Bioinformatics Methods for Prediction of Gene Families Encoding Extracellular Peptides

Loup Tran van Canh, Sébastien Aubourg

Methods in Molecular Biology 2731

Springer Protocols

Andreas Schaller
Editor

Plant Peptide Hormones and Growth Factors

MOREMEDIA

Humana Press

Chapter 1

Bioinformatics Methods for Prediction of Gene Families Encoding Extracellular Peptides

Loup Tran Van Canh and Sébastien Aubourg

Abstract

Genes encoding small secreted peptides are widely distributed among plant genomes but their detection and annotation remains challenging. The bioinformatics protocol described here aims to identify as exhaustively as possible secreted peptide precursors belonging to a family of interest. First, homology searches are performed at the protein and genome levels. Next, multiple sequence alignments and predictions of a secretion signal are used to define a set of homologous proteins sharing features of secreted peptide precursors. These protein sequences are then used as input of motif detection and profile-based tools to build representative matrices and profiles that are used iteratively as guides to scan again the proteome and genome until family completion.

Key words Phytocytokine, Conserved motif, Secretion signal, Data mining, SCOOP, PIP

1 Introduction

Small secreted peptides (SSPs) are important players in the extracellular space of plants and are known to regulate a large diversity of biological processes. In the last decade, the increasing number of publications describing such peptides has revealed an unexpected diversity of structures and biological functions. Their actions range from antimicrobial to signaling properties, controlling development, growth, reproduction, and defense against biotic and abiotic stresses [1]. In most cases the precursors of these extracellular peptides, named prepropeptides, possess a signal sequence in their N-terminal part, directing them to the endoplasmic reticulum for vesicle-based transport out of the cell into the apoplast. The secreted peptides include phytocytokines that are recognized by transmembrane receptors to trigger signal transduction and immune responses [2, 3]. On the basis of structural features and their mode of maturation, SSPs have been classified into two main groups by Matsubayashi [4]: The post-translationally modified

Andreas Schaller (ed.), *Plant Peptide Hormones and Growth Factors*, Methods in Molecular Biology, vol. 2731,
https://doi.org/10.1007/978-1-0716-3511-7_1,

peptides (PTMPs) and the cysteine-rich peptides (CRPs). The maturation of PTMPs involves frequent proline hydroxylation (often followed by arabinosylation) and tyrosine sulfation, as well as the proteolytic action of one or several proteases to release the final short functional peptides [5]. A large majority of phytocytokines belong to this class. The CRPs are characterized by an even number of cysteines involved in disulfide bridges defining the final structure of the bioactive peptides.

Bioinformatics approaches for structural annotation of genes encoding precursors of SSPs are limited by their small size and the low level of sequence conservation of the coding regions, thus impairing their detection along genomic sequences by hidden Markov models and similarity searches. Furthermore, functional annotation of SSPs, mainly based on function inference by similarity, also suffers from very low conservation levels, especially for PTMPs for which the conserved sequences are restricted to the short region encoding the mature peptide. These particular features and the limited sensitivity of classical detection methods have slowed down the correct prediction of genes encoding secreted peptides and their correct clustering and classification into gene families [6]. Even within well-known peptide families in species with frequently updated whole genome annotation, new members are still being discovered, and characterizing the full scope of the families requires careful analyses and expert assessments [7, 8].

The only common features that can be used for the prediction of genes encoding extracellular peptides are small size (usually less than 200 amino acids (aa) for the encoded protein) and the presence of a sequence coding for an N-terminal secretion signal [9, 10]. Beyond these simple filters, the annotation of PTMP precursors requires detection and definition of the protein motif corresponding to the mature functional peptide. This key step relies on the sequence conservation of such motifs and, therefore, on the identification of homologous proteins. The protocol that we propose aims to detect potentially related precursors of secreted peptides starting from a single candidate sequence using a highly supervised strategy. Proceeding step by step, through detection of homologs, prediction of secretion signal, and search of shared motifs, the iterative process aims to retrieve lowly conserved SSP sequences until family completion. We also examine some atypical situations that require special expertise to correct automatic annotation errors or ambiguous detection of signal peptides and provide guidelines to predict shared motifs.

To illustrate this protocol, we applied it in two different contexts: the identification of the PAMP-Induced secreted Peptide (PIP) family in *Solanum lycopersicum*, and the extension of the Serine-riCh endOgenOus Peptides (SCOOP) family in *Arabidopsis thaliana*. These two PTMP families represent distinct situations with medium and low levels of sequence conservation between homologs, respectively. The PIP/PIP-like family is well described and comprises 11 genes in *A. thaliana* [11]. Our protocol identified 19 putative homologs in *S. lycopersicum*, only four of which were previously reported [12]. The *Brassicaceae*-specific SCOOP family has first been described as a 14-membered gene family in *A. thaliana* [13]. However, new sequence analyses with lower stringencies have recently extended and questioned the scope of the family [14, 15]. The exploration of the Arabidopsis genome and proteome using the protocol described here highlights 48 putative members. This includes seven genes annotated as non-coding RNA genes in Araport11, two other genes for which the position of the start codon had to be corrected, and two previously unpredicted genes. The results support our belief that the plurality of tools combined with a thorough curated analysis of each detected sequence make our approach sensitive and reliable.

2 Materials

2.1 Omic Datasets

This protocol requires genome and proteome datasets for each investigated species. For the proposed examples, datasets of the latest *A. thaliana* (Col-0) and *S. lycopersicum* (cv. Heinz 1706) genomes and their gene/protein annotations (FASTA and GFF files) were downloaded from TAIR (https://www.arabidopsis.org; Araport11) and Phytozome (https://phytozome-next.jgi.doe.gov; ITAG4.0), respectively. In theory, a file containing the protein sequences deduced from the structural annotation process of the whole genome would be sufficient to detect genes encoding secreted peptides. However, jointly exploring the whole genome sequence allows us to ease our dependence on the gene prediction process that is known to be poorly effective for these types of genes. In this way, we can detect candidate genes which are underpredicted or erroneously annotated as either pseudogenes or non-coding RNA genes. As an alternative method, if the genomic sequence is partial or of poor quality (low sequence coverage), a de novo transcriptome assembly can be used as nucleic acid sequence input.

