE-2011-IEF (300582), and the Portuguese Foundation for Science and Technology (FCT), European Union, QREN, FEDER and COMPETE for funding RV (IF/00286/2015), the iBiMED (UID/BIM/04501/2013), UnIC (UID/IC/00051/2013) and QOPNA (UID/QUI/UI0062/2013) research units.

REFERENCES

1. Good DM, Zürbig P, Argilés A, Bauer HW, Behrens G, Coon JJ, et al. Naturally occurring human urinary peptides for use in diagnosis of chronic kidney disease. Mol. Cell. Proteomics. 2010;9:2424–37.

2. Stalmach A, Johnsson H, McInnes IB, Husi H, Klein J, Dakna M, et al. Identification of urinary Peptide biomarkers associated with rheumatoid arthritis. PLoS One. 2014;9:e104625.

3. Kistler AD, Serra AL, Siwy J, Poster D, Krauer F, Torres VE, et al. Urinary proteomic biomarkers for diagnosis and risk stratification of autosomal dominant polycystic kidney disease: a multicentric study. PLoS One. 2013;8:e53016.

4. Bjerkan L, Schreurs O, Engen S a, Jahnsen FL, Baekkevold ES, Blix IJ, et al. The short form of TSLP is constitutively translated in human keratinocytes and has characteristics of an antimicrobial peptide. Mucosal Immunol. 2014;8:1–8.

5. Gela A, Kasetty G, Jovic S, Ekoff M, Nilsson G, Morgelin M, et al. Eotaxin-3 (CCL26) exerts innate host defense activities that are modulated by mast cell proteases. Allergy Eur. J. Allergy Clin. Immunol. 2015;70:161–70.

6. Rydengård V, Shannon O, Lundqvist K, Kacprzyk L, Chalupka A, Olsson AK, et al. Histidine-rich glycoprotein protects from systemic Candida infection. PLoS Pathog. 2008;4.

7. Hong SW, Seo D-G, Baik JE, Cho K, Yun C-H, Han SH. Differential profiles of salivary proteins with affinity to Streptococcus mutans lipoteichoic acid in caries-free and caries-positive human subjects. Mol. Oral Microbiol. 2014;29:208–18.

8. Allakhverdi Z, Comeau MR, Jessup HK, Yoon B-RP, Brewer A, Chartier S, et al. Thymic stromal lymphopoietin is released by human epithelial cells in response to microbes, trauma, or inflammation and potently activates mast cells. J. Exp. Med. 2007;204:253–8.

9. Ostaff MJ, Stange EF, Wehkamp J. Antimicrobial peptides and gut microbiota in homeostasis and pathology. EMBO Mol. Med. 2013;5:1465–83.

10. Gusman H, Travis J, Helmerhorst EJ, Potempa J, Troxler RF, Oppenheim FG. Salivary histatin 5 is an inhibitor of both host and bacterial enzymes implicated in periodontal disease. Infect. Immun. 2001;69:1402–8.

11. Osaki T, Sasaki K, Minamino N. Peptidomics-based discovery of an antimicrobial peptide derived from insulin-like growth factor-binding protein 5. J. Proteome Res. 2011;10:1870–80.

12. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015;43:D447–52.

13. The UniProt Consortium. UniProt: a hub for protein information. Nucleic Acids Res. 2014;43:D204–12.

14. Klein J, Eales J, Zürbig P, Vlahou A, Mischak H, Stevens R. PROTEASIX: a tool for automated and large-scale prediction of proteases involved in naturally occurring peptide generation. Proteomics. 2013;13:1077–82.

15. Arguello Casteleiro M, Klein J, Stevens R. The PROTEASIX Ontology. J. Biomed. Semantics. 2016;7:33.

16. Turk BE, Huang LL, Piro ET, Cantley LC. Determination of protease cleavage site motifs using mixture-based oriented peptide libraries. Nat. Biotechnol. 2001;19:661–7.

17. Ng NM, Pike RN, Boyd SE. Subsite cooperativity in protease specificity. Biol. Chem. 2009. p. 401–7.

18. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13:2498–504.

19. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, et al. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics. 2009;25:1091–3.

20. Khokha R, Murthy A, Weiss A. Metalloproteinases and their natural inhibitors in inflammation and immunity. Nat. Rev. Immunol. 2013;13:649–65.

21. Van Lint P, Libert C. Chemokine and cytokine processing by matrix metalloproteinases and its effect on leukocyte migration and inflammation. J. Leukoc. Biol. 2007;82:1375–81.

22. Parks WC, Shapiro SD. Matrix metalloproteinases in lung biology. Respir. Res. 2001;2:10–9.

23. Alnemri ES, Fernandes-Alnemri T, Litwack G. Cloning and expression of four novel isoforms of human interleukin-1 beta converting enzyme with different apoptotic activities. J. Biol. Chem. 1995;270:4312–7.

