Antimicrobial peptides in the seedling transcriptome of the tree legume *Peltophorum dubium*

Susana Rodríguez-Decuadro[a]; Gabriela da Rosa[b]; Santiago Radío[c]; Mariana Barraco-Vega[b]; Ana Maria Benko-Iseppon[d]; Pablo D. Dans[e], Pablo Smircich[c]; Gianna Cecchetto[b,f]*.

[a]Departamento de Biología Vegetal, Facultad de Agronomía, Universidad de la República, Garzón 780, Montevideo 12900, Uruguay; surodriguez9@gmail.com

[b]Departamento de Biociencias, Facultad de Química, Universidad de la República, General Flores 2124, Montevideo 11800, Uruguay; gabodrc@gmail.com; mariveba@gmail.com

[c]Departamento de Genómica, Instituto de Investigaciones Biológicas Clemente Estable. MEC – Laboratorio de Interacciones Moleculares, Facultad de Ciencias, Universidad de la República, Iguá 4225, Montevideo 11400, Uruguay; sradio91@gmail.com; psmircich@fcien.edu.uy

[d]Universidade Federal de Pernambuco, Centro de Biociências, Av. Prof. Moraes Rego, 1235. CEP 50.670-420 - Recife – PE, Brazil; ana.iseppon@gmail.com

[e] Departamento de Ciencias Biológicas, CENUR Litoral Norte, Universidad de la República. General Fructuoso Rivera 1350, Salto 50000, Uruguay - Laboratorio Genética Funcional, Institut Pasteur de Montevideo. Mataojo 2020, Montevideo 11400, Uruguay; pablo.dans@unorte.edu.uy

[f]Instituto de Química Biológica, Facultad de Ciencias - Facultad de Química, Universidad de la República, General Flores 2124, Montevideo 11800, Uruguay; gianna.cecchetto@gmail.com

* Corresponding author: gianna.cecchetto@gmail.com; Tel.: +(598) 29244209

RNA sequence data from this article have been deposited with the GenBank Data Libraries under ID: PRJNA625609

Abstract

Antimicrobial peptides (AMPs) play an essential role in plant defense against invading pathogens. Due to their biological properties, these molecules have been considered useful for drug development, as novel agents in disease therapeutics, applicable to both agriculture and medicine. New technologies of massive sequencing open opportunities to discover novel AMP encoding genes in wild plant species. This work aimed to identify cysteine-rich AMPs from *Peltophorum dubium*, a legume tree from South America. We performed whole-transcriptome sequencing of *P. dubium* seedlings followed by *de novo* transcriptome assembly, uncovering 78 AMP transcripts classified into four families: hevein-like, lipid-transfer proteins (LTPs), alpha hairpinins, defensins, and snakin/GASA (Giberellic Acid Stimulated in Arabidopsis) peptides. No transcripts with similarity to cyclotide, alpha-hairpin, or thionin genes were identified. Genomic DNA analysis by PCR confirmed the presence of 18 genes encoding six putative defensins and 12 snakin/GASA peptides and allowed the characterization of their exon-intron structure. The present work demonstrates that AMP prediction from a wild species is possible using RNA sequencing and *de novo* transcriptome assembly, regarding a starting point for studies focused on AMP gene evolution and expression. Moreover, this study allowed the detection of strong AMP candidates for drug development and novel biotechnological products.

Keywords: Fabaceae, Antimicrobial peptides, RNA-Seq, Cysteine motifs.

1. Introduction

Plants are continuously being attacked by pathogens. However, few invaders can produce a systemic infection and disease development due to both constitutive and inducible defense mechanisms. After the perception of a pathogen, a signaling cascade is triggered, inducing the strengthening of the cell wall, the production of secondary metabolites, and the synthesis of pathogenesis-related (PR) proteins [1]. Among the different PR protein classes, the so-called PR peptides stand out, due to their size (lower than 10 kDa; [2]), besides some features in common like net positive charge at physiological pH and an even number of cysteine (Cys) residues [3]. In turn, AMPs differ in size, amino acid (aa) composition, and molecular structure. Based on homology of aa sequences, cysteine motifs, and the three-dimensional structures, a number of distinct groups have been defined, including defensins, thionins, lipid-transfer proteins, snakins, cyclotides, and hevein-like proteins [4,5].

Playing essential roles in plant defense, AMPs have high potential as therapeutic agents. Therefore, they could be included in products used for plant defense, disease control and management in agricultural production, reducing the use of agrochemicals [6]. Further, previous reports show that plant defensins, for example, show activity not only against phytopathogens but also against human bacterial pathogens, being considered promising molecules for the development of new compounds with antibiotic action [7]. The isolation, characterization, and synthesis of a wide range of effective AMPs would then be essential for the continued development of mandatory products used in medicine and plant protection.

Gene isolation strategies are being increasingly used for the characterization of several AMPs, mainly owing to the high number of AMP nucleotide (nt) sequences available coming from "omics" data [8]. Analysis of the sequenced plant genomes revealed that AMPs have been under-predicted in model plants like *Arabidopsis thaliana* and *Oryza sativa*, accounting for 2–3% of the gene repertoire of each model species [9]. The sequenced *Medicago truncatula* and *A. thaliana* genomes, for example, possess >300 defensin-like genes,

present as multigene families [10,11]. In spite of a large number of results available concerning plant AMPs, there is little information on such peptides derived from native non-model species that constitute the most significant plant biodiversity [12]. With the advent of the new technologies of massive sequencing (NGS-Next Generation Sequencing), the complete characterization and the global analysis of genomes and transcriptomes are now possible at reasonable costs, even without any previous genomic information [13]. Therefore, genome-wide sequencing technologies open new opportunities to discover novel AMPs from wild plant species. Koehbach et al. [14] have combined transcriptome mining and mass spectrometry to identify and characterize new cyclotides from Rubiaceae family. A whole-transcriptome sequencing and *de novo* assembly was performed to identify transcripts encoding AMPs from seedlings of the wild-growing Poaceae *Leymus arenarius* [15] and the weed *Stellaria media* (L.) Vill (Caryophyllaceae) [16].

Peltophorum dubium (Fabaceae) is a native tree from South America with pharmaceutical use in folk medicine, with reported applications against respiratory, gastrointestinal, and skin diseases [17]. However, it remains poorly explored for drug development. Interestingly, some works have isolated trypsin inhibitors from *P. dubium* seeds, with activity against insects and rat lymphoma cells [18,19]. In addition, our group has recently isolated a snakin gene (PdSN1) from *P. dubium* genome. PdSN1 is expressed in leaves and seedlings and presents *in vitro* antimicrobial activity when produced in *E. coli* [20]. Until now, there is no transcriptomic or genomic data for this legume. Nevertheless, transcriptome sequencing for this non-model plant is now accessible using a *de novo* assembly, allowing AMP prospection from AMP-like transcripts. In this work, we performed the first RNA-Seq and *de novo* assembly analysis of *P. dubium* seedlings for AMP prediction, focusing on defensin and snakin/GASA genes.

2. Material and methods

2.1. Biological Material

Seeds and leaves of *P. dubium* were obtained from the gardens of Facultad de Agronomía (Montevideo, Uruguay). Seeds were immersed in H_2SO_4 (concentrated grade) for 15 min for scarification. Surface sterilized seeds were germinated on Whatman paper soaked with distilled water in Petri dishes at 28 ºC. Five days old seedlings (ca. 6 cm long) were frozen in liquid nitrogen and stored at -70 ºC until further processing.

2.2. RNA and DNA isolation

Total RNA was extracted from five days old seedlings using Qiagen RNeasy Plant Mini Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. Genomic DNA was extracted from leaves using the standard cetyl-trimethylammonium bromide (CTAB) method [21].

2.3. Library preparation and Illumina sequencing

The quality of total RNA was checked using an Agilent 2100 Bioanalyzer (Agilent, USA). The RNA Integrity Number (RIN) of the extracted RNA was 7.8. Library preparation and sequencing were carried out by Macrogen Inc. (Seoul, Korea). 10 μg of total RNA was used for cDNA library construction using the TruSeq stranded mRNA LT Sample Prep. Kit according to manufacturer's instructions. The library was sequenced using Illumina HiSeq 2000 sequencer in a paired-end 100 bp run.

2.4. Transcriptome assembly

Raw reads were filtered with Trimmomatic [22], and the resulting good quality read pairs were assembled using Trinity [23]. The quality of the assembled transcriptome was checked with Transrate [24]. For each transcript, all open reading frames longer than 90 nt were obtained using getorf [25]. Functional annotation was assigned by performing BLASTp against SwissProt and TrEMBL databases (maximum expected value of $10e^{-5}$).

2.5. AMP prediction in *P. dubium* seedlings transcriptome

A list of motifs that define distinct types of AMPs was obtained from Slavokhotova et al. [15] and used to search the *in silico* translated ORFs by using regular expression using in house python script. Positive peptides were grouped using cd-hit, and for each cluster of identical peptides, only one was chosen for further analysis. The presence of a signal peptide in the sequences was analyzed using SignalP program [26]. Subcellular localization was predicted with DeepLoc [27]. Multiple alignments with *P. dubium* deduced peptides (Table S2) and representative AMP family members from other plants (Table S3) were performed in ClustalW from BioEdit program (Hall 1999). Consensus sequences were visualized using Jalview v2.11.0 [28]. The deduced aa sequences encoding for the mature *P. dubium* snakin/GASA peptides were aligned against a set of mature plant proteins with GASA domain. All sequences used were taken from the UniProt database [29] (Table S3). An unrooted tree was generated using the Neighbor-Joining method of the MEGA package version 5 [30], with 5000 bootstrap replicates.

To predict antimicrobial capacity of the putative defensin and snakin peptides, AMPA (Torrent et al. 2012) and CAMP$_{R3}$ (Waghu et al. 2016) computational tools were used. For CAMP$_{R3}$ database, the Support Vector Machine, Random Forest, Artificial Neural Network and Discriminant Analysis algorithms were selected. An *Erythrina crista-galli* defensin (EcgDf1) with confirmed antimicrobial activity [31] was included in the analysis.

2.6. 3D structure determination

A protein-protein analysis was achieved against the Protein Data Bank (PDB). All the templates used are described in Table S4. Three alignment algorithms (BLAST, HHBLIST, and MODELLER) were tested to obtain the most satisfactory matching between our sequences and each template. The program Modeller 9.20 (with standard options) was used to build 100 structures of each peptide [32]. To produce a refined model, the structure with lower "internal energy" (DOPE score in MODELLER) was simulated at near-physiological conditions through short, unbiased, Molecular Dynamics (MD) simulations. The protonated state for each model was predicted at pH 6.5 with H++ 3.2 software

(http://biophysics.cs.vt.edu/H++) [33]. The models were minimized *in vacuo*, neutralized with 16 Cl-, solvated (with explicit water and 0.15 M of K+Cl-), and minimized in solution with harmonic positional restraints on the peptide. To produce an ensemble of "relaxed" conformations, the minimized structures were thermalized to 298 °C at NVT, pre-equilibrated for 10 ns at NPT (P = 1 atm), and then simulated during 50 ns using our well-established MD multi-step protocol [34]. The final structure, representative of the 3D prediction done to each native sequence, was chosen as the frame from the production MD trajectory nearest to the average conformation (obtained from averaging 50,000 conformations).

