Доклады Международной конференции "Математическая биология и биоинформатика". *Под ред. В.Д. Лахно. Том 7. Пущино: ИМПБ РАН, 2018. Статья № e63. doi: 10.17537/icmbb18.13*

Identification of antimicrobial peptides in *Lycopersicon esculentum* genome

Kovtun A.S.[1], Istomina E.A.[2], Slezina M.P.[2], Odintsova T.I.[2]

[1]*Moscow Institute of Physics and Technology, Dolgoprudny, Moscow Region, Russia*
[2]*Vavilov Institute of General Genetics Russian Academy of Science, Moscow, Russia*

mer06@yandex.ru

Antimicrobial peptides (AMPs) constitute an important part of the plant immune system protecting plants from the invading pathogens. Some antimicrobial peptides are also active against human pathogenic microorganisms, including antibiotic-resistant strains that makes these molecules promising candidates for the design of next-generation drugs to treat infections. Plants represent a valuable source of effective yet poorly explored antimicrobial peptides. An efficient pipeline for high-throughput *in silico* detection of antimicrobial peptides in *Lycopersicon esculentum* genome has been developed. As many as 66 putative AMPs were revealed in *L. esculentum* genome. The discovered AMPs belong to four AMP families: defensins, thionins, lipid-transfer proteins, and hevein-type peptides. The vast majority of newly discovered peptides have not been annotated in *L. esculentum* genome so far. Further functional analysis of detected AMPs will evaluate their potential as novel drug leads and biopesticides for practical application in agriculture and medicine.

Key words: in silico, cysteine-rich peptides, antimicrobial peptides, bioinformatics analysis, Lycopersicon esculentum.

Идентификация антимикробных пептидов в геноме *Lycopersicon esculentum*

Ковтун А.С.[1], Истомина Е.А.[2], Слезина М.П.[2], Одинцова Т.И.[2]

[1]*Московский физико-технический институт, Московская область, Россия*
[2]*Институт общей генетики им. Н.И. Вавилова Российской академии наук, Москва, Россия*

Антимикробные пептиды (АМП) составляют важную часть иммунной системы, защищая растения от патогенов. Некоторые антимикробные пептиды также активны против патогенных микроорганизмов человека, включая устойчивые к антибиотикам штаммы, что делает эти молекулы перспективными кандидатами для разработки лекарств нового поколения для лечения инфекций. Растения представляют собой ценный источник эффективных, но плохо изученных антимикробных пептидов. Нами был разработан эффективный метод широкомасштабного поиска *in silico* антимикробных пептидов в геноме *L. esculentum*. Было выявлено 66 предполагаемых АМП. Обнаруженные АМП принадлежат к четырем семействам АМП растений: дефензины, тионины, липид-переносящие белки и гевеиноподобные пептиды. Подавляющее большинство недавно открытых пептидов до сих пор не были аннотированы в геноме *L. esculentum*. Дальнейший функциональный анализ выявленных АМП позволит оценить их потенциал в качестве новых лекарственных средств и биопестицидов для практического применения в сельском хозяйстве и медицине.

Ключевые слова: in silico, цистеин-богатые пептиды, антимикробные пептиды, биоинформатический анализ, Lycopersicon esculentum.

Introduction

In recent years, due to the rising antibiotic resistance in pathogens, natural antibiotics – antimicrobial peptides (AMPs), have gained considerable interest in novel drug design and biopesticide development. AMPs represent versatile biologically active molecules that occur in a great number of organisms ranging from microorganisms to animals and plants. AMPs share similar structural characteristics: (i) low molecular weight (≤ 10 kDa), (ii) positive charge of the molecule, (iii) amphiphilic nature. Plant AMPs are cysteine-rich peptides, whose structure is stabilized by 2–6 disulphide bridges [1]. Based on structure similarity, plant AMPs are classified into several families: defensins, thionins, lipid-transfer proteins, hevein- and knottin-type peptides, haipinins, and cyclotides [2]. AMPs have a wide range of functions in defense and developmental processes and thus present a vast, but still poorly explored pool of molecules with important practical application in medicine and agriculture.

In the search for novel peptides with promising properties, the approaches of classical peptidomics and bioinformatics-based *in silico* mining strategies are

being currently used. However, the peptidomic approaches are presently facing serious challenges, especially for plants [3]. Isolation and characterization of peptides from plant tissues using analytical methods, such as liquid chromatography and mass spectrometry (MS), can be both time-consuming and inefficient. This is due to extreme complexity of peptide pools in plant tissues, often low quantities of biologically active compounds, and insufficient resolution of chromatographic methods [4]. With the advent of next-generation sequencing technologies, the abundance of genome and transcriptome datasets increases, providing a new approach to peptide discovery. *In silico* mining has become not only an alternative, but a complementary strategy for detecting novel biologically active peptides [5]. It expands our knowledge of plant peptidomes and makes high-throughput peptide discovery more feasible. Another advantage of *in silico* mining is that it allows the identification of precursor proteins.

