ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

Unveiling Encrypted Antimicrobial Peptides from Cephalopods' Salivary Glands: A Proteolysis-Driven Virtual Approach

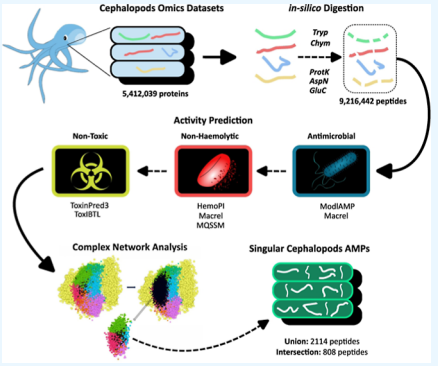
Guillermin Agüero-Chapin,* Dany Domínguez-Pérez, Yovani Marrero-Ponce,* Kevin Castillo-Mendieta, and Agostinho Antunes*

Cite This: *ACS Omega* 2024, 9, 43353−43367

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Antimicrobial peptides (AMPs) have potential against antimicrobial resistance and serve as templates for novel therapeutic agents. While most AMP databases focus on terrestrial eukaryotes, marine cephalopods represent a promising yet underexplored source. This study reveals the putative reservoir of AMPs encrypted within the proteomes of cephalopod salivary glands via in silico proteolysis. A composite protein database comprising 5,412,039 canonical and noncanonical proteins from salivary apparatus of 14 cephalopod species was subjected to digestion by 5 proteases under three protocols, yielding over 9 million of nonredundant peptides. These peptides were effectively screened by a selection of 8 prediction and sequence comparative tools, including machine learning, deep learning, multiquery similarity-based models, and complex networks. The screening prioritized the antimicrobial activity while ensuring the absence of hemolytic and toxic properties, and structural uniqueness compared to known AMPs. Five relevant AMP datasets were released, ranging from a comprehensive collection of 542,485 AMPs to a refined dataset of 68,694 nonhemolytic and nontoxic AMPs. Further comparative analyses and application of network science principles helped identify 5466 unique and 808 representative nonhemolytic and nontoxic AMPs. These datasets, along with the selected mining tools, provide valuable resources for peptide drug developers.



1. INTRODUCTION

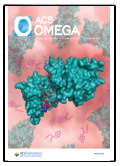
Antimicrobial resistance (AMR) poses a significant global public health threat, prompting the urgent need for novel antimicrobial agents. The diminishing effectiveness of conventional antibiotics against a wide range of resistant pathogens has driven the search for alternative solutions.[1] Antimicrobial peptides (AMPs) have emerged as promising candidates to address this crisis, offering versatile antimicrobial activities and diverse modes of action. Their therapeutic potential extends beyond the development of new antibiotics to combat multidrug-resistant bacteria.[2,3] AMPs also hold promise in the creation of agents with antitumoral, antiviral, antifungal, and other therapeutic properties.[4]

To fully harness the potential of AMPs, extensive efforts have been made to compile and organize AMP-related information into specialized databases. Notable among these are databases like the antimicrobial peptide database,[5] collection of antimicrobial peptides,[6] and database of antimicrobial activity and structure of peptides,[7] which have been continuously updated. In addition to these, the StarPep database (StarPepDB) stands out as one of the most comprehensive curated repositories of AMPs, integrating unique entries from 42 AMP databases with their metadata.[8] These databases facilitate the study of AMP sequences, structures, activities, and other relevant information, significantly enhancing their potential translation into therapeutic interventions.

The origin distribution of AMPs has also been facilitated by databases. Most AMPs reported to date stem from eukaryotic origins, notably plants, animals, and fungi.[9] Antimicrobial properties have been attributed to various bodily fluids since 1885, including blood, sweat, saliva, plasma, white blood cell secretions, and granule extracts.[10] Historically, terrestrial eukaryotes have been a primary source of AMPs. However, more recently, marine organisms, particularly invertebrates, have gained prominence due to their robust and effective innate immune systems, enabling their survival for over 450 million years in diverse ecological niches.[11,12] The immense ecological diversity of marine environments provides a promising landscape for the discovery of AMPs with unique structures and potent antimicrobial activities. Notably, AMPs from marine invertebrates constitute a significant proportion, approximately 67% of all marine AMPs (statistics as of December 2022).[12]

Marine invertebrates, including shrimp, oysters, and horseshoe crabs, are known to consistently express AMPs.[13,14] For instance, horseshoe crabs produce highly effective AMPs like

Received: February 28, 2024
Revised: April 26, 2024
Accepted: April 30, 2024
Published: October 14, 2024



tachyplesin and polyphemusin, exhibiting antibacterial and antifungal properties at low micromolar levels.[11] Notably, polyphemusin, similar to several other AMPs, demonstrates antiviral activity against human immunodeficiency virus.[15] More recently, the exploration of marine invertebrates has expanded through omics techniques, offering greater sensitivity in detecting the presence of AMPs.[16] In this context, our research group identified AMPs within the ascidian's tunic and the salivary glands of *Octopus vulgaris* through shotgun proteomics analyses,[17,18] and more recently others found AMPs common within octopus skin mucus proteome.[19] The comprehensive discovery of AMPs in *O. vulgaris* was made possible by applying an optimized methodological workflow and utilizing a composite protein database constructed from proteomic and transcriptomic data of the cephalopods' salivary apparatus.[18,20,21] This database included 16,990 characterized AMPs from StarPepDB,[8] notably excluding the only 14 peptides registered for the Cephalopoda class, which are neuropeptides with no reported antimicrobial activity. Despite this, the proposed proteome-wide exploration predominantly detected histone-derived AMPs, ubiquitin-like AMPs, and bovine pancreatic trypsin inhibitor (BPTI)-related AMPs.[18] In order of discovery, these included analogues of well-known AMPs such as Buforin,[22] cgUbiquitin,[23] and BTP1.[24]

Cephalopods, known for their efficient predatory tactics involving a diverse array of substances, predominantly cephalotoxins and neurotoxins to immobilize prey,[25] possess omics data characterizing their salivary apparatus that holds the potential for venom-related proteins, toxins, and AMPs, as substantiated in previous research.[18,21] However, AMPs with encrypted sequences within longer transcripts or proteins, exemplified by cases such as histones,[26] those not constitutively expressed, or potentially disregarded by the computational omics workflow [e.g., small-size transcripts or protein fragments less than 100 amino acids (AAs)][16] can be unveiled by a comprehensive examination of a composite protein database sourced from the cephalopods' salivary apparatus.[20] This composite protein database was purposefully built for a proteome-wide AMPs discovery by including "noncanonical" proteins, exploring all the open reading frames (ORFs) from cephalopods' salivary glands transcriptomes, and proteins shorter than the TransDecoder default minimum protein length threshold of 100 AAs.[20]

In this context, this study focused on a privileged marine source represented by the salivary glands of cephalopods, where a potentially abundant reservoir of hidden AMPs is believed to exist. Our approach to unveil these cryptic AMPs involves the in silico proteolysis of the composite protein database originating from cephalopods' salivary apparatus. This enzymatic digestion is performed using proteases commonly employed in proteomics. Subsequently, the resulting peptide libraries, comprising millions of peptides, were subjected to in silico screening. During this screening, we consider essential AMP characteristics relevant for drug development and pay particular attention to their structural distinctiveness within chemical space.

The resulting mining workflow yields various AMP datasets, catering to a spectrum of research needs. These datasets range from those solely focusing on antimicrobial activity to a refined, distinct dataset consisting of nonhemolytic AMPs devoid of toxic attributes. These datasets are publicly accessible and offer valuable resources for peptide drug developers, adaptable to their specific requirements.

2. DATASETS AND METHODS

2.1. Omics Data as a Substrate for in Silico Proteolysis.

A version of the composite protein database, comprising various omics datasets sourced from the cephalopods' salivary apparatus, as reported in ref 20, served as the substrate for the in silico proteolysis. This composite database includes five distinct datasets originally labeled and referenced as follows:

- *Database A*—19,087 proteins derived from proteogenomic analyses of the *O. vulgaris* salivary apparatus, as reported by Fingerhut et al. (2018).[21]
- *Database C*—2427 proteins corresponding to the postsalivary glands (PSGs) of three *O. vulgaris* specimens, as detailed by Almeida et al. (2020).[18]
- *Database D*—84,778 proteins identified by TransDecoder across 16 publicly available transcriptomes from PSGs of 13 different cephalopod species.[18,20]
- *Database E*—5,106,635 six-frame translated proteins shorter than the TransDecoder default minimum protein length threshold of 100 AAs, which were not included in Database D.[18,20]
- *Database F*—720,910 six-frame translated proteins extracted from the ORFs from *O. vulgaris* PSGs transcriptomes that were not part of Database A.[20,21]

Database B was not considered for proteolysis because it contained characterized AMPs from the StarPepDB.[8] Putative duplicates in each database were removed, and then the databases were fused into a composite protein database, followed by a redundancy removal process with the cd-hit tool at 0.98 sequence identity (https://github.com/weizhongli/cdhit).[27] The seqkit tool (https://bioinf.shenwei.me/seqkit/) was used to assist in both duplicates removal and finding common sequences between two databases,[28] which allowed for an *all-vs-all* comparison among databases. The Jaccard index was used as a pairwise similarity metric.[29]

2.2. In Silico Proteolysis and Peptidomes Characterization. Five main proteases commonly used in proteomics: trypsin, chymotrypsin, proteinase K, AspN, and GluC were applied.[30] Peptidomes were generated using 13 distinct proteolysis protocols involving the action of one enzyme or two enzymes, which could be applied in sequential (S) or concurrent (C) mode. We performed the in silico enzymatic digestion using the Rapid Peptides Generator (RPG) tool (https://rapid-peptide-generator.readthedocs.io/en/latest/index.html).[31] The previously mentioned proteases were involved in the three digestion protocols (Table 1).

Table 1

One Enzyme	Two Enzyme Sequential Mode	Two Enzyme Concurrent Mode
Tryp	Tryp-Chym	Tryp-Chym
Chym	Tryp-Proteinase-K	Tryp-Proteinase-K
Proteinase-K	Tryp-GluC	Tryp-GluC
GluC	Tryp-AspN	Tryp-AspN
AspN		

The peptide libraries or peptidomes resulting from each proteolysis protocol were filtered following these steps: (i) retaining only peptides that were 6−40 AAs in length, (ii) removing duplicates, (iii) removing peptides sharing above 0.98 of sequence identity, (iv) leaving out peptides with nonstandard AAs. The seqkit and cd-hit tools were used to perform this

https://doi.org/10.1021/acsomega.4c01959
ACS Omega 2024, 9, 43353−43367

prescreening. Then, each peptide library was characterized based on its global peptide features, such as sequence length, AA frequency, isoelectric point (pI), global charge, global hydrophobicity, and global hydrophobic moment. The PDAUG package (https://github.com/jaidevjoshi83/pdaug) was used to calculate these features.[32]

2.3. Antimicrobial and Toxicity Screening. Each resulting peptide library is subsequently screened for promising AMPs, which had been revealed by the proteolysis step. To ensure accurate detections, we determined the final prediction output by the consensus agreement of three prediction models/tools. The screening of the 13 peptidomes started by the prediction of the antimicrobial activity using one model implemented in Macrel: (Meta)genomic AMP Classification and Retrieval[33] and other two from modlAMP.[34] The subcommand "macrel peptides" were used to run marcel on peptide libraries (https://github.com/BigDataBiology/macrel) while modlAMP used the data "AMPvsUniProt" for training its two implemented machine learning (ML)-based classifiers: modlAMP_Random Forest and modlAMP_Support Vector Machine (https://modlamp.org).