24. Sealy L, Mota F, Rayment N, Tatnell P, Kay J, Chain B. Regulation of cathepsin E expression during human B cell differentiation in vitro. Eur. J. Immunol. 1996;26:1838–43.

25. Baki L, Marambaud P, Efthimiopoulos S, Georgakopoulos A, Wen P, Cui W, et al. Presenilin-1 binds cytoplasmic epithelial cadherin, inhibits cadherin/p120 association, and regulates stability and function of the cadherin/catenin adhesion complex. Proc. Natl. Acad. Sci. U. S. A. 2001;98:2381–6.

26. Bennett BD, Babu-Khan S, Loeloff R, Louis JC, Curran E, Citron M, et al. Expression analysis of BACE2 in brain and

peripheral tissues. J. Biol. Chem. 2000;275:20647–51.

27. Im H, Manolopoulou M, Malito E, Shen Y, Zhao J, Neant-Fery M, et al. Structure of substrate-free human insulin-degrading enzyme (IDE) and biophysical analysis of ATP-induced conformational switch of IDE. J. Biol. Chem. 2007;282:25453–63.

28. Li Q, Krogmann T, Ali MA, Tang W-J, Cohen JI. The amino terminus of varicella-zoster virus (VZV) glycoprotein E is required for binding to insulin-degrading enzyme, a VZV receptor. J. Virol. 2007;81:8525–32.

Figure legends

Figure 1- Screenshots of the step-by-step tutorial on how to use PROTEASIX from Input data to Download phase: A) Upper-left, step 3; B) Upper-right, step 5; C) Lower-left, step 6; D) Lower-right, step 8.

Figure 2- Screenshot of the output results in table format as generated by PROTEASIX.

Figure 3- Screenshots of the step-by-step tutorial on how to use PROTEASIX from Curated data to network establishment: A) Upper-left, data arrangement in table format; B) Upper-right, new network creation; C) Lower-left, data importation; D) Lower-right, source, target and interactor selection.

Figure 4- Bar chart depicting the contribution/frequency distribution of antimicrobial peptide-generating human peptidases for the human antimicrobial peptidome. Each number above the bar corresponds to all possible cleavage sequences on all targeted substrates of each enzyme. If a given antimicrobial peptide results from the combined activity of more than one enzyme, it contributes for the bar height of more than one enzyme.

Figure 5- Network depicting extracellular human antimocribial peptides (red nodes), generating proteases (blue nodes) and cleavage sequences (edges, line labels). Network displayed by inputing on Cytoscape the corresponding protease-antimicrobial peptide combinations provided by PROTEASIX as a circular layout. Arrows originate from proteases and point to the corresponding cleavage peptides.

Figure 6- Network depicting betweenness centrality between extracellular human antimicrobial peptides (inner circle) and corresponding generating proteases (outer circle). In this weighted network, a node's strength is given by its capacity, influence, frequency and connectivity, which are in turn supported by the sum of the weights of its adjacent edges. Bigger nodes and larger edges have a stronger influence over the whole network.

Figure 7- Network depicting antimicrobial peptide-generating proteases (inner small pie charts) and corresponding molecular function or activity. Edges (lines) are grouped together (bundled) so that functions (nodes) shared by a larger number of proteases are connected to these by stronger lines (bundles).

Figure 8- Pie charts depicting the distribution of A) Immune Functions and B) Biological processes of antimicrobial peptide-generating human proteases.

Figure 9- Paired bar chart depicting the contribution/frequency distribution of antimicrobial peptide-generating human peptidases vs chronic kidney disease-associated peptide-generating

proteases considering the top 20 most frequently observed proteases. Each number above the bar corresponds to all possible cleavage sequences on all targeted substrates of each enzyme. If a given antimicrobial peptide results from the combined activity of more than one enzyme, it contributes for the bar height of more than one enzyme. If a given protease does not generate a peptide which is in turn observed among the top 20 of the other dataset, it gets attributed the value of zero.

Supplementary Table S1- Input Sheet: Input data inputted into PROTEASIX search engine. **Output Sheet:** Non-curated output raw results table generated by PROTEASIX from the starting inputted data. **Curated Output Sheet**: Curated outputted results with all information provided by PROTEASIX. **Simplified Table Sheet**: Curated outputted results containing only information relevant for downstream bioinformatics analysis. **SimplifTableMergedCells(=to S2) Sheet**: Curated outputted results containing only information relevant for downstream bioinformatics analysis grouped by similar/merged cells (same as Supplemental Table S2 provided in word format).

Supplemental Table S2- Manually curated list of antimicrobial peptide-generating proteases retrieved using PROTEASIX, corresponding targeted peptides and recognition sequences with start and stop amino acids.