To date, identification of peptides through *in silico* methods has relied mainly on similarity searches for sequence homology to annotated genomes. However, plant AMPs display low sequence similarity, therefore searches for homology may be inefficient in novel AMP discovery. To cope with this problem, in addition to searching for conserved sequences, special characteristics of peptides of interest can also be used as search criteria, e.g. cysteine motifs, namely the arrangement of cysteine residues in the polypeptide chain.

In current work, *in silico* methods were applied for detection of the genes of antimicrobial peptides that belong to the families of defensins (DEFL), lipid-transfer proteins (LTP), hevein-type peptides (HEV) and thionines (THI) in the genome of tomato, *Solanum lycopersicum* str. Heinz 1706. For the analysis, the latest version of the genome assembly available in GenBank by accession number GCA_000188115.2 was used. Tomatoes (*Lycopersicon esculentum*, Solanaceae family) belong to one of the most consumed and important crops worldwide (FAOSTAT; http://faostat.fao.org/). Tomato diseases decrease crop yield and storage time of the fruits. Understanding of the tomato defense arsenal will form the basis for controlling diseases and production of high-quality fruits. The release of the tomato genome sequence [6] opened new avenues for *in silico* mining of defense molecules including AMPs in this species.

Results

Pipelines for *in silico* AMP mining

Two pipelines for revealing putative AMPs in genome and transcriptome datasets were developed in Perl. Both pipelines can be run in Linux command line and at the input accept files with amino acid sequences in FASTA format.

The first pipeline is based on the method of hidden Markov models [7]. The ready models of AMP precursors were obtained from SPADA [8]. The pipeline works in several steps. First, the hidden Markov models are aligned against the genome with hmmsearch from HMMER package [9]. Amino acid sequences in FASTA-like formats are required as input. Next, the detected sequences are filtered by the Perl scripts. The first script filters the hits by E-value (E-value < 10^{-3}). The second script detects signal peptides in the remaining sequences of AMP precursors using SignalP v4.1 for this purpose [10]. Sequences without signal peptides are discarded. After that, the third script checks the discovered sequences to match the structure MZ..Z{C}*m*{X}*n*{C}*l*{X}*k*..*, where MZ..Z is a signal peptide; M, methionine; Z, any amino acid; C, cysteine; X, any amino acid residue except cysteine; *m*, *n*, *l*, *k* = 1, 2, 3…; * is a stop codon. After all quality control processes, the nucleotide sequences of peptides are detected by a specific script written in Perl. As a result, a collection of predicted amino acid and nucleotide sequences of identified putative AMP precursors is obtained.

The second pipeline uses the method of regular expressions to detect sequences of putative defensin precursors. This pipeline consists of scripts that scan transcriptome for sequences that match certain regular expressions. The general structure of regular expressions was as shown above. They include known structures of cysteine motifs found in AMP-like peptides [11]. After obtaining a set of sequences that satisfy the structure of constructed regular expressions, the identified sequences are filtered by the presence of a signal peptide. At this step, the script from the first pipeline with the corresponding function is used. Finally, the nucleotide sequences are obtained using the same approach as above.

Defensins

Defensins are small (~ 5 kDa) cationic peptides that form an important component of the plant immune system [12]. They possess 8–10 cysteine residues that form 4–5 disulphide bridges providing the molecule with high structural stability. The structure of defensins is characterized by the so-called cysteine-stabilized αβ-motif also present in insect defensins and scorpion toxins. Defensins are found in virtually all multicellular organisms. All plant defensins are synthesized as precursor proteins containing a signal peptide for secretion into endoplasmic reticulum and the mature peptide domain (class I defensins). A small group of plant defensins possess an additional C-terminal prodomain (CTPP) (class II defensins). Silverstein discriminated 4 cysteine motifs in defensin-like peptides: two with 8, one with six and one with 4 cysteine residues [13].