Subsequently, the toxicity, which is the most undesired property of AMPs for drug development, was assessed by the prediction of their hemolytic potential and their content of toxic signatures. The hemolysis prediction was also performed by macrel,[33] HemoPI,[35] and by a multiquery similarity searching model (MQSSM) developed in ref[36]. Since macrel output also provides hemolytic predictions for detected AMPs, "macrel peptides" were run as before (https://github.com/BigDataBiology/macrel). The standalone version of HemoPI was used (https://webs.iiitd.edu.in/raghava/hemopi/standalone.php), particularly its virtual screening option where the hybrid model is selected. The hybrid model considers the integration of motif- and SVM-based predictions. The MQSSM **I1**, the best model reported in ref.,[36] was constructed using the half-space proximal network (HSPN) projecting the chemical space of 2004 hemolytic peptides from StarPepDB. The HSPN was constructed without similarity cutoff, and the angular separation was used as a pairwise similarity metric. Subsequently, a representative hemolytic subset was extracted from the HSPN using the following parameters: hub-bridge centrality, global alignment, and a similarity cutoff of 0.8. This representative subset was further improved as described in ref.,[36] and it was finally used to build a MQSSM model using global alignment and a similarity cutoff of 0.40.

The detection of toxic signatures was performed by two models from ToxinPred3[37] and by ToxIBTL.[38] ToxinPred3 has implemented two prediction model types, a ML-based classifier trained with compositional features of peptides and the other a hybrid model combining two or more models including motif- and ML-based predictions (https://github.com/raghavagps/toxinpred3). ToxIBTL is a deep learning approach based on the integration of evolutionary information and physicochemical properties of peptides into the information bottleneck principle, and transfer learning to predict the toxicity of peptides (https://server.wei-group.net/ToxIBTL/Server.html). Venn diagrams were used for identifying consensus predictions among the outputs of the three prediction models.

In summary, the 13 proteolysis protocols rendered the following datasets: (i) peptide libraries (peptidomes), (ii) AMP consensus, (iii) nonhemolytic AMPs, and (iv) nonhemolytic/nontoxic AMPs. Peptide subsets corresponding to the 13 digestion protocols within each of the four datasets were concatenated and sequence redundancy was removed at 0.98 of identity with cd-hit.

2.4. Selection of Cephalopods Singular AMPs. A comparison of the nonredundant nonhemolytic and nontoxic AMPs to StarPepDB,[8] one of the most comprehensively reported peptide databases, was performed using the cd-hit-2d tool at 0.40, 0.50, 0.60, 0.70, 0.80, and 0.90 identity cutoffs. This was done to identify new peptide representations encoded in the cephalopods' proteome that differ from the previously reported peptides. Cephalopods singular peptides (CSPs) are considered those sharing sequence identities below the 0.40 threshold with StarPepDB members, while peptides with an equal or higher threshold were considered similar. Prior to the comparison, the StarPepDB's original space of 45,120 peptides was reduced to 32,863 by applying the cd-hit tool at 0.98 identity and retaining only peptides that ranged from 5 to 100 AAs in length and contained standard AAs.

2.4.1. Validating the Singularity of Cephalopods' AMPs Using Complex Networks. To validate the no relatedness of CSPs with respect to the known chemical space from StarPepDB,[8] both chemical spaces were represented as HSPNs.[39] The nonredundant nonhemolytic and nontoxic AMPs from cephalopods were divided into a set of CSPs (identity <0.40 with StarPepDB space) and a more-closely related set to StarPepDB (identity >0.40). These sets were then used together with the 32,863 peptides from StarPepDB to build HSPNs using the StarPep Toolbox.[40] Each peptide/node was represented by an optimized set of molecular descriptors, and the Euclidean distance metric with min−max normalization were applied to determine the pairwise similarity relationships among them. AMPs within the HSPNs were clustered using the modularity optimization algorithm based on the Louvain method.[41] Peptides sharing similar features are grouped together, thus, occupying the same chemical space in the network.

2.5. Representativeness from Cephalopods Singular AMPs by Complex Networks. To further reduce the CSPs at selecting the most representative ones, centrality analyses were performed. A HSPN was constructed using only the CSPs (identity <0.40 with StarPepDB space). The HSPN construction followed the same procedure described above, but a similarity cutoff of 0.75 was applied to improve network topology for mining information. Clusters, also known as communities, were identified using the Louvain method,[41] and then two centrality measures were calculated: hub-bridge centrality (HB)[42] and harmonic centrality (HC).[43] Centrality values measure the importance of a node in a network. Additionally, pairwise similarity comparisons were performed using the Smith-Waterman method.[44] Using the peptide's centrality and a sequence similarity cutoff of 0.30, the least redundant yet most important peptides in the network were identified. This process is described in ref[39]. Afterward, two datasets were recovered: the union and the intersection of the sets recovered using both HB and HC centralities.

2.6. Computer Resources. The in silico proteolysis of 5,412,039 proteins and the subsequent screening of resulting peptidomes were managed using a high-performance desktop computer with the following specifications: CPU: Dual 20-core Intel Xeon Gold 6148 processors with (min/max) speed 1010/1000/3700 MHz, RAM: 256 GB, SSD: NVMe KINGSTON SNV2S/2000G (2 TB - M.2-3500 MB/s), Operating System: Linux kernel 5.15.0−72-generic for x86_64 architecture, Processors: 880.

https://doi.org/10.1021/acsomega.4c01959
ACS Omega 2024, 9, 43353−43367

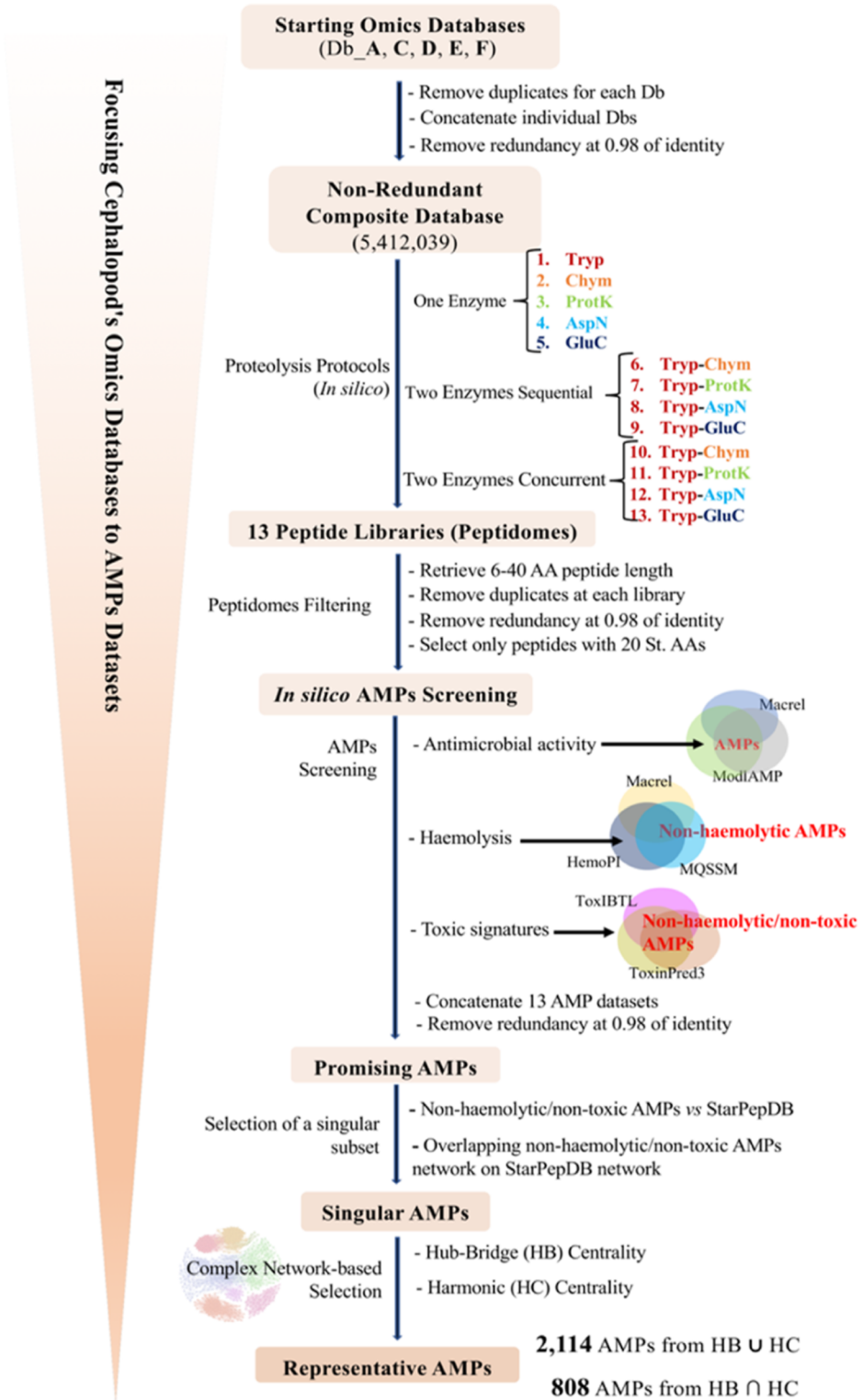


Figure 1. Workflow proposed to uncover several AMPs datasets encrypted in cephalopod salivary glands. This scheme shows how millions of proteins that characterize the cephalopod salivary apparatus are focused into several AMP databases by using proteolysis and a rational screening strategy.

2.7. Workflow Focusing Cephalopods' Omics Data to AMP DataSets.

The diagram representing how cephalopods' omics data have been focused to different AMP datasets to cater to a spectrum of research needs is displayed in Figure 1.

3. RESULTS

3.1. Construction of the Starting Composite Database from Cephalopods Salivary Glands.

The composite protein database integrating transcriptomic and proteomic data used for the wide-proteome discovery of AMPs in *O. vulgaris*,[18,20] is reutilized here, for uncovering AMPs encrypted within the salivary apparatus of cephalopods. The scheme for building such

https://doi.org/10.1021/acsomega.4c01959
ACS Omega 2024, 9, 43353−43367

composite database is depicted in Figure 2, where it is evident that characterized AMPs originally integrated as database B are

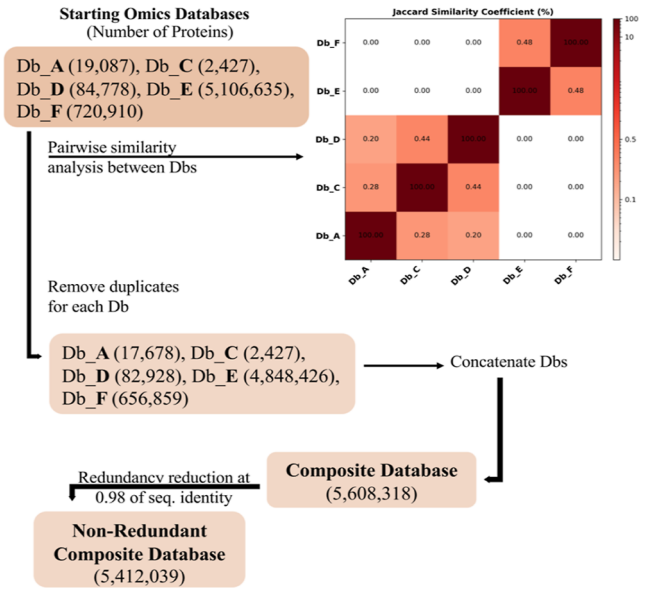


Figure 2. Building a nonredundant composite database with 5,412,039 proteins for the in silico proteolysis. This figure illustrates the scheme followed for concatenating and analyzing the starting omics database from cephalopods' salivary apparatus to generate the final composite protein database.

leaving out. The original smaller databases (A, C, D, E, and F) that integrated the composite were analyzed by considering pairwise similarities based on common sequences, which are encoded by the Jaccard index (Figure 2). The individual databases exhibited significant uniqueness. Although databases C and D could be related, since the latter was used to detect the 2427 proteins registered in database C, their similarity, as measured by a Jaccard index, is only 0.48%. Likewise, despite being derived from non-standard ORFs in cephalopod PSGs transcriptomes, databases E and F also show considerable divergence, with a Jaccard index of just 0.44% (Figure 2). Redundancy was also explored within each individual database, and duplicates were found in 4 out of 5 databases. The resulting individual databases after removing duplicates can be found at doi:10.17632/hgwkkmms3h.1 (Dataset 1). Such individual databases were concatenated, and a more stringent redundancy reduction was carried out at 0.98 sequence identity on the resulting composite database. Thus, a nonredundant composite database made up of 5,412,039 proteins was created (doi:10.17632/gxmkytwdhx.1, Dataset 2), to be used as substrate at the in silico proteolysis.