2.2 Starting Sequences

At least one protein sequence of a putative secreted peptide precursor is required as input to start the workflow. As starting sequences, we used two datasets, the first composed of the *A. thaliana* precursors of three PIP and eight PIPL (PIP-like) proteins according to Vie et al. [11], and the second being the *A. thaliana* PROSCOOP12 protein, precursor of the predicted SCOOP12 peptide [13]. Beyond these examples, any small protein exhibiting an N-terminal signal peptide may be chosen as an interesting candidate to start the proposed workflow (*see* Subheading 3.2.2 for prediction of such candidates). With a cut-off size of 300 aa, the

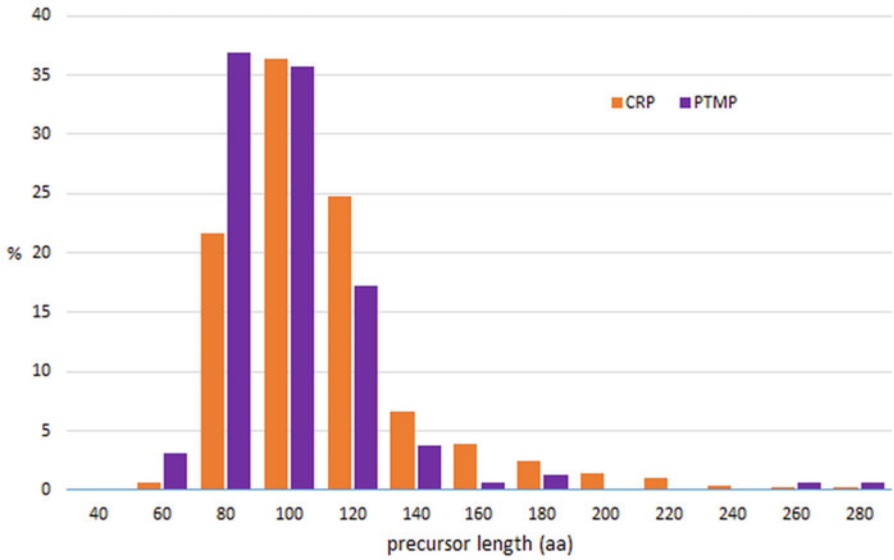

Fig. 1 Distribution of SSP precursor lengths (in aa). Data are based on 672 CRP (Cystein-Rich Peptide) and 157 PTMP (Post-Translationally Modified Peptide) precursors previously described in *A. thaliana*. Their average size is 103 and 90 aa, respectively

diversity of currently known peptide precursors characteristics is covered (Fig. 1).

2.3 Operating System, Hardware, and Software Requirements

All the software used in this protocol are run through a web browser or shell commands on a Debian GNU/Linux 11 bullseye (×86–64) Operating System, but should work under other Unix systems (e.g., MacOs) as well. All software must be available in the executive path. When available, web-based alternative versions are mentioned.

Hardware requirements depend on the size of the dataset. Here we used an Intel© Xeon© CPU E3–1240 v5 @ 3.50GHz × 4; 15.6 Go RAM.

All the tools used to investigate secreted peptide families in this protocol are freely available. The *BLAST+ v2.11.0* package (https://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST/) comprising *makeblastdb*, *blastp*, and *tblastn* is used to retrieve sequences presenting local similarities. Motif predictions and iterative searches are performed using *HMMER v3.3.2* (https://github.com/EddyRivasLab/hmmer) tools (*HMMbuild*, *HMMsearch*, *jackhmmer*) and *MEME suite 5.4.1* (https://meme-suite.org/meme/index.html) including *MEME*, *MAST*, *GLAM2*, and *GLAM2SCAN*. Multiple alignments are performed using *MUSCLE v3.8.1551* (https://github.com/rcedgar/muscle) and displayed using *AliView v1.28* (https://github.com/AliView/AliView). Predictions of signal peptides, topological features, and cellular localization of proteins are performed using *SignalP v5.0* (https://services.healthtech.dtu.dk/service.php?SignalP-5.0), *DeepLoc v2.0* (https://services.healthtech.dtu.dk/service.php?DeepLoc-2.0), *DeepTMHMM v1.0.12* (https://dtu.biolib.com/DeepTMHMM), and *Predotar v1.04* (https://urgi.versailles.inra.fr/predotar/). The manual annotation step is facilitated by the use of *ORFfinder v0.4.3* (https://ftp.ncbi.nlm.nih.gov/genomes/TOOLS/ORFfinder/linux-i64/), *Netgene2 v2.42* (https://services.healthtech.dtu.dk/service.php?NetGene2-2.42), and *Artemis v18.0.0* (http://sanger-pathogens.github.io/Artemis/Artemis/).

3 Methods

The protocol is divided into three main parts (Fig. 2). The first part (Subheading 3.1) aims to detect sequences similar to the starting sequence(s); the second part (Subheading 3.2) integrates sequence analyses for inspection and selection of candidate peptide precursors; the last part (Subheading 3.3) consists in building position weight matrix (PWM) and/or hidden Markov models (HMM) as signature sequences for peptide families that represent the variability of the selected sequences and allow to re-screen sequence libraries through an iterative process. This workflow is reinforced by control and inspection steps ensuring the quality of the results of each part. In our examples, we apply it in order to explore (i) the *S. lycopersicum* genome to identify the PIP/PIPL gene family from the known Arabidopsis members, and (ii) the Arabidopsis genome to identify highly divergent PROSCOOP12 homologs.

3.1 Homolog Search

This protocol aims to detect and retrieve a maximum of sequences similar to the starting sequence(s) of interest (i.e., putative homologs). Since the secreted peptide genes are relatively short and poorly conserved, this search targets not only the annotated protein database but also genome sequences to bypass annotation errors. Similarities are observed at the protein level for higher sensitivity. Therefore, the tools blastp and jackhmmer are used to scan the proteome, and tblastn is applied to scan nucleic sequences (genome if available, transcriptome assembly if not). Albeit slower, jackhmmer has the advantage to run efficient iterative searches [16].

Hereafter, <protein_query_file> is the file containing the starting protein sequence of interest, <proteome_file> contains all the protein sequences deduced from the whole genome annotation, and < genome_file> contains the genomic sequences (Araport11 or ITAG4.0 in our study case). All these files must be in FASTA format.

