UNIVERSIDAD REGIONAL AMAZÓNICA IKIAM

Facultad de Ciencias de la Vida
Ingeniería en Biotecnología

Discovery of antimicrobial peptides in spider silk glands using Expressed Sequence Tag data

AUTOR: ALEX FABRICIO SANCHEZ YUMBO
TUTOR: PATRICIA ELENA SALERNO DOMÍNGUEZ

Napo - Ecuador. 2022

Declaración de derecho de autor, autenticidad y responsabilidad

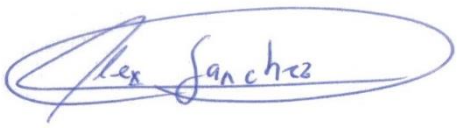
Tena, 23 de marzo de 2021

Yo, Alex Fabricio Sánchez Yumbo con documento de identidad N° 0106114606, declaro que los resultados obtenidos en la investigación que presento en este documento final, previo a la obtención del título de Ingeniero en Biotecnología son absolutamente inéditos, originales, auténticos y personales.

En virtud de lo cual, el contenido, criterios, opiniones, resultados, análisis, interpretaciones, conclusiones, recomendaciones y todos los demás aspectos vertidos en la presente investigación son de mi autoría y de mi absoluta responsabilidad.

Por la favorable atención a la presente, suscribo de usted,

Atentamente,



Alex Fabricio Sánchez Yumbo

Certificado de dirección de trabajo de integración curricular

Certifico que el trabajo de integración curricular titulado: “Discovery of antimicrobial peptides in spider silk glands using Expressed Sequence Tag data”, en la modalidad de: artículo original, fue realizado por: Alex Fabricio Sánchez Yumbo, bajo mi dirección.

El mismo ha sido revisado en su totalidad y analizado con respecto a similitud de contenido; por lo tanto, cumple con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad Regional Amazónica Ikiam, para su entrega y defensa.

Tena, 23 de marzo de 2021

Firma:

……

Patricia Elena Salerno Domínguez

C.I: 1759267857

Dedicatoria

Este trabajo está dedicado a mi madre María por su apoyo emocional en toda mi formación personal y profesional, a mi padre Luis quien tuvo un papel clave en la formación de mi carácter. A mis hermanos Byron y Juan quienes siempre me incentivaron a seguir mis estudios a pesar de estar lejos de toda mi familia.

Dedico este trabajo a mi profesor Juan Francisco Tlapanco por todas sus enseñanzas y consejos que me ayudaron a no rendirme en la vida académica y personal, a la profesora Jennifer Guevara quien fue la primera en apoyar mi tema de estudio y me guio hasta antes del inicio de la pandemia.

A mis tutores Moisés Gualapuro y Patricia Salerno. Ambos me apoyaron desde el inicio con este proyecto y otros que aparecieron en el camino. En todo este transcurso compartimos trabajos, risas, anécdotas y consejos para la vida académica y personal.

Finalmente, dedico mi trabajo a todos mis amigos quienes siempre me apoyaron cada vez que lo necesitaba incondicionalmente, espero algún día poder retribuírselos.

Contenido

Índice de Tablas

Índice de Figuras

Resumen

Los péptidos antimicrobianos (AMP) emergen como una solución novedosa a la creciente problemática de la resistencia a antibióticos. Los AMPs han sido descritos en varios organismos por métodos experimentales o *in sillico*, pero pocos esfuerzos se han hecho para explorar AMP en biomateriales prometedores como la seda de araña. La seda de araña está en un alto riesgo de infección por microorganismos al estar expuesta directamente al ambiente y varios reportes muestran que este biomaterial puede inhibir su crecimiento sugiriendo la presencia de compuestos antimicrobianos. En este estudio, diseñé e implementé un pipeline para extraer AMP a partir de marcadores de secuencia expresada en glándulas de seda de araña basado en alineaciones con bases de datos de AMP y perfiles de Modelos Markov Escondidos de familias de AMPs. Usando seis sets de datos EST, descubrí cinco péptidos descritos y tres nuevos que son expresados en glándulas de seda. Se sugiere que los péptidos conocidos son parte del sistema inmunológico humoral. Uno de los tres péptidos novedosos tiene posible actividad antimicrobiana debido a su estructura anfipática, demostrando el potencial de las glándulas de la seda como fuente de AMPs. A pesar de que la mayoría de péptidos fueron descubiertos en glándulas de seda, deben ser exploradas en la seda una vez que esté en el ambiente.

Palabras clave: péptidos antimicrobianos, bioinformática, marcador de secuencia expresada, seda de araña, glándulas de seda

Abstract

Antimicrobial peptides (AMP) emerge as a novel solution to the increasing problem of antibiotic resistance. AMPs have been described in several organisms by experimental or i*n silico* analysis, but little efforts have been made to screen AMP in promising biomaterials such as spider silk. Spider silk is at high risk of microorganism infections from being exposed directly to the environment, and several reports show that this biomaterial can inhibit their growth. Although the mechanisms behind antimicrobial defense of spider silk is still unknow, studies suggest the presence of antimicrobial compounds. Here, I designed and implemented a pipeline to mine AMP from Expressed Sequence Tag (EST) data of spider silk gland tissues based on alignments with AMP databases and Hidden Markov Models profiles of AMP families. Using six EST datasets, we discovered five known and three novel peptides that are expressed in spider silk glands. Known peptides are suggested to be part of the humoral immune system of spiders. One of the three novel peptides has the potential to be antimicrobial due to its amphipathic structure, demonstrating the potential of spider silk glands as a source of AMPs. Although most peptides were discovered from silk glands, they need to be screened in silk once it is in the environment.

Keywords: antimicrobial peptide, bioinformatics, Expressed Sequence Tag, spider silk, silk glands

Abstract

Antimicrobial peptides (AMP) emerge as a novel solution to the increasing problem of antibiotic resistance. AMPs have been described in several organisms by experimental or i*n silico* analysis, but little efforts have been made to screen AMP in promising biomaterials such as spider silk. Spider silk is at high risk of microorganism infections from being exposed directly to the environment, and several reports show that this biomaterial can inhibit their growth suggesting the presence of antimicrobial compounds. Here, I designed and implemented a pipeline to mine AMP from Expressed Sequence Tag (EST) data of spider silk gland tissues based on alignments with AMP databases and Hidden Markov Models profiles of AMP families. Using six EST datasets, I discovered five known and three novel peptides that are expressed in spider silk glands. Known peptides are suggested to be part of the humoral immune system of spiders. One of the three novel peptides has the potential to be antimicrobial due to its amphipathic structure, demonstrating the potential of spider silk glands as a source of AMPs. Although most peptides were discovered from silk glands, they need to be screened in silk that have been exposed to the environment.

Introduction

Increasing antibiotic resistant infections is an important and severe problem that affects humans worldwide [1]. Bad practices in disease control and antibiotic abuse have made many antibiotics obsolete, increasing the vulnerability to antibiotic-resistance pathogens and decreasing available treatments [2]. Antimicrobial peptides (AMP) emerge as novel candidates of antibiotics because they target the membrane of microorganisms, while traditional antibiotics act on specific target proteins that are altered, resulting in a less susceptible receptor [3]. AMPs have a high potency and selectivity, can inhibit a wide spectrum of microorganisms, and have low toxicity and low tissue accumulation [4]. However, few drugs are based on AMPs, so it is of paramount importance to increase screening and testing of AMP as antibiotic treatments. These compounds had been isolated and described in several organisms of different taxa, but few efforts had been made to describe AMPs in promising biomaterials such as spider silk glands.

Spider silk is a versatile biomaterial with several biological functions such as spider web construction for catching and wrapping prey, communication through vibrations, building egg sacs, and offspring protection [5]. It has enormous potential in the development of biomedical materials [6] due to its mechanical properties [7] and biocompatibility in mammal cell lines [8]. Spiders of the super family Araneoidea produce silk from seven specialized glands: flagelliform, aggregated, major ampullate, minor ampullate, aciniform, piriform and tubilifrom [9]. These glands, except when aggregated, produce auto assembly proteins called spidroins that are the structural base of silk fibers [10]. Aggregated gland produce glue cover, which is composed of organic molecules such as glycoproteins, amino acids, fatty acids, amides and others [11]. A combination of different fibers and glue cover allow silk to

be produced with different properties and functions. Although silk may confer a versatile material for spiders, it also poses an infection risk from its surrounding environment. Several studies report antimicrobial activity in silk of different spider species suggesting antimicrobial compounds as the mechanisms for this feature (Table 1). Some of the tested microorganisms listed in Table 1 are in the World Health Organization list of priorities for new drugs resistant pathogens [12].