In silico analysis revealed 10 putative defensin-like peptides (pDEFLs) in tomato genome (Fig. 1,A), which have not been annotated so far. Three of tomato

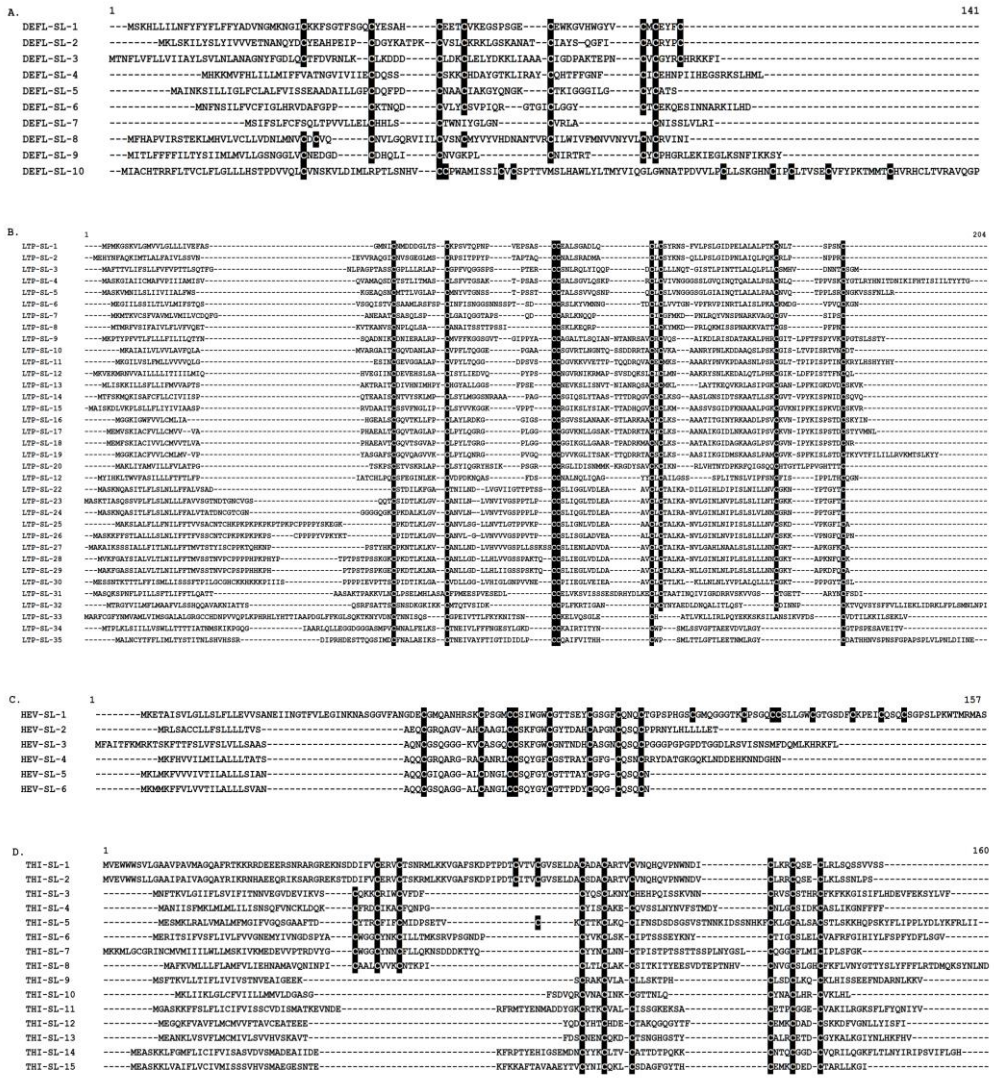

Fig. 1. Sequence alignment AMP: (A) – defensins; (B) – lipid-transfer proteins; (C) – hevein-type AMPs; (D) – thionins.

pDEFLs (DEFL-SL-1÷DEFL-SL-3) possess a typical 8-Cys motif CX{4, 25}CX{2, 12}CX{3, 4}CX{3, 17}CX {4, 32}CXCX{1, 6}C characteristic of defensins. Other peptides have a variable number of cysteine residues ranging from 4 to 11. pDEFLs with a classical 8-Cys motif show sequence similarity to DEFLs from *L. esculentum* with identity score from 40 to 46 %. One of the remaining tomato peptides (DEFL-SL-3) have low sequence similarity (26 %) to a cysteine-rich low-molecular-weight protein from *Arabidopsis thaliana*, three peptides (DEFL-SL-4, 5, 8) show similarity to "uncharacterized" proteins from different plant species, and three (DEFL-SL-7, 9, 10) discovered peptides have no BAST hits, and are, therefore, new *L. esculentum* peptides.