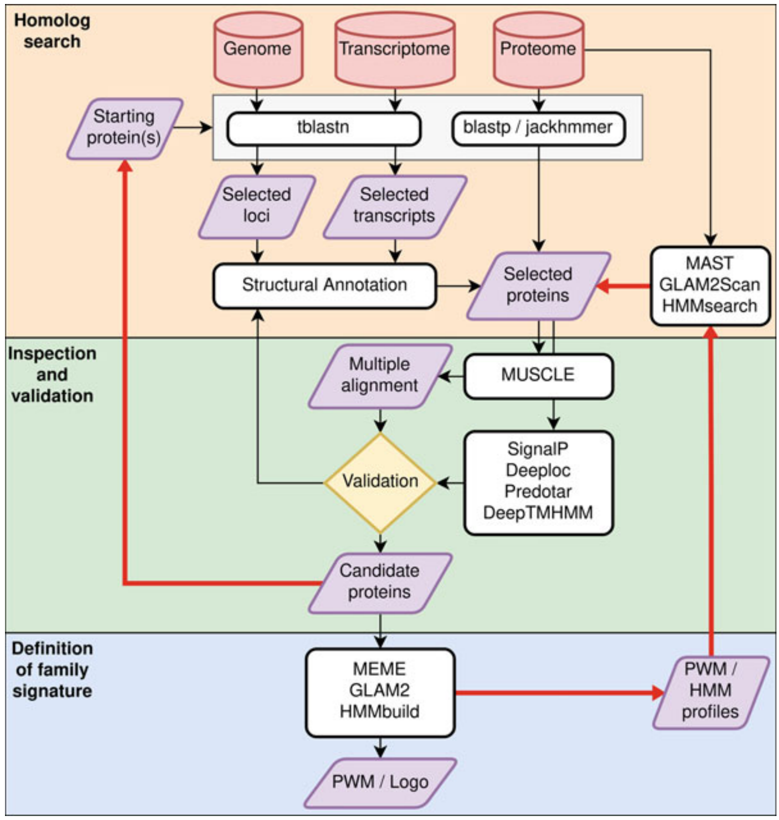

Fig. 2 Protocol for the definition of gene families encoding putative secreted peptide precursors. Software and files are represented by white and purple shapes, respectively. Red arrows illustrate the iterative parts of the method

3.1.1 Commands and Parameters

Strictly following the proposed parameters ensures a search of high sensitivity but reduces its specificity, thus requiring more hands-on expertise to eliminate false positives, especially in the first and second parts of the workflow. Parameters should be adapted to each situation accordingly.

Prior to the use of blastp and tblastn (BLAST+ package) [17], set up a database using makeblastdb:

```
> makeblastdb -dbtype prot -in <proteome_file> -out <proteo-
me_db>
```

where <proteome_db > is the name of the protein database defined by the user.

```
> makeblastdb -dbtype nucl -in <genome_file> -out <genome_db>
```

where <genome_db > is the name of the nucleic database defined by the user.

Run blastp with the following command:

```
> blastp -query <protein_query_file> -db <proteome_db> -evalue
10 -outfmt 0 -out <output_file>
```

The local alignments of the proteins to the proteome are displayed in the defined output file. Start by setting the *-evalue* at 10 to ensure a low selectivity, then, after inspection of the results, lower it to increase the stringency and reduce non-significant alignments in the next runs. Use the *-outfmt 0* option to format results as detailed pairwise sequence alignments to facilitate examination and eventually to remove false positives.

Run tblastn with the following command:

```
> tblastn -query <protein_query_file> -db <genome_db> -evalue
10 -outfmt 0 -out <output_file>
```

The local alignments of the proteins to the 6 frames-translated genome are displayed in the defined output file (*see* **Note 1**).

To perform an iterative protein search, use jackhammer, either with the web-based version (https://www.ebi.ac.uk/Tools/hmmer/search/jackhmmer), or with the following command (the order of arguments must be respected):

```
>jackhmmer -o <output_file> <protein_query_file> <proteome_-
file>
```

The local alignments with the detected similar proteins are displayed in the output file for each iteration. The advantage of this method is that jackhmmer builds a HMM profile after each iteration to improve the following one. We recommend to start with default settings (low threshold and a maximum of 5 iterations) to avoid getting excessively noisy results.

3.1.2 Result Integration and Gene (Re)annotation

The homolog search provides a list of predicted proteins similar to the starting sequences (blastp and jackhmmer results). This list must be completed by adding the results of tblastn which provides hit positions relative to genome sequences tagging candidate regions. For this purpose, it is necessary to identify only the genomic regions for which no gene/protein has been predicted (comparison of hit positions and GFF files describing the position of all annotated genes, *see* **Note 2**). These selected regions probably contain genes of interest that were missed or considered as non-coding RNA genes by gene predictors (false negatives of the whole genome automatic annotation) and should be analyzed manually (*see* **Note 3**).

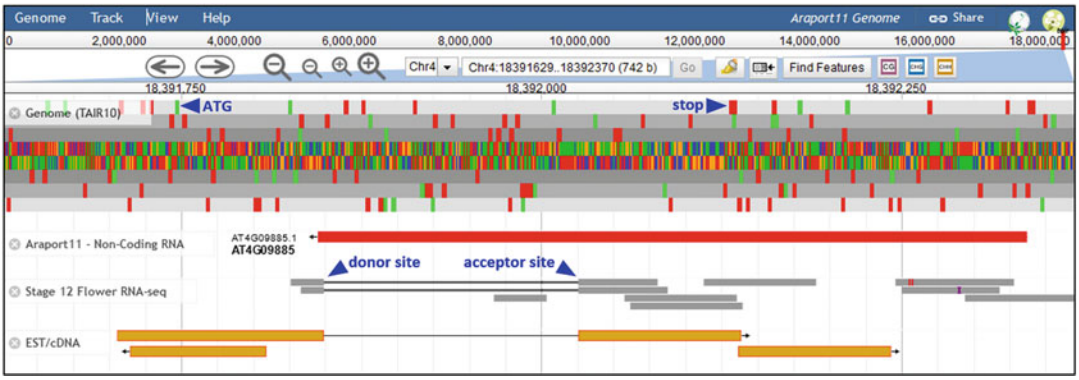

Fig. 3 Screenshot of a genome browser showing the integration of transcript sequences for manual (re)-annotation of a locus of interest. In this example (JBrowse at TAIR, https:/www.arabidopsis.org/), PROSCOOP similarities detected with tblastn partially overlap a non-coding RNA gene predicted in Araport11 (AT4G09885, red track). The display of mapped transcript sequences (RNA-seq reads in grey and EST/cDNA in orange) highlights the presence of an intron. The joint consideration of the 6-frames translated genomic sequences allows the selection of a start codon compatible with both the intron position and the conserved ORF. If several ATG codons have the right properties, signal peptide prediction for each possible N-terminal sequence can be used to select the most likely ATG (*see* Subheading 3.2.2). Start/stop codons and splicing sites selected to predict the final gene/CDS structure are indicated in blue in this example