3.2. In Silico Proteolysis of Cephalopods Omics Data and Filtering of Virtual Peptidomes. This composite protein database was used as the substrate for the intended in silico digestion. The digestion used the five main proteases used in proteomics: trypsin, chymotrypsin, proteinase K, AspN, and GluC. Since trypsin is the most commonly used protease in proteomics, it was combined with the remaining four proteases in sequential and concurrent mode. This combination was aimed at complementing the trypsin action with other cutting sites in order to obtain a higher diversity within the virtual peptidomes.

As previously mentioned, three main digestion protocols were applied involving five enzymes, so a total of 13 distinct enzymatic digestions were performed on the nonredundant composite database (Figure 3). Consequently, 13 virtual peptide libraries were generated, offering a wide peptide diversity from cephalopods to explore in the field of peptide science. Such peptidomes are publicly available at doi:10.17632/c3zhzgwsnw.1 (Dataset 3).

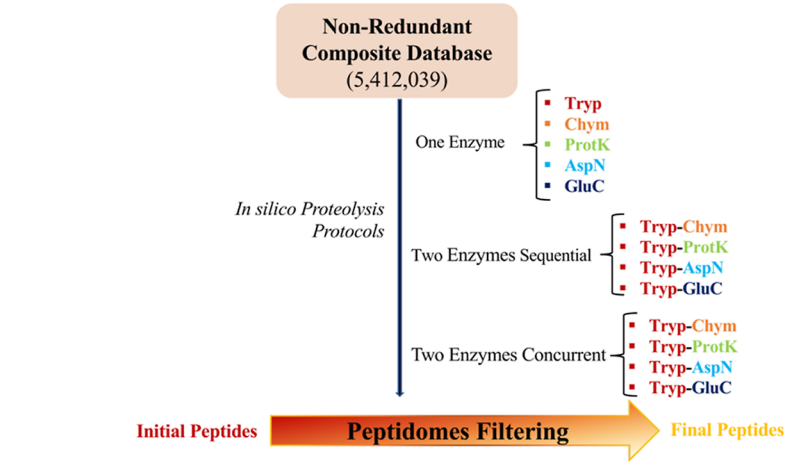
Each resulting peptidome was filtered to approach them to AMPs features. In this sense, only peptides ranging from 6 to 40 AAs in length were initially selected, followed by the removal of duplicates and a more stringent redundancy reduction at 0.98 sequence identity using cd-hit. At this stage, the length range for the peptides varied from 11 to 40 AA and nonstandard AAs were also removed to facilitate further screenings.

Figure 3 shows how much each peptide library varied at each filtering step, arriving at the final libraries containing peptides with standard AAs ranging from 11 to 40 AAs in length (doi:10.17632/6fjsdnvygb.1, Dataset 4). These 13 final peptide libraries were concatenated to give a total of 52,488,742 peptides, subsequently reduced to 9,216,442 peptides when applying redundancy removal at 0.98 sequence identity (doi:10.17632/v67g7r8nf2.1, Dataset 5). This extensive but nonredundant peptidome sourced from cephalopods' salivary glands will be of great utility for those researchers who want to discover new bioactive peptides by computational and in vitro screenings.

3.3. Focusing Cephalopods Peptidomes to Several AMP DataSets. The final peptidomes corresponding to each proteolysis protocol (last column of the table shown in Figure 3) were screened individually against antimicrobial activity. To determine whether a query peptide is an AMP, consensus prediction agreement among three models was considered. The Figure 1SA contains 13 Venn diagrams corresponding to the screened peptidomes, showing the AMPs detected solely by macrel, modelAMP_RF, and modlAMP_SVM, respectively, as well as the agreement/intersection among the three prediction tools. The FASTA files containing AMPs libraries, identified by consensus across three prediction models for each proteolysis protocol, can be accessed freely at doi:10.17632/wwk7zzcfhv.1 (Dataset 6). Additionally, the results of predictions on the 13 individual peptidomes by the three models are available in File 1S (available at 10.26434/chemrxiv-2023-rqqqb).

The consensus AMP libraries were then filtered by considering their toxicity potential, expressed by their hemolytic activity and the presence of toxic signatures. The hemolytic activity was first evaluated by three prediction tools. The Venn diagram representing nonhemolytic predictions from Macrel, HemoPI, and MQSSM is illustrated in Figure 1SB. The definitive predictions for nonhemolytic AMPs are found at the intersections. The corresponding FASTA files for the 13 nonhemolytic AMP consensus libraries can be accessed at doi:10.17632/pvptjh7kmv.1 (Dataset 7). Additionally, the raw predictions made by each individual model are available in File 2S (available at 10.26434/chemrxiv-2023-rqqqb). Subsequently, these consensus nonhemolytic AMPs were screened against toxic signatures using ML-based and hybrid models implemented in ToxinPred3 and the deep learning tool ToxIBTL. Similarly, Venn diagrams were employed to establish consensus predictions for nonhemolytic/nontoxic AMPs, as depicted in Figure 1SC. The libraries containing nonhemolytic/nontoxic AMPs, identified through the agreement of the three models, can be accessed publicly at doi:10.17632/ccp94tgcp2.1 (Dataset 8). Furthermore, the raw predictions from each model are available for consultation in File 3S. The tracking of this

https://doi.org/10.1021/acsomega.4c01959
ACS Omega 2024, 9, 43353−43367



	Protease	Total Peptides	6-40AA length	Non-Duplicated Peptides	Non-Redundant Peptides_0.98 id	11-40AA length (20 AA Alphabet)
One Enz	Tryp	23,325,454	10,326,244	9,203,883	4,811,218	4,785,875
	Chym	49,733,690	8,278,781	6,463,396	1,591,585	1,574,181
	ProtK	86,266,471	2,478,822	1,626,459	121,454	112,387
	AspN	10,168,230	7,503,076	7,126,125	5,667,005	5,643,298
	GluC	16,340,411	9,614,878	8,780,707	5,636,064	5,609,322
Two Enz Seq	Tryp-Chym	73,059,144	18,605,025	15,642,454	6,098,130	6,061,301
	Tryp-ProtK	109,591,925	12,805,066	10,827,193	4,915,100	4,883,567
	Tryp-AspN	33,493,684	17,829,320	15,700,559	7,618,700	7,583,442
	Tryp-GluC	39,665,865	19,941,122	17,611,912	8,085,798	8,047,367
Two Enz Conc	Tryp-Chym	67,713,515	5,261,016	719,996	526,570	515,994
	Tryp-ProtK	104,246,296	974,563	587,639	31,323	26,425
	Tryp-AspN	27,587,032	10,645,211	9,273,776	4,226,789	4,201,827
	Tryp-GluC	34,332,230	10,163,230	8,663,851	3,467,440	3,443,756
					Total Peptides	52,488,742
					Non-R_0.98_id.	**9,216,442**

Figure 3. Tracking the filtering of the peptidomes resulting from each proteolysis protocol. The scheme illustrates how the number of peptides decreased as the number of screening steps increased. Initial peptides were produced by directly applying proteases and were filtered to satisfy mainly sequence length (6−40 AAs) and redundancy (no duplicates and representative peptides at 0.98 sequence identity) criteria.

Table 2. Focusing Peptidomes Resulting from Each Proteolysis Protocol to AMP Datasets[a]

Proteolysis protocol	Peptidomes (no. peptides)	AMPs_Consensus	Non-Hem. AMPs_Cons	Non-Hem/Non-Tox. AMPs_Cons
Tryp	4,785,875	46,615	9897	7478
Chym	1,574,181	21,801	3970	2604
ProtK	112,387	775	310	157
AspN	5,643,298	294,959	33,108	22,978
GluC	5,609,322	404,990	43,955	31,756
Tryp-Chym_S	6,061,301	67,811	13,558	9875
Tryp-ProtK_S	4,883,567	47,316	10,142	7599
Tryp-AspN_S	7,583,442	307,767	36,455	25,454
Tryp-GluC_S	8,047,367	413,108	70,043	42,514
Tryp-Chym_C	515,994	1179	600	430
Tryp-ProtK_C	26,425	168	148	52
Tryp-AspN_C	4,201,827	46,713	9670	7270
Tryp-GluC_C	3,443,756	57,691	10,335	7696
total	52,488,742	1,710,893	242,191	165,863
NonR_0.98_SeqId	**9,216,442**	**542,485**	**104,242**	**68,694**

[a]The table illustrates how the peptide libraries are rationally reduced by the robust detection of AMPs, non-haemolytic AMPs and non-haemolytic/non-toxic AMPs by three prediction tools at each screening step.

https://doi.org/10.1021/acsomega.4c01959
ACS Omega 2024, 9, 43353−43367

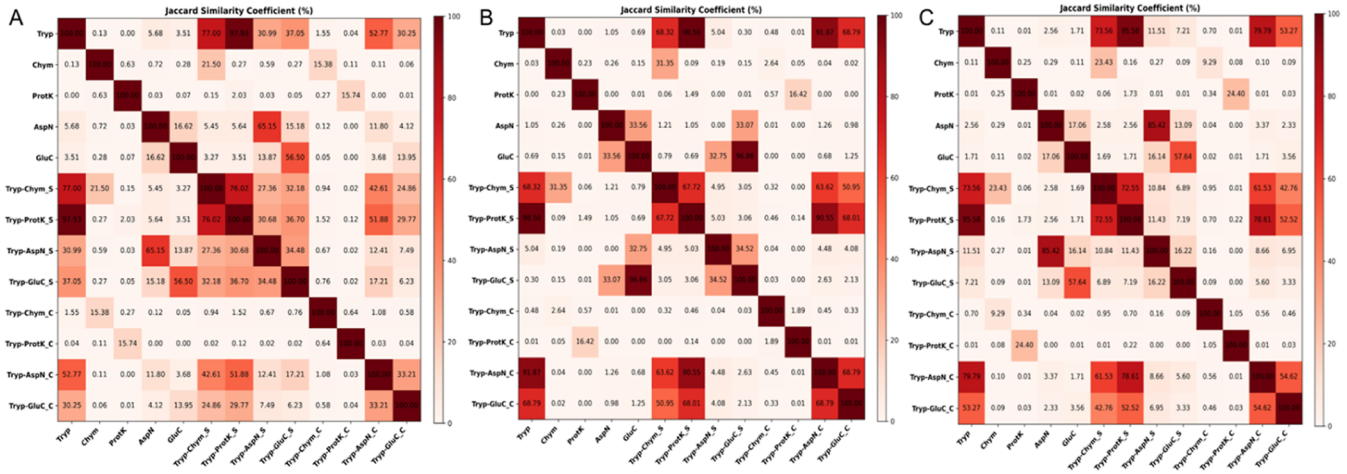


Figure 4. Peptide diversity among 13 proteolysis protocols in three steps of AMP mining on cephalopod salivary glands. (A) Virtual peptidomes were generated by 13 proteolysis protocols. (B) AMPs detected by the consensus of three prediction models from peptidomes are shown in (A). (C) Nonhemolytic/nontoxic AMPs detected by the consensus of three prediction models from AMPs libraries shown in B. Jaccard index is used as a pairwise similarity metric.

screening process from the peptidomes generated by the 13 proteolysis protocols to the generation of the datasets corresponding to nonhemolytic/nontoxic AMPs is summarized in Table 2.