Table 1. Microorganisms inhibition assays with spider silk

Family	Specie	Silk type	Microorganisms	Reference
Agelenidae	*Tegenaria domestica*	Dragline and capture	*Bacillus subtilis,*	[8]
			Escherichia coli	
Araneidae	*Nephila pilipes*	Dragline	*E. coli*	[13]
			Staphylococcus aureus	
			Pseudomonas aureginosa	
Pholcidae	*Pholcus phalangioides*	Cobweb	*E. coli*	[14]
			Listeria monocytogenes	
Eresidae	*Stegodyohus dumicola*	Capture and refuge	*Bacillus thuringiensis*	[15]
Araneidae	*Cyclosa confraga*	Not specified	*Streptococcus* sp.	[6]
			Acinetobacter sp.	
Lycosidae	*Pardosa brevivulva*	Not specified	*B. megaterium*	[16]
			Salmonella typhi	
			Klebsiella pneumoniae	
			Aspergillus flavus	
			Candida albicans	
			Ustilago maydis	
			Alternaria solan	
Theridiidae	*Latrodectus hesperus*	Gumfoot	*E. coli*	[17]

Although antimicrobial activity of spider silk is still unknown, it has potential to be a source of novel AMPs against pathogenic microorganisms. For example, *Pseudomonas aeruginosa* and *Staphylococcus aureus* were used in inhibition assays with dragline silk of *Nephila pilipes*; results showed a significant decrease of growth capacity of these bacteria [13]. In the same way, *Streptococcus pneumoniae*

was tested against *Cyclosa congrafa* silk with similar results [6]. These antimicrobial properties can be explored using several approaches, but by far the most used are *in silico* analyses since there are many available databases that can be explored and mined bioinformatically.

Bioinformatic approaches offer a strong research tool as the enormous amount of transcriptomic data in databases allows us to easily discover many potential compounds from existing data. Some studies focus on the discovery of novel compounds from expressed sequence tags (ESTs) of mammals, plants and insects using tools such as Open Reading Frame (ORF) predictors, alignment tools and Hidden Markov Models profiles (HMM). ESTs are short sequence reads (200-800 bases) derived from 5' or 3' ends of cDNA libraries that provide information about expressed genes [18]. These reads are analyzed to get ORFs that are translated into amino acid (AA) sequences and then align with known AMPs. Missed reads are found applying HMM profiles of AMP families; these profiles are statistical models that describe the evolution of events (transitions, indels, deletions, mutations) in a sequence [19,20]. Biomining of EST data has been successfully applied to discover AMPs in birds [21], plants [22], and even arthropods [23,24]. Thus, bioinformatic mining of EST data on spider silk offers a unique opportunity to explore this biomaterial.

The process of mining EST data for potential AMPs can be divided in three main steps: curation, mining and analyzing. Curation of EST involves trimming adapter sequences of the used cloning vector in order to avoid clustering heterologous sequences in further steps. This process can be done with tools like Figaro, which calculates oligo frequencies in the 5' region [25], and cross_match [26] that compares reads with the cloning vector sequence. Trimmed sequences are then clustered to avoid analyzing individual sequences that are homologous to others with tools such as MeShClust that implements an optimized algorithm of clustering process [27].

The mining step of EST data involves ORF prediction, database alignments and HMM profiles. Conventional ORF prediction programs, which are specialized for proteins, are not useful for mining AMPs with varying sequence length from less than 10 to hundreds of AA. MiPepid (Micro-Peptid Identification) is a state of the art Machine Learning tool developed to predict short ORFs that code for small peptides (<= 100 AA) making it ideal for mining ORFs from ESTs [28] Predicted AA sequences are then compared with AMP databases to know if they had been described. LAMP and CAMP are the two largest and most complete AMP databases, with more than 23,000 and 8,000 sequences respectively [29,30], which allows us to explore AMPs previously described in various taxa . If no-hit sequences are found, then the AMPs present in them can also be explored using HMM profiles. CAMPSign is a tool to predict AMPs comparing AA sequences against 75 HMM profiles of 45 AMP families. This tool implements more than one profile by AMP family based on their size, allowing them to be classified by family and length [31].

The last step in the pipeline, analyzing, involves determining physicochemical and secondary structure properties of the candidate AMPs found in the previous step. DBAASP v3.30-Property calculator is used to determine the physicochemical properties such as net charge and hydrophobicity [32]. Secondary structure can be determined by Pep-Fold, a peptide structure specialized tool [33]. This information is used to analyze the position of AA residues and later visualize the peptide. For example, HeliQuest server allows to plot helical wheels that highlight polar and hydrophobic regions [34]. All these tools for curating, mining and analyzing EST sequences are essential for finding candidate AMPs from any organism or tissue.

In this study, I designed a pipeline based on the above-mentioned process for AMP discovery from EST data. We implement this pipeline to explore the potential of spider silk glands as a source of AMPs. Using spider silk gland libraries obtained from the GenBank-EST database I found eight peptide sequences both known and unknown and evaluated their potential antimicrobial activity.

Materials and methods

EST (Expressed Sequence Tag) sequences were obtained from the GenBank database (www.ncbi.nlm.nih.gov/genbank/dbest/). Samples that contain at least one spider silk gland tissue were selected and sequences were downloaded in FASTA format. I obtained and analyzed six EST datasets from five species that include samples with all silk gland tissues, only aggregated silk gland tissue, and whole spider tissue. Figaro v1.05 [25] and cross_match v1.090518 software were used to trim adapter sequences (Fig 1, step 2). These programs revealed that EST sequences were already trimmed for adapter sequences.

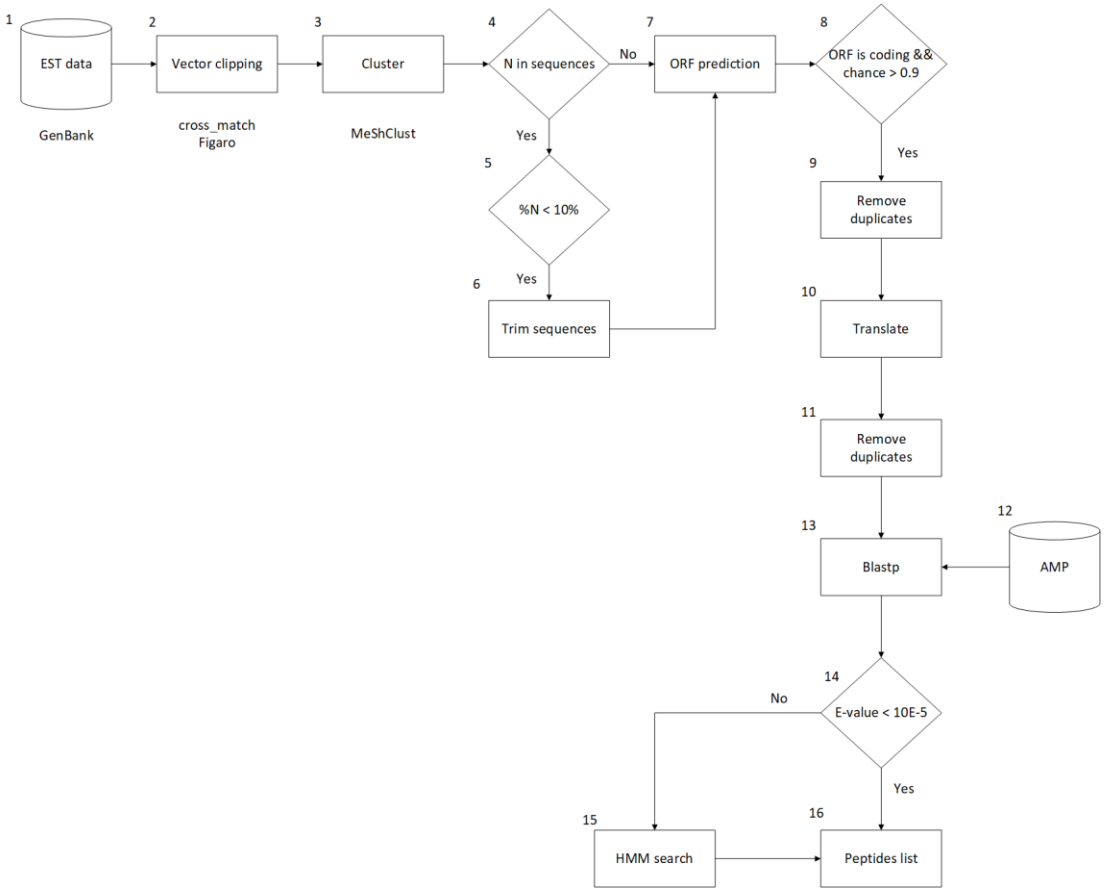


Fig 1. Diagram for AMP discovery using silk gland EST sequences.