The peptides were treated as described previously [20]. Briefly, we used the ff14SB force field [35] surrounded by a truncated octahedral box of TIP3P water molecules (10 Å), and Dang parameters for ions. The Berendsen algorithm was used to control the temperature and the pressure with a coupling constant of 5 ps, removing the center-of-mass motion every 10 ps. SHAKE was applied to all bonds involving hydrogen, fixing at 2 fs the integration step of Newton equations of motion. Electrostatic interactions were treated in the framework of the Particle Mesh Ewald method with a real-space cutoff of 9 Å. All simulations were carried out using the PMEMD CUDA code [36] of AMBER 20 [37] and were analyzed with CPPTRAJ from AMBERTOOLS 20 [37]. Structures and MD trajectories were visually analyzed using VMD 1.9.

2.7. PCR validation of defensin and snakin/GASA genes

Six defensin and 12 snakin/GASA genes were PCR amplified from *P. dubium* genomic DNA by using primers designed from transcript contigs with the aid of Primer3 program. The primers targeted the region corresponding to the mature peptide and/or in 5′ and 3′ UTRs when the transcript was complete (Table S5). The PCR reactions were performed in a 20 µl reaction containing: 1X buffer with 2 mM $MgCl_2$, 0.2 mM dNTPs, 2 µM of each primer, 50 ng template DNA, and 0.5 U Taq DNA polymerase (Invitrogen, Carlsbad, USA). The PCR program was as follows: 94°C for 3 min; followed by 35 cycles of 94°C for 30 sec, 52-58°C for 40 sec and 72°C for 40 sec. The amplified fragments were visualized by agarose gel

electrophoresis and purified according to Richero et al. [38] and cloned into a pGEM-T easy vector (Promega Corporation, Madison, USA). AMP gene sequences were confirmed by Sanger sequencing at Macrogen Inc. (Seoul, Korea).

3. Results

3.1. Sequencing and *de novo* assembly of *P. dubium* transcriptome

Illumina sequencing of cDNA library obtained from *P. dubium* seedlings mRNA generated 52 million 100 bp reads (Table 1). Good quality reads were assembled with Trinity into approx. 127K transcripts. This assembly was evaluated and filtered with the Transrate tool that estimates individual contig quality. Its use reduced the final number of non-redundant transcripts to approx. 108K with an N50 statistic of 1347 (see Table S1).

Table 1. Raw data statistics of Illumina sequencing

Total # read bases (pb)	Total # reads	GC(%)	AT(%)	Q20(%)	Q30(%)
5,268,156,364	52,159,964	49.12	50.88	94.84	89.17

3.2. AMP identification

According to the arrangement of cysteine residues (cysteine motifs), AMPs can be divided into several families, including defensins, snakins, lipid-transfer proteins, thionins, alpha-hairpinins, hevein-like, and cyclotides. As a result of our search, 2644 transcripts with one or more cysteine motifs were identified. From these, only 441 were annotated against SwissProt and TrEMBL databases (maximum expected value of 10 e^{-5}). In 63 transcripts, the annotation corresponded to known AMPs, in 101 cases to uncharacterized proteins, whereas 277 were annotated as other protein families. Of the 441 transcripts with annotation, only 152 deduced peptides presented less than 200 aa. Predicted proteins that did not include the complete CDS were kept for further analysis if they included a cysteine motif.

In summary, a total of 78 transcripts were classified into five families: hevein-like (8), LTPs (28), alpha-hairpinin (10), defensins (14) and snakin/GASA peptides (18). No transcripts with similarity to thionin genes were identified; neither transcripts with cyclotides cysteine motifs were found in our seedling transcriptome. We only found a transcript with 61 % similarity with cliotide T18 (a cyclotide from *C. terneata*) but with only five cysteines instead of the six commonly found in this family of peptides (Fig. S1a). We also found one transcript with similarity to an albumin 2S, a seed storage protein; however, *P. dubium* putative 2S albumin lacks an SFTI-1 (cyclic peptide from sunflower) like domain (Fig. S1b).

ALPHA-HAIRPININS, HEVEIN-LIKE and LIPID-TRANSFER PROTEINS (LTPs)

We found more than 100 transcripts with alpha hairpinin motif ($CX_3CX_{1-20}CX_3C$); 65 of them with annotation. Eighteen had similarities to snakin/GASA genes and were therefore classified within this family in this work. Of the rest, 36 have similarities with nuclear, chloroplastic and membrane proteins; all of them without signal peptide, so they were not clasified as alpha-hairpinins. Finally, only 10 transcripts were considered, having similarities with uncharacterized proteins (3) (Fig 1a), with cysteine protease like papain (5) (Fig 1b), and also with miraculin and Kunitz trypsin inhibitors (2) (Fig 1c). The cysteine motifs are $CX_3CX_{9-12}CX_3C$, $CX_3CX_7CX_3C$, and $CX_3CX_2CX_3C$, respectively. All transcripts presented a single putative alpha-hairpinin motif.

Hevein-like peptides contain a conserved chitin-binding motif. We found eight transcripts with similarity with pro-hevein, whereas five of them have a hevein motif (with eight cysteines) (Fig. 2), while the remaining three have only the C-terminal domain named Barwin domain. These last have similarities with proteins annotated as PR-4 proteins or hevein-like.

We identify 28 transcripts with LTP cysteine motif, 18 with similarity with subfamily 1 (LTP1; Fig. 3a), and 10 with subfamily 2 (LTP2; Fig. 3b), all with eight cysteines conserved at specific positions. We observed three groups within subfamily 1. The first (I) presents similarities with typical LTP1 (like A0AT29 from *Lens culinaris* and Q42589 from *A. thaliana*). The second (group II), with 13 members, have similarities with EARLI1, a putative

Arabidopsis LTP. This protein has a putative signal peptide of 25 aa at the N-terminus, a hydrophilic proline-rich domain in the middle, and a hydrophobic C-terminus (with eight Cys and high similarity to plant LTPs) [39]. In group III, there is only one protein which (like the others) has 8 Cys, where the third and fourth Cys are consecutive in the polypeptide chain. In turn, the fifth and sixth Cys are separated by only one residue, but the number of residues among the remaining cysteines differs from the other two groups. This protein has a high similarity with a putative LTP (A0A1S3TFR4) of *Vigna radiata*. Within subfamily 2, we detected a typical LTP2 (group I) with five members and another (group II), that exhibits a variable number of residues in the C-terminus (not shown in the alignment; see Supplementary Table S2), following the conserved motif of 8 Cys.

DEFENSINS

We found 142 transcripts with defensin-like cysteine motifs of which 127 did not have functional annotation. Because our focus was on defensin-like proteins, and since an underestimation of these types of genes has been proposed [11], all transcripts were analyzed in detail, even those without annotation. In several cases, the ORF that included the cysteine motif was not the longest for the transcript and/or had an annotation in another reading frame, so they were discarded. In cases when the ORF with the cysteine motif was the longest but had a different annotation (not correspond with defensins, generally encoding proteins with more than 200 aa) the sequences were discarded. Many transcripts without annotation contained some kind of disperse repeat, like transposons. Some of these transcripts could be pseudogenes or assembly errors, so they were not taken into account. The remaining transcripts were further analyzed.

A list of 14 defensin-like genes was finally considered. The deduced peptides were separated into three subgroups. The first (Fig. 4a) includes 11 peptides with similarity with "true defensins", having eight Cys at conserved positions, with a cysteine-stabilized motif like the CS$\alpha\beta$ motif; this motif contains the residues CX_n**CXX**CX_nCX_nCXC (C=Cys, X=any amino acid, where n indicates a non-conserved number of amino acids [11]). These peptides also have a γ-core motif ($GXCX_{3\text{-}9}C$), distinctive of defensins. Eight of them had other

conserved aa: a glycine (G), a glutamic acid (E), and a serine (S). The second group includes a single transcript with the defensin cysteine motif $CX_5\mathbf{CX_3}CX_{10}CX_8\mathbf{CCC}$ (Fig. 4b), keeping cysteine residues like the conserved positions of the CS$\alpha\beta$ motif ($CX_nCXXXCX_nCX_nCXC$). The predicted peptide has similarity with some defensin-like proteins from *A. thaliana* that have the sixth, seventh, and eighth cysteine located consecutively. This peptide maintains the γ-core motif $GXCX_{3\text{-}9}C$. The third group (Fig 4c) includes two peptides with eight Cys, six of them ordered like the $\alpha\beta$ motif ($CX_{7\text{-}8}\mathbf{CX_3}CX_6CX_8\mathbf{CXC}$), whereas the γ-core motif is absent. These two putative defensins were aligned with floral defensins identified in *Nicotiana alata* and *Petunia hybrida*, characterized by having a prodomain at the C-terminus. Of the 14 defensin-like transcripts, two are not complete at 5' and, of these, no signal peptide was detected for PdDf11predicted peptide.

Antimicrobial capacity of the putative defensin-like peptides, using AMPA and $CAMP_{R3}$ predictors showed that nine defensin-like peptides (PdDf1; PdDf3-9; PdDf12) would have antimicrobial activity with the five algorithms used, two with four algorithms (PdDf2; PdDf13-14) and two with three algorithms (PdDf10-11). Prediction of antimicrobial regions within peptides showed one or two regions with more probability (Table S6).

Taking advantage of the relatively high number of 3D structures of defensin peptides determined experimentally and available in the Protein Data Bank (PDB), a search for adequate templates was conducted based on protein-protein sequence alignment. We found that all of our sequences matched to some extend with defensin peptides in the PDB (Fig. S2), 12 of them (PdDf1-12) showing high sequence identity/similarity and wide coverage (Table S4). Using homology modeling techniques and Molecular Dynamics simulations for further structural refinement, we predicted the 3D structures of 12 native defensin peptides in their monomeric form (Fig. 5). Albeit the given sequence variation within the native defensins and the sequence variation found with those deposited in the PDB, all models, based on 10 different templates, displayed a clear and typical cysteine-stabilized α/β motif (CSα/β).

The characteristic secondary structure, which remained stable along the simulations (Table S4), is composed of three antiparallel β-sheets and one α-helix (α1) which is connected by two disulfide bonds to the β3-sheet. Usually, an extra disulfide bond connects the loop between β1

and α1 to the second beta-sheet exposing the β2-loop-β3 region which is thought to be the "reactive region" of the peptide toward biological membranes (Fig. 5).

SNAKINS

In this work, we found 18 transcripts with a snakin cysteine motif, including PdSN1, previously isolated for our group from *P. dubium* genome and similar to potato snakin-1 (StSN1). The alignment of mature predicted peptides is shown in Fig. 6, following their classification in subfamilies I, II and III (according to [40]). Of the two predicted peptides with high similarity with StSN1 one of them is PdSN1. Ten peptides have similarities with potato snakin-2 (StSN2) and six with potato snakin-3 (RSI-1; StSN3[41]). Some conserved residues within each subfamily were shaded in gray. Two members of subfamily II (PdSN5 and PdSN13) have a proline-rich variable region. All predicted snakin/GASA members have a conserved GASA domain and a signal peptide at the N-terminus (with 19 to 27 aa) according to SignalP software, and are predicted to be soluble according to DeepLoc.

Alignment analysis of *P. dubium* snakin/GASA mature sequences and other plant mature proteins with the GASA domain revealed three groups that coincide with the three subfamilies previously described [40] (Fig. 7). For this analysis, sequences with reported experimental evidence such as antimicrobial activity were used, including a snakin-2 from tomato (|E5KBY0|E5KBY0_SOLLC; [42], a snakin-1 from alfalfa (|H9D2D5|H9D2D5_MEDSA; [43], a snakin-1 from pepper (|B2ZAW4|B2ZAW4_CAPAN; [44], a snakin-3 from potato (|M1BA38|M1BA38_SOLTU; [41], and sequences obtained from the curated UniProt database.