Table 2 also displays in bold the total number of non-redundant AMPs after all libraries within each column were concatenated and sequence redundancy was removed with cd-hit at 0.98 of sequence identity. The resulting 542,485 AMP sequences from cephalopods, which represent a potential reservoir of novel AMPs, are promising for additional screenings to uncover peptide candidates for drug development (doi:10.17632/tr7xbp2pyt.1, Dataset 9). This subset was further filtered by extracting 104,242 nonhemolytic AMPs, which may have an increased relevance for drug development (doi:10.17632/6gsdfj9876.1, Dataset 10). However, the most promising dataset, made up of privileged AMPs, was obtained after toxic signatures were removed from nonhemolytic AMPs, rendering 68,694 nonhemolytic/nontoxic AMPs (doi:10.17632/8mttp4pvmc.1, Dataset 11).

The evolution of the 13 virtual peptidomes at the key AMPs mining points, which are shaded in Table 2, is monitored by changes in the distribution of six global peptide features, such as length, AA (AA) frequency, isoelectric point (pI), global charge, global hydrophobicity, and global hydrophobic moment, within each peptide library class (Figure 2S). Changes in the distribution of the global peptide feature values can be observed from the peptidomes (Figure 2SA) to the nonhemolytic/nontoxic AMPs (Figure 2C). While median peptide length at the peptidomes are generally below 15 AAs, there is a shift to higher than 15 AAs with a top around 28 AAs in mostly of the nonhemolytic/nontoxic AMPs libraries. A similar shift to increased values is shown for the distribution of the pI and global charge. The median pI values distribution at peptidomes changed to be roughly around 8 to be consistently distributed around 10 at intermediate AMPs (Figure 2SB) and final AMPs datasets (Figure 2SC). Similarly, the global charge is completely shifted to the right at the AMPs and nonhemolytic/nontoxic AMPs libraries, where most of the AMPs take charges above 0. On the other hand, the hydrophobicity holds its values in a range from −1 to 1 for the peptidomes, intermediate, and final AMPs libraries, while the hydrophobic moment values slightly moved from a median of 0.35 at the peptidomes to higher values than 0.4 in the nonhemolytic/nontoxic AMPs libraries. The AA frequency did not change significantly from the peptidomes to AMP libraries, even focusing attention on the positively charged AAs, which are key for antimicrobial activity.

The singularity of the peptide libraries generated at the AMPs mining points highlighted in Table 2, is also inspected by all-vs-all comparison using the Jaccard index. The Jaccard index quantifies how many peptides are shared by two libraries, namely, the intersection of two sets. Thus, it is used as a pairwise similarity metric to evaluate the diversity among peptide libraries from each proteolysis protocol at three key AMPs mining steps. The Jaccard index heatmaps corresponding to each proteolysis protocol for the generated peptidomes, predicted AMPs consensus, and predicted nonhemolytic/nontoxic AMPs are shown in Figure 4.

Generally, the heatmaps show a striking singularity among the digestion protocols at each of the evaluated mining steps (Figure 4). The Jaccard index only reached values above 60% among the peptidomes when trypsin was compared to a combination of trypsin-chymotrypsin and trypsin-proteinase K in a sequential mode or when these last proteolysis protocols were compared to each other. A significant library redundancy is also observed when comparing the proteolysis with AspN to its sequential action after trypsin (Figure 4A).

Similarly, redundancy among AMP libraries is mostly observed between trypsin and its sequential counterparts, trypsin-chymotrypsin and trypsin-proteinase K. However, additional pairs from the concurrent mode, such as trypsin-AspN and trypsin-GluC, also show significant similarities with trypsin proteolysis. GluC proteolysis shows a high AMP redundancy with its trypsin-GluC sequential counterpart. The same sequential pairs that shared redundancy with trypsin, trypsin-chymotrypsin and trypsin-proteinase K, also share redundancy with the concurrent action of trypsin-AspN and trypsin-GluC (Figure 4B).

Finally, a similar redundancy pattern is displayed for the nonhemolytic/nontoxic AMPs (Figure 4C), including the high peptide redundancy derived from the action of AspN and the sequential proteolysis of trypsin-AspN.

https://doi.org/10.1021/acsomega.4c01959
ACS Omega 2024, 9, 43353−43367

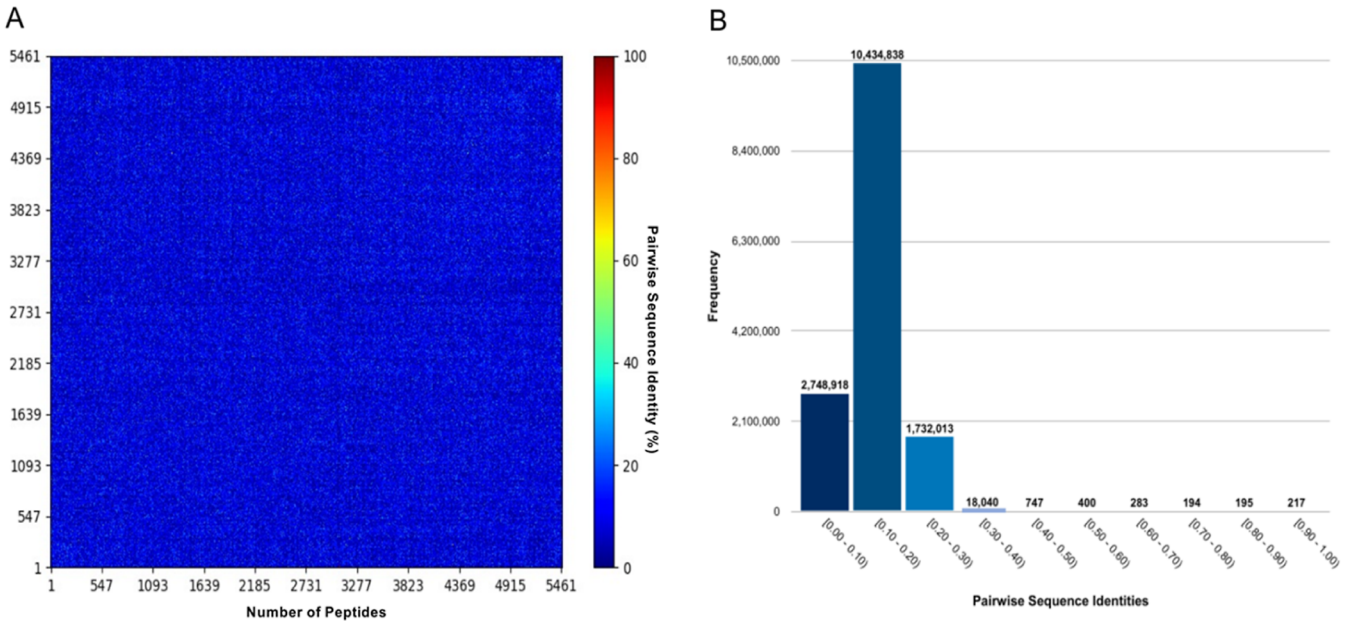


Figure 5. (A) Heat map and (B) histogram of pairwise sequence identity of the 5466 CSPs. The heat map and histogram were built with in-house tools SeqDivA (https://github.com/eancedeg/SeqDivA)[45] and Dover Analyzer (http://mobiosd-hub.com/doveranalyzer/).[46]

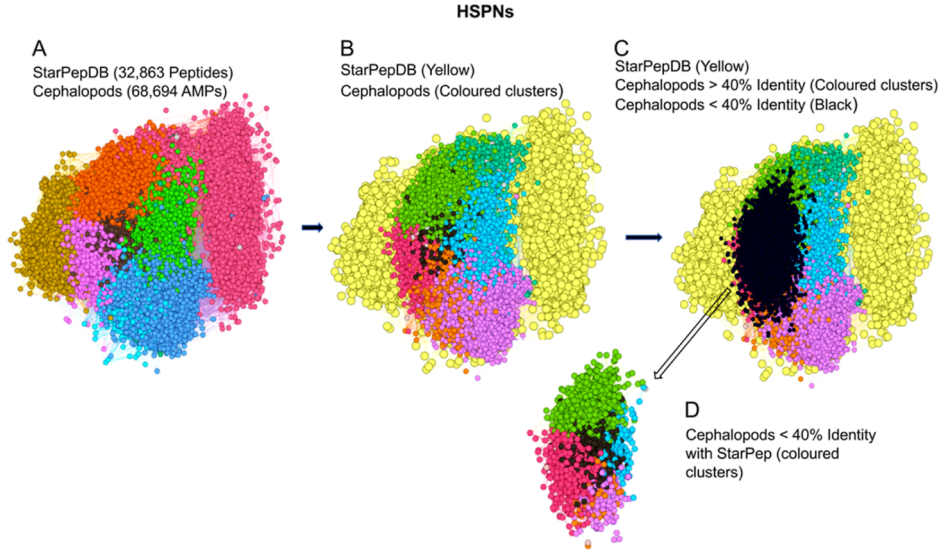


Figure 6. Superposition of the 68,694 nonhemolytic/nontoxic AMPs from cephalopods on the known sequence space represented by 32,863 peptides from StarPepDB, projected through Half-Proximal Similarity Networks (HSPNs). (A) HSPN constructed with cephalopod and StarPep datasets. Clusters are delineated using different colors. (B) HSPN projecting the superposition of cephalopods AMPs on StarPepDB members in yellow. (C) HSPN projecting the overlapping of three AMP datasets: (i) StarPepDB in yellow, (ii) cephalopod AMPs sharing higher than 40% of sequence identity with StarPepDB, colored by clusters, and (iii) cephalopod AMPs sharing less than 40% of sequence identity (black) with StarPepDB. (D) HSPN projecting cephalopod AMPs sharing less than 40% of sequence identity with StarPepDB, highlighting the network clusters or communities.

While the heatmaps allowed for comparative analyses even between proteolysis protocol pairs not originally intended in the primary design, this analysis suggests that sequential application of chymotrypsin and proteinase K after trypsin leads to high peptide redundancy at all mining stages. Similarly, but in a less consistent manner, this is observed for the concurrent action of trypsin with AspN and GluC, respectively.

Therefore, for future proteolysis-driven virtual mining efforts aimed at AMP discovery using the proposed enzymatic digestion protocols, it is not recommended to employ the sequential application of chymotrypsin and proteinase K following trypsin. Similarly, although with less emphasis, the concurrent action of trypsin with AspN and GluC should be avoided. These two recommendations are further supported by the observed recurrent similarity in the distribution pattern of global peptide features between trypsin and its sequential action with chymotrypsin and proteinase K, as well as between trypsin and its concurrent action with AspN and GluC at the same mining AMP stages (Figure 2SA−C).