Lipid-transfer proteins

Lipid-transfer proteins (LTPs) comprise a large protein family present in all land plants [14]. They are small (< 10 kDa) cysteine-rich polypeptides containing 8 cysteine residues. They possess 4–5 α-helices stabilized by disulphide bridges. LTPs are classified into two families LTP1 and LTP2 on the basis of molecular size: LTP1s have about 90 amino acid residues, and LTP2s have about 70 residues. The molecules of LTPs are characterized by the presence of a tunnel-like hydrophobic cavity for binding and transporting different lipids. They are produced as precursors with a signal peptide, which targets them to the apoplast. LTPs have a role in the synthesis of lipid barrier polymers, such as cuticular waxes and suberin. They are also involved in signaling during pathogen attack, pollen and defense response to biotic and abiotic stress.

In silico mining in *L. esculentum* genome revealed 35 putative LTPs (Fig. 1,B) with a characteristic cysteine motif: CX{6, 15}CX{9, 31}CCX{8, 21}CXC X{13, 35}CX{5, 18}C. Thirty-one peptides were 100 % identical to tomato proteins from the NCBI database. Of them, 15 peptides were annotated as *L. esculentum* LTPs, 11 peptides as proline-rich

peptides, and 5 peptides were annotated as "uncharacterized" or "predicted" proteins. Four belonged to the LTP family according to their cysteine motifs and similarity to LTPs from other plants. One tomato peptide (LTP-SL-12) was similar to "hypothetical" protein CQW3_2361 (69 % identity); one (LTP-SL-33) to an "uncharacterized" protein (65 % identity), and one (LTP-SL-35) to an egg cell-secreted protein from *Capsicum baccatum* (77 % identity).

Hevein-type AMPs

Hevein-type AMPs received their name from hevein, the AMP from the rubber tree latex [15]. They are cysteine-rich peptides with all 6, 8 or 10 cysteines forming disulphide bridges. All hevein-type AMPs possess a chitin-binding site composed of several conserved hydrophobic residues that bind chitin of the cell walls of pathogens. Binding to the tips of growing fungal hyphae is supposed to inhibit their elongation and subsequent expansion to plant tissues making hevein-type AMPs active members of the plant defense machinery. Hevein-type AMPs are synthesized as precursor proteins, consisting of a signal peptide, the mature peptide domain and the C-terminal prodomain. Some other hevein-type AMPs are produced by post-translational proteolytic degradation of class 1 chitinases.

Genome analysis revealed 6 putative hevein-type peptides in tomato (Fig. 1,C). Five of them possess a cysteine motif characteristic of 8-Cys hevein-type peptides: CX{1, 8}CX{4, 5}CCX{5}CX{6}CX{3, 5}C X{3, 4}C. One peptide (HEV-SL-1) has two 8-Cys motifs, and is, therefore, a two cysteine-domain protein. Three tomato putative hevein-type peptides (HEV-SL-2, 3, 4) possess a C-terminal prodomain found in all characterized peptides of this family, while the remaining two peptides (HEV-SL-5, 6) lack this domain. All identified putative hevein-type peptides show high sequence similarity to chitin-binding domains of chitinases from the Solanaceae species. It should be specifically noted that *L. esculentum* hevein-type peptides show considerable sequence similarity with each other that might indicate origin from a common ancestor.

Thionins

Thionins are short toxic AMPs found in a wide range of mono- and dicotyledonous plants [16]. They have either 6 or 8 cysteine residues that form disulphide bonds. Thionins are active against fungi and bacteria causing diseases not only in plants but in humans as well. They also inhibit growth of cancer cells that makes them promising leads for the development of anti-cancer agents. The mode of antimicrobial action of thionins includes formation of pores in membranes culminating in membrane disruption. Thionins are synthesized as precursor proteins containing a signal peptide, the mature peptide region and the conserved cysteine-rich C-terminal prodomain, which is supposed to mask the lytic properties of the mature peptide during its transport to the vacuole.

Mining of *L. esculentum* genome for thionin sequences revealed 15 peptides (Fig. 1,D). All of them possessed a characteristic 6-Cys motif in the C-terminal prodomain typical for thionins: CX{3}CX{3, 4}CX{4, 32}CX{2, 3}CX{3, 4}C. The number of cysteine residues in the predicted mature peptide region varied from 0 to 8. However, none of the sequences had a CC-motif characteristic of classical thionins. Five of tomato thionin-like peptides were identical to the proteins from the NCBI database, four of them were annotated as "uncharacterized" proteins, and one as "major pollen allergen Ole e 6". One identified peptide (THI-SL-5) had 99 % sequence identity to thionin-like peptide from *L. esculentum*.