For snakin-like peptides, CAMP predictor showed that all the 18 transcripts could have regions with credible activity (with the four algorithms). AMPA predicted one or two regions with antimicrobial activity for 15 of the 18 putative snakins (see Table S7).

3.3. Defensin and snakin/GASA genes validation

Eighteen transcripts encoding six putative defensin and 12 snakin/GASA genes were selected for gene validation using PCR amplification from genomic DNA. Of the six defensin genes examined, we obtained four genes from the start codon to stop codon (Fig

8a). For the remaining two (PdDf1 and PdDf2), only the part corresponding to the mature peptide was amplified.

For snakin/GASA subfamily I, we identified two genes (from which one was previously reported and named PdSN1; [20]). Besides, other snakin/GASA members were observed, being seven from subfamily II and three from subfamily III (Fig. 8b). Comparison of the *de novo* assembled transcript sequences with the corresponding genomic DNA sequences showed that all snakin/GASA coding sequences are interrupted by one to three introns. The last exon of all verified snakin/GASA genes have a very conserved size, with 182 or 185 nt; this exon encodes the GASA domain. According to the SignalP analysis, the first exon at subfamilies I and II encodes the signal peptide and a small fragment of the mature peptide; however, for family III, it encodes only the signal peptide. The length of the introns varied from ~100 to ~250 nt.

Some of the snakin/GASA predicted genes are very similar and may have arisen by duplication. PdSN1 and PdSN2 are structurally similar, with an ORF with five differences at the signal peptide and five substitutions at the mature peptide. PdSN8 and PdSN9 are highly similar even in their UTR regions (differing in 12 nt at the ORF, that in PdSN9 encodes 4 additional aa). Thus, we could not amplify both genes, being able to verify only PdSN9. PdSN14 and PdSN15, which encode proteins of 115 and 111 aa, respectively, present an identical GASA domain but are flexible at the signal peptide and the variable region. The 3' UTRs were very similar, so we could not design two specific R primers.

4. Discussion

In this study, we performed for the first sequencing of *P. dubium* seedling transcriptome and *de novo* assembly, aiming to identify putative novel AMP coding genes. Considering the arrangement of cysteine residues (Cys motifs), AMPs can be divided into several families, being the most recognized: defensins, snakins, LTPs, thionins, alpha-hairpinins, hevein-like, and cyclotides. The conserved cysteine motifs, characteristic of each family, was used to predict potential AMPs in *P. dubium* transcriptome using previously described "cysteine

motifs" obtained from Slavokhotova et al. [15]. A total of 78 transcripts could be classified into five families: hevein-like (8), LTPs (28), alpha-hairpinin (10), defensins (14) and snakin/GASA peptides (18). No transcripts with similarity to thionin or cyclotides cysteine motifs were found in our seedling transcriptome.

Thionins comprise small AMPs of the PR-13 protein family [2]. Thionin genes have been identified in 15 different plant species (reviewed by [45]) of the families Santalaceae, Brassicaceae, Poaceae, Ranunculaceae, and Liliaceae, but to date, no reports are available describing thionins in Fabaceae and the only thionin-like proteins from this family annotated in the Uniprot database (https://www.uniprot.org/; 09-01-2020) are gamma thionins, the name previously used for plant defensins. This protein group might not be present in legumes, as is the case of *P. dubium*. Alternatively, these genes may have a discrete expression or, still, can be absent in seedlings, or the *de novo* assembly did not allow them to be detected.

Cyclotides (short cyclic peptides that have a head-to-tail cyclized backbone and three conserved disulfide-bonds in a knotted arrangement [46]) were found in several species of the Rubiaceae, Cucurbitaceae, Violaceae, Solanaceae, Poaceae, and Fabaceae families [14]. Poth et al. [47] reported the isolation of novel cyclotides from *Clitoria ternatea* seeds, being the first report of cyclotides in Fabaceae. In legumes, according to the findings in *C. terneata*, the biosynthetic origin of a mature cyclotide is an albumin precursor protein, unlike all previously reported cyclotides [47]. Even though we expected to find cyclotides in *P. dubium*, we could not identify transcripts with similarity to their protein precursor albumin-1. This protein is present in other legumes such as *Glycine max*, *Pisum sativum*, and *Vigna radiata*. These Fabaceae, as well as *C. terneata,* belong to Papilionoideae subfamily, while *P. dubium* belongs to the subfamily Caesalpinioideae. Albumin-1 could be specific to subfamily Papilionoideae. The only transcript (with 61 % similarity with cliotide T18 from *C. terneata* presented only five cysteines. This transcript was not complete at 5' and it was scarcely assembled. Alternatively, it could be a pseudogene whose expression may be due to insufficient time for the complete degeneration of regulatory signals [48]. We also found one transcript with similarity to an albumin 2S. The sunflower trypsin inhibitor peptide SFTI-1,

a 14 aa cyclic peptide with a single disulfide bond (GRCTKSIPPICFPD), is embedded within a 2S albumin [49]. However, *P. dubium* putative 2S albumin lacks an SFTI-1 like domain (see Fig. S1b). 2S albumin large chain seems conserved in *P. dubium*, *Helianthus annuus*, and *G. max*, but the small chain seems to be more variable. *P. dubium* putative albumin exhibits three cysteines between the signal peptide and the Large Chain while the 2S sunflower albumin presents four Cys (two within SFTI-1 and two within the Small chain). In turn, *G. max* 2S albumin has only two Cys (within Small Chain). Thus a cyclic peptide between the signal peptide and the short-chain of the 2S albumin seems to be absent in both species (*G. max* and *P. dubium*). Finally, it cannot be discarded that *P. dubium* seedling transcriptome could have cyclic peptides without similarity with those previously reported.

A new trypsin inhibitor with a helical hairpin structure was reported as belonging to a new plant AMP family, named alpha-hairpinins [50]. Although variable in aa sequences, this family comprises peptides with the same cysteine motif ($CX_3CX_{1-20}CX_3C$) that share a helix-loop-helix fold stabilized by two disulfide bridges C1–C4 and C2–C3 [51,52]. The three transcripts with similarity to uncharacterized proteins and the $CX_3CX_{9-12}CX_3C$ motif showed very low sequence similarity with other reported alpha hairpinins like Ec-AMP1, Sm-AMP-X, VhT1, and Luffin P1, peptides representing three types of biological activity: antimicrobial, trypsin inhibitor, and ribosome-inactivating [52]. In future works, it should be studied whether these peptides have any of these biological activities. Two of the transcripts found (with the cysteine motif $CX_3CX_2CX_3C$) had similarities with miraculin, a taste-modifying protein from *Richadella dulcifica*. Miraculin has homology with soybean Kunitz trypsin inhibitors [53]. The remaining five transcripts showed similarity to papain, a plant cysteine protease enzyme, isolated from papaya (*Carica papaya* L.) latex [54]. However, in papain, an alpha hairpinin motif is absent (see Fig 1b). Visas-Villamil et al. [55] have suggested that papain-like cysteine proteases play a key role in plant immunity, inducing broad spectrum defense responses, including plant cell death. None of the genes with a cysteine motif considered here appear to have a modular structure with several alpha hairpinin domains like those reported for other proteins classified as alpha-hairpinins [52].

Hevein-like peptides received this name from hevein, a protein from the rubber tree (*Hevea brasiliensis).* This protein is formed from a larger protein called prohevein, with a chitin-binding N-terminal 43 amino-acid hevein [56]. Members of this AMP family contain conserved chitin-binding motif involved in peptide-carbohydrate interactions with different pathogens, mainly chitin-containing fungi. Most of them possess eight cysteine residues forming four disulfide bridges. However, variants with six and ten cysteine residues also occur [4]. Of the eight transcripts with similarity with pro-hevein, only five have a hevein motif (with eight cysteines), while the remaining three have only the C-terminal Barwin domain. These last have similarities with proteins annotated as PR-4 proteins or hevein-like. PR-4 proteins are classified as chitinases type I and II [57]; therefore, the three identified transcripts could codify class II chitinases since this type of chitinase does not have the amino-terminal hevein motif [58].

LTPs are subdivided into two subfamilies, which present molecular masses of around 10 (LTP1) and seven (LTP2) kDa, both with eight cysteines conserved at specific positions [40, 41]. Due to their possible role in plant defense, LTPs are recognized as PR proteins and are classified in the PR-14 family. Their ability to bind a wide range of lipids could explain their inhibitory activity against fungi and some bacterial pathogens [2]. Of the 28 transcripts found with LTP cysteine motif, 18 have similarity with subfamily 1 and 10 with subfamily 2. We observed three groups within subfamily 1, the first (I) presents similarities with typical LTP1 while the second (group II), have similarities with a putative Arabidopsis LTP, a protein with a putative signal peptide at the N-terminus, a hydrophilic proline-rich domain in the middle, and a hydrophobic C-terminus, with eight Cys and high similarity to plant LTPs) [39]. The only protein in group III differs from the other two groups in the number of residues among cysteines, except for the third, fourth, fifth and sixth Cys. Within subfamily 2, we detected two groups, a typical first LTP2 group and another (group II), that exhibits a variable number of residues in the C-terminus following the conserved motif of 8 Cys. The number of putative LTP transcripts in the seedlings of *P. dubium* (28) was similar to that of the weed *S. media* (31) [16]. LTPs are widely distributed in the plant kingdom and seem to be abundant in plant genomes forming multigenic families of related proteins.

Genome-wide analysis of the LTP gene family identified 52 members in rice (1C=0.5) and 49 in Arabidopsis (1C=0.3 pg), many of them arranged in tandem duplication repeats [61]. If the number of LTP genes can be proportional to the genome size of *P. dubium* (1C=0.9 pg; [62]), then half of the genes would be expressed in seedlings.

Defensins have a widespread distribution throughout the plant kingdom [63]. The primary structure of these peptides generally consists of a signal peptide at the amino-terminal and a basic, cysteine-rich mature peptide (45 to 54 aa). However, defensin-like peptides with unusual structures, such as extra N and C-terminal domains, have been identified [64]. Of the 14 defensin-like transcripts considered in this work, two are not complete at 5′ and, of these, no signal peptide was detected for PdDf11 predicted peptide. The remaining 13 defensin-like peptides present a signal peptide detected by SignalP tool and two of them could have a C-terminal prodomain. They were separated into three subgroups. The first includes peptides with similarity with "true defensins", that is, those defensins whose three-dimensional structure has been elucidated by NMR spectroscopy. Despite limited amino acid sequence identities, these defensins have similar overall fold features with one α-helix and three antiparallel β-sheets that are stabilized by four intramolecular disulfide bonds formed by eight conservative cysteine residues [3,65]. This cysteine-stabilized $\alpha\beta$ motif includes three disulfide bridges and is remarkably similar to those of insect defensins [66]. These peptides are also characterized by the occurrence of a γ-core motif $GXCX_{3-9}C$ [67]. Our structural analysis showed that PdDf1-11 structures are composed of three antiparallel β-sheets and one α-helix (α1), connected by three disulfide bonds. This particular secondary structure orientation, cysteine motifs, and disulfide bond connectivity is typical for the Cys-defensin superfamily [68]. An extra disulfide bond connects the N-terminal with the C-terminal. This conformation, also shared by other defensins [69], is thought to be stabilized by four disulfide bonds due to the lack of a stabilizing hydrophobic core. The second group includes only one transcript and the predicted peptide (PdDf12) has similarity with some defensin-like proteins from *A. thaliana* that have the sixth, seventh, and eighth cysteine located consecutively. The predicted structure of PdDf12 shows only three disulfide bonds.

However, the cysteine-stabilized αβ motif stayed completely stable during the simulation. This peptide maintains the γ-core motif $GXCX_{3-9}C$.