Original EST datasets were obtained from the GenBank-EST database considering tissues with at least one silk gland (Step 1), then sequences were clustered using MeShClust program (Step 3). Clustered sequences that contain 'N' char on them were trimmed (Step 4-6), and complementary sequences were obtained for each read (Step 7). MiPepid program predicted all ORF and the chance of being coding (Step 8), duplicated sequences were deleted (Step 9). Coding and non-duplicated ORFs were translated into AA sequences (Step 10). AA duplicated sequences were deleted (Step 11) and were BLAST against the LAMP database (Step 12-14). Non-aligned sequences were compared against HMM profiles of CAMPSign server (Step 15).

The ESTs were clustered using MeShClust [27] with the following parameters: id 0.75, kmer 5, delta 5, iterations 20, align, sample 3000, pivot 40, and threads 1. Then, clustered sequences were merged using multiple alignment to obtain consensus sequences. Both consensus and non-clustered sequences with less than 10% of "N's" were selected and if N's were present, they were trimmed (Fig 1, step 4-6). ORFs were determined with the MiPepid program [28], and coding sequences with a chance greater than 90% were translated into AA sequences.

AA sequences were blasted against the LAMP database with an e-value of 1E-4. No-hit sequences were then compared with HMM profiles of 45 AMP families using CAMPSing [30,31]. Match blast sequences were compared with the CAMPR3 database. Matches were compared against the non-redundant protein sequence database. The secondary structure of novel sequences was modelled with Pep-Fold server [33] and graphed with BIOVIA Discovery Studio Visualizer software. To determine hydrophobic and polar regions of novel peptides I plotted helical wheels using HeliQuest server [34]. Net charge and mean hydrophobic moment were calculated with DBAASP v3.30-Property calculator tool using the Moon and Fleming scale [32]. Further details on data, methods, and pipeline can be found in the GitHub repository for this project (github.com/alexsanyum/AS_Thesis_AMPSpiderSilk)

Results

I obtained a wide range of retained reads, ORFs and AA translated sequences for each of the EST dataset (Fig 2). For aggregated silk gland data, I obtained 66 reads from *Leucauge venusta* dataset1 [35] and 59 reads from *Parasteatoda tepidariorum* [36]. For all silk glands, I obtained 86 reads from *L. venusta* dataset2 [37], 512 reads from *Nephila antipodiana* [38], and 51 reads from *Steatoda grossa* [39]. For whole spider, 245 reads were obtained from *Latrodectus hesperus* [40]. After screening for vector residues in the sequences and not finding any residues, I obtained a total of 832 clusters for all datasets. After, sequences containing Ns were filtered and trimmed (Fig 1), resulting in a total of 847 sequences. Forward and complementary sequences (1,692 reads) were used for predicting ORFs with MiPepid obtaining 16,114 ORFs, of which 8,810 were classified as coding. After removing duplicate sequences, I translated each ORF into AA sequences and further removed translated duplicates (synonymous codons), retaining a total of 6,969 reads that were used in downstream analyses (blast and comparison with HMM profiles).

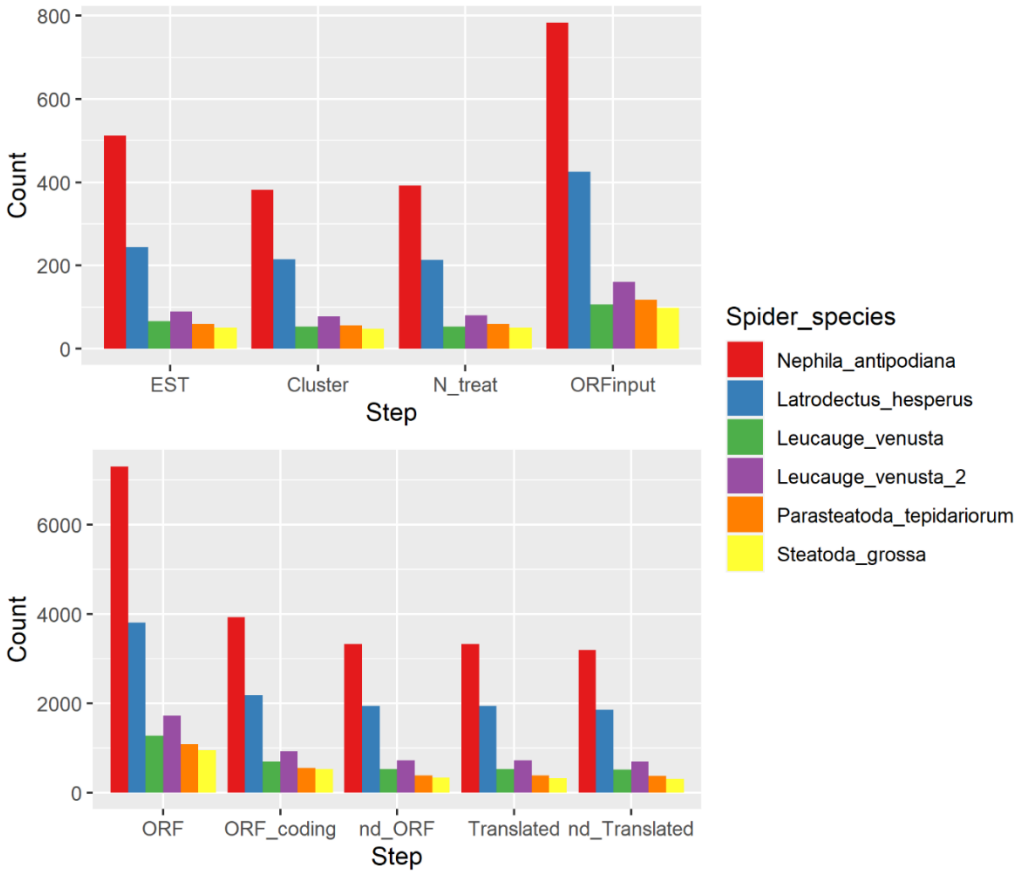


Fig 2. Summary of sequences retained at each step of the pipeline from the six datasets used.

I obtained BLAST matches for five AA sequences against known AMPs (using the CAMP database), and three matches using predictive HMM profiles (Table 2). For *L. hesperus* I found two matches: LhB_seq1 and LhB_seq2; for *N. antipodiana* I found two matches NaB_seq1 and NaB_seq2; for *S. grossa* I found one match SgB_seq1; I did not find known peptides in the *L. venusta* and *P. tepidariorum* datasets. Half of the peptides matches have positive net charges, while the rest have neutral charge. Mean hydrophobic moments vary from 0.26 to 1.97.

Table 2. Blast and HMM matches that produced significant alignments, their respective net charge, mean hydrophobic moment and their CAMP accession ID.

Dataset	Name	Method used	Match fragment sequence	Net charge	Mean Hydrophobic moment	Accession ID
L. hesperus	LhB_seq1	Blastp	KVHGSLARAGKVKGQTPKVEKQEKKKRKTGRAKRRMQFNRRFVNVVVTFGRKKGPNSNS	+18	0.32	CAMPSQ3754
L. hesperus	LhB_seq2	Blastp	MQFNRRFVNVVVTFGRKKGPNSNS	+5	0.45	CAMPSQ3754
L. hesperus	LhH_seq1	HMM	SPTGLNTVYASLT	0	0.82	CAMPBacH32
L. venusta 2	Lv2H_seq1	HMM	LWKTLLK	+2	1.97	CAMPDerH28
N. antipodiana	NaB_seq1	Blastp	MQIFVKTLTGKTITLESEPSDTIENVKTKIQTKKASPQ	+1	0.52	CAMPSQ3702
N. antipodiana	NaB_seq2	Blastp	GGKAGQDACKGDGGGPLVCFRSCAGGKAGQDACKGDGGGPLVCFRSDNSYTVAGLVSWGIDCGQEGIPGVYVNVKKYNDWIVSKTQKPIENY	0	0.264	CAMPSQ3345
S. grossa	SgB_seq1	Blastp	FYYNDVAKKCEIFYYGGCKGNENNFPSEDHCKEAGG	-2	0.26	CAMPSQ2798
S. grossa	SgH_seq1	HMM	AAGNAAKGVASDA	0	0.93	CAMPDerH

Although CAMP BLAST hits of known peptides matched with AMPs from non-spider or arachnid species, BLAST using the non redundant protein sequence database matched with the same peptide sequences found in spider and other

species (Table 3). LhB_seq1 and LhBseq2 matched with the same sequence in CAMP and in the non redundant protein sequence database. NaB_seq1 aligned against Ubiquitin of bacteria, parasite, and a rodent species, while NaB_seq2 matched against phenoloxidase activating factors of different spider species. SgB_seq1 matches against Kunitz protein of parasite, crab and a rodent species.