The singularity of the sequence space represented by the 68,694 nonhemolytic/nontoxic AMPs from cephalopods' salivary glands was evaluated against the 32,863 characterized AMPs registered in StarPepDB. To achieve this, both databases were compared using cd-hit-2d to identify how many and which

https://doi.org/10.1021/acsomega.4c01959
ACS Omega 2024, 9, 43353−43367

AMPs from cephalopods were clustered to StarPepDB's members above identity cutoffs of 0.40, 0.50, 0.60, 0.70, and 0.80. The similarity clusters resulting from the comparison were parsed to extract the cephalopod AMP sequences satisfying the previously mentioned identity cutoffs. Both the similarity clusters and the FASTA files corresponding to the extracted subsets are shown in File 4S (available at 10.26434/chemrxiv-2023-rqqqb). Out of 68,694 cephalopod AMPs, 63,228 were clustered to StarPepDB members above the threshold of 0.40 sequence identity, suggesting that these AMPs are more closely related to the known chemical space of characterized AMPs.

The remaining 5466 nonhemolytic/nontoxic AMPs are denoted by the acronym CSPs, as explained in the Materials and Methods section. Both sets of AMPs are accessible at doi:10.17632/8mttp4pvmc.1 (Dataset 12), along with additional datasets that categorize the similarity with StarPepDB based on identity percentages within the following ranges: 40−50, 50−60, 60−70, 70−80, and greater than 80. These datasets consist of 26,744, 30,217, 5,716, 453, and 98 AMPs, respectively.

Given that the 5466 CSPs share less than 40% of sequence identity with the characterized chemical space of AMPs, their internal diversity was also explored by all-vs-all global alignments (Figure 5). Figure 5A illustrates the heatmap of the pairwise sequence identities from all-vs-all global alignments above, while Figure 5B shows the distribution/frequency of the peptide pairs satisfying sequence identities at ranges increasing by 0.10 units. This analysis was also applied to characterize the 63,228 cephalopods AMPs displaying similarities above 40% of identity with StarPepDB, using the datasets discretized by identity ranges (Figure 3S).

As shown in Figure 5, the internal sequence diversity among the 5466 CSPs is high, also indicating the structural singularity among its members. This singularity among these virtual scaffolds bearing privileged antimicrobial potentials is a strong point for peptide drug development.

3.4. Singularity of Cephalopods' AMPs from the Outlook of Complex Networks. Based on the previous comparison, where 63,228 out of 68,694 promising cephalopod AMPs were identified as more closely related to characterized StarPepDB members, while the remaining 5466 appear to be unique with respect to the known chemical space, the relatedness of the cephalopod AMPs with the characterized chemical space of AMPs can also be demonstrated using HSPNs, which are less computationally demanding at considering all peptides but not all pairwise similarity relationships.[39,47] A HSPN was constructed from the 32,863 StarPepDB peptides and the 68,694 cephalopod AMPs, including the two subsets with different degrees of relatedness to the StarPepDB chemical space. A clustering algorithm was then performed over the network topology to delineate network communities that should group peptides with similar features. Figure 6 illustrates how the cephalopod chemical space represented by 68,694 promising AMPs are overlapped on the known sequence space represented by the 32,863 peptides from StarPepDB. HSPNs were used to project such chemical/sequence spaces.

Figure 6C supports the findings of the comparison of the 68,694 nonhemolytic/nontoxic AMPs from cephalopods to StarPepDB using the cd-hit-2d tool. The chemical space corresponding to the cephalopod AMPs sharing higher than 40% of sequence identity with StarPepDB is closer to the known sequence space of StarPepDB (colored in yellow), while the sequence space occupied by CSPs sharing less than 40% of sequence identity (black) with StarPepDB is somewhat spatially disconnected from the yellow zone.

3.5. Singularity of Cephalopods AMPs as Seen from Physicochemical Characterization of Network Clusters. The 5466 CSPs were studied in the context of the 32,863 peptides from StarPepDB. The HSPN consisting of both peptide datasets revealed nine clusters (Figure 7).

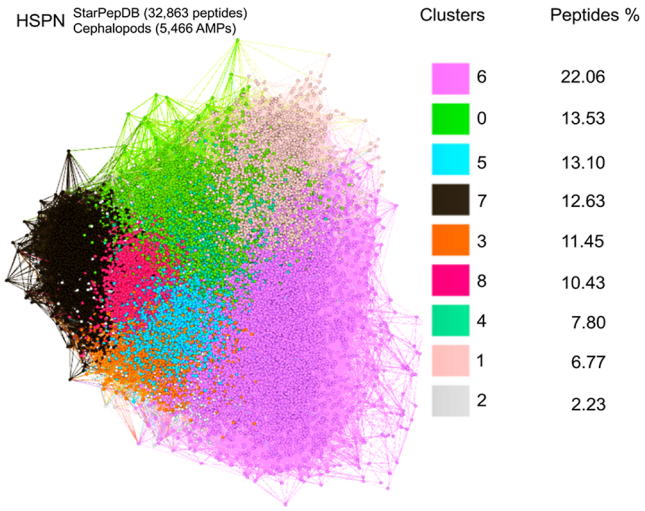


Figure 7. HSPN corresponding to the clustering of 32,863 peptides from StarPepDB and the 5466 nonhemolytic/nontoxic AMPs from cephalopods. Nine clusters (0−8) were identified, and their peptide contents are displayed as a percentage.

The peptides from each cluster were identified and physicochemically characterized. The detailed composition of peptide clusters and their physicochemical characterization can be found in File 5S (available at 10.26434/chemrxiv-2023-rqqqb). Of the nine clusters, two were highly represented by CSPs: cluster 8 (50.7%) and cluster 5 (44.0%). The remaining clusters were only represented by 0.04−18.40% CSPs (Figure 8).

Cluster 8 is characterized by having an intermediate peptide length (∼36 AAs), low hydrophobicity (−0.21), high net charge (4.03), intermediate amphiphilicity (1.02), high isoelectric point (9.65), and high Boman index (1.99). On the other hand, peptides from cluster 5 are shorter (∼28 AAs) and more hydrophobic (−0.07), but they are also less charged (1.54), with a lower amphiphilicity (0.82), isoelectric point (8.54), and Boman index (0.95).

Overall, peptides from clusters 8 and 5 differ from other peptide clusters in their sequence length, as they are neither as long as in cluster 7 (∼73 AAs) nor as short as those in cluster 1 (∼11 AAs). Additionally, they tend to have higher net charge, hydrophobicity, and isoelectric point values. These findings provide more evidence that CSPs are novel peptide representations.

3.6. Complex Networks for Extracting Representativeness from the CSPs. The 5466 CSPs were further reduced by extracting the most representative ones using network science. First, an HSPN projecting the chemical space of the CSPs was constructed. However, to achieve effective extraction of representative CSPs, HSPN projecting the most informative topology should be used. This HSPN was found by applying an optimal similarity cutoff of 0.75 to produce a reasonable trade-off between the number of communities and singletons,

https://doi.org/10.1021/acsomega.4c01959
ACS Omega 2024, 9, 43353−43367

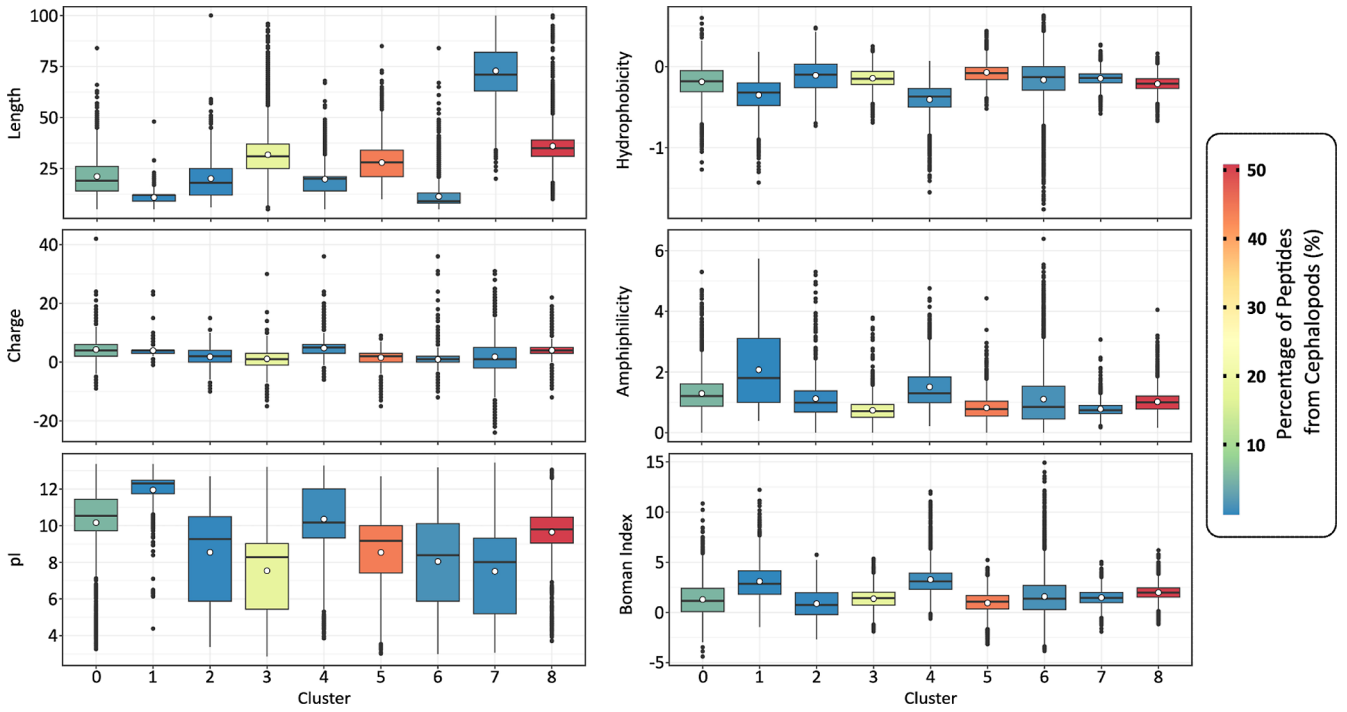


Figure 8. Physicochemical characterization of the peptide clusters. This figure shows the distribution of the physicochemical properties of the peptides belonging to different network clusters or communities. The color of each cluster represents the percentage of CSPs that it contains. Only clusters 8 and 5 are mostly represented by CSPs. The clusters were obtained after building a HSPN with the StarPepDB peptides and the CSPs.

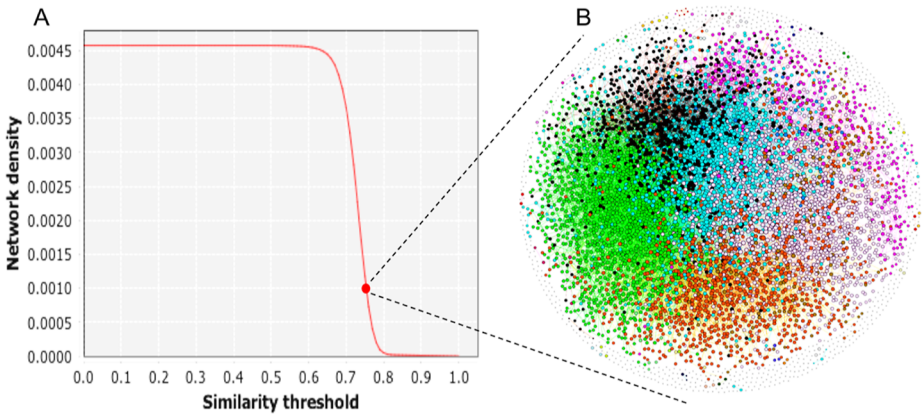


Figure 9. Selection of the most informative HSPN projecting the chemical space of the 5466 CSPs, by applying an optimal similarity cutoff. (A) Network density plots at different similarity thresholds. The similarity cutoff of 0.75, indicated in the plot, was selected as optimal. (B) HSPN topology results from applying the optimal similarity threshold. The HSPN topology is formatted according to the Fruchterman-Reingold layout.

considering the diversity of the CSPs set. A community or cluster within the network is considered when at least two nodes or peptides are connected, and the singletons are those that are not connected with any other in the network. Singletons are atypical peptides with singular structures that may represent privileged scaffolds for designing peptide drugs.

