De novo transcriptome assembly of the extremophile plant *Calotropis procera*

Rivas, Rebeca[1]; Bezerra-Neto, João Pacífico[2]; Pandolfi, Valesca[2]; Coêlho, Maria Reis Velois[1]; Santos, Mauro Guida[1]; Benko-Iseppon, Ana Maria[2]

[1]Universidade Federal de Pernambuco, Departamento de Botânica, Laboratório de Fisiologia Vegetal, Recife, PE, Brazil; [2]Universidade Federal de Pernambuco, Departamento de Genética, Laboratório Genética e Biotecnologia Vegetal, Recife, PE, Brazil

Calotropis procera (Apocynaceae) is an evergreen shrub found in arid and semiarid environments, whose parts (shoot, leaf, root, flower and especially latex) are widely used in phytotherapy. Its anti-inflammatory properties make this species an excellent candidate for the search of antimicrobial peptides (AMPs). The objective of this work was to analyse the *C. procera* transcriptome under environmental stress and to identify AMPs, focusing on the thaumatin PR-5 family. High throughput sequencing was performed using Illumina Hi-seq 2500 2 x 100 bp reads, from six libraries of the *C. procera,* treated (30min, 2h, 8h and 45 days under 100 mM NaCl stress imposition) and untreated (0h and 45 days, controls). The sequencing data analysis and assembly were performed using Trinity platform. Assembled sequences were compared against the Universal Protein Resource (UNIPROT). To evaluate the AMPs, these sequences were submitted to the tool AMP-Identifier v.1.0. We obtained 284 million reads including 26, 29, 62, 26, 46 and 95 million for 0 h, 30 min, 2 h, 8 h, and 45 days after salinity stress and 45 days control, respectively. De novo assembled reads generated 224,652 transcripts with a mean of 745 bp in length (224 and 31,249 bp for minimum and maximum, respectively), N50 of 2,525 bp and GC content 39.33%. The transcripts comprise 134,461 unigenes with a mean length of 417 bp and N50 of 1,606 bp. After sequence annotation (UNIPROT database), 80.64% (181,149 transcripts) were predicted and 19.36% (43,503 transcripts) were annotated as unknown. The expression analysis showed 50% of the transcripts with fold change (FC) < 2 for each comparison. FC >2 and <10 were represented by 25% and 37% of up and down-regulated transcripts, respectively. In the search for AMPs, via AMP-Identifier 880 candidates were obtained, distributed in 40 AMPs families. After AMP prediction ($CAMP_{R3}$) and conserved domain search, we obtained 95 AMPs distributed in six families: cecropin (40), cystatin (2), defensin (7), moricin (2), thaumatin (42) and transferrin (2). All 42 possible thaumatins presented the complete expected domain, 32 (76%) thaumatin candidates presented signal peptide and 40 (95%) presented 16 cysteines conserved with eight disulphide bonds, typical of thaumatins. After expression analysis, nine thaumatins were upregulated (FC > 2) in early times of salinity (30 min, 2 h, and 8 h) and two were down-regulated (FC < -2) in the late times after stress (45 days). Thaumatin expression data indicate that increased expression occurs in the first hours after stress in response to salinity.

Financial support: CAPES/CNPq.