This (re)annotation aims to predict the correct intron-exon structure and the coding region (CDS) of the gene. To do this correctly, take advantage of available RNA-seq resources. Genome browsers that aggregate and display such resources (i.e., JBrowse at TAIR/Phytozome, Ensembl Plants, or Artemis) are powerful for manual annotation (Fig. 3). If no transcript data are available to guide intron-exon structure annotation, de novo splice sites prediction of the concerned regions can be achieved using software such as NetGene2 [18]. Once the predicted transcript/CDS is recovered (after in silico intron splicing), use ORFfinder to check the integrity of the Open Reading Frame (ORF) and obtain the corresponding translated protein (*see* **Note 4**). Use ORFfinder online (https://www.ncbi.nlm.nih.gov/orffinder/) or with the following command:

```
> ORFfinder -in <predicted_transcript_file.fasta> -strand both
-out <output_file>
```

The protein sequences deduced from each ORF are generated in the defined output file. Select those containing the conserved region(s) previously detected by tblastn.

3.2 Inspection and Validation of the Gene Family

The second part of this protocol uses a FASTA file containing all previously selected homologous proteins (named <selected_proteins_file> hereafter). It contains proteins tagged by blastp and jackhmmer (after removing redundancy) in the whole annotated proteome and those resulting from the (re)annotation tasks.

3.2.1 Multiple Sequence Alignment

The visualization of all the aligned proteins helps to decide which proteins are relevant and which are not. Indeed, the multiple sequence alignment allows the user to consider similarities at the gene family scale and not only locally between two sequences, facilitating the examination of the selected proteins. If false-positive proteins were selected during the search for homologs (Subheading 3.1), they will appear as aberrant in the result of the multiple sequence alignment and should be removed.

For the multiple sequence alignment, run MUSCLE [19] using this command:

```
> muscle -in <selected_proteins_file> -out <alignment_file>
```

where <alignment_file> contains the multiple sequence alignment in aln format.

Because the SSP precursors are often poorly conserved, the multiple alignment may need to be improved locally and manually [20]. Graphical application such as AliView [21] can be used to visualize and edit the multiple alignment file using the following command:

```
> aliview <alignment_file>
```

The multiple sequence alignment provides a first visual overview of the conserved regions(s) shared by the selected proteins. The analysis of this result allows to detect and to remove dissimilar and too divergent proteins wrongly selected at the previous stage. In addition to doubtful similarities, protein length greater than 300 aa (Fig. 1) can be a filtering criterion but it should be employed with caution. Indeed, an unusual protein size (compared to the homologs) can result from errors in the genome annotation pipeline (e.g., erroneous gene structure, gene merging...). Therefore, manual (re)annotation of the respective locus (Subheading 3.1.2) is recommended before eliminating the sequence.

Any modification of the selected protein list requires a new multiple sequence alignment using MUSCLE.

3.2.2 Signal Peptide Prediction

The main feature shared by all SSP precursors, with only rare exceptions such as the PEP family [22], is the presence of an N-terminal signal peptide (SP) addressing them to the endoplasmic reticulum for secretion into the extracellular space. Several

a

SignalP-5.0 Organism: euk

# ID	Prediction	SP(Sec/SPI)	OTHER	CS Position
Solyc02g090610.1.1	SP(Sec/SPI)	0.763568	0.236432	27-28 IEG-RH
Solyc07g062330.2.1	SP(Sec/SPI)	0.993543	0.006457	23-24 FEA-RT
Solyc02g090590.1.1	SP(Sec/SPI)	0.947082	0.052918	23-24 SQG-RN
Solyc03g044530.1.1	OTHER	0.433999	0.566001	
Solyc02g090600.2.1	OTHER	0.010717	0.989283	
Solyc02g090600.2.1_newATG	SP(Sec/SPI)	0.984710	0.015290	24-25 AEG-RQ