Conclusion

An efficient pipeline for identification of putative AMPs in genome or transcriptome datasets has been developed. As many as 66 putative AMPs were revealed in *L. esculentum* genome. The discovered AMPs belong to four AMP families: defensins, thionins, LTPs, and hevein-type peptides. Thus, the pipeline developed is suitable for high-throughput detection of AMPs in plant genomes. However, it should be specifically noted that although *in silico* mining is a powerful tool for the discovery of novel AMPs, it cannot provide information concerning the actual presence of the gene product at the peptide level, the length of the mature peptide, or post-translational modifications that might be necessary for peptide functioning. Therefore, further functional analysis of detected AMPs is necessary, which will evaluate their potential as novel drug leads and biopesticides to combat infection for practical application in agriculture and medicine.

Acknowledgments

This work was supported by the Russian Science Foundation (grant № 16-16-00032).

References

1. Manners J.M. Hidden weapons of microbial destruction in plant genomes. *Genome Biology.* 2007. V. 8. № 9. P. 225–228. doi: 10.1186/gb-2007-8-9-225.
2. Tam J.P., Wang S., Wong K.H., Tan W.L. Antimicrobial peptides from plants. *Pharmaceuticals (Basel).* 2015. V. 8. № 4. P. 711–757. doi: 10.3390/ph8040711.
3. Koehbach J., Jackson K.A.V. Unravelling peptidomes by *in silico* mining. *Peptidomics.* 2015. V. 2. P. 17–25.
4. Finoulst I., Pinkse M., Van Dongen W., Verhaert P. Sample preparation techniques for the untargeted LC-MS-based discovery of peptides in complex biological matrices. *J. Biomed.*

Biotechnol. 2011. V. 2011. P. 245–291. doi: 10.1155/2011/245291.
5. Gruber C.W., Muttenthaler M. Discovery of defense- and neuropeptides in social ants by genome mining. *PLoS ONE*. 2012. doi: 10.1371/journal.pone.0032559.
6. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*. 2012. V. 485. № 7400. P. 635–641. doi: 10.1111/tpj.12616.
7. Eddy S.R., Mitchison G., Durbin R. Maximum discrimination hidden Markov models of sequence consensus. *J. Comput. Biol*. 1995. V. 2. № 1. P. 9–23. doi: 10.1089/cmb.1995.2.9.
8. Zhou P., Silverstein K.A., Gao L., Walton J.D., Nallu S., Guhlin J., Young N.D. Detecting small plant peptides using SPADA (Small Peptide Alignment Discovery Application). *BMC Bioinformatics*. 2013. V. 14. P. 355–370. doi: 10.1186/1471-2105-14-335.
9. Durbin R., Eddy S.R., Krogh A., Mitchison G. *Biological sequence analysis: probabilitic models of proteins and nucleic acids*. Cambridge, UK: Cambridge University Press, 1998.
10. Bendtsen J.D., Nielsen H., von Heijne G., Brunak S. Improved prediction of signal peptides: SignalP 3.0. *Journal of Molecular Biology*. 2004. V. 340. № 4. P. 783–795. doi: 10.1016/j.jmb.2004.05.028.
11. Silverstein K.A., Moskal W.A.Jr., Wu H.C., Underwood B.A., Graham M.A., Town C.D., Vanden Bosch K.A. Small cysteine-rich peptides resembling antimicrobial peptides have been under-predicted in plants. *The Plant Journal*. 2007. V. 51. № 2. P. 262–280. doi: 10.1111/j.1365-313X.2007.03136.x.
12. Lay F.T., Anderson M.A. Defensins – components of the innate immune system in plants. *Current Protein and Peptide Science*. 2005. V. 6. № 1. P. 85–101. doi: 10.2174/1389203053027575.
13. Silverstein K.A., Graham M.A., Paape T.D., Vanden Bosch K.A. Genome organization of more than 300 defensin-like genes in Arabidopsis. *Plant Physiology*. 2005. V. 138. № 2. P. 600–610. doi: 10.1104/pp.105.060079.
14. Salminen T.A., Blomqvist K., Edqvist J. Lipid transfer proteins: classification, nomenclature, structure, and function. *Planta*. 2016. V. 244. № 5. P. 971–997. doi: 10.1007/s00425-016-2585-4.
15. Slavokhotova A.A., Shelenkov A.A., Andreev Y.A., Odintsova T.I. Plant hevein-type peptides. *Biochemistry (Mosc)*. 2017. V. 82. № 13. P. 1659–1674. doi: 10.1134/S0006297917130065.
16. Stec B. Plant thionins-the structural perspective. *Cell Mol. Life Sci.* 2006. V. 63. № 12. P. 1370–1385. doi: 10.1007/s00018-005-5574-5.