AMP-Identifier: A Unix shell script for antimicrobial peptide identification

Bezerra-Neto, João Pacifico[1,2]; Santos, Mauro Guida[2]; Benko-Iseppon, Ana Maria[1]

[1]Laboratory of Plant Genetics and Biotechnology, Genetics Department, Universidade Federal de Pernambuco, Av. Prof. Moraes Rego, 1235, 50.670-423, Recife, PE, Brazil; [2]Laboratory of Plant Physiology, Department of Botany, Universidade Federal de Pernambuco, Av. Prof. Moraes Rego, 1235, 50.670-423, Recife, PE, Brazil

With the advent of experimental high-performance platforms, in particular next-generation sequencing (NGS) optimized assay systems, advanced bioinformatics approaches have enabled comprehensive and maximized studies of eukaryotic genomes in a quick and economically viable manner. In plants, for example, transcriptome studies have been used for quantitative analysis of thousands of expressed genes related to germination, growth and development, flowering, and conditions of biotic and abiotic stresses, allowing the understanding plant response mechanisms against stresses. The identification of gene families in the huge amount of new data has been facilitated by bioinformatic methods and by the availability of several online repositories. The main computational methods developed for identifying AMPs (Antimicrobial Peptides) on a genome-wide scale involved in silico approaches to evaluate their amino acid composition and structure. Scripts, written in Unix shell or other scripting languages such as Perl, can be seen as the most basic form of pipeline framework. The AMPs generally present between 12 and 50 amino acids, including the presence of disulfide bond and/or cyclization of the peptide chain. These peptides have a variety of antimicrobial activities ranging from membrane permeabilization to action on a range of cytoplasmic targets. To improve the identification of AMPs at omic level, we developed a Unix shell script to integrate other analysis tools to find AMPs from HMM models. This automated pipeline adopts classification based on HMM models cataloged in CAMP (www.camp3.bicnirrh.res.in) database, search based om HMMER tool in some cases translation from nucleotide to amino acids for genomic data input using TransDecoder tool. This script gathers all tools and HMM database, building all commands to execute the analysis, asking the user just about input and parameters of AMP search. To evaluate the script efficiency we downloaded the Arabidopsis thaliana genome to run as input, using a cutoff of 0.00001. We obtained 32 AMP families identified on A. thaliana genome under the selected cutoff, using 89 HMM models. For predicted Defensin sequences we found that against GenBank, only 50% were confirmed by BLAST and CD-search tools, indicating that our tools may identify sequences not classified as AMPs in conventional alignment approaches. The here presented tool facilitates AMP global identification and its usability, once Unix environment may be a challenge for most biologists, since the implementations are based on Linux command lines, often requiring some knowledge.

Financial support: CNPq, CAPES.