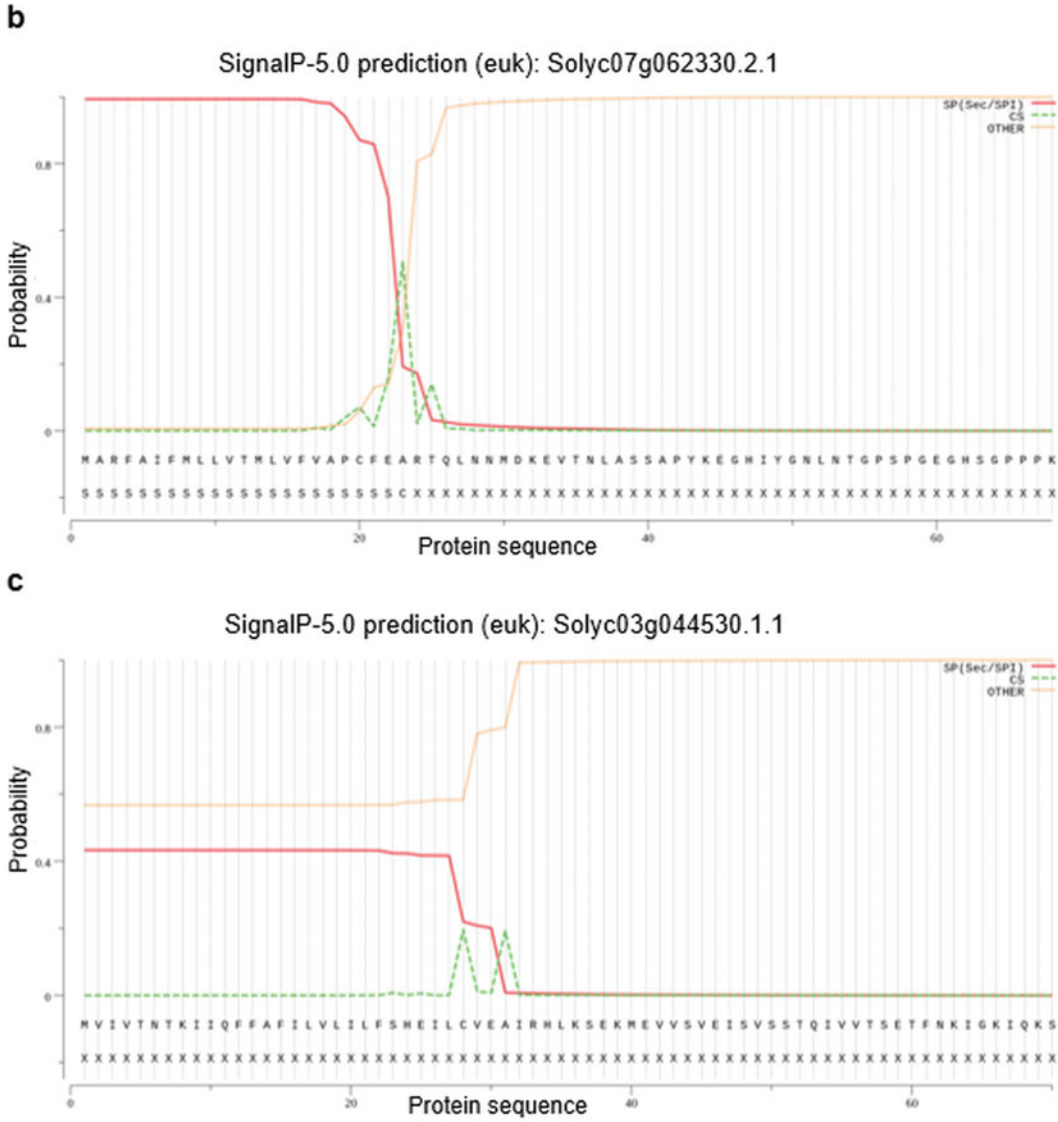

Fig. 4 Example of SignalP5.0 results obtained for *S. lycopersicum* proteins similar to *A. thaliana* PIP/PIPL precursors. (**a**) Tabulated results describing the SIGNALP5.0 predictions. Input sequence names are listed in the first column, associated prediction is indicated in the second column, “SP” corresponds to secreted peptide, whereas “OTHER” indicates that the peptide is not secreted; associated probabilities are displayed in the third and fourth columns, respectively; the fifth column indicates the predicted cleavage site position, its

complementary tools are used to predict the secretion signals of the considered protein sequences.

SignalP5.0 is a reference [23] and should be used first, via its web-based version (https://services.healthtech.dtu.dk/service.php?SignalP-5.0). Alternatively, run it locally with the following command:

```
> signalp -fasta <selected_proteins_file> -format long -mature
```

The optional parameter *-mature* produces a FASTA file containing exclusively the protein sequences lacking the predicted SPs. The option *-format long* generates graphs in png format relevant to assess the predictions (Fig. 4).

To finalize this step, submit the protein sequences for which SignalP5.0 gave unclear results to alternative tools. We propose to use DeepLoc2.0 [24], DeepTMHMM [25], and Predotar [26] that differ in sensitivity. Use DeepLoc2.0 through the web-based application (https://services.healthtech.dtu.dk/service.php?DeepLoc-2.0) or call it using the following command:

```
> deeploc2 -f <selected_proteins_file> -o <output_file> -p -m
Accurate
```

Results are summarized in the defined output file and are completed with graphs if the option *-p* is used. The argument *-m Accurate* uses a high-quality model instead of the default fast high-throughput model.

DeepTMHMM is available online at https://dtu.biolib.com/DeepTMHMM where the <selected_proteins_file> can be submitted as input. Results detail the position of SP and transmembrane segments (*see* **Note 5**). Predotar is available on a web server at https://urgi.versailles.inra.fr/predotar/ for the prediction of subcellular localization. These tools have similar objectives but differ in their algorithm, settings, and training set. In some situations, they give slightly different results and are therefore complementary (see for example Fig. 4c).

The absence of an expected predicted SP should primarily question the protein annotation quality. Indeed, the selection of a wrong start codon can mask the presence of an SP. Therefore, it is necessary to check alternative upstream or downstream start codon

Fig. 4 (continued) 3 upstream and 2 downstream residues, and its associated probability. (**b**) Graphical output corresponding to the N-terminus of the protein Solyc07g062330 for which a clear SP (score 0.99, position 1–23) has been predicted. (**c**) SignalP5.0 concludes that there is no SP in Solyc03g044530 but the graphical output relativizes this conclusion as its N-terminus has SP properties with unclear cleavage site around the 30th aa. For this protein, DeepLoc2.0 predicts extracellular localization, DeepTMHMM predicts an SP (region 1–31), and Predotar localization in the endoplasmic reticulum

(s) in the same reading frame and if present, to test again the SP prediction with the modified protein sequence(s). The protein Solyc02g090600 (putative PIPL) illustrates this situation (Fig. 4a): no peptide signal is detected with the initial protein (ITAG4.0, 173 aa) but the selection of a downstream start codon results in the identification of a new shorter protein of 145 aa with a clear SP (SignalP5.0 score of 0.98).

Finally, proteins for which the absence of an SP is confirmed and for which similarities with the starting protein are doubtful should be removed from the selection (*see* **Notes 5** and **6**).

3.3 Definition of a Family Signature

The third part of this protocol focuses on the characterization of a signature sequence specific to the studied SSP precursor family. Because all SP sequences have similar features (stretch of hydrophobic residues, mainly Leucine) shared by almost all the secreted proteins, we strongly advise you to generate a new file (in FASTA format) containing the sequences of the previously selected homologous proteins excluding the predicted signal peptides (Subheading 3.2.2). This file is named <selected_proteins_withoutSP_file> in the following steps.