The optimal cutoff of 0.75 was determined by exploring the network density at different similarity cutoffs. From 0.70 to 0.80, a significant change in network density is observed, reaching the desired value of 0.001 for HSPNs at a similarity cutoff of 0.75 (Figure 9A). At this similarity cutoff of 0.75, the number of communities/clusters increased to 60, while the network density decreased to 0.001, as mentioned before. Additionally, the number of disconnected peptides increased to 763, the so-called singletons, with a degree of 0 (File 6S, available at 10.26434/chemrxiv-2023-rqqqb). The HSPN representing this topology is visualized in Figure 9B, after applying the Fruchterman-Reingold layout.

From this optimal HSPN topology, the most representative peptides were extracted using the procedure described in ref 39. Two subsets of nonredundant and representative peptides were extracted based on their harmonic (HC) and hub-bridge (HB) centrality measures. HC centrality weights the relevance or popularity of each peptide in the entire network, while HB centrality measures the relevance at the community level. Thus, two subsets of 1469 and 1453 CSPs were extracted using HC and HB centralities, respectively. File 6S (available at 10.26434/chemrxiv-2023-rqqqb) contains the sequences corresponding to these two subsets, the HSPN characterization at a 0.75 similarity cutoff, and the properties of its 5466 nodes (CSPs), including the HC and HB values.

https://doi.org/10.1021/acsomega.4c01959
ACS Omega 2024, 9, 43353−43367

Finally, the union and intersection of these two subsets resulted in 2114 and 808 nonhemolytic/nontoxic AMPs, respectively (Figure 10). These two final datasets are freely

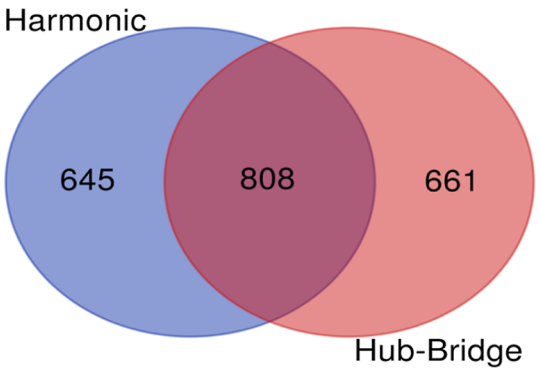


Figure 10. Venn diagram illustrating the union and intersection of the 1469 and 1453 nonhemolytic/nontoxic AMPs that were extracted using harmonic and hub-bridge centralities, respectively. From the union and intersection of these two subsets, the final AMP datasets from this study were obtained: 2114 and 808 nonhemolytic/nontoxic AMPs from cephalopods.

available at doi:10.17632/vv5fcxk5rn.2 (Dataset 13). The larger final dataset is a nonredundant but comprehensive representative subset of CSPs, while the smaller one is composed of the representative CSPs commonly identified by each centrality metric.

To validate the overall mining process, we assessed whether the final 2114 CSPs were unique or shared homology with known AMPs. We utilized a nonredundant dataset of 19,456 AMPs from StarPepDB, selected for their antimicrobial functions and ranging from 10 to 100 AAs in length, with redundancy removed at 98% sequence identity. The CSPs were compared against this dataset using Smith-Waterman alignment with the BLOSUM62 matrix at the StarPep Toolbox.[40] Out of the CSPs, 18 exhibited similarity scores above 0.6, with alignment lengths of at least 10 AA, sharing similarity regions of 10−14 AAs with 14 known AMPs, 13 of which are synthetic (File 7S). This small number of matches (0.09%), along with the synthetic origin of the corresponding AMPs, highlights the novelty of the CSPs.

4. DISCUSSION

In silico proteolysis has been mostly applied to protein families from plants to identify promising bioactive with clinical potential.[48−52] However, this approach has not been extended to omics data for the same purpose. This proteolysis-based exploration has been limited to small protein datasets, likely due to the high dimensionality and diversity of peptides resulting from protease application, despite trypsin being the most commonly used protease and targeting 10.7% of the AAs.[30] Trypsin is the preferred protease for (MS)-based proteomics. It cleaves the carboxy-terminal to arginine and lysine residues, resulting in a positive charge at the peptide C-terminus, which is beneficial for MS analysis. Nonetheless, other proteases are frequently used to gather Supporting Information, such as AspN

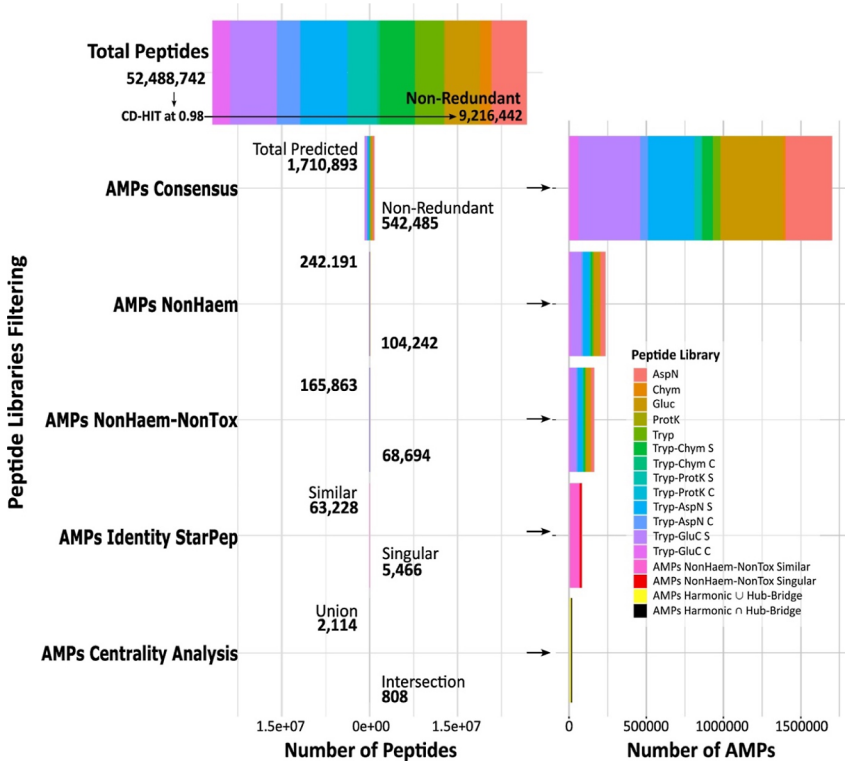


Figure 11. Tracking the screening of nonredundant cephalopods peptides (9,216,442) derived from the application of 13 proteolysis protocols. Different AMPs libraries were generated, considering the (i) antimicrobial activity (AMPs consensus), (ii) nonhemolytic potential (AMPs NoHaem), (iii) no presence of toxic signatures (AMPs NoHaem-NonTox), (iv) AMPs singularity regarding the known sequence space of StarPep (similar and singular NoHaem-NonTox AMPs), and (v) Representative subsets according to network centrality analyses (union and intersection of the subsets extracted with HC and HB centralities).

https://doi.org/10.1021/acsomega.4c01959
ACS Omega 2024, 9, 43353−43367

and GluC, which target acidic AAs, and chymotrypsin, which primarily targets aromatic AAs.[30]

The sequential use of these proteases following trypsin has recently been shown to enhance the identification of proteins and peptides by MS, even encompassing less commonly used proteases in proteomics like proteinase K due to its broad specificity, targeting 53.3% of AAs.[30] Inspired by these findings and the growing need to utilize omics data to identify new AMPs, we evaluated trypsin, chymotrypsin, AspN, GluC, and proteinase K in silico, as well as the activity of these last four proteases following trypsin in a sequential and concurrent manner, using a composite protein database that incorporates all proteomic and transcriptomic data from cephalopods salivary glands (CSGs).[20]

One of the primary challenges of this work was addressing the "curse of dimensionality", which is exacerbated when generating peptidomes through in silico proteolysis of 5,412,039 proteins representing a comprehensive proteome characterizing the cephalopods' salivary apparatus. The total number of non-redundant peptides (9,216,442) from the 13 proteolysis protocols significantly exceeded the initial number of proteins (5,412,039). The selection of appropriate proteolysis and AMPs mining tools, capable of exploiting high-performance computing resources and integrated into a rational screening strategy combining machine learning, deep learning, multiquery similarity searches, and complex networks for AMP discovery, enabled the processing of millions of proteins/peptides until manageable AMP datasets were obtained. The RPG tool played a pivotal role in the AMPs mining process by facilitating the execution of the intended proteolysis protocols involving five proteases and, crucially, enabling the processing of millions of protein sequences from the composite database.[31]

The use of an encompassing omics database characterizing the salivary apparatus of cephalopods for the proteolysis-based AMPs exploration is a strong point of the study. This comprehensive database integrates 16 translated transcriptomes from cephalopods' PSGs using six ORF translations, considering noncanonical transcripts. Additionally, proteins shorter than 100 AAs, often disregarded by the TransDecoder coding-region identifier tool, are included. Therefore, the in silico proteolysis not only revealed encrypted AMPs from existing proteins but also brought to light potential AMPs hidden in noncanonical proteins or in those typically methodologically discarded.

The rational in silico reduction from 9,216,442 unannotated peptides to various AMP datasets/libraries with varying relevance for further screenings aimed at discovering/developing new peptide drugs is depicted in Figure 11. This rational mining strategy yielded AMPs datasets that were subsequently narrowed down to a privileged subset of 5466 CSPs, which could be represented by either 2114 or 808 nonhemolytic/nontoxic AMPs according to network centralities. These AMPs datasets are publicly accessible and can be utilized by drug developers according to their specific requirements.

The sequential application of chymotrypsin, AspN, GluC, and proteinase K after trypsin has been demonstrated to enhance peptide detection by MS.[30] However, our study shows that the sequential use of chymotrypsin and proteinase K following trypsin does not significantly increase peptide diversity compared with trypsin alone, despite both enzymes having more cleavage sites than trypsin. Furthermore, the concurrent action of AspN and GluC with trypsin does not significantly contribute to the diversity of the resulting libraries compared with trypsin alone. This is evident as AspN and GluC proteases are highly specific, targeting only aspartic (D) and glutamic (E) acids, which represent only 5.4 and 6.8% of AAs in proteins.[30]

5. CONCLUSIONS

Cephalopod salivary glands harbor a remarkable reservoir of AMPs, including nonhemolytic and nontoxic AMPs, underscoring the remarkable biological diversity of these marine invertebrates and their potential as antimicrobial agents. A significant portion of these AMPs exhibit unique sequences that expand the chemical space for exploration beyond existing databases.

Omics data and advanced in silico analyses provide a powerful strategy for AMP identification. This multifaceted strategy has the potential to uncover a vast array of AMPs, including those encrypted within existing and noncanonical proteins, as well as those present in smaller proteins often overlooked by standard translation tools. The proteolysis-driven mining strategy, coupled with rigorous virtual screening steps aimed to effectively identify promising AMPs based on their characteristic signatures, nontoxic nature, and sequence singularity, expands the potential for AMP discovery in proteogenomic data.

