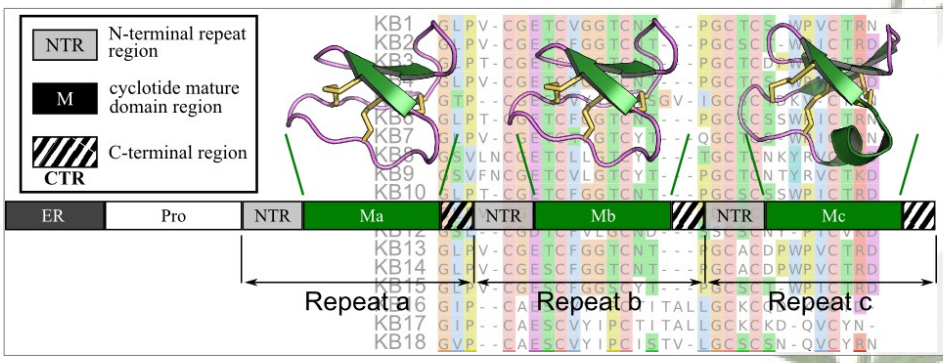
Assembling large sets of similar transcripts from short reads: cyclotide identification from the *Oldenlandia affinis* transcriptome

Quentin Kaas, Husen Jia, Joshua S. Mylne and David J. Craik

Institute for Molecular Bioscience, The University of Queensland, St Lucia, Brisbane, Queensland 4072 Australia

Cyclotides are one of the largest families of plant peptides, and species producing cyclotides belong to several families, including Rubiaceae, Solanaceae and Violaceae.[1–3] Cyclotides have a similar three-dimensional structure, which is characterized by a cyclic peptide backbone and a knotted arrangement of three disulfide bonds that together render them highly resistant to thermal and enzymatic degradation.[4] Three cyclotides have been shown to have insecticidal activities but the biological function for the great majority of them is still unknown.[4] A single plant species could express over 100 different cyclotides, but the difficulty to isolate these peptides at the protein level as prevented the comprehensive study of cyclotide content of any one species. Here, three transcriptomes of a cyclotide producing species, *Oldenlandia affinis*, were assembled *de novo* from 75 bp and 90 bp paired-end reads. *O. affinis* is of African origin and its cyclotides were originally discovered in the 1970s and it remains the best studied species for its cyclotide content.[5,6] Assembling cyclotide transcripts in the absence of a reference genome is challenging because cyclotide precursors contain very similar and at times internally repeating domain sequences. A set of ~500 EST sequences[7] was used to identify the best strategy to assemble cyclotide precursor transcripts as well as other transcripts. Several factors were tested, including read cleaning parameters, assembly algorithms, k-mer lengths, and two different methods to merge multiple assemblies. Using the EST sequences, potential assembly mistakes were evaluated. The most common assembly errors were the artificial introduction of frameshifts and the merging of unrelated transcript fragments. Two strategies were employed to improve the initial assembly; reassembling contigs with the cap3 algorithm[8] and combining knowledge on the sequencing depth and on alignments with a sequence database to excise low reliability sequence segments. The accuracy of cyclotide transcript assembly was assessed by Sanger sequencing of some newly discovered cyclotides. A web based interface was created to visualize for each transcript: its sequencing depth, open reading frame, alignments with UniProt-KB,[9] and predicted function.



Cyclotide precursors display several cyclotide domains, which mature into cyclic peptides. At the start of this project, ~20 cyclotides were already known in O. affinis (illustrated in background).

1. Craik, D.J.; Daly, N.L.; Bond, T.; Waine, C. J. Mol. Biol. 1999, 294, 1327–1336.
2. Craik, D.J.; Cemazar, M.; Wang, C.K.L.; Daly, N.L. Biopolymers 2006, 84, 250–266.
3. Poth, A.G.; Mylne, J.S.; Grassl, J.; Lyons, R.E.; *et al.* J. Biol. Chem. 2012, 27, 27033-46
4. Craik, D.J. Toxins 2012, 4, 139–156.
5. Kaas, Q.; Craik, D.J. Biopolymers 2010, 94, 584–591.
6. Mylne, J.S.; Wang, C.K.; van der Weerden, N.L.; Craik, D.J. Biopolymers 2010,
7. Qin, Q.; McCallum, E.J.; Kaas, Q.; Suda, J.; Saska, I.; *et al.* BMC Genomics 2010, 11, 111.
8. Huang, X.; Madan, A. Genome Res. 1999, 9, 868–877.
9. UniProt Consortium Nucleic Acids Res. 2012, 40, D71–75.