3.3.1 Motif and Logo Construction

The definition of conserved motifs, which may correspond to the mature secreted peptides, in a set of sequences can be performed with the MEME tool from the MEME suite v5.4.1 [27]. MEME has the advantage of searching motifs on unaligned sequences and therefore of detecting a variable number of motifs on each input sequence. This may be of interest since precursor proteins may be processed into different SSPs, as described for some members of PTMP families [9, 11, 28, 29]. Use MEME as a web-based application (https://meme-suite.org/meme/tools/meme) or locally with the following command:

```
> meme <selected_proteins_withoutSP_file> -o <output_folder>
-minw <min> -maxw <max> -nmotifs <nmotifs>
```

All results (XML, html, png, and txt files) are saved in the output folder defined by the user. By default, the size of the searched motif is between 8 and 50 aa, but you can adjust it with the *-minw* and *-maxw* parameters according to the first results and the previous multiple sequence alignment (Subheading 3.2.1). If short conserved regions are expected, especially for PTMPs, the motifs can be sized from 5 to 25 aa. The number of searched motifs (1 per default) can also be changed with the *-nmotifs* option if relevant (secondary motifs defining subgroups of proteins can be found). MEME describes the detected motifs with Position Weight Matrix (PWM), also called Position-Specific Scoring Matrix (PSSM) as well as their representative sequence logo. The html file displays interactive graphical results with logo, sequence

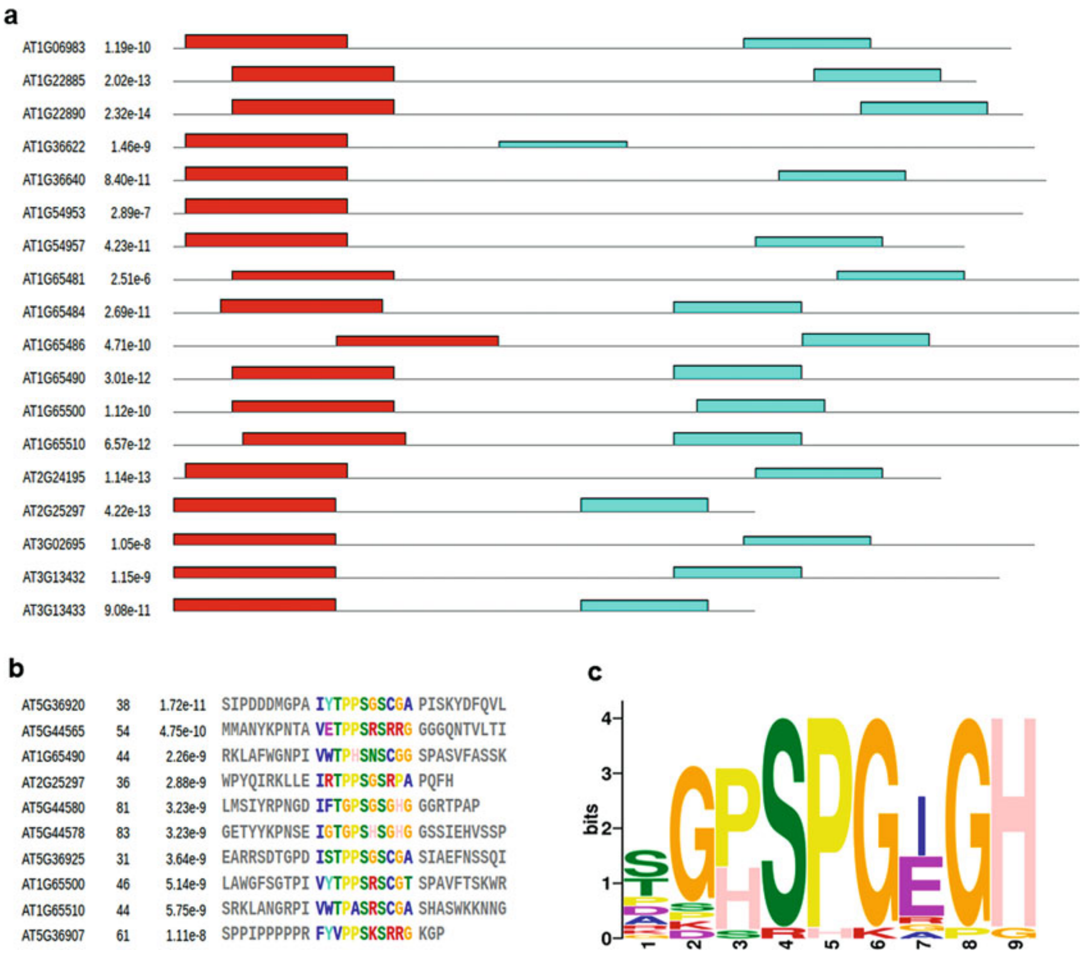

Fig. 5 Illustration of MEME html outputs for motif detection. (**a**) This result has been obtained with two searched motifs ranging from 8 to 50 aa and Arabidopsis candidate PROSCOOP proteins as input. The predicted signal peptides have not been removed before MEME analysis to highlight their biased composition. Motif 1 (red box) overlaps with the SP, and Motif 2 (cyan box) matches with the active SCOOP peptide [13]. (**b**) Sequence alignment provided for each detected motif (here the SCOOP motif) with start positions and p-values. (**c**) Sequence logo representing the sequence signature of the detected motif, here a result obtained with one searched motif ranging from 5 to 9 aa with the 19*S. lycopersicum* candidate PIP/PIP-like proteins

alignment, and motif locations relative to the protein sequences (Fig. 5). Input sequences in which the conserved motif would not be detected deserve to be checked for the robustness of their selection.

The MEME suite proposes an alternative tool named GLAM2 [30] allowing for insertions and/or deletions in the search motifs. Run GLAM2 online (https://meme-suite.org/meme/tools/glam2) or locally using the following command:

```
> glam2 p <selected_proteins_withoutSP_file> -o <output_-
```

```
folder>
```

The output folder defined by the user contains result files (html, png, and txt) including sequence motif alignment, motif description in logo, regular expression, and PSSM matrix.

Another powerful way to describe a protein family is to use a Hidden Markov Model (HMM). The package HMMER3 contains the HMMbuild tool [31] which uses a multiple sequence alignment as input. To generate this alignment file in the required aln format, use MUSCLE and AliView for visualization and optimization, if necessary:

```
> muscle -in <selected_proteins_withoutSP_file> -out <align-
ment_file>
> hmmbuild --amino <output_file> <alignment_file>
```

The output of hmmbuild is a text file corresponding to the HMM profile.

3.3.2 Iterative Search with PWM and HMM Profile

This last step of the protocol exploits the previously generated PWM and HMM profiles to re-scan the entire proteome with greater sensitivity. Indeed, this new proteome screening takes into account the sequence degeneracy observed and tolerated within the signature motif. For this purpose, the results from MEME, GLAM2, and HMMbuild are used as inputs for the tools MAST [32], GLAM2Scan [30], and HMMsearch [31], respectively.