Table 3. Description and source organism of blastp matches of CAMP and top three non-redundant protein sequence database hits.

Sequence	CAMP	Source organism	Source Organisms | Non redundant database
LhB_seq1	Ubiquicidin	*Mus musculus*	• *S. dumicola* | Ubiquitin-like protein 1-40S ribosomal protein S27a • *Dolomedes sulfureus* | Ubiquitin-like 40S ribosomal S30 protein fusion • *Araneus ventricosus* | 40S ribosomal protein S30
LhB_seq2	Ubiquicidin	*Mus musculus*	• *S. dumicola* | Ubiquitin-like protein 1-40S ribosomal protein S27a • *P. tepidariorium* | Ubiquitin-like protein FUBI • *S. mimosarum* | 40S ribosomal protein S30
NaB_seq1	CgUbiquitin	*Crassostrea gigas*	• *Flavobacterium sp.* | Ubiquitin • *Paragonimus kellicotti* | Hypothetical protein AH37_07978 • *Octodon degus* | Ubiquitin-like
NaB_seq2	TCP	*Homo sapiens*	• *Araneus ventricosus* | Phenoloxidase-activating factor 2 • *Argiope bruennichi* | Phenoloxidase-activating factor 2 like protein • *P. tepidariorum* | Serine proteinase stubble-like
SgB_seq1	Luxuriosin*	*Acalolepta luxuriosa*	• *Schistosoma haematobium* | Putative kunitz-type protease inhibitor • *Chionoecetes opilio* | Amyloid-like protein 2 • *Mus caroli* | kunitz-type protease inhibitor 3

*Non significant alignment

In order to find AMPs that could be missed by blastp, I used the CAMPSign tool. MM profiles of CAMPSign returned two sequences classified as dermaseptin for *L. venusta* dataset2 named as Lv2H_seq1 and for *S. grossa* named as SgH_seq1. Lv2H_seq was classified as a dermaseptin of a fixed length of 28 AAs,

while SgH_seq1 has variable AA length. I also discovered one bacteriocin in *L. hesperus* LhH_seq1 that was aligned against a profile of 32 AA in length. All novel peptides show helical secondary structures (Fig 3). Helical wheel plots show that LhH_seq1 has polar and hydrophobic regions while for SgH_seq1 there are not. Lv2H_seq1 helical wheel was not plotted because its short length of 7 AA is not supported by the used server.

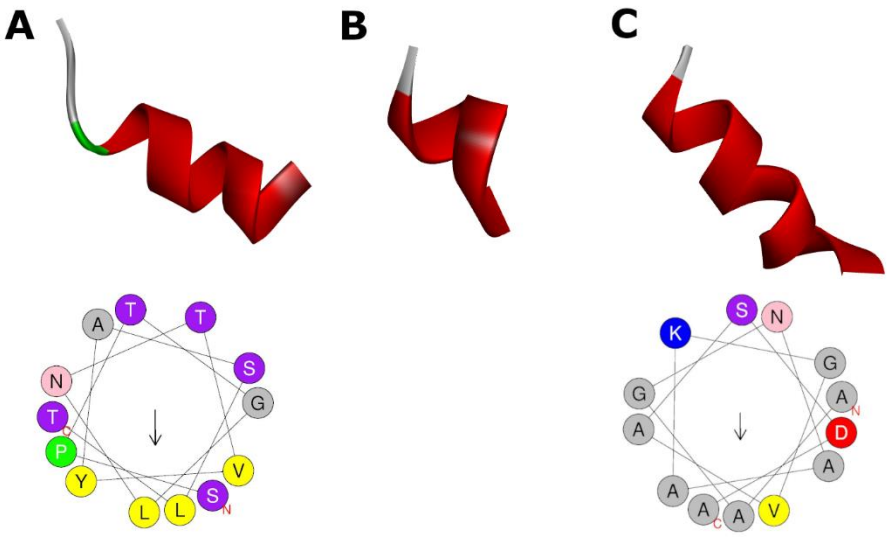


Fig 3. Secondary structure models (top) and helical wheel plots (bottom) of novel peptides.

Purple: polar residues, yellow: hydrophobic residues. (A) LhH_seq1, (B) Lv2H_seq1, (C) SgH_seq1.

Discussion

Here, I developed and implemented a pipeline to discover AMPs from EST data, and found a total of five described and three novel peptides from spider silk gland tissues that have not been previously reported. Peptide discovery represents the first approach to investigating the potential of AMPs in silk glands. Previous research into exploring properties of silk have mainly focused on understanding mechanical properties [7,41,42], silk proteins and their properties [5,40,43] and

synthesis [44–46], but no previous research had examined antimicrobial compounds in silk glands.

The analysis includes three important achievements used separately in previous studies: MiPepid ORF finder program, blastp and CAMPSign. MiPepid is able to predict potential coding ORF [28], blastp can compare those translated ORF against known peptides [47], and CAMPSign can find missed peptides through HMM profiles [30,31]. Although only five known peptides and three novel peptides sequences were discovered with the pipeline, this number is high compared to similar studies on different species where thousands of EST reads were analyzed resulting in a small number peptides. For example, mining more than 420,000 reads (compared to 1,022 reads analyzed in this study) resulted in nine novel peptides in *Gallus gallus* [21]; further, in a *Brassica napus* analysis of more than 810,000 ESTs, 972 genes matched against known AMP [22]. The analysis of 1,305 sequences of *Phoneutria nigriventer* venom glands resulted in a discovery of 51 cysteine-rich peptides [23]. It can be suggested that while the number of reads increase, the chance to find peptides also does. AMPs in the two datasets derived from aggregated gland tissue were not found, suggesting their potential absence in this gland. However, it is likely that we did not find any AMPs due to the much smaller datasets (Fig 2). The largest datasets used (*N. antipodiana*, *L. hesperus*, and *L. venusta* 2) resulted in six of the eight peptides found. However, the smallest dataset (*S. grossa*), that was obtained from all silk gland tissues, resulted in a total of two peptide matches, suggesting other factors may have also caused the absence of peptides sequences in the two aggregated gland tissue samples.

Many factors could affect results in aggregated silk gland datasets (*L. venusta* and *P. tepidariorum*), such as the nature of the gland and the software parameters that were used. Aggregated silk glands do not produce silk fibers proteins as the other six types. Instead, this gland produces the glue cover of silk that is conformed by a complex mix of inorganic salts, proteins, peptides, lipids and aromatic toxins [48]. Current knowledge about aggregated silk gland comes only from orb-web and

cobweb spiders, and very few peptides have been associated to this gland, such as bradykinin in *Trichonephila clavipes* [49] and SCP-1, and SCP-2 in *L. hesperus* [11] (none of which have been tested for antimicrobial activity). Despite the low number of peptides described for aggregated silk glands, it is the only gland that has been related to peptides.

Various factors in the pipeline, such as trimming sequences and MiPepid filter methods used instead of the nature of the gland itself could potentially have significant effects in the results and their interpretation. For example, the process of trimming sequences in order to eliminate unknown bases (Ns) (Fig 1. Step 4-6) may directly affect the amount of ORFs that can be found especially in datasets where the amount of Ns was big. Trimming sequences each time that a N was found reduces readable ORF (it is possible to trim sequences after the beginning of an ORF) affecting further process of the pipeline. For example, *P. tepidoriorum* dataset, that did not return any AMP, have a great amount of Ns specially in the 3' regions. In contrast, *N. antipodiana* dataset did not contain any N in their reads, and was the dataset with more AMPs found.

In the same way, MiPepid program predicts ORFs and estimates the probability of it being coding or not. Here we apply a random high filter and choose only genes with a chance of coding greater than 90% (Fig 1, step 8). This filter dramatically reduces the number of sequences in further steps of the pipeline. For example, in the troubleshooting phase, we accidentally analyzed ORFs with a probability less than 90%, finding peptides in all datasets. The lack of AMPs found in aggregated silk gland is likely to be more associated with the quality data and the pipeline instead of nature of the gland. To confirm this, it will be needed to analyze no ambiguous aggregated silk gland EST data and reduce MiPepid results filter. However, MiPepid filter must be permuted to obtain an optimized parameter.