Unlike de first and second, in the third group that includes two peptides (PdDf13-14), the γ-core motif is absent. These proteins have similarities with uncharacterized proteins and with Kazal-type serine protease inhibitors. However, this kind of protein has six Cys with a cysteine motif $CX_{6-9}CX_7CX_6YX_3CX_{2-3}CX_{17}C$ [70], where CXC is not present. These two putative defensins aligned with floral defensins, characterized by having a prodomain at the C-terminus [71], belonging to the class II defensins, according to its classification in solanaceous species. In the first class, the precursor protein is composed of an endoplasmic reticulum signal sequence and a mature defensin domain. These proteins enter the secretory pathway and have no obvious signals for post-translational modification or subcellular targeting. The second class is produced as larger precursors with a C-terminal prodomain of about 33 aa [3]. Our two putative defensins could have a C-ter prodomain. However, only six cysteines exhibited conserved positions corresponding to the mature peptide of *N. alata* and *P. hybrida* defensins. Homology modeling cannot be performed for these two peptides. Even though these peptides have six Cys ordered like the αβ motif ($CX_{7-8}CX_3CX_6CX_8CXC$); to confirm this αβ motif, an *ab initio* modeling should be performed. These *P. dubium* peptides have some characteristics in common with defensins and some in common with Kazal-type serine protease inhibitors. Further studies are necessary to define their classification and function.

Evaluation of antimicrobial capacity of the putative defensin-like peptides, using AMPA and $CAMP_{R3}$ predictors showed that all putative defensins would have antimicrobial activity with at least three algorithms. For PdDf1, 4-6, predicted antimicrobial regions within peptides were similar to the region found for EcgDf1, a defensin with confirmed activity. The highest predicted antimicrobial activity within EcgDf1 include CSαβ and γ-core motifs [31].

According to Silverstein et al. [11], the number of defensins in plant genomes has been underestimated. These authors found more than 300 defensin-like genes in Arabidopsis. Of

these, about 51 are expressed in seedlings, while a similar number (56) was found to be expressed in seeds of the *M. truncatula* legume [72]. The number of putative defensins genes found by our group in the *P. dubium* seedling transcriptome is approximately a quarter of those expressed in seeds of *M. truncatula*, but is similar, however, to that found from the weed *Stellaria media* seedling transcriptome (16) [16]. These discrepancies might be explained by methodological differences in each work. For example, in our study, the initial number of proteins with defensin motifs was reduced by performing a manual inspection of each sequence. Alternatively, *P. dubium* and *S. media* seedlings express a lower number of defensin genes in the analyzed conditions.

Snakins are encoded by snakin/GASA genes and involve a group of widely distributed peptides among higher plants. They are characterized by having a GASA domain of approximately 60 aa, with 12 cysteine residues in highly conserved positions, that may be involved in the formation of up to six disulfide bonds [73]. They constitute a family of AMPs defined from potato snakin-1 (StSN1, [74]) and snakin-2 (StSN2, [40]), that correspond to the first peptides isolated of this family. Our group has previously isolated a snakin gene (PdSN1) from *P. dubium* genome, similar to StSN1 [20]. Accordingly with Berrocal-Lobo et al. [40] snakins can be classified in subfamilies I, II, and III. The alignment of mature 18 predicted peptides with a snakin cysteine motif (including PdSN1), found in this work revealed three groups that coincide with the three subfamilies described for Berrocal-Lobo et al. [40].

Snakin/GASA proteins contain three distinct domains: a signal peptide with 18-29 residues, a variable region that is highly divergent between family members, both in sequence length and amino acid composition, and a conserved C-terminal region: the GASA domain [75]. In our work all predicted snakin/GASA members have a signal peptide at the N-terminus (confirmed by SignalP software analysis) and all have a conserved GASA domain. Even though all are predicted to be soluble and with extracellular location according to DeepLoc, two members of subfamily II have a proline-rich variable region like the Arabidopsis protein GASAE [76], The peptides with a proline-rich domain could be important components of cell wall proteins. Proline-rich proteins (PRPs) seem to play an important role in cell wall

signal transduction cascades, plant development and stress tolerance, contributing to cell wall modification under stress [77]. A subclass of PRPs is the hybrid PRPs, that represent putative cell wall proteins consisting of a repetitive proline-rich N-terminal domain and a conserved C-terminal domain [77], like PdSN5 and PdSN13. These putative snakin/GASA peptides could fulfill their function at cell wall level.

A genome-wide search for new snakin/GASA members in potato uncovered 16 snakin/GASA genes, in addition to StSN1 and StSN2, a number equal to that found by our group in the *P. dubium* seedling transcriptome. However, the number within each subfamily was different (subfamily I: four genes; subfamily II: seven genes; subfamily III: seven genes) [41]. Fourteen GASA proteins have been described in Arabidopsis [76]. If genomes of different plant species have a similar number of snakin/GASA genes, then the seedlings of *P. dubium* could be expressing all or almost all of the snakin/GASA genes. Moreover, the number of *P. dubium* snakin/GASA transcripts was similar to that found from the weed *S. media* seedling transcriptome (16) [16]. The identification of a large number of these genes in several species with a highly conserved domain suggests that they play an essential biological function.

To confirm the reliability of the transcriptome assembly we decided to use PCR amplification from genomic DNA because it gave us the advantage of detecting their exon-intron structure. Of the six defensin genes examined, we obtained four genes from the start codon to stop codon. For the remaining two, only the part corresponding to the mature peptide was amplified because only primers designed from the mature peptide functioned under the conditions tested. As for other reported defensins [65], PdDf gene structure is comprised of two exons and one intron where the first exon entirely encodes the signal peptide, and the second exon encodes the last aa of the signal peptide and the mature defensin peptide. We believe that for this reason, we could not detect the presence of an intron in the genes PdDf2 and PdDf1, only verifying part of the second exon. Introns in defensins have been reported to be variable in size [65,78]. Our work revealed three defensin genes with a similar size while the fourth has about 200 nt more.

For snakin/GASA genes, comparison of the *de novo* assembled transcript sequences with the corresponding genomic DNA sequences showed that all snakin/GASA coding sequences are interrupted by one to three introns. Considering the number of introns, subfamily II seems to be the most variable with one, two or three introns, while subfamily I seems to have only one intron. These results are similar to those found by Nahirñak et al. [41] in potato but with the difference that the most variable subfamily was III. The number of introns within subfamily I was one for all the genes reported in potato.

5. Conclusions

In this work, the first survey of *P. dubium* antimicrobial peptides was carried out from a *de novo* seedlings transcriptome. Considering that AMPs have low primary structure similarity, they could not be revealed by performing only homology-based search. However, they share conserved cysteine motifs so we could identify putative AMPs from *P. dubium* native legume by searching for those motifs, and found eight hevein-like, 28 lipid-transfer proteins, 14 defensins, and 18 snakin/GASA gene candidates. Although no reference genome is available for this South American tree legume, NGS-technologies allowed inferring on the abundance and diversity of AMPs in *P. dubium* seedlings transcriptome. This is a good starting point to deepen the knowledge of their gene organization and evolution and role in the plant defense system. Validated genes deserve more focused investigations (expression profiling and functional analysis), regarding the biological roles of defensin and snakin/GASA genes during plant development and defense response. Also, these genes could be the first targets for production in heterologous systems to evaluate their potential as antimicrobial agents.

Author contributions

SRD and GC conceived and designed the research. SRD, GdR, SR, PS, GC, PDD and MBV conducted the experiments. SRD drafted the manuscript. PS, AMBI, and GC critically revised the manuscript. All authors read and approved the final version of the manuscript.

Declaration of interest

The authors declare that they have no competing interests.

Funding

This research was funded by CSIC (Comisión Sectorial de Investigación Científica, Universidad de la República, Uruguay) (Grant #267)

Acknowledgments

The authors thank CSIC (Comisión Sectorial de Investigación Científica, Universidad de la República, Uruguay), PEDECIBA (Programa de Desarrollo de las Ciencias Básicas, Uruguay), CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brazil) and CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico, Brazil). SRD was supported by a doctoral fellowship from CSIC. PDD, PS and GC are PEDECIBA and SNI (Sistema Nacional de Investigadores, Uruguay) researchers.

References

[1] A.M. Benko-Iseppon, S.L. Galdino, T. Calsa, E.A. Kido, A. Tossi, L.C. Belarmino, S. Crovella, Overview on plant antimicrobial peptides, Curr. Protein Pept. Sci. 11 (2010) 181–8. https://doi.org/10.2174/138920310791112075.

[2] J. Sels, J. Mathys, B.M.A. De Coninck, B.P.A. Cammue, M.F.C. De Bolle, Plant pathogenesis-related (PR) proteins: A focus on PR peptides, Plant Physiol. Biochem. 46 (2008) 941–950. https://doi.org/10.1016/j.plaphy.2008.06.011.

[3] F.T. Lay, M.A. Anderson, Defensins-Components of the innate immune system in plants, Curr. Protein Pept. Sci. 6 (2005) 85–101. https://doi.org/10.2174/1389203053027575.

[4] T. Odintsova, T. Egorov, Plant Antimicrobial Peptides, in: H.R.I. and C. Gehring (Ed.), Plant Signal. Pept., Berlín, 2012: pp. 107–133. https://doi.org/10.1007/978-3-642-27603-3.

[5] R. Nawrot, J. Baryiski, G. Nowicki, B. Justyna, W. Buchwald, A. Gozdzicka-Jozefiak, Plant antimicrobial peptides, Folia Microbiol. (Praha). 59 (2014) 181–196. https://doi.org/10.1007/978-3-319-32949-9_5.

[6] E. Montesinos, Antimicrobial peptides and plant disease control, FEMS Microbiol. Lett. 270 (2007) 1–11. https://doi.org/10.1111/j.1574-6968.2007.00683.x.

[7] A.E. Sathoff, S. Velivelli, D.M. Shah, D.A. Samac, Plant defensin peptides have antifungal and antibacterial activity against human and plant pathogens, Phytopathology. 109 (2019) 402–408. https://doi.org/10.1094/PHYTO-09-18-0331-R.

[8] C.A. Santos-silva, L. Zupin, M. Oliveira-lima, L. Maria, B. Vilela, J.P. Bezerra-neto, J.R. Ferreira-neto, J. Diogo, C. Ferreira, R.L. De Oliveira-silva, C.D.J. Pires, F.F. Aburjaile, M.F. De Oliveira, E.A. Kido, S. Crovella, A.M. Benko-iseppon, Plant antimicrobial peptides : State of

the art , in silico prediction and perspectives in the omics era, Bioinform. Biol. Insights. 14 (2020) 1–22. https://doi.org/10.1177/1177932220952739.

[9] K.A.T. Silverstein, W.A. Moskal, H.C. Wu, B.A. Underwood, M.A. Graham, C.D. Town, K.A. VandenBosch, Small cysteine-rich peptides resembling antimicrobial peptides have been under-predicted in plants, Plant J. 51 (2007) 262–280. https://doi.org/10.1111/j.1365-313X.2007.03136.x.

[10] M.A. Graham, K.A. Silverstein, S.B. Cannon, K.A. Vandenbosch, Computational Identification and Characterization of Novel Genes from Legumes, Plant Physiol. 135 (2004) 1179–1197. https://doi.org/10.1104/pp.104.037531.exchange.

[11] K. a T. Silverstein, M. a Graham, T.D. Paape, K. a VandenBosch, Genome organization of more than 300 defensin-like genes in Arabidopsis, Plant Physiol. 138 (2005) 600–610. https://doi.org/10.1104/pp.105.060079.

