Prototheca wickerhamii genome sequencing project – a preliminary report

Zofia Bakuła[1], Paweł Siedlecki[2,3], Jan Gawor[4], Albert Bogdanowicz[2], Robert Gromadka[4], and Tomasz Jagielski[1]

[1]Department of Applied Microbiology, Institute of Microbiology, Faculty of Biology, University of Warsaw, Poland
[2]Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Warsaw, Poland;
[3]Department of Systems Biology, Institute of Experimental Plant Biology and Biotechnology, University of Warsaw, Poland
[4]DNA 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, unicellular, colourless, 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.

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.

Materials and methods

The strain used in our study is *P. wickerhamii* PL1, originally isolated from the first case of human protothecosis in Poland. The preparation of the strain for the purpose of genome sequencing involved its revitalization on culture media and large-scale nuclear DNA extraction using in-house isolation method specially designed for disruption of *P. wickerhamii* cell walls. The sequencing was performed on next generation sequencing instrument MiSeq (Illumina).

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.

MAKER was used as the main annotation pipeline for gene prediction and annotation. Scaffold sequences were masked with RepeatMasker. Expressed sequence tags (EST) and protein sequences from related organisms (*Chlorella* sp., *Chlamydomonas* sp.) were used as gene evidence source. Within MAKER, AUGUSTUS was trained with *Chlorella* sp. genome for *ab initio* gene prediction with default parameters. Sequence similarity was assessed using BLAST and sequence alignments were prepared using Exonerate. tRNAscan was used to predict tRNA genes. Functions were assigned to genes based on the best alignments using Blastx (E-value < = 1e-10) against the NR database. Domain annotations were carried out using InterProScan against publicly available databases, including ProDom, PRINTS, Pfam, SMART, and PROSITE. Gene Ontology (GO) descriptions for the individual genes were obtained using BLAST2GO.