Thus, the peptide datasets provided lay the foundation for further exploration of cephalopod salivary glands as a rich source of novel AMPs with therapeutic potential. These findings contribute significantly to the field of AMP research, our approach being extensive to other organisms, which holds promise for combating AMR and promoting peptide-based drug development.

■ ASSOCIATED CONTENT

Data Availability Statement

The starting omic data and peptide datasets generated in this study are publicly available through Mendeley Data Repository. The datasets are listed in the following order, corresponding to their citation order in the text: Dataset 1. Omics Datasets to Create a Composite Protein Database from Cephalopods Salivary Glands (CSGs) for in silico Enzymatic Digestion and Peptide Library Generation (Mendeley Data, V1, doi: 10.17632/hgwkkmms3h.1). Dataset 2. Composite Protein Database (nr) from CSGs for in silico Enzymatic Digestion and Peptide Library Generation (Mendeley Data, V1, doi: 10.17632/gxmkytwdhx.1). Dataset 3. Generation of Peptide Libraries: Applying Various in silico Enzymatic Digestion Protocols on the Composite Protein Database (nr) (Mendeley Data, V1, doi: 10.17632/c3zhzgwsnw.1). Dataset 4. Peptide Libraries from CSGs for Potential Antimicrobial Peptides (Mendeley Data, V1, doi: 10.17632/6fjsdnvygb.1). Dataset 5. An Extensive and Non-Redundant Peptide Library Derived from Omics Data of CSGs (Mendeley Data, V1, doi: 10.17632/v67g7r8nf2.1). Dataset 6. Consensus Antimicrobial Peptides Identified via Three Prediction Models from Peptide Libraries of CSGs (Mendeley Data, V1, doi: 10.17632/wwk7zzcfhv.1). Dataset 7. Non-Hemolytic AMPs Commonly Identified from CSGs by Three Prediction Models (Mendeley Data, V1, doi: 10.17632/pvptjh7kmv.1). Dataset 8. Non-Hemolytic and Non-Toxic Antimicrobial Peptides Commonly Identified from CSGs by Three Prediction Models (Mendeley Data, V1, doi: 10.17632/ccp94tgcp2.1). Dataset 9. Antimicrobial Peptides Predicted from Omics Data of CSGs (Mendeley Data, V1, doi: 10.17632/tr7xbp2pyt.1). Dataset 10. Non-Hemolytic Antimicrobial Peptides Predicted from Omics Data of CSGs (Mendeley Data, V1, doi: 10.17632/6gsdfj9876.1). Dataset 11. Non-Hemolytic and Non-Toxic AMPs from CSGs: Singular

https://doi.org/10.1021/acsomega.4c01959
ACS Omega 2024, 9, 43353−43367

Set and Sets with Similarity to Characterized AMPs (Mendeley Data, V1, doi: 10.17632/8mttp4pvmc.1). Dataset 12. Cephalopods Singular Non-Hemolytic/Non-Toxic AMPs (Mendeley Data, V1, doi: 10.17632/8mttp4pvmc.1). Dataset 13. Representative Sets of Singular Non-Hemolytic/Non-Toxic AMPs from CSGs Extracted via Network Centralities (Mendeley Data, V2, doi: 10.17632/vv5fcxk5rn.2). All standalone software and Web servers used for mining AMPs from omic data are freely available as indicated in the text. Specifically, our in-house (StarPep) software for complex network analyses and visualization is publicly available at https://github.com/Grupo-Medicina-Molecular-y-Traslacional/StarPepand the online documentation is available at https://grupo-medicina-molecular-y-traslacional.github.io/StarPep_doc.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.4c01959.

Venn diagrams representing the prediction results from the three evaluated models; consensus prediction for 1SA—AMPs detection, 1SB—Nonhemolytic AMPs and 1SC—Nonhemolytic/nontoxic AMPs; distribution of global peptide features (length, AA frequency, isoelectric point (pI), global charge, global hydrophobicity, and global hydrophobic moment) within each peptide library class; 2SA—AMPs consensus, 2SB—Nonhemolytic AMPs, 2SC—Nonhemolytic/nontoxic AMPs; histograms of pairwise sequence identity for datasets sharing similarity with StarPepDB at following identity percentage ranges: 40−50, 50−60, 60−70, 70−80, and greater than 80 (PDF)

Raw prediction results for AMPs detection on the 13 individual peptidomes by each of the three models (File S1) (ZIP)

Raw prediction results for nonhemolytic AMPs detection on the 13 individual peptidomes by each of the three models (File S2) (ZIP)

Raw prediction results for nonhemolytic AMPs deprived of toxic signatures detection (nonhemolytic/nontoxic AMPs) on the 13 individual peptidomes by each of the three models (File S3) (ZIP)

Similarity clusters resulting from the comparison between 68,694 nonhemolytic/nontoxic AMPs from cephalopods versus StarPepDB members at different identity cutoffs; FASTA sequences from cephalopods extracted from similarity clusters at different identity cutoffs (File S4) (ZIP)

HSPN projecting the clustering of CSPs with StarPepDB; clusters composition and their characterization through peptide length, charge, pI, hydrophobicity, amphiphilicity, Boman index (File 5S) (ZIP)

HSPN that projects the chemical/sequence space of the 5466 CSPs at 0.75 of similarity cutoff; HSPN properties and CSPs' representative subsets extracted with network centralities (File S6) (ZIP)

Homology assessment of the 2114 CSPs vs 19,456 known AMPs from StarPepDB by local alignment (File S7) (ZIP)

AUTHOR INFORMATION

Corresponding Authors

Guillermin Agüero-Chapin − *CIIMAR—Centro Interdisciplinar de Investigação Marinha e Ambiental, Universidade do Porto, Terminal de Cruzeiros do Porto de Leixões, Porto 4450-208, Portugal; Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Porto 4169-007, Portugal;* orcid.org/0000-0002-9908-2418; Email: gchapin@ciimar.up.pt

Yovani Marrero-Ponce − *Universidad San Francisco de Quito (USFQ), Grupo de Medicina Molecular y Traslacional (MeM&T), Colegio de Ciencias de la Salud (COCSA), Escuela de Medicina, Edificio de Especialidades Médicas; and Instituto de Simulación Computacional (ISC-USFQ), Diego de Robles y vía Interoceánica, Quito 170157 Pichincha, Ecuador; Facultad de Ingeniería, Universidad Panamericana, Benito Juárez 03920 Ciudad de México, Mexico;* Email: ymarrero@usfq.edu.ec

Agostinho Antunes − *CIIMAR—Centro Interdisciplinar de Investigação Marinha e Ambiental, Universidade do Porto, Terminal de Cruzeiros do Porto de Leixões, Porto 4450-208, Portugal; Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Porto 4169-007, Portugal;* Email: aantunes@ciimar.up.pt

Authors

Dany Domínguez-Pérez − *Department of Biology and Evolution of Marine Organisms (BEOM), Stazione Zoologica Anton Dohrn, 87071 Amendolara, Italy; PagBiOmicS—Personalised Academic Guidance and Biodiscovery-integrated OMICs Solutions, Porto 4200-603, Portugal*

Kevin Castillo-Mendieta − *School of Biological Sciences and Engineering, Yachay Tech University, Urcuquí 100119, Ecuador;* orcid.org/0000-0002-0383-8285

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.4c01959

Author Contributions

G.A.-C. and D.D.-P. contributed equally to this work. G.A.-C and D.D.-P were involved in the design and conduction of all peptide mining analyses. K.C.-M worked mainly on the MQSS for hemolysis prediction and on HSPN clusters characterization. G.A.-C and Y.M.-P worked on the conceptualization, supervision, writing, and reviewing of the manuscript. A.A participated in funding acquisition, writing, and reviewing the manuscript. All authors have read and agreed to the published version of the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

G.A.-C. and A.A. acknowledge the support of the FCT - Foundation for Science and Technology under UIDB/04423/2020 and UIDP/04423/2020, and to the CIIMAR Out of the Box project (UIDB-COB-31711). D.D.-P thanks the support provided by the CRIMAC - Fondo FSC 2014-2020 - Piano Stralcio "Ricerca e Innovazione 2015-2017" − PNIR, CUP C64I20000320001. Y.M.P. acknowledges support from USFQ "MED Grant 2023-4 (Project ID23234).

REFERENCES

(1) Murray, C. J. L.; Ikuta, K. S.; Sharara, F.; Swetschinski, L.; Robles Aguilar, G.; Gray, A.; Han, C.; Bisignano, C.; Rao, P.; Wool, E.; et al. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet* **2022**, *399* (10325), 629−655.

(2) Miethke, M.; Pieroni, M.; Weber, T.; Bronstrup, M.; Hammann, P.; Halby, L.; Arimondo, P. B.; Glaser, P.; Aigle, B.; Bode, H. B.; et al.