Run MAST and GLAM2Scan online (https://meme-suite.org/meme/tools/mast and https://meme-suite.org/meme/tools/glam2scan) or locally using the following commands:

```
> mast <MEME_output> <proteome_file> -o <output_folder>
> glam2scan p <GLAM2_output> <proteome_file> -o <output_-
folder>
```

The inputs <MEME_output> (xml format) and < GLAM2_output> (txt format) are the files describing the motif (s) previously obtained with MEME and GLAM2 respectively. For both tools, the results are presented in html files listing the proteins in which the motif has been detected with its relative position. Files are generated in the output folders defined by the user.

Run HMMsearch using the following command:

```
> hmmsearch --incE 10 --max -o <output_file> <HMMbuild_output>
<proteome_file>
```

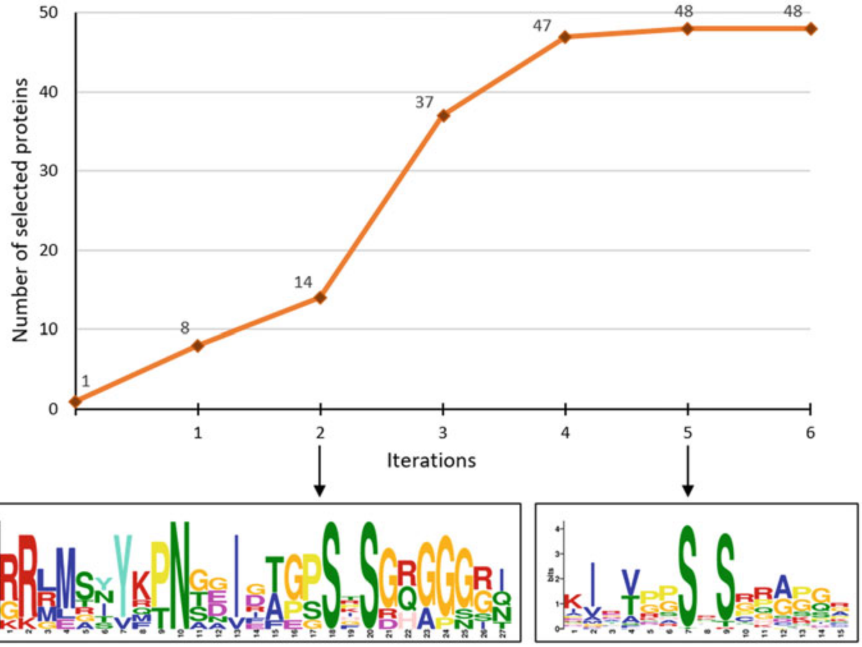

Fig. 6 Progress of the number of selected homologous proteins according to the protocol iterations. The PROSCOOP12 protein is used as starting sequence. The conserved motif (PWM/logo) defined by the MEME tool with default parameters is shown after iterations 2 and 5

The file <HMMbuild_output> contains the HMM profile describing the protein family generated by HMMbuild. The defined output is a txt file containing local alignments between profile and tagged proteins. The optional argument *--incE* is set at 10 to retrieve more proteins than the default settings, but it can be lowered down to gain stringency according to the first results. The optional argument *--max* produces better results at the expense of speed.

The results of MAST, GLAM2Scan, and HMMsearch allow the identification of new proteins that probably belong to the studied family. To verify this, these new sequences should be added to the <selected_proteins_file> file for inspection in comparison with the previously selected proteins (part 2 of the protocol, Subheading 3.2). After validation, these additional sequences will allow the definition of new and more relevant matrices and profiles that can be used again to scan the proteome in an iterative way (red arrows in Fig. 2). For completeness, the newly identified proteins should also be used as new input of jackhmmer and tblastn to re-scan the proteome and the genome and identify new candidates (Subheading 3.1.2). New iterations have to be performed until no more candidates are detected (Fig. 6). The PWM obtained after the last iteration, the final result of the proposed protocol, is an informative signature sequence, limited to the most highly conserved residues, diagnostic of the SSP family studied.

As a final guideline, gradually expanding the omics dataset to other species can also help to construct more representative and pondered PWM and HMM profiles that can recursively feed the workflow to strengthen its efficiency. Although the notion of homology remains questionable with such low similarities, the conserved motif finally defined is a robust prediction of what the functional mature peptide may be. Of course, experimental tests (e.g., with synthetic peptides) remain necessary to confirm the identification of these extracellular peptides.

4 Notes

1. If no whole genome sequence is available for the species of interest, RNA-seq data can also be used as omic resource. In such case, tblastn can be used in the same way by replacing the genome sequence file with transcriptome de novo assembly (FASTA format):

```
makeblastdb -dbtype nucl -in <RNAseq_assembly_file> -out <RNAseq_assembly_db>
```

```
tblastn -query <protein_query_file> -db <RNAseq_assembly_db> -evalue 10 -outfmt 0 -out <output_file>
```

2. To avoid the subtraction of loci tagged with tblastn hits with those corresponding to annotated genes (and also tagged at the protein level), you may want to apply tblastn only against all intergenic regions (instead of the whole genome). Such a file can be generated from the genome sequence and gene feature annotations (GFF file).
3. This manual curation and reannotation is time-consuming, but it should be required only for a limited number of loci. In situations where the number of unannotated loci is too high, automatic gene prediction pipelines could be considered with specific software such as SPADA [33]. However, previously wrongly annotated regions risk to be wrongly annotated again unless additional RNA-seq libraries are supplied to the pipeline.
4. In the same way, transcript sequences tagged as similar to the query sequence by tblastn in a transcriptome assembly (gathered in <selected_transcript_file> in FASTA format) can be analyzed by ORFfinder to retrieve the protein sequences.

```
ORFfinder -in <selected_transcript_file> -strand both -out
```

```
<output_file>
```

Depending on the quality of the assembly, eventual frameshifts (short indels) have to be considered by comparison with the tblastn results.

5. There are certain sequence features that may cast doubt on the secretion of the candidate SSP precursor and therefore justify their exclusion: (i) the presence of transmembrane segment (outside the SP which has similar properties) predicted with DeepTMHMM; (ii) the presence of a C-terminal endoplasmic reticulum-retention signal that can be suspected if a positive match with the motif PS00014 is obtained using ScanProsite (https://prosite.expasy.org/scanprosite/) [34]; (iii) the presence of Glycosylphosphatidylinositol (GPI)-anchor that can be predicted online with PredGPI (http://gpcr.biocomp.unibo.it/predgpi) [35].
6. The spreading of false positives through iterative homolog searches is a concern inherent to the procedure. We advise users to carefully select their protein candidates. Proteins that do not fulfill requirements (e.g., relative position of the conserved regions along the sequence, presence of large insertion, atypical N- or C-termini, and/or even rare intron/exon structure) should be discarded and stored independently until more insights about the family have been obtained. Note that false positives will tend to exclude themselves during the multiple alignment process, producing visual subgroups separating them from the correctly discovered candidates.