[12] M.C. Pestana-Calsa, I.L.A.C. Ribeiro, T.C. Jr, Bioinformatics-Coupled Molecular Approaches for Unravelling Potential Antimicrobial Peptides Coding Genes in Brazilian Native and Crop Plant Species, Curr. Protein Pept. Sci. (2010) 199–209.

[13] J.A. Martin, Z. Wang, Next-generation transcriptome assembly, Nat. Rev. Genet. 12 (2011) 671–682. https://doi.org/10.1038/nrg3068.

[14] J. Koehbach, A.F. Attah, A. Berger, R. Hellinger, T.M. Kutchan, E.J. Carpenter, M. Rolf, M.A. Sonibare, J.O. Moody, G.K.S. Wong, S. Dessein, H. Greger, C.W. Gruber, Cyclotide discovery in Gentianales revisited--identification and characterization of cyclic cystine-knot peptides and their phylogenetic distribution in Rubiaceae plants, Biopolymers. 100 (2013) 438–452. https://doi.org/10.1002/bip.22328.

[15] A.A. Slavokhotova, A.A. Shelenkov, T.I. Odintsova, Prediction of *Leymus arenarius* (L.) antimicrobial peptides based on de novo transcriptome assembly, Plant Mol. Biol. 89 (2015) 203–214. https://doi.org/10.1007/s11103-015-0346-6.

[16] A.A. Slavokhotova, A.A. Shelenkov, T. V. Korostyleva, E.A. Rogozhin, N. V. Melnikova, A. V. Kudryavtseva, T.I. Odintsova, Defense peptide repertoire of *Stellaria media* predicted by high throughput next generation sequencing, Biochimie. 135 (2017) 15–27. https://doi.org/10.1016/j.biochi.2016.12.017.

[17] M. Bolson, S.R. Hefler, E.I. Dall'Oglio Chaves, A. Gasparotto Junior, E.L. Cardozo Junior, Ethno-medicinal study of plants used for treatment of human ailments, with residents of the surrounding region of forest fragments of Paraná, Brazil, J. Ethnopharmacol. 161 (2015) 1–10. https://doi.org/10.1016/j.jep.2014.11.045.

[18] M. Macedo, M. Freire, E. Cristina, M. Lı, M.H. Toyama, A trypsin inhibitor from *Peltophorum dubium* seeds active against pest proteases and its effect on the survival of *Anagasta kuehniella* (Lepidoptera: Pyralidae), Biochim. Biophys. Acta. 1621 (2003) 170–182. https://doi.org/10.1016/S0304-4165(03)00055-2.

[19] M.F. Troncoso, P. Zolezzi, U. Hellman, C. Wolfenstein-todel, A novel trypsin inhibitor from *Peltophorum dubium* seeds, with lectin-like properties, triggers rat lymphoma cell apoptosis, Arch. Biochem. Biophys. 411 (2003) 93–104. https://doi.org/10.1016/S0003-9861(02)00726-9.

[20] S. Rodríguez-Decuadro, M. Barraco-Vega, P.D. Dans, V. Pandolfi, A.M. Benko-iseppon, G. Cecchetto, Antimicrobial and structural insights of a new snakin-like peptide isolated from

Peltophorum dubium (Fabaceae), Amino Acids. 50 (2018) 1245–1259. https://doi.org/10.1007/s00726-018-2598-3.

[21] J. Doyle, DNA protocols for plants-CTAB total DNA isolation, in: Mol. Tech. Taxon., 1991: pp. 283–293.

[22] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, Bioin. 30 (2014) 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

[23] M.G. Grabherr, B.J. Haas, M. Yassour, J.Z. Levin, D.A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. Palma, B.W. Birren, C. Nusbaum, K. Lindblad-toh, N. Friedman, A. Regev, Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data, Nat. Biotechnol. 29 (2013) 644–652. https://doi.org/10.1038/nbt.1883.Trinity.

[24] R. Smith-Unna, C. Boursnell, R. Patro, J.M. Hibberd, S. Kelly, TransRate: reference-free quality assessment of de novo transcriptome assemblies, Genome Res. 26 (2016) 1134–1144. https://doi.org/10.1101/gr.196469.115.Freely.

[25] P. Rice, EMBOSS: The European Molecular Biology Open Software Suite, Trends Genet. 16 (2000) 276–277.

[26] H. Nielsen, Predicting secretory proteins with SignalP, in: K. D. (Ed.), Methods Mol. Biol., New York, 2017: pp. 59–73. https://doi.org/10.1007/978-1-4939-7015-5.

[27] J.J. Almagro Armenteros, C.K. Sønderby, S.K. Sønderby, H. Nielsen, O. Winther, Sequence analysis DeepLoc: prediction of protein subcellular localization using deep learning, Bioinformatics. 33 (2017) 3387–3395. https://doi.org/10.1093/bioinformatics/btx431.

[28] A.M. Waterhouse, J.B. Procter, D.M.A. Martin, M. Clamp, G.J. Barton, Jalview Version 2- a multiple sequence alignment editor and analysis workbench, Bioinformatics. 25 (2009) 1189–1191. https://doi.org/10.1093/bioinformatics/btp033.

[29] R. Apweiler, A. Bairoch, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M.J. Martin, D.A. Natale, C.O. Donovan, N. Redaschi, L.L. Yeh, UniProt: the Universal Protein knowledgebase, Nucleic Acids Res. 32 (2004) D115–D119. https://doi.org/10.1093/nar/gkh131.

[30] K. Tamura, D. Peterson, N. Peterson, G. Stecher, M. Nei, S. Kumar, MEGA5 : Molecular Evolutionary Genetics Analysis Using Maximum Likelihood , Evolutionary Distance , and Maximum Parsimony Methods Research resource, Mol. Biol. Evol. 28 (2011) 2731–2739. https://doi.org/10.1093/molbev/msr121.

[31] S. Rodríguez-Decuadro, P.D. Dans, M.A. Borba, A.M. Benko-Iseppon, G. Cecchetto, Gene isolation and structural characterization of a legume tree defensin with a broad spectrum of antimicrobial activity, Planta. 250 (2019) 1757–1772. https://doi.org/10.1007/s00425-019-03260-w.

[32] B. Webb, A. Sali, Comparative protein structure modeling using MODELLER, Curr. Protoc. Bioinforma. 54 (2017) 5.6.1–5.6.37. https://doi.org/10.1002/cpbi.3.Comparative.

[33] R. Anandakrishnan, B. Aguilar, A. V Onufriev, H + + 3 . 0 : automating p K prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations, Nucleic Acids Res. 40 (2012) 537–541. https://doi.org/10.1093/nar/gks375.

[34] P.D. Dans, L. Danil, F. Lanka, A. Hospital, R.I. Pujagut, F. Battistini, L. Gelp, R. Lavery, M. Orozco, Long-timescale dynamics of the Drew – Dickerson dodecamer, Nucleic Acids Res. 44 (2016) 4052–4066. https://doi.org/10.1093/nar/gkw264.

[35] J.A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K.E. Hauser, C. Simmerling, ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99SB, J. Chem. Theory Comput. 11 (2015) 3696–3713. https://doi.org/10.1021/acs.jctc.5b00255.ff14SB.

[36] R. Salomon-ferrer, A.W. Gotz, D. Poole, S. Le Grand, R.C. Walker, Routine microsecond molecular dynamics simulations with AMBER on GPUs . 2 . Explicit Solvent Particle Mesh Ewald, J. Chem. Theory Comput. 9 (2013) 3878–3888. https://doi.org/10.1021/ct200909j.

[37] D.A. Case, K. Belfon, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, G. Giambasu, M.K. Gilson, H. Gohlke, A.W. Goetz, R. Harris, S. Izadi, S.A. Izmailov, K. Kasavajhala, A. Kovalenko, R. Krasny, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, V. Man, K.M. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, A. Onufriev, F.Pan, S. Pantano, R. Qi, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, N.R.Skrynnikov, J. Smith, J. Swails, R.C. Walker, J. Wang, L. Wilson, R.M. Wolf, X. Wu, Y. Xiong, Y. Xue, D.M. York and P.A. Kollman (2020), AMBER 2020, University of California, San Francisco.

[38] M. Richero, M. Barraco-Vega, M.P. Cerdeiras, G. Cecchetto, Development of SCAR molecular markers for early and late differentiation of *Eucalyptus globulus* ssp globulus from *E. globulus* ssp maidenii, Trees. 27 (2013) 249–257. https://doi.org/10.1007/s00468-012-0792-6.

[39] L. Li, C. Zhang, D. Xu, M. Schläppi, Z.Q. Xu, Expression of recombinant EARLI1, a hybrid proline-rich protein of Arabidopsis, in *Escherichia coli* and its inhibition effect to the growth of fungal pathogens and *Saccharomyces cerevisia*e, Gene. 506 (2012) 50–61. https://doi.org/10.1016/j.gene.2012.06.070.

[40] M. Berrocal-Lobo, A. Segura, M. Moreno, G. López, F. García-Olmedo, A. Molina, Snakin-2 , an antimicrobial peptide from potato whose gene is locally induced by wounding and responds to pathogen infection, Plant Physiol. 128 (2002) 951–961. https://doi.org/10.1104/pp.010685.1.

[41] V. Nahirñak, M. Rivarola, M. Gonzalez de Urreta, N. Paniego, H.E. Hopp, N.I. Almasia, C. Vazquez-Rovere, Genome-wide analysis of the Snakin/GASA gene family in Solanum tuberosum cv. Kennebec, Am. J. Potato Res. (2016). https://doi.org/10.1007/s12230-016-9494-8.

[42] V. Herbel, H. Schäfer, M. Wink, Recombinant production of snakin-2 (an antimicrobial peptide from tomato) in *E. Coli* and analysis of its bioactivity, Molecules. 20 (2015) 14889–14901. https://doi.org/10.3390/molecules200814889.

[43] A.N. García, N.D. Ayub, A.R. Fox, M.C. Gómez, M.J. Diéguez, E.M. Pagano, C.A. Berini, J.P. Muschietti, G. Soto, Alfalfa snakin-1 prevents fungal colonization and probably coevolved with rhizobia, BMC Plant Biol. 14 (2014) 248.

[44] Z. Mao, J. Zheng, Y. Wang, G. Chen, Y. Yang, D. Feng, B. Xie, The new CaSn gene belonging to the snakin family induces resistance against root-knot nematode infection in pepper, Phytoparasitica. 39 (2011) 151–164. https://doi.org/10.1007/s12600-011-0149-5.

[45] B. Stec, Plant thionins – the structural perspective, Cell. Mol. Life Sci. 63 (2006) 1370–1385. https://doi.org/10.1007/s00018-005-5574-5.

[46] D.J. Craik, N.L. Daly, T. Bond, C. Waine, Plant cyclotides: A unique family of cyclic and knotted proteins that defines the cyclic cystine knot structural motif, J. Mol. Biol. 294 (1999) 1327–1336. https://doi.org/10.1006/jmbi.1999.3383.

[47] A.G. Poth, M.L. Colgrave, R.E. Lyons, N.L. Daly, D.J. Craik, Discovery of an unusual biosynthetic origin for circular proteins in legumes, Proc. Natl. Acad. Sci. 108 (2011) 10127–10132. https://doi.org/10.1073/pnas.1103660108.

[48] C. Zou, M.D. Lehti-shiu, T. Prakash, Evolutionary and Expression Signatures of Pseudogenes, Plant Physiol. 151 (2009) 3–15. https://doi.org/10.1104/pp.109.140632.

[49] J.S. Mylne, M.L. Colgrave, N.L. Daly, A.H. Chanson, A.G. Elliott, E.J. Mccallum, A. Jones, D.J. Craik, Albumins and their processing machinery are hijacked for cyclic peptides in sunflower, Nat. Chem. Biol. 7 (2011) 1–3. https://doi.org/10.1038/nchembio.542.