Towards the sustainable discovery and development of new antibiotics. *Nat. Rev. Chem* **2021**, *5* (10), 726−749.
(3) Magana, M.; Pushpanathan, M.; Santos, A. L.; Leanse, L.; Fernandez, M.; Ioannidis, A.; Giulianotti, M. A.; Apidianakis, Y.; Bradfute, S.; Ferguson, A. L.; et al. The value of antimicrobial peptides in the age of resistance. *Lancet Infect. Dis.* **2020**, *20* (9), e216−e230.
(4) Lei, J.; Sun, L.; Huang, S.; Zhu, C.; Li, P.; He, J.; Mackey, V.; Coy, D. H.; He, Q. The antimicrobial peptides and their potential clinical applications. *Am. J. Transl. Res.* **2019**, *11* (7), 3919−3931.
(5) Wang, G.; Li, X.; Wang, Z. APD3: the antimicrobial peptide database as a tool for research and education. *Nucleic Acids Res.* **2016**, *44* (D1), D1087−D1093.
(6) Waghu, F. H.; Barai, R. S.; Gurung, P.; Idicula-Thomas, S. $CAMP_{R3}$: a database on sequences, structures and signatures of antimicrobial peptides: Table 1. *Nucleic Acids Res.* **2016**, *44* (D1), D1094−D1097.
(7) Pirtskhalava, M.; Amstrong, A. A.; Grigolava, M.; Chubinidze, M.; Alimbarashvili, E.; Vishnepolsky, B.; Gabrielian, A.; Rosenthal, A.; Hurt, D. E.; Tartakovsky, M. DBAASP v3: database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics. *Nucleic Acids Res.* **2021**, *49* (D1), D288−D297.
(8) Aguilera-Mendoza, L.; Marrero-Ponce, Y.; Beltran, J. A.; Tellez Ibarra, R.; Guillen-Ramirez, H. A.; Brizuela, C. A. Graph-based data integration from bioactive peptide databases of pharmaceutical interest: toward an organized collection enabling visual network analysis. *Bioinformatics* **2019**, *35* (22), 4739−4747.
(9) Kumar, P.; Kizhakkedathu, J. N.; Straus, S. K. Antimicrobial Peptides: Diversity, Mechanism of Action and Strategies to Improve the Activity and Biocompatibility In Vivo. *Biomolecules* **2018**, *8* (1), 4.
(10) Skarnes, R. C.; Watson, D. W. Antimicrobial factors of normal tissues and fluids. *Bacteriol. Rev.* **1957**, *21* (4), 273−294.
(11) Tincu, J. A.; Taylor, S. W. Antimicrobial peptides from marine invertebrates. *Antimicrob. Agents Chemother.* **2004**, *48* (10), 3645−3654.
(12) Wang, S.; Fan, L.; Pan, H.; Li, Y.; Qiu, Y.; Lu, Y. Antimicrobial peptides from marine animals: Sources, structures, mechanisms and the potential for drug development. *Front. Mar. Sci.* **2023**, *9*, 1112595.
(13) Bachere, E.; Gueguen, Y.; Gonzalez, M.; de Lorgeril, J.; Garnier, J.; Romestand, B. Insights into the anti-microbial defense of marine invertebrates: the penaeid shrimps and the oyster Crassostrea gigas. *Immunol. Rev.* **2004**, *198*, 149−168.
(14) Kawabata, S. I. Evolution and phylogeny of defense molecules associated with innate immunity in horseshoe crab. *Front. Biosci.* **1998**, *3*, D973−D984.
(15) Masuda, M.; Nakashima, H.; Ueda, T.; Naba, H.; Ikoma, R.; Otaka, A.; Terakawa, Y.; Tamamura, H.; Ibuka, T.; Murakami, T.; et al. A novel anti-HIV synthetic peptide, T-22 ([Tyr5,12,Lys7]-polyphemusin II). *Biochem. Biophys. Res. Commun.* **1992**, *189* (2), 845−850.
(16) Aguero-Chapin, G.; Galpert-Canizares, D.; Dominguez-Perez, D.; Marrero-Ponce, Y.; Perez-Machado, G.; Teijeira, M.; Antunes, A. Emerging Computational Approaches for Antimicrobial Peptide Discovery. *Antibiotics* **2022**, *11* (7), 936.
(17) Matos, A.; Dominguez-Perez, D.; Almeida, D.; Aguero-Chapin, G.; Campos, A.; Osorio, H.; Vasconcelos, V.; Antunes, A. Shotgun Proteomics of Ascidians Tunic Gives New Insights on Host-Microbe Interactions by Revealing Diverse Antimicrobial Peptides. *Mar. Drugs* **2020**, *18* (7), 362.
(18) Almeida, D.; Dominguez-Perez, D.; Matos, A.; Aguero-Chapin, G.; Osorio, H.; Vasconcelos, V.; Campos, A.; Antunes, A. Putative Antimicrobial Peptides of the Posterior Salivary Glands from the Cephalopod Octopus vulgaris Revealed by Exploring a Composite Protein Database. *Antibiotics* **2020**, *9* (11), 757.
(19) Perez-Polo, S.; Imran, M. A. S.; Dios, S.; Perez, J.; Barros, L.; Carrera, M.; Gestal, C. Identifying Natural Bioactive Peptides from the Common Octopus (Octopus vulgaris Cuvier, 1797) Skin Mucus By-Products Using Proteogenomic Analysis. *Int. J. Mol. Sci.* **2023**, *24* (8), 7145.
(20) Almeida, D.; Domínguez-Pérez, D.; Matos, A.; Agüero-Chapin, G.; Castaño, Y.; Vasconcelos, V.; Campos, A.; Antunes, A. Data Employed in the Construction of a Composite Protein Database for Proteogenomic Analyses of Cephalopods Salivary Apparatus. *Data* **2020**, *5* (4), 110.
(21) Fingerhut, L.; Strugnell, J. M.; Faou, P.; Labiaga, A. R.; Zhang, J.; Cooke, I. R. Shotgun Proteomics Analysis of Saliva and Salivary Gland Tissue from the Common Octopus Octopus vulgaris. *J. Proteome Res.* **2018**, *17* (11), 3866−3876.
(22) Park, C. B.; Kim, M. S.; Kim, S. C. A novel antimicrobial peptide from Bufo bufo gargarizans. *Biochem. Biophys. Res. Commun.* **1996**, *218* (1), 408−413.
(23) Seo, J. K.; Lee, M. J.; Go, H. J.; Kim, G. D.; Jeong, H. D.; Nam, B. H.; Park, N. G. Purification and antimicrobial function of ubiquitin isolated from the gill of Pacific oyster, Crassostrea gigas. *Mol. Immunol.* **2013**, *53* (1−2), 88−98.
(24) Bleackley, M. R.; Hayes, B. M.; Parisi, K.; Saiyed, T.; Traven, A.; Potter, I. D.; van der Weerden, N. L.; Anderson, M. A. Bovine pancreatic trypsin inhibitor is a new antifungal peptide that inhibits cellular magnesium uptake. *Mol. Microbiol.* **2014**, *92* (6), 1188−1197.
(25) Gonçalves, C.; Costa, P. M. Cephalotoxins: A Hotspot for Marine Bioprospecting? *Front. Mar. Sci.* **2021**, *8*, 647344.
(26) Cho, J. H.; Sung, B. H.; Kim, S. C. Buforins: histone H2A-derived antimicrobial peptides from toad stomach. *Biochim. Biophys. Acta* **2009**, *1788* (8), 1564−1569.
(27) Li, W.; Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **2006**, *22* (13), 1658−1659.
(28) Shen, W.; Le, S.; Li, Y.; Hu, F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLoS One* **2016**, *11* (10), No. e0163962.
(29) Reina, D.; Toral, S.; Johnson, P.; Barrero, F. Improving discovery phase of reactive ad hoc routing protocols using Jaccard distance. *J. Supercomput.* **2014**, *67*, 131−152.
(30) Dau, T.; Bartolomucci, G.; Rappsilber, J. Proteomics Using Protease Alternatives to Trypsin Benefits from Sequential Digestion with Trypsin. *Anal. Chem.* **2020**, *92* (14), 9523−9527.
(31) Maillet, N. Rapid Peptides Generator: fast and efficient in silico protein digestion. *NAR: Genomics Bioinf.* **2020**, *2* (1), lqz004.
(32) Joshi, J.; Blankenberg, D. PDAUG: a Galaxy based toolset for peptide library analysis, visualization, and machine learning modeling. *BMC Bioinf.* **2022**, *23* (1), 197.
(33) Santos-Junior, C. D.; Pan, S.; Zhao, X. M.; Coelho, L. P. Macrel: antimicrobial peptide screening in genomes and metagenomes. *PeerJ* **2020**, *8*, No. e10555.
(34) Muller, A. T.; Gabernet, G.; Hiss, J. A.; Schneider, G. modlAMP: Python for antimicrobial peptides. *Bioinformatics* **2017**, *33* (17), 2753−2755.
(35) Chaudhary, K.; Kumar, R.; Singh, S.; Tuknait, A.; Gautam, A.; Mathur, D.; Anand, P.; Varshney, G. C.; Raghava, G. P. A Web Server and Mobile App for Computing Hemolytic Potency of Peptides. *Sci. Rep.* **2016**, *6*, 22843.
(36) Castillo-Mendieta, K.; Aguero-Chapin, G.; Marquez, E.; Perez-Castillo, Y.; Barigye, S. J.; Perez-Cardenas, M.; Perez-Gimenez, F.; Marrero-Ponce, Y. Multiquery Similarity Searching Models: An Alternative Approach for Predicting Hemolytic Activity from Peptide Sequence. *Chem. Res. Toxicol.* **2024**, *37* (4), 580−589.
(37) Rathore, A. S.; Arora, A.; Choudhury, S.; Tijare, P.; Raghava, G. P. S. ToxinPred 3.0: An improved method for predicting the toxicity of peptides. *bioRxiv* **2023**, 2023.2008.2011.552911.
(38) Wei, L.; Ye, X.; Sakurai, T.; Mu, Z.; Wei, L. ToxIBTL: prediction of peptide toxicity based on information bottleneck and transfer learning. *Bioinformatics* **2022**, *38* (6), 1514−1524.
(39) Aguilera-Mendoza, L.; Marrero-Ponce, Y.; Garcia-Jacas, C. R.; Chavez, E.; Beltran, J. A.; Guillen-Ramirez, H. A.; Brizuela, C. A. Automatic construction of molecular similarity networks for visual graph mining in chemical space of bioactive peptides: an unsupervised learning approach. *Sci. Rep.* **2020**, *10* (1), 18074.

https://doi.org/10.1021/acsomega.4c01959
ACS Omega 2024, 9, 43353−43367

(40) Aguilera-Mendoza, L.; Ayala-Ruano, S.; Martinez-Rios, F.; Chavez, E.; Garcia-Jacas, C. R.; Brizuela, C. A.; Marrero-Ponce, Y. StarPep Toolbox: an open-source software to assist chemical space analysis of bioactive peptides and their functions using complex networks. *Bioinformatics* **2023**, *39* (8), btad506.
(41) Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.: Theory Exp.* **2008**, *2008* (10), P10008.
(42) Ghalmane, Z.; Hassouni, M. E.; Cherifi, H. Immunization of networks with non-overlapping community structure. *Soc. Netw. Anal. Min.* **2019**, *9* (1), 45.
(43) Boldi, P.; Vigna, S. Axioms for centrality. *Internet Math.* **2014**, *10* (3−4), 222−262.
(44) Smith, T. F.; Waterman, M. S. Identification of common molecular subsequences. *J. Mol. Biol.* **1981**, *147* (1), 195−197.
(45) Aguero-Chapin, G.; Galpert, D.; Molina-Ruiz, R.; Ancede-Gallardo, E.; Perez-Machado, G.; de la Riva, G. A.; Antunes, A. Graph Theory-Based Sequence Descriptors as Remote Homology Predictors. *Biomolecules* **2019**, *10* (1), 26.
(46) Aguilera-Mendoza, L.; Marrero-Ponce, Y.; Tellez-Ibarra, R.; Llorente-Quesada, M. T.; Salgado, J.; Barigye, S. J.; Liu, J. Overlap and diversity in antimicrobial peptide databases: compiling a non-redundant set of sequences. *Bioinformatics* **2015**, *31* (15), 2553−2559.
(47) Aguero-Chapin, G.; Antunes, A.; Mora, J. R.; Perez, N.; Contreras-Torres, E.; Valdes-Martini, J. R.; Martinez-Rios, F.; Zambrano, C. H.; Marrero-Ponce, Y. Complex Networks Analyses of Antibiofilm Peptides: An Emerging Tool for Next-Generation Antimicrobials' Discovery. *Antibiotics* **2023**, *12* (4), 747.
(48) Prasertsuk, K.; Prongfa, K.; Suttiwanich, P.; Harnkit, N.; Sangkhawasi, M.; Promta, P.; Chumnanpuen, P. Computer-Aided Screening for Potential Coronavirus 3-Chymotrypsin-like Protease (3CLpro) Inhibitory Peptides from Putative Hemp Seed Trypsinized Peptidome. *Molecules* **2022**, *28* (1), 50.
(49) Qiao, L.; Li, B.; Chen, Y.; Li, L.; Chen, X.; Wang, L.; Lu, F.; Luo, G.; Li, G.; Zhang, Y. Discovery of Anti-Hypertensive Oligopeptides from Adlay Based on In Silico Proteolysis and Virtual Screening. *Int. J. Mol. Sci.* **2016**, *17* (12), 2099.
(50) Guo, H.; Richel, A.; Hao, Y.; Fan, X.; Everaert, N.; Yang, X.; Ren, G. Novel dipeptidyl peptidase-IV and angiotensin-I-converting enzyme inhibitory peptides released from quinoa protein by in silico proteolysis. *Food Sci. Nutr.* **2020**, *8* (3), 1415−1422.
(51) Udenigwe, C. C. Towards rice bran protein utilization: In silico insight on the role of oryzacystatins in biologically-active peptide production. *Food Chem.* **2016**, *191*, 135−138.
(52) Langyan, S.; Khan, F. N.; Yadava, P.; Alhazmi, A.; Mahmoud, S. F.; Saleh, D. I.; Zuan, A. T. K.; Kumar, A. In silico proteolysis and analysis of bioactive peptides from sequences of fatty acid desaturase 3 (FAD3) of flaxseed protein. *Saudi J. Biol. Sci.* **2021**, *28* (10), 5480−5489.

https://doi.org/10.1021/acsomega.4c01959
ACS Omega 2024, 9, 43353−43367