Known peptides

Known peptides found in the silk gland ESTs had never been reported as AMPs for spiders nor other arachnids. CAMP queries revealed that AMP matches are from non-spider species such as rodents, parasites, bacteria, insects, and mollusks (Table 3). These results are due to the use of LAMP and CAMP, specialized databases. These databases are specific for AMPs that have been validated by experimental or in *silico* methods [29,30]. Thus, like a few spiders and arachnids peptides are screened for antimicrobial properties, it is not surprising that all hit results were against peptides where sources are from non-spider species. In order to contrast this, I analyze the same peptides against the non redundant protein database (Table 3).

LhB_seq1 and LhB_seq2, which originally matched against Ubiquicidin, hit with spiders 40S ribosomal proteins, suggesting that these AMPs can be generate through protein proteolysis. Peptides generated by proteolysis are common in the humoral immune system of many species [50]. Given that these samples were from while body samples, it is likely that these peptides are part of *L. hesperus* innate inmune system instead of silk-specific microbial defense.

NaB_seq1 was matched against ubiquitin of different taxa (Table 3). Ubiquitin is a peptide with proteolytic functions and antimicrobial activity, and is part of the humoral immune system of many taxa. It acts as a signal in processes such as bacterial infection defense, DNA damage repair, gene regulation [51], and innate antimicrobial activities [52,53]. Ubiquitin isolated from *C. gigas* shows activity against *Streptococcus iniae* and *Vibrio parahaemolyticus* [53]. N and C terminal fragments of bovine ubiquitin have activity against Gram positive and Gram negative bacteria and filamentous fungi [52]. The presence of ubiquitin in ESTs derived from spider silk glands suggests two mechanism of actions, one as a signal against bacterial infection and other as an AMP.

NaB_seq2 matched against TCP (Thrombin-derived C terminal peptide) in CAMP and phenoloxidase (PO) activating factors in the non redundant protein sequence database. Both TCP and PO activating factors are related to serine protease cascade, an immune response against microorganism infection on vertebrates and invertebrates [54]. In spiders, humoral immune system activates due to the recognition of pathogen-associated molecular patterns (PAMPs). PAMP recognition can deliver melanin in the hemolymph [55]. Serine proteases and PO plays a key role on that process. First, PAMPs are known to activate a series of serine proteases that activate proPO-activating enzymes which activate PO enzymes. PO catalyses the oxidation of phenolic compounds into quinone, which is polymerized by generating melanin [56]. Despite PO activation factors playing a key-role in the humoral immune response against microorganisms [56], the same molecule has antimicrobial activity. Thrombin, a human serine protease, is amphipathic, cationic and has a helical structure, all common properties in helical AMPs that have been found to have antibacterial and antifungal activity [57]. This suggest NaB_seq2 is part of the humoral immune system in the serine protease cascade and potentially also an AMP.

SgB_seq1 matched non-significantly to the insect peptide luxuriosin, and Kunitz protease inhibitors of parasites, arthropods and rodents. Luxuriosin has a Kunitz domain in its structure. Kunitz family inhibitors regulated the serine protease cascade in arthropods [58]. Thus, like previous described AMPs, SgB_seq1 is suggested to be part of the humoral immune system of spiders as a regulating factor of the serine protease cascade.

Novel peptides

Novel peptides found in this study all have common properties of the α-helix AMP group. LhH_seq1 shows helical and amphipathic structure hence the chance to have antimicrobial activity is high despite its neutral charge (Fig 3.A, Table 2). The positively charged side allows it to interact with the negatively charged membrane

and then insert itself into the non-polar region disrupting it, forming pores, barrels or breaking down the bilayer structure [59]. It was described that positive net charge is not a critical parameter for antimicrobial activity but it is for hemolytic activity. While net charge and number of positive charged increase, hemolytic activity also does [60]. In this context, LhH_seq1 likely has high antimicrobial activity and also low hemolytic activity making it an ideal candidate for further testing.

In contrast, SgH_seq1 did not show AMP properties and Lv2H_seq1 was too short to analyze AA residues positions. Despite SgH_seq1 having a helical structure, it does not have polar and hydrophobic sides (Fig 3.A), impeding electrostatics interaction with membranes. However, it is possible to modify AA sequence content in order to get amphipathic structure [61], yet this method requires various forms of validation.

All novel peptides found in this study should be validated by *in silico* or experimental methods. Here, I implemented HMM profiles to obtain previously unknown AMPs in spider silk glands, yet further research is needed. Although other bioinformatics tools exist for predicting AMP activity, such as molecular docking (a tool that allows to simulate drug-membrane interactions and predict membrane disrupting mechanisms [62]) experimental methods are more widely accepted as validation of AMP candidates. Predicted AMPs are produced through chemical synthesis and then used in antagonist assays [24]. Target microorganisms must be focus on those present in Table 1 which have been proved to be inhibited by spider silk.

Analyzed data shows that spider silk glands express genes that contain fragments with potential antimicrobial activity. However, this does not imply that silk will contain those AMPs once it is in the environment. Two peptides SCP-1 and SCP-2 have already been isolated from silk produced by *L. hesperus* and have been proposed to have antimicrobial activity due to structural properties that allow peptides to bind metal ions and releasing them in microorganism infections inhibiting

their growth [11]. Other known peptides are related to the humoral immune system, which is in the hemolymph of spiders [55], but not necessarily in the silk. Thus, further research is needed to confirm the presence of these and other AMPs in silk itself.

Conclusions

I developed a pipeline to discover AMPs in silk glands using ORF predictors, blast and HMM profiles using EST data. This method allowed me to find five known and three novel peptide sequences with potential antimicrobial activity. Known peptides are suggested to be a part of the humoral immune system of spiders but also have antimicrobial properties on their own. On the other hand, only one novel peptide LhH_seq1 is predicted to be an AMP with low hemolytic activity. All these sequences must be validated using molecular docking tools or preferably via synthesis and antagonist experimental assays. Despite our findings of six peptides mined from silk gland ESTs, further research is needed to confirm their presence in spider silk that has been exposed to the environment.

Acknowledgments

Research in this article was completed at Universidad Regional Amazónica Ikiam, Ecuador, under the supervision of Dr. Patricia Salerno. I also thanks to M.Sc Moises Gualapuro for his support in the learning, designing, implementing and writing process.

References

1. Périchon B, Courvalin P, Stratton CW. Antibiotic resistance. Encyclopedia of Microbiology. 2019. pp. 127–139. doi:10.1016/B978-0-12-801238-3.02385-0
2. Rather IA, Kim BC, Bajpai VK, Park YH. Self-medication and antibiotic resistance: Crisis, current challenges, and prevention. Saudi Journal of Biological Sciences. 2017. pp. 808–812. doi:10.1016/j.sjbs.2017.01.004
3. Baltzer SA, Brown MH. Antimicrobial peptides-promising alternatives to conventional antibiotics. Journal of Molecular Microbiology and Biotechnology. 2011. doi:10.1159/000331009
4. E. Greber K, Dawgul M. Antimicrobial Peptides Under Clinical Trials. Curr Top Med Chem. 2017;17: 620–628. doi:10.2174/1568026616666160713143331
5. Römer L, Scheibel T. The elaborate structure of spider silk: structure and function of a natural high performance fiber. Prion. 2008. pp. 154–161. doi:10.4161/pri.2.4.7490
6. Tahir HM, Qamar S, Sattar A, Shaheen N, Samiullah K. Evidence for the antimicrobial potential of silk of cyclosa confraga (Thorell, 1892) (Araneae: Araneidae). Acta Zool Bulg. 2017;69: 593–595.
7. Porter D, Guan J, Vollrath F. Spider silk: Super material or thin fibre? Adv Mater. 2013;25: 1275–1279. doi:10.1002/adma.201204158
8. Wright S, Goodacre SL. Evidence for antimicrobial activity associated with common house spider silk. BMC Res Notes. 2012;5. doi:10.1186/1756-0500-5-326
9. Tillinghast EK, Townley MA. Silk Glands of Araneid Spiders. 1993. pp. 29–44. doi:10.1021/bk-1994-0544.ch003
10. dos Santos-Pinto JRA, Lamprecht G, Chen WQ, Heo S, Hardy JG, Priewalder H, et al. Structure and post-translational modifications of the web silk protein spidroin-1 from Nephila spiders. Journal of Proteomics. 2014. pp. 174–185. doi:10.1016/j.jprot.2014.01.002
11. Hu X, Yuan J, Wang X, Vasanthavada K, Falick AM, Jones PR, et al. Analysis of aqueous glue coating proteins on the silk fibers of the cob weaver, Latrodectus hesperus. Biochemistry. 2007;46: 3294–3303. doi:10.1021/bi602507e
12. Tacconelli E, Carrara E, Savoldi A, Harbarth S, Mendelson M, Monnet DL, et al.

Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. Lancet Infect Dis. 2018;18: 318–327. doi:10.1016/S1473-3099(17)30753-3

13. Amaley A, Gawali AA, Akarte SR. Antibacterial nature of dragline silk of *Nephila pilipes* (Fabricius, 1793). Indian Soc Arachnol. 2014;3: 1–8. Available: http://indianarachnology.com/ija/wp-content/uploads/ija_2014_v3_n1_p2_8_11.pdf
14. Roozbahani H, Asmar M, Ghaemi N, Issazadeh K. Evaluation of Antimicrobial Activity of Spider Silk Pholcus Phalangioides Against Two Bacterial Pathogens in Food Borne. Int J Adv Biol Biomed Res. 2014;2: 2197–2199.
15. Keiser CN, DeMarco AE, Shearer TA, Robertson JA, Pruitt JN. Putative microbial defenses in a social spider: immune variation and antibacterial properties of colony silk. J Arachnol. 2015;43: 394–399. doi:10.1636/arac-43-03-394-399
16. Phartale NN, Kadam TA, Bhosale HJ, Karale MA, Garimella G. Exploring the antimicrobial potential of Pardosa brevivulva silk. J Basic Appl Zool. 2019;80. doi:10.1186/s41936-019-0102-6
17. Alicea-Serrano AM, Bender K, Jurestovsky D. Not all spider silks are antimicrobial. J Arachnol. 2020;48: 84–89. doi:10.1636/0161-8202-48.1.84
18. Nagaraj SH, Gasser RB, Ranganathan S. A hitchhiker's guide to expressed sequence tag (EST) analysis. Briefings in Bioinformatics. 2007. pp. 6–21. doi:10.1093/bib/bbl015
19. Yoon B-J. Hidden Markov Models and their Applications in Biological Sequence Analysis. Curr Genomics. 2009. doi:10.2174/138920209789177575
20. Franzese M, Iuliano A. Hidden markov models. Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics. 2018. doi:10.1016/B978-0-12-809633-8.20488-3
21. Lynn DJ, Higgs R, Gaines S, Tierney J, James T, Lloyd AT, et al. Bioinformatic discovery and initial characterisation of nine novel antimicrobial peptide genes in the chicken. Immunogenetics. 2004;56: 170–177. doi:10.1007/s00251-004-0675-0
22. Ke T, Cao H, Huang J, Hu F, Huang J, Dong C, et al. EST-based in silico identification and in vitro test of antimicrobial peptides in Brassica napus. BMC Genomics. 2015;16. doi:10.1186/s12864-015-1849-x

23. Paiva ALB, Mudadu MA, Pereira EHT, Marri CA, Guerra-Duarte C, Diniz MRV. Transcriptome analysis of the spider Phoneutria pertyi venom glands reveals novel venom components for the genus Phoneutria. Toxicon. 2019;163: 59–69. doi:10.1016/j.toxicon.2019.03.014
24. Yoo WG, Lee JH, Shin Y, Shim JY, Jung M, Kang BC, et al. Antimicrobial peptides in the centipede Scolopendra subspinipes mutilans. Funct Integr Genomics. 2014;14: 275–283. doi:10.1007/s10142-014-0366-3
25. White JR, Roberts M, Yorke JA, Pop M. Figaro: A novel statistical method for vector sequence removal. Bioinformatics. 2008;24: 462–467. doi:10.1093/bioinformatics/btm632
26. Gordon D. Viewing and editing assembled sequences using Consed. Curr Protoc Bioinforma. 2003;Chapter 11: Unit11.2. doi:10.1002/0471250953.bi1102s02
27. James BT, Luczak BB, Girgis HZ. MeShClust: an intelligent tool for clustering DNA sequences. Nucleic Acids Res. 2018;46: e83. doi:10.1093/nar/gky315
28. Zhu M, Gribskov M. MiPepid: MicroPeptide identification tool using machine learning. BMC Bioinformatics. 2019;20. doi:10.1186/s12859-019-3033-9
29. Ye G, Wu H, Huang J, Wang W, Ge K, Li G, et al. LAMP2: A major update of the database linking antimicrobial peptides. Database. 2020. doi:10.1093/database/baaa061
30. Waghu FH, Idicula-Thomas S. Collection of antimicrobial peptides database and its derivatives: Applications and beyond. Protein Sci. 2020;29: 36–42. doi:10.1002/pro.3714
31. Waghu FH, Barai RS, Idicula-Thomas S. Leveraging family-specific signatures for AMP discovery and high-throughput annotation. Sci Rep. 2016;6. doi:10.1038/srep24684
32. Pirtskhalava M, Amstrong AA, Grigolava M, Chubinidze M, Alimbarashvili E, Vishnepolsky B, et al. DBAASP v3: Database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics. Nucleic Acids Res. 2021. doi:10.1093/nar/gkaa991
33. Maupetit J, Derreumaux P, Tuffery P. PEP-FOLD: An online resource for de novo peptide structure prediction. Nucleic Acids Res. 2009. doi:10.1093/nar/gkp323

34. Gautier R, Douguet D, Antonny B, Drin G. HELIQUEST: A web server to screen sequences with specific α-helical properties. Bioinformatics. 2008. doi:10.1093/bioinformatics/btn392
35. Gold C, Skinner F, Hoff L, N.Ayoub A, Hester T, Powers D, et al. Expressed sequence tags from an orchard spider aggregate silk gland cDNA library. Unpublished. 2018.
36. Bowman J., North M., Lavette L., Lee D., Pogrebna VV, Zachry J., et al. Expressed sequence tags from a house spider aggregate silk glands cDNA library. Unpublished. 2018.
37. Salchert D, Higgins C, Telese R, Barham W, Strosnider K, Grebas C, et al. Expressed sequence tags from an orchard spider silk gland cDNA library. Unpublished. 2018.
38. Huang W, Lin Z, Yang D. From EST to potentially novel spider silk gene identification. Unpublished. 2014.
39. Fazzone ME, Gannett JS, Hayashi CY, Ayoub NA. Expressed sequence tags from a false black widow silk gland cDNA library (2010b). Unpublished. 2013.
40. Lane AK, Hayashi CY, Whitworth GB, Ayoub NA. Complex gene expression in the dragline silk producing glands of the Western black widow (Latrodectus hesperus). BMC Genomics. 2013. doi:10.1186/1471-2164-14-846
41. Lepore E, Bosia F, Bonaccorso F, Bruna M, Taioli S, Garberoglio G, et al. Spider silk reinforced by graphene or carbon nanotubes. 2D Mater. 2017;4. doi:10.1088/2053-1583/aa7cd3
42. Tahir HM, Zahra K, Zaheer A, Samiullah K. Spider silk: An excellent biomaterial for medical science and industry. Punjab Univ J Zool. 2017;32: 143–154.
43. Babb PL, Lahens NF, Correa-Garhwal SM, Nicholson DN, Kim EJ, Hogenesch JB, et al. The Nephila clavipes genome highlights the diversity of spider silk genes and their complex expression. Nat Genet. 2017;49: 895–903. doi:10.1038/ng.3852
44. Tokareva O, Michalczechen-Lacerda VA, Rech EL, Kaplan DL. Recombinant DNA production of spider silk proteins. Microb Biotechnol. 2013;6: 651–663. doi:10.1111/1751-7915.12081
45. Scheller J, Henggeler D, Viviani A, Conrad U. Purification of spider silk-elastin from

transgenic plants and application for human chondrocyte proliferation. Transgenic Res. 2004;13: 51–57. doi:10.1023/B:TRAG.0000017175.78809.7a