[50] P.B. Oparin, K.S. Mineev, Y.E. Dunaevsky, A.S. Arseniev, M.A. Belozersky, E.V. Grishin, T.A. Egorov, A.A. Vassilevski, Buckwheat trypsin inhibitor with helical hairpin structure belongs to a new family of plant defence peptides, Biochem. J. 446 (2012) 69–77. https://doi.org/10.1042/BJ20120548.

[51] A.A. Slavokhotova, E.A. Rogozhin, A.K. Musolyamov, Y.A. Andreev, P.B. Oparin, A.A. Berkut, A.A. Vassilevski, T.A. Egorov, E. V. Grishin, T.I. Odintsova, Novel antifungal α-hairpinin peptide from *Stellaria medi*a seeds: Structure, biosynthesis, gene structure and evolution, Plant Mol. Biol. 84 (2014) 189–202. https://doi.org/10.1007/s11103-013-0127-z.

[52] A.A. Slavokhotova, E.A. Rogozhin, Defense peptides from the α -Hairpinin family are components of plant innate immunity, Front. Plant Sci. 11 (2020) 1–12. https://doi.org/10.3389/fpls.2020.00465.

[53] S. Theerasilps, H. Hitotsuya, S. Nakajoq, K. Nakayaq, Y. Nakamuraq, Y. Kuriharall, Complete amino acid sequence and structure characterization of the taste-modifying proteina, Miraculin, J. Biol. Chem. 264 (1989) 6655–6659.

[54] E. Amri, F. Mamboya, Papain, a plant enzyme of biological importance: a review, Am. Jorunal Biochem. Biotechnol. 8 (2012) 99–104. https://doi.org/10.3844/ajbbsp.2012.99.104.

[55] G.D. Misas-villamil, Johana C, Renier A L Van Der Hoorn, Papain-like cysteine proteases as hubs in plant immunity, New Phytol. 212 (2016) 902–907.

[56] H.I. Lee, W.F. Broekaert, N. V. Raikhel, Co- and post-translational processing of the hevein preproprotein of latex of the rubber tree (Hevea brasiliensis), J. Biol. Chem. 266 (1991) 15944–15948.

[57] S. Ali, B.A. Ganai, A.N. Kamili, A.A. Bhat, Z.A. Mir, J.A. Bhat, A. Tyagi, S.T. Islam, M. Mushtaq, P. Yadav, S. Rawat, A. Grover, Pathogenesis-related proteins and peptides as promising tools for engineering plants with multiple stress tolerance, Microbiol. Res. 212–213 (2018) 29–37. https://doi.org/10.1016/j.micres.2018.04.008.

[58] T. Araki, T. Torikata, Structural classification of plant chitinases: two subclasses in Class I and Class II chitinases, Biosci. Biotech. Biochem. 59 (1995) 336–338. https://doi.org/10.1248/cpb.37.3229.

[59] J. Kader, Lipid-Transfer Proteins in plants, Annu. Rev. Plant Physiol. Plant Mol. Biol. 47 (1996) 627–654. https://doi.org/10.1146/annurev.arplant.47.1.627.

[60] A.O. Carvalho, V.M. Gomes, Role of plant lipid transfer proteins in plant cell physiology-A concise review, Peptides. 28 (2007) 1144–1153. https://doi.org/10.1016/j.peptides.2007.03.004.

[61] F. Boutrot, N. Chantret, M.F. Gautier, Genome-wide analysis of the rice and arabidopsis non-specific lipid transfer protein (nsLtp) gene families and identification of wheat nsLtp genes by EST data mining, BMC Genomics. 9 (2008) 1–19. https://doi.org/10.1186/1471-2164-9-86.

[62] B. Van-lume, G. Souza, Cytomolecular analysis of species in the Peltophorum clade, Brazilian J. Bot. 41 (2018) 385–392. https://doi.org/10.1007/s40415-018-0449-9.

[63] B.P.H.J. Thomma, B.P.A. Cammue, K. Thevissen, Plant defensins, Planta. 216 (2002) 193–202. https://doi.org/10.1007/s00425-002-0902-6.

[64] B. De Coninck, B.P.A. Cammue, K. Thevissen, Modes of antifungal action and in planta functions of plant defensins and defensin-like peptides, Fungal Biol. Rev. 26 (2013) 109–120. https://doi.org/10.1016/j.fbr.2012.10.002.

[65] D.O. Carvalho, V.M. Gomes, Peptides Plant defensins — Prospects for the biological functions and biotechnological properties, Peptides. 30 (2009) 1007–1020. https://doi.org/10.1016/j.peptides.2009.01.018.

[66] B. Cornet, J. Bonmatin, C. Hetru, J.A. Hoffmann, M. Ptak, F. Vovelle, Refined three-dimensional solution structure of insect defensin A, Structure. 3 (1995) 435–448.

[67] N.Y. Yount, M.R. Yeaman, Multidimensional signatures in antimicrobial peptides, Proc. Natl. Acad. Sci. 101 (2004) 7363–7368. https://doi.org/10.1073/pnas.0401567101.

[68] T.M.A. Shafee, F.T. Lay, T.K. Phan, M.A. Anderson, M.D. Hulett, Convergent evolution of defensin sequence, structure and function, Cell. Mol. Life Sci. 74 (2017) 663–682. https://doi.org/10.1007/s00018-016-2344-5.

[69] U.S. Sagaram, K. El-Mounadi, G.W. Buchko, H.R. Berg, J. Kaur, R.S. Pandurangi, T.J. Smith, D.M. Shah, Structural and functional studies of a phosphatidic acid-binding antifungal plant defensin MtDef4: Identification of an RGFRRR motif governing fungal cell entry, PLoS One. 8 (2013) 1–22. https://doi.org/10.1371/journal.pone.0082485.

[70] S. Pariani, M. Contreras, F.R. Rossi, V. Sander, M.G. Corigliano, F. Simón, M. V. Busi, D.F. Gomez-Casati, F.L. Pieckenstain, V.G. Duschak, M. Clemente, Characterization of a novel Kazal-type serine proteinase inhibitor of *Arabidopsis thaliana*, Biochimie. 123 (2016) 85–94. https://doi.org/10.1016/j.biochi.2016.02.002.

[71] F.T. Lay, F. Brugliera, M.A. Anderson, Isolation and Properties of Floral Defensins from Ornamental Tobacco and Petunia, Plant Physiol. 131 (2003) 1283–1293. https://doi.org/10.1104/pp.102.016626.).

[72] M. Tesfaye, K.A.T. Silverstein, S. Nallu, L. Wang, C.J. Botanga, S.K. Gomez, L.M. Costa, M.J. Harrison, D.A. Samac, J. Glazebrook, F. Katagiri, J.F. Gutierrez-Marcos, K.A. VandenBosch, Spatio-Temporal Expression Patterns of *Arabidopsis thaliana* and *Medicago truncatula* Defensin-Like Genes, PLoS One. 8 (2013). https://doi.org/10.1371/journal.pone.0058992.

[73] M. Oliveira-Lima, A.M. Benko-Iseppon, R.J. Ferreira, Costa, S. Rodríguez-Decuadro, E.A. Kido, S. Crovella, V. Pandolfi, Snakin : Structure , Roles and Applications of a Plant Antimicrobial Peptide, Curr. Protein Pept. Sci. 18 (2017) 1–7. https://doi.org/10.2174/1389203717666160619l.

[74] A. Segura, M. Moreno, F. Madueño, A. Molina, F. García-Olmedo, Snakin-1, a peptide from potato that is active against plant pathogens., Mol. Plant. Microbe. Interact. 12 (1999) 16–23. https://doi.org/10.1094/MPMI.1999.12.1.16.

[75] V. Nahirñak, N.I. Almasia, H.E. Hopp, C. Vazquez-Rovere, Involvement in hormone crosstalk and redox homeostasis Snakin / GASA proteins, Plant Signal. Behav. 7 (2012) 1004–1008.

[76] I. Roxrud, S.E. Lid, J.C. Fletcher, E.D.L. Schmidt, H.-G. Opsahl-Sorteberg, GASA4 , one of the 14-member Arabidopsis GASA family of small polypeptides, regulates Flowering and Seed Development, Plant Cell Physiol. 48 (2007) 471–483. https://doi.org/10.1093/pcp/pcm016.

[77] Kavi Kishor, Polavarapu B., H. Kumari, M. Sunita, N. Sreenivasulu, Role of proline in cell wall synthesis and plant development and its implications in plant ontogeny, Front. Plant Sci. 6 (2015) 1–17. https://doi.org/10.3389/fpls.2015.00544.

[78] L. Padovan, L. Segat, A. Tossi, T. Calsa, A.K. Ederson, L. Brandao, R.L. Guimarães, V. Pandolfi, M.C. Pestana-Calsa, L.C. Belarmino, A.M. Benko-Iseppon, S. Crovella, Characterization of a new defensin from cowpea (*Vigna unguiculata* (L.) Walp.)., Protein Pept. Lett. 17 (2010) 297–304. https://doi.org/10.2174/092986610790780350.

Figure

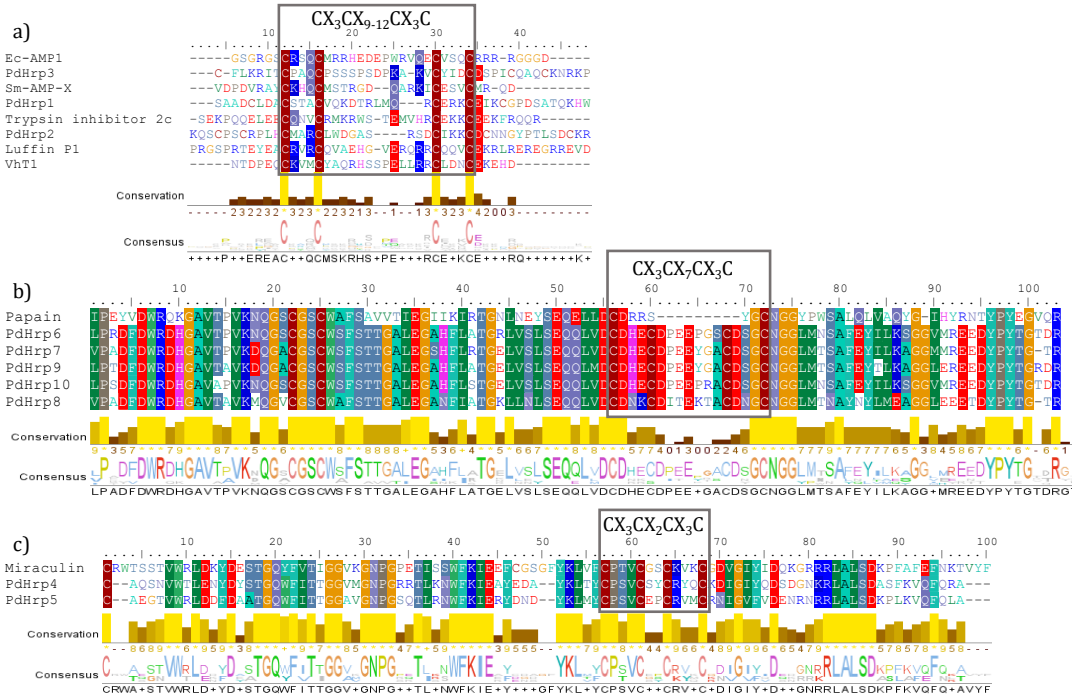

Fig. 1. Sequence comparison of putative *P. dubium* alpha-hairpinin with corresponding members from well-studied and model plants. a) Alignment of *P. dubium* predicted mature alpha-hairpinins like peptides (PdHrp1-13) with known alpha hairpinins peptides. b) Alignment of *P. dubium* mature peptides (PdHrp6-10) with an alpha-hairpinin motif with Papain, a cysteine protease from *Carica papaya*. c) Alignment of *P. dubium* mature peptides (PdHrp4-5) with an alpha-hairpinin motif with Miraculin, a taste-modifing protein from *Richadella dulcifica* that has homology with soybean Kunitz trypsin inhibitors (the last 100 aa are shown). The descriptions of reference family members from other plants are given in Table S3. Threshold (%) for Identity/Similarity shading was 75 %.