Acknowledgments

Authors are grateful to Jean-Marc Celton, Marie-Charlotte Guillou, and Jean-Pierre Renou for the critical reading of the manuscript, and to ANR (ANR-20-CE20-0025), INRAE and French Region Pays de la Loire for funding.

References

1. Tavormina P, De Coninck B, Nikonorova N et al (2015) The plant Peptidome: an expanding repertoire of structural features and biological functions. Plant Cell 27:2095–2118
2. Luo L (2012) Plant cytokine or phytocytokine. Plant Signal Behav 7:1513–1514
3. Gust AA, Pruitt R, Nürnberger T (2017) Sensing danger: key to activating plant immunity. Trends Plant Sci 22:779–791
4. Matsubayashi Y (2011) Post-translational modifications in secreted peptide hormones in plants. Plant Cell Physiol 52:5–13
5. Stintzi A, Schaller A (2022) Biogenesis of post-translationally modified peptide signals for plant reproductive development. Curr Opin Plant Biol 69:102274
6. Takahashi F, Hanada K, Kondo T et al (2019) Hormone-like peptides and small coding genes in plant stress signaling and development. Curr Opin Plant Biol 51:88–95

7. Abarca A, Franck CM, Zipfel C (2021) Family-wide evaluation of RAPID ALKALINIZATION FACTOR peptides. Plant Physiol 187: 996–1010
8. Carbonnel S, Falquet L, Hazak O (2022) Deeper genomic insights into tomato CLE genes repertoire identify new active peptides. BMC Genom 23:756
9. Murphy E, Smith S, De Smet I (2012) Small signaling peptides in Arabidopsis development: how cells communicate over a short distance. Plant Cell 24:3198–3217
10. Boschiero C, Lundquist PK, Roy S et al (2019) Identification and functional investigation of genome-encoded, small, secreted peptides in plants. Curr Protoc Plant Biol 4:e20098
11. Vie AK, Najafi J, Liu B et al (2015) The IDA/IDA-LIKE and PIP/PIP-LIKE gene families in Arabidopsis: phylogenetic relationship, expression patterns, and transcriptional effect of the PIPL3 peptide. J Exp Bot 66:5351–5365
12. Combest MM, Moroz N, Tanaka K et al (2021) StPIP1, a PAMP-induced peptide in potato, elicits plant defenses and is associated with disease symptom severity in a compatible interaction with potato virus Y. J Exp Bot 72: 4472–4488
13. Gully K, Pelletier S, Guillou M-C et al (2019) The SCOOP12 peptide regulates defense response and root elongation in Arabidopsis thaliana. J Exp Bot 70:1349–1365
14. Hou S, Liu D, Huang S et al (2021) The Arabidopsis MIK2 receptor elicits immunity by sensing a conserved signature from phytocytokines and microbes. Nat Commun 12: 5494
15. Zhang J, Zhao J, Yang Y et al (2022) EWR1 as a SCOOP peptide activates MIK2-dependent immunity in Arabidopsis. J Plant Interact 17: 562–568
16. Potter SC, Luciani A, Eddy SR et al (2018) HMMER web server: 2018 update. Nucleic Acids Res 46:W200–W204
17. Camacho C, Coulouris G, Avagyan V et al (2009) BLAST+: architecture and applications. BMC Bioinfor 10:421
18. Hebsgaard SM, Korning PG, Tolstrup N et al (1996) Splice site prediction in Arabidopsis thaliana pre-mRNA by combining local and global sequence information. Nucleic Acids Res 24:3439–3452
19. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797
20. Ranwez V, Chantret NN (2020) Strengths and limits of multiple sequence alignment and filtering methods. In Scornavacca C, Delsuc F, Galtier N (eds) Phylogenetics in the genomic era. No Commercial Publisher, pp 2.2:1–2.2:36
21. Larsson A (2014) AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics 30:3276–3278
22. Jing Y, Shen N, Zheng X et al (2020) Danger-associated peptide regulates root immune responses and root growth by affecting ROS formation in Arabidopsis. Int J Mol Sci 21: 4590
23. Almagro Armenteros JJ, Tsirigos KD, Sønderby CK et al (2019) SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat Biotechnol 37:420–423
24. Thumuluri V, Almagro Armenteros JJ, Johansen AR et al (2022) DeepLoc 2.0: multi-label subcellular localization prediction using protein language models. Nucleic Acids Res 50: W228–W234
25. Hallgren J, Tsirigos KD, Pedersen MD et al (2022), DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks. https://www.biorxiv.org/content/10.1101/2022.04.08.487609v1
26. Small I, Peeters N, Legeai F et al (2004) Predotar: a tool for rapidly screening proteomes for N-terminal targeting sequences. Proteomics 4:1581–1590
27. Bailey TL, Johnson J, Grant CE et al (2015) The MEME suite. Nucleic Acids Res 43:W39–W49
28. Roberts I, Smith S, De Rybel B et al (2013) The CEP family in land plants: evolutionary analyses, expression studies, and role in Arabidopsis shoot development. J Exp Bot 64: 5371–5381
29. Guillou MC, Balliau T, Vergne E et al (2022) The *PROSCOOP10* gene encodes two extracellular hydroxylated peptides and impacts flowering time in Arabidopsis. Plan Theory 11:3554
30. Frith MC, Saunders NFW, Kobe B et al (2008) Discovering sequence motifs with arbitrary insertions and deletions. PLoS Comput Biol 4:e1000071
31. Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 39:W29–W37
32. Bailey TL, Gribskov M (1998) Combining evidence using p-values: application to sequence homology searches. Bioinformatics 14:48–54

33. Zhou P, Silverstein KA, Gao L et al (2013) Detecting small plant peptides using SPADA (Small Peptide Alignment Discovery Application). BMC Bioinfor 14:335

34. Sigrist CJA, de Castro E, Cerutti L et al (2013) New and continuing developments at PROSITE. Nucleic Acids Res 41:D344–D347

35. Pierleoni A, Martelli PL, Casadio R (2008) PredGPI: a GPI anchor predictor. BMC Bioinfor 9:392