Prototheca wickerhamii genome sequencing project – a preliminary report

Zofia Bakula*,1, Paweł Siedlecki2, Jan Gawor3, Robert Gromadka3, Tomasz Jagielski1

1University of Warsaw, Faculty of Biology, Department of Applied Microbiology, Warsaw, Poland
2Institute of Biochemistry and Biophysics, Polish Academy of Sciences; Department of Systems Biology, Institute of Experimental Plant Biology and Biotechnology, University of Warsaw, Warsaw, Poland
3DNA Sequencing and Oligonucleotides Synthesis Unit at the Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Warsaw, Poland

Background: Prototheca is a genus of aerobic, colorless, yeast-like algae widely distributed in the environment. Although normally saprophytic, these organisms may, under certain conditions, produce infections in humans and different animal species. Prototheca algae are thus the only known plants with pathogenic ability for humans and animals. Of the six currently recognized Prototheca species, three have been implicated in human disease: P. wickerhamii, P. zopfii, and P. cutis. The aim of the project is to perform a preliminary sequencing analysis of the whole genome of P. wickerhamii, a major etiological agent of human protothecosis.

Material/methods: The strain used in our study is P. wickerhamii PL1, originally isolated from the first case of human protothecosis in Poland. The sequencing was performed on next generation sequencing instrument MiSeq (Illumina). The long paired end and mate pair reads were assembled into contigs and ordered into large scaffolds using Newbler (v3.0) de novo assembler. Two versions of Prototheca wickerhamii genome assembly were screened for regions to be masked (v3 and v4). In order to ensure a reference free initial detection, inverted repeat finder (IRF) and RepeatModeler were used to generate transposon candidates de novo. Since IRF & RepeatModeler produce multiple overlapping hits, CD-HIT was utilized for sequence clustering with similarity threshold set at 100% and query coverage set at 99% of the shorter sequence. Pfam & CDD protein domain profiles were used to elucidate motifs typical for transposons. The above procedure resulted in a custom RepeatMasker library construction. With this tool coordinates of detected transposable elements and coordinates of detected simple repeats were estimated.

Results: Genome sequencing generated 27.2 million reads and 4.65 Gb of sequence data. Sequence reads assembly yielded 2,860 chromosomal scaffolds with the total length of 29 Mbps and sequencing depth of approximately 100x. The size of the chromosomal fraction of genome was estimated at about 35 Mbps. We have also obtained two complete organellar circular genomes: 52.1-kbp mitochondrial genome and 48.2-kbp plastid genome. The study shows that P. wickerhamii is characterized by low abundance of transposable elements (TE). Using a custom, curated database of TE, developed in-house, we found 20 LTR_Gypsy elements and 1 DIRS-Ngaro element in v3 genome results (24 LTR_Gypsy and 1 DIRS-Ngaro for v4 genome). Sequencing data showed sporadic occurrence of simple repeats (SR) in the P. wickerhamii genome (v3 had 52, v4 had 62). Most of them were of size ranging between 100 and 300 bps.

Conclusions: The analyses so far performed within the project provided preliminary information about the overall genome organization of P. wickerhamii. In the next step, gene content (with special focus
on virulence genes) and metabolic capacities of the pathogen will be investigated. The study was financed by the National Science Centre (2013/09/N/NZ2/00248)