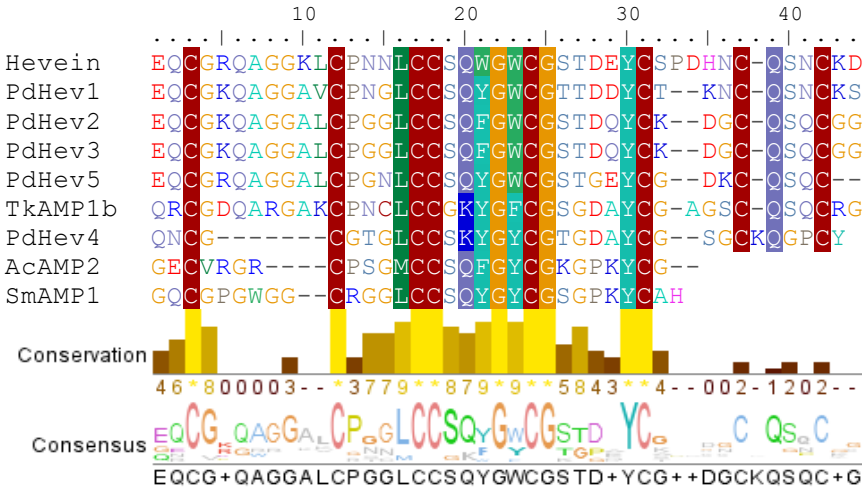

Fig. 2. Sequence comparison of putative *P. dubium* hevein-like peptides with corresponding sequences from well-studied plants. Hevein motif alignment of *P. dubium* predicted hevein-like peptides (PdHev1-5) with known hevein-like peptides. The descriptions of reference family members from other plants are given in Table S3. Threshold (%) for Identity/Similarity shading was 90 %.

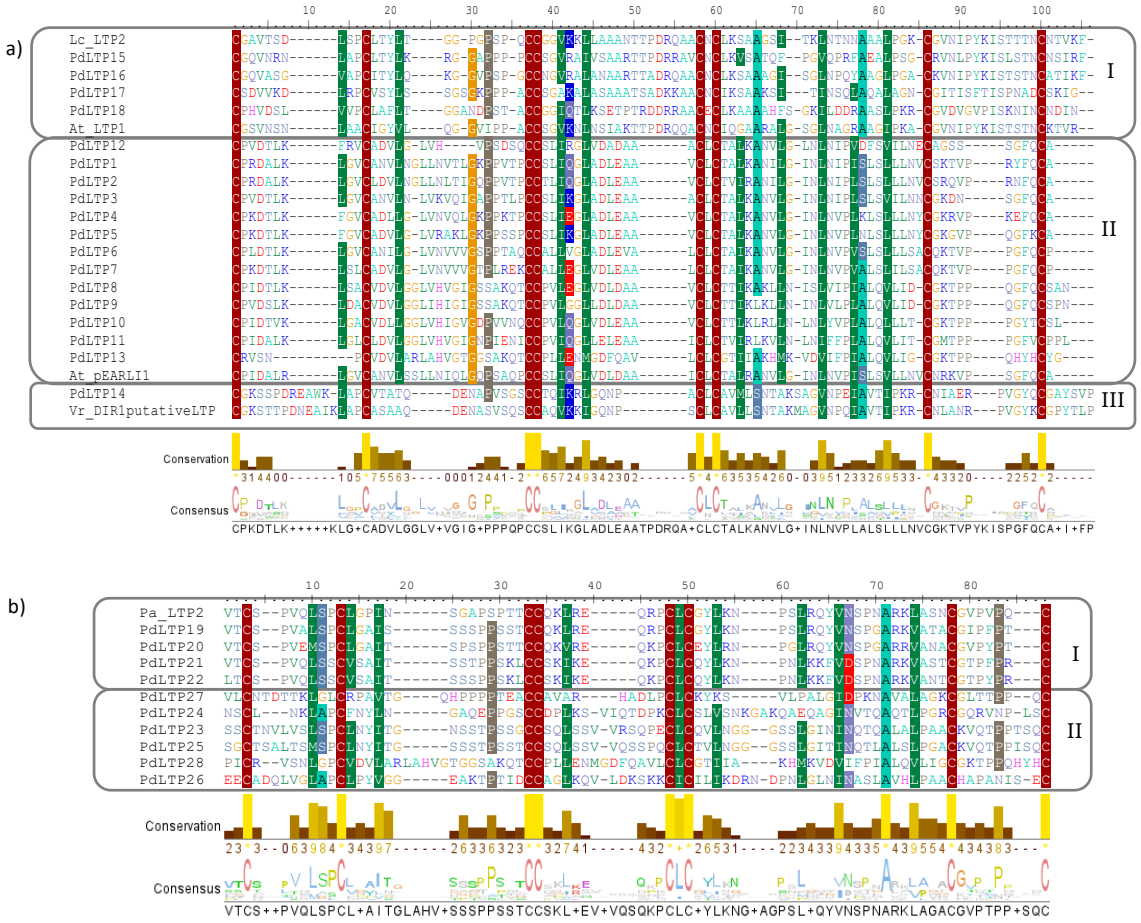

Fig. 3. Sequence comparison of putative *P. dubium* LTP peptides with corresponding sequences from well-studied and model plants. a) Alignment of *P. dubium* predicted mature LTP1 peptides (PdLTP1-18) with known LTP subfamily 1 peptides. Three groups were identified within this subfamily, showing similarity to different LTP1 members from other plants. b) Alignment of *P. dubium* predicted mature LTP2 (PdLTP19-28) and known LTP subfamily 2 peptides. Two groups were identified within this subfamily. The descriptions of reference family members from other plants are given in Table S3. Threshold (%) for Identity/Similarity shading was 75 %.

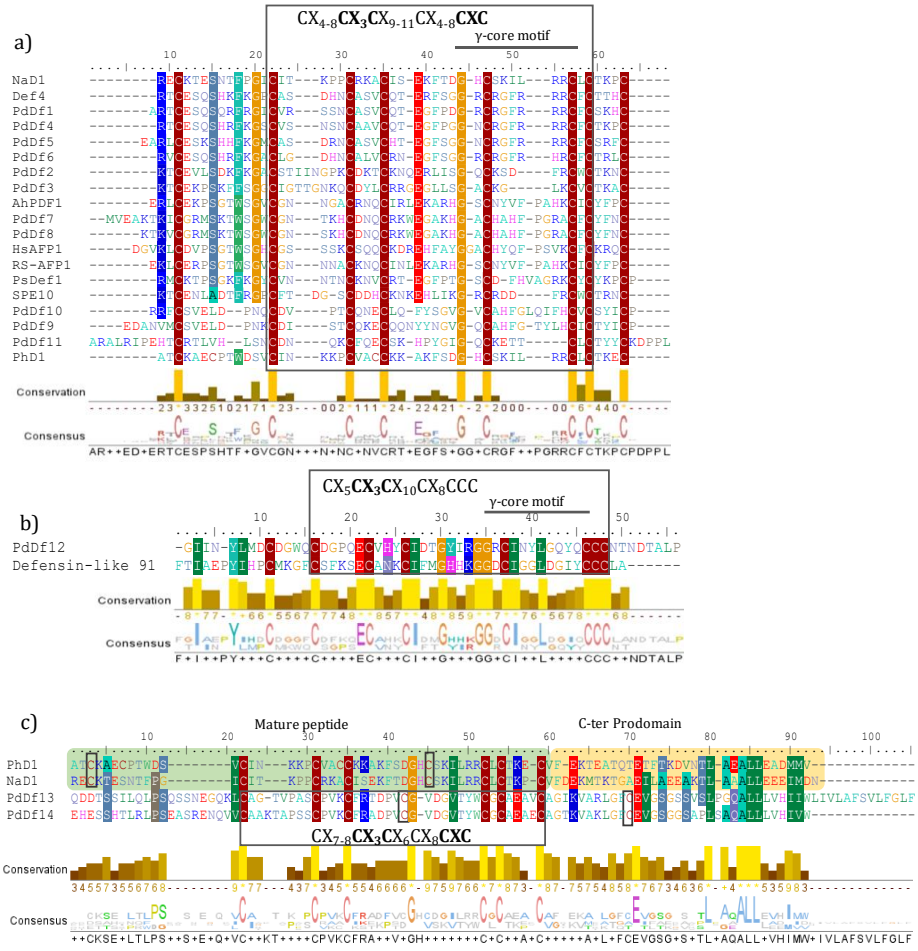

Fig. 4. Sequence comparison of putative *P. dubium* defensin-like peptides with corresponding members of well-studied and model plants. a) Alignment of *P. dubium* predicted "true" mature defensin peptides (named PdDf1 to PdDf11) and known mature defensin peptides. All the sequences used in this alignment correspond to defensins, whose 3D structure has been determined.. b) Alignment of a *P. dubium* peptide (PdDf12) with a defensin-like sequence from *A. thaliana* that have cysteines located consecutively at positions six, seven, and eight. c) Alignment of *P. dubium* predicted defensin-like peptides (PdDf13,14) and known defensin peptides with a C-ter Prodomain. Cysteines that differ in location with respect to previously described defensins are highlighted with gray rectangles. The descriptions of reference family members from other plants are given in Table S3. Threshold (%) for Identity/Similarity shading was 75 %.