46. Gomes SC, Leonor IB, Mano JF, Reis RL, Kaplan DL. Antimicrobial functionalized genetically engineered spider silk. Biomaterials. 2011;32: 4255–4266. doi:10.1016/j.biomaterials.2011.02.040
47. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990. doi:10.1016/S0022-2836(05)80360-2
48. Townley MA, Tillinghast EK. Aggregate silk gland secretions of araneoid spiders. Spider Ecophysiology. 2013. doi:10.1007/978-3-642-33989-9_21
49. Volsi ECFR, Mendes MA, Marques MR, dos Santos LD, Santos KS, de Souza BM, et al. Multiple bradykinin-related peptides from the capture web of the spider Nephila clavipes (Araneae, Tetragnatidae). Peptides. 2006;27: 690–697. doi:10.1016/j.peptides.2005.08.011
50. Pasupuleti M, Schmidtchen A, Malmsten M. Antimicrobial peptides: Key components of the innate immune system. Critical Reviews in Biotechnology. 2012. doi:10.3109/07388551.2011.594423
51. Zinngrebe J, Montinaro A, Peltzer N, Walczak H. Ubiquitin in the immune system. EMBO Reports. 2014. doi:10.1002/embr.201338025
52. Kieffer AE, Goumon Y, Ruh O, Chasserot-Golaz S, Nullans G, Gasnier C, et al. The N- and C-terminal fragments of ubiquitin are important for the antimicrobial activities. FASEB J. 2003. doi:10.1096/fj.02-0699fje
53. Seo JK, Lee MJ, Go HJ, Kim G Do, Jeong H Do, Nam BH, et al. Purification and antimicrobial function of ubiquitin isolated from the gill of Pacific oyster, Crassostrea gigas. Mol Immunol. 2013. doi:10.1016/j.molimm.2012.07.003
54. Loof TG, Mörgelin M, Johansson L, Oehmcke S, Olin AI, Dickneite G, et al. Coagulation, an ancestral serine protease cascade, exerts a novel function in early immune defense. Blood. 2011. doi:10.1182/blood-2011-02-337568
55. Kuhn-Nentwig L, Nentwig W. The immune system of spiders. Spider Ecophysiology. 2013. doi:10.1007/978-3-642-33989-9_7
56. Amparyup P, Charoensapsri W, Tassanakajon A. Prophenoloxidase system and its role in shrimp immune responses against major pathogens. Fish Shellfish Immunol.

2013. doi:10.1016/j.fsi.2012.08.019

57. Papareddy P, Rydengård V, Pasupuleti M, Walse B, Mörgelin M, Chalupka A, et al. Proteolysis of human thrombin generates novel host defense peptides. PLoS Pathog. 2010. doi:10.1371/journal.ppat.1000857
58. Kanost MR. Serine proteinase inhibitors in arthropod immunity. Dev Comp Immunol. 1999. doi:10.1016/S0145-305X(99)00012-9
59. Le CF, Fang CM, Sekaran SD. Intracellular targeting mechanisms by antimicrobial peptides. Antimicrobial Agents and Chemotherapy. 2017. doi:10.1128/AAC.02340-16
60. Jiang Z, Vasil AI, Hale JD, Hancock REW, Vasil ML, Hodges RS. Effects of net charge and the number of positively charged residues on the biological activity of amphipathic α-helical cationic antimicrobial peptides. Biopolym - Pept Sci Sect. 2008. doi:10.1002/bip.20911
61. Bahar AA, Ren D. Antimicrobial peptides. Pharmaceuticals. 2013. doi:10.3390/ph6121543
62. Kaur K, Kaur P, Mittal A, Nayak SK, Khatik GL. Design and molecular docking studies of novel antimicrobial peptides using autodock molecular docking software. Asian J Pharm Clin Res. 2017. doi:10.22159/ajpcr.2017.v10s4.21332

Figuras

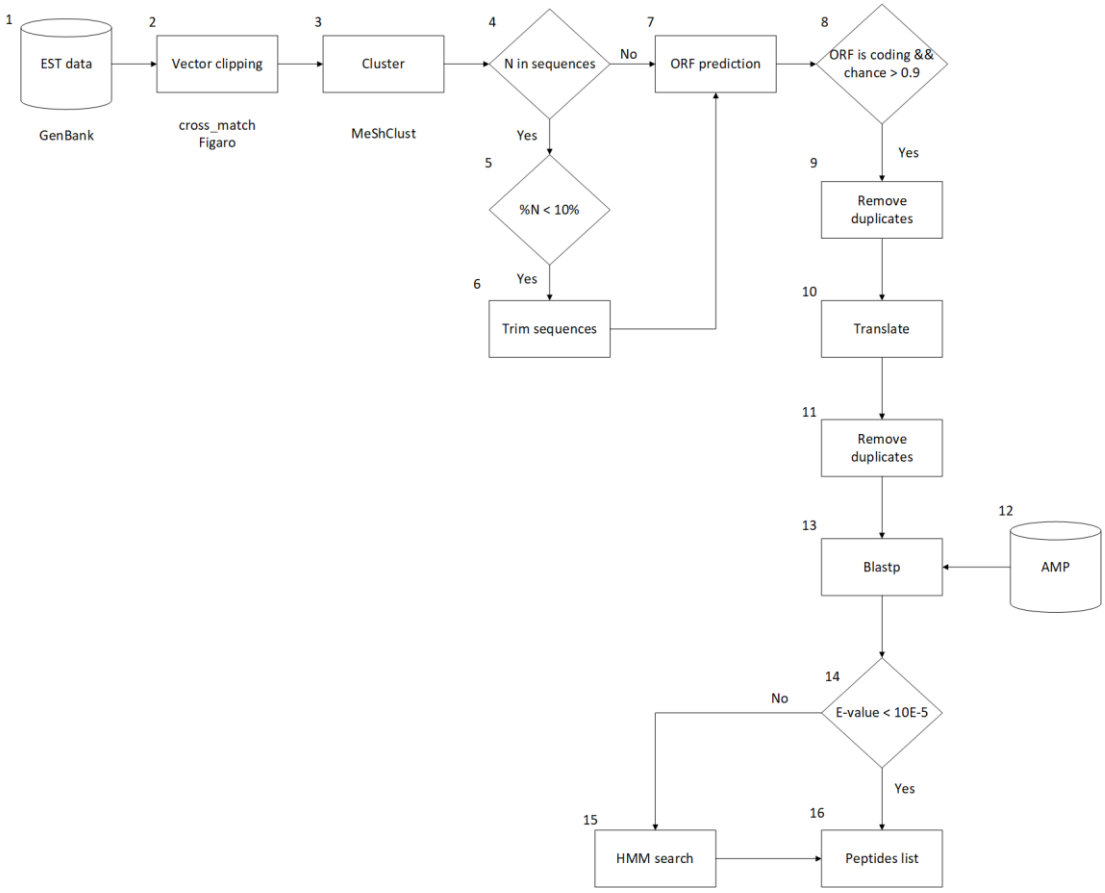


Fig 1. Diagram for AMP discovery using silk gland EST sequences.

Original EST datasets were obtained from the GenBank-EST database considering tissues with at least one silk gland (Step 1), then sequences were clustered using MeShClust program (Step 3). Clustered sequences that contain 'N' char on them were trimmed (Step 4-6), and complementary sequences were obtained for each read (Step 7). MiPepid program predicted all ORF and the chance of being coding (Step 8), duplicated sequences were deleted (Step 9). Coding and non-duplicated ORFs were translated into AA sequences (Step 10). AA duplicated sequences were deleted (Step 11) and were BLAST against the LAMP database (Step 12-14). Non-aligned sequences were compared against HMM profiles of CAMPSign server (Step 15).

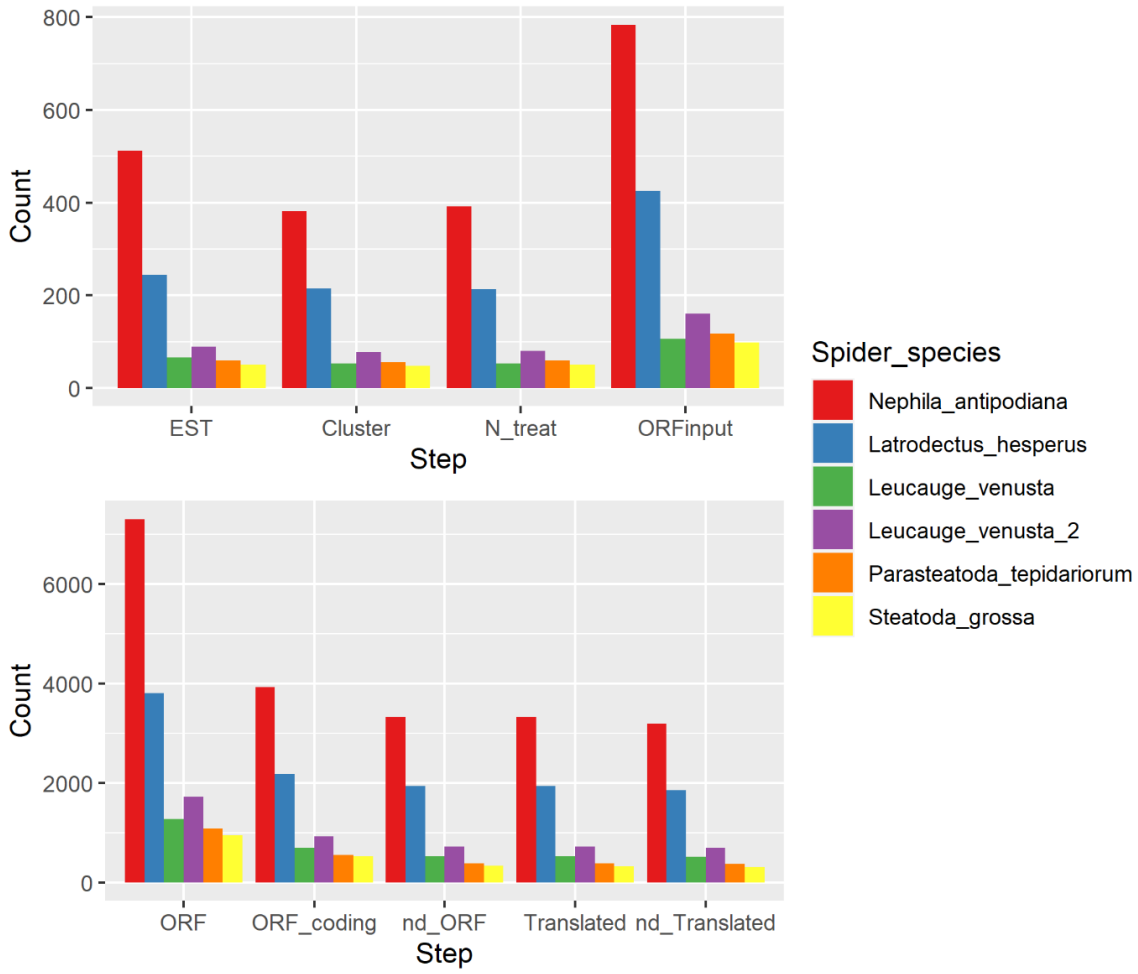


Fig 2. Summary of sequences retained at each step of the pipeline from the six datasets used.

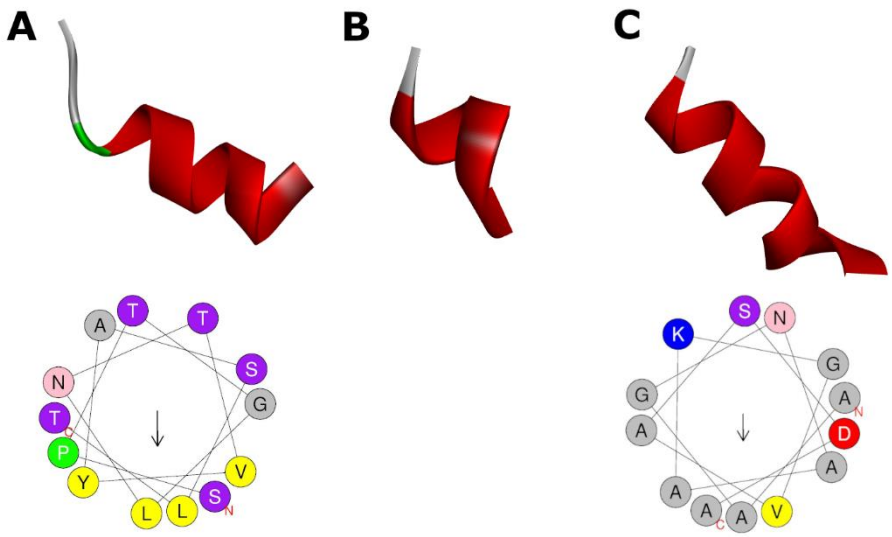


Fig 3. Secondary structure models (top) and helical wheel plots (bottom) of novel peptides.

Purple: polar residues, yellow: hydrophobic residues. (A) LhH_seq1, (B) Lv2H_seq1, (C) SgH_seq1.

Tablas

Table 1. Microorganisms inhibition assays with spider silk

Family	Specie	Silk type	Microorganisms	Reference
Agelenidae	*Tegenaria domestica*	Dragline and capture	*Bacillus subtilis,*	[8]
			Escherichia coli	
Araneidae	*Nephila pilipes*	Dragline	*E. coli*	[13]
			Staphylococcus aureus	
			Pseudomonas aureginosa	
Pholcidae	*Pholcus phalangioides*	Cobweb	*E. coli*	[14]
			Listeria monocytogenes	
Eresidae	*Stegodyohus dumicola*	Capture and refuge	*Bacillus thuringiensis*	[15]
Araneidae	*Cyclosa confraga*	Not specified	*Streptococcus* sp.	[6]
			Acinetobacter sp.	
Lycosidae	*Pardosa brevivulva*	Not specified	*B. megaterium*	[16]
			Salmonella typhi	
			Klebsiella pneumoniae	
			Aspergillus flavus	
			Candida albicans	
			Ustilago maydis	
			Alternaria solan	
Theridiidae	*Latrodectus hesperus*	Gumfoot	*E. coli*	[17]

Table 2. Blast and HMM matches that produced significant alignments, their respective net charge, mean hydrophobic moment and their CAMP accession ID.

Dataset	Name	Method used	Match fragment sequence	Net charge	Mean Hydrophobic moment	Accession ID
L. hesperus	LhB_seq1	Blastp	KVHGSLARAGKVKGQTPKVEKQEKKKRKTGRAKRRMQFNRRFVNVVVTFGRKKGPNSNS	+18	0.32	CAMPSQ3754
L. hesperus	LhB_seq2	Blastp	MQFNRRFVNVVVTFGRKKGPNSNS	+5	0.45	CAMPSQ3754
L. hesperus	LhH_seq1	HMM	SPTGLNTVYASLT	0	0.82	CAMPBacH32
L. venusta 2	Lv2H_seq1	HMM	LWKTLLK	+2	1.97	CAMPDerH28
N. antipodiana	NaB_seq1	Blastp	MQIFVKTLTGKTITLESEPSDTIENVKTKIQTKKASPQ	+1	0.52	CAMPSQ3702
N. antipodiana	NaB_seq2	Blastp	GGKAGQDACKGDGGGPLVCFRSCAGGKAGQDACKGDGGGPLVCFRSDNSYTVAGLVSWGIDCGQEGIPGVYVNVKKYNDWIVSKTQKPIENY	0	0.264	CAMPSQ3345
S. grossa	SgB_seq1	Blastp	FYYNDVAKKCEIFYYGGCKGNENNFPSEDHCKEAGG	-2	0.26	CAMPSQ2798
S. grossa	SgH_seq1	HMM	AAGNAAKGVASDA	0	0.93	CAMPDerH

Table 3. Description and source organism of blastp matches of CAMP and top three non-redundant protein sequence database hits.

Sequence	CAMP	Source organism	Source Organisms | Non redundant database
LhB_seq1	Ubiquicidin	*Mus musculus*	• *S. dumicola* | Ubiquitin-like protein 1-40S ribosomal protein S27a • *Dolomedes sulfureus* | Ubiquitin-like 40S ribosomal S30 protein fusion • *Araneus ventricosus* | 40S ribosomal protein S30
LhB_seq2	Ubiquicidin	*Mus musculus*	• *S. dumicola* | Ubiquitin-like protein 1-40S ribosomal protein S27a • *P. tepidariorium* | Ubiquitin-like protein FUBI • *S. mimosarum* | 40S ribosomal protein S30
NaB_seq1	CgUbiquitin	*Crassostrea gigas*	• *Flavobacterium sp.* | Ubiquitin • *Paragonimus kellicotti* | Hypothetical protein AH37_07978 • *Octodon degus* | Ubiquitin-like
NaB_seq2	TCP	*Homo sapiens*	• *Araneus ventricosus* | Phenoloxidase-activating factor 2 • *Argiope bruennichi* | Phenoloxidase-activating factor 2 like protein • *P. tepidariorum* | Serine proteinase stubble-like
SgB_seq1	Luxuriosin*	*Acalolepta luxuriosa*	• *Schistosoma haematobium* | Putative kunitz-type protease inhibitor • *Chionoecetes opilio* | Amyloid-like protein 2 • *Mus caroli* | kunitz-type protease inhibitor 3

*Non significant alignment