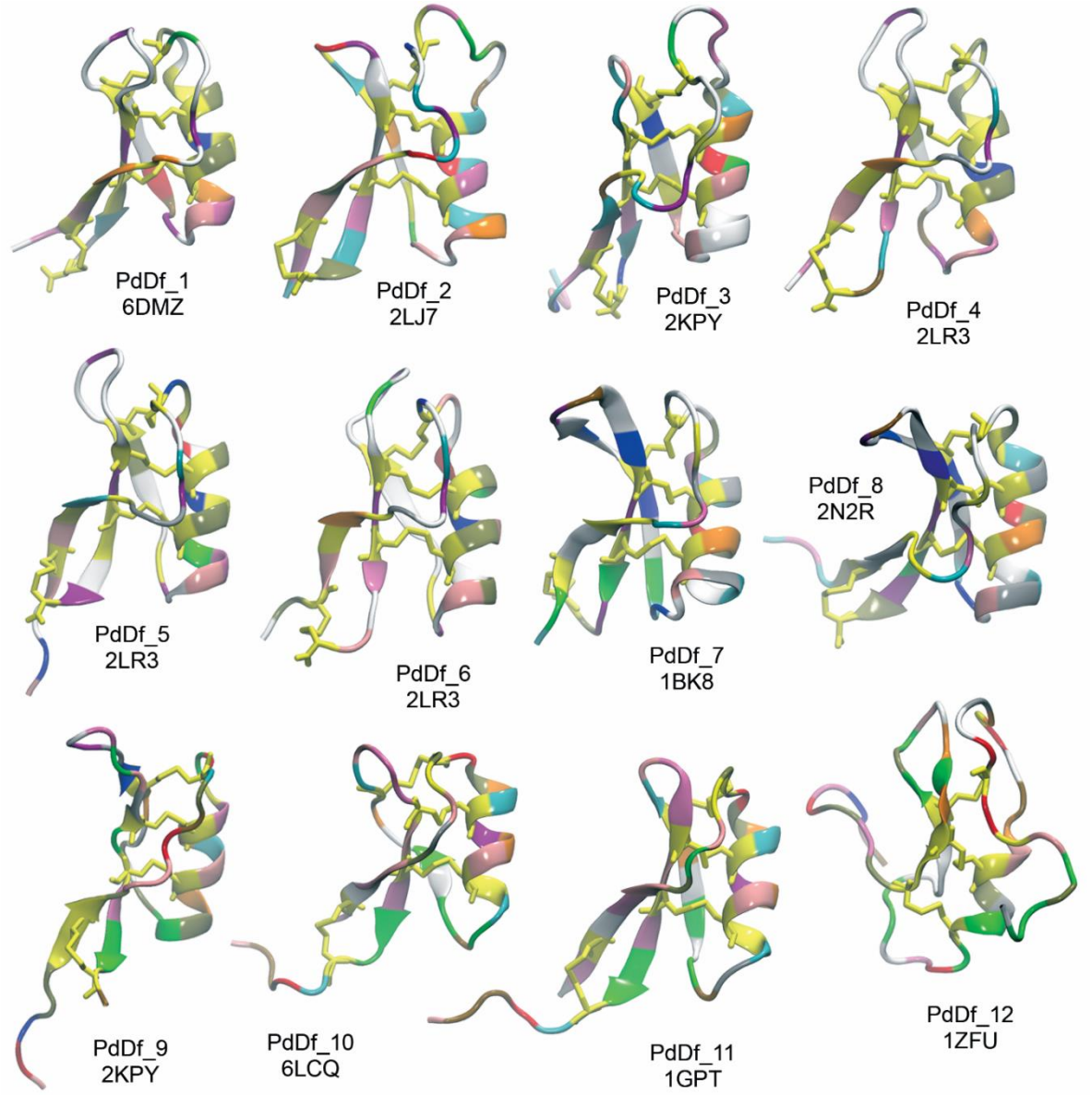

Fig 5. Predicted 3D structure for 12 native Defensin peptides. One representative structure (see Methods) for each sequence is represented by its secondary structure and colored according to the amino acid sequence. Cysteine residues and disulfide bonds are shown in yellow. The PDB id used as a template is provided below each sequence label.

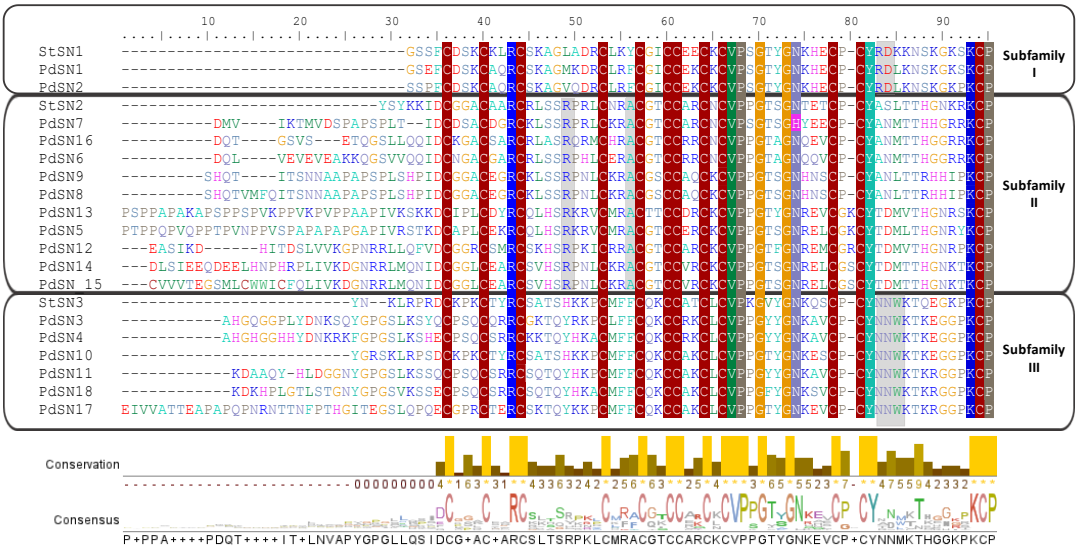

Fig. 6. Sequence comparison of putative *P. dubium* snakin/GASA peptides with corresponding members from *Solanum tuberosum*. Alignment of *P. dubium* predicted mature snakin/GASA peptides (PdSN1-18) and snakin peptides from potato. Alignment revealed three groups that coincide with the three subfamilies described for Berrocal-Lobo et al. [40]. Conserved residues exclusively within each subfamily were shaded in gray. sequence names: StSN1, StSN2, StSN3. The descriptions of reference family members from *S. tuberosum* are given in Table S3. Threshold (%) for Identity/Similarity shading was 100 %.

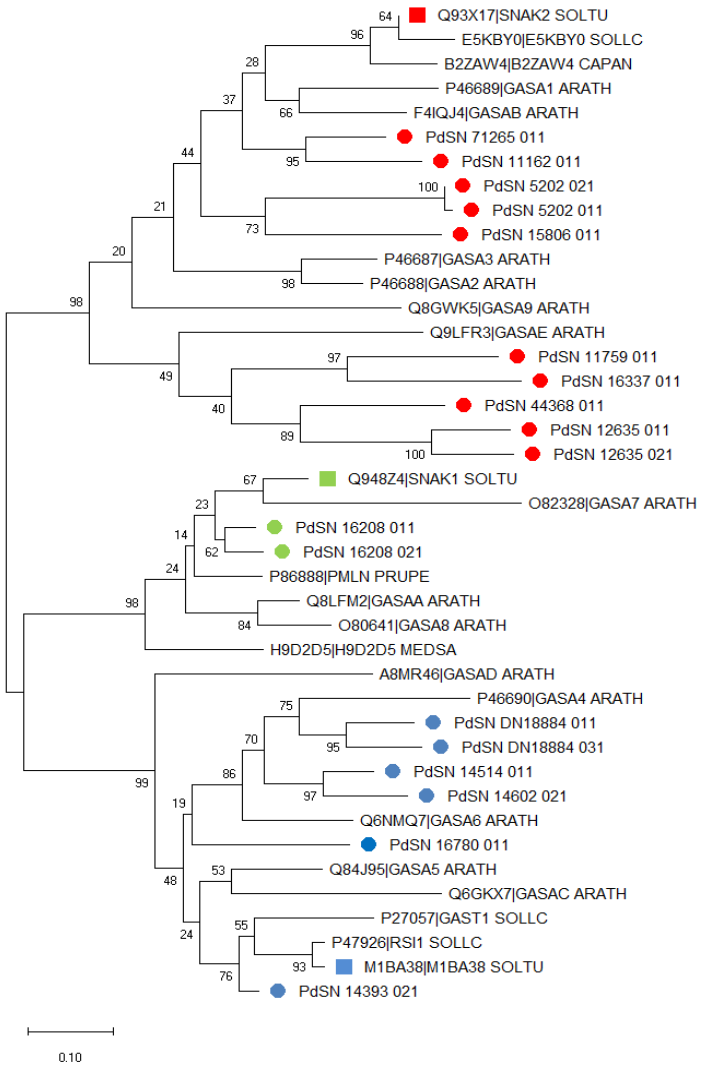

Fig. 7 Neighbor-Joining unrooted tree of predicted *P. dubium* snakin/GASA mature peptides with other members of this protein family. The deduced amino acid sequences of the predicted genes were aligned with a set of mature plant proteins with GASA domains. Sequences used were manually annotated and retrieved from the UniProt database (http://www.uniprot.org/). Four snakins with reported experimental evidence were also included: E5KBY0_SOLLC, H9D2D5_MEDSA, B2ZAW4_CAPAN, | M1BA38_SOLTU (StSN3). Squares indicate snakin-1 (StSN1) snakin-2 (StSN2) and snakin-3 (StSN3) from *S. tuberosum,* and circles indicate snakin/GASA predicted peptides from *P. dubium* (PdSN1-18). In green, members from subfamily I; in red, members of subfamily II, and in blue members from subfamily III. Values in the nodes regard bootstrap values (5000 replicates). Each sequence was named according to the Mnemonic identifier of a UniProtKB entry. ARATH: *A. thaliana*, SOLTU: *S. tuberosum*, SOLLC: *Solanum lycopersicum*, PRUPE: *Prunus persica*. MEDSA: *Medicago sativa*, CAPAN: *Capsicum annuum*.

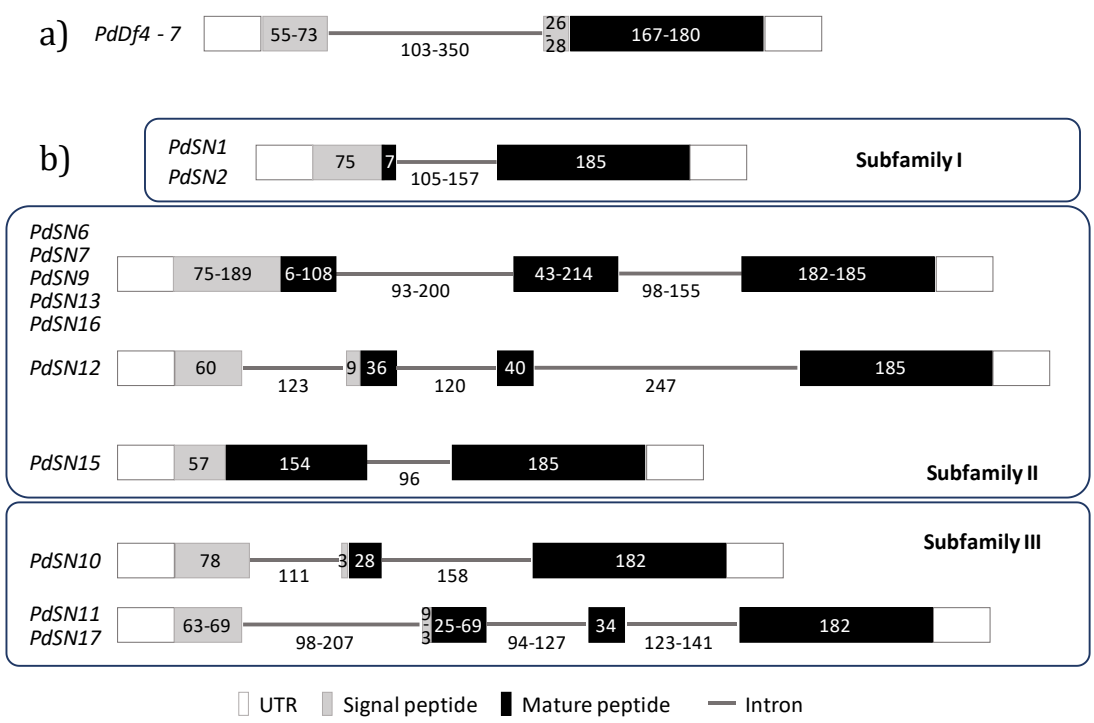

Fig. 8. Schematic representation of the structure of 16 *P. dubium* genes. a) Schematic representation of the structure of four *P. dubium* defensin genes. b) Schematic representation of the structure of 12 *P. dubium* snakin/GASA genes. Snakin/GASA genes were grouped according to the classification into subfamilies I, II, and III. The corresponding exons and introns and their respective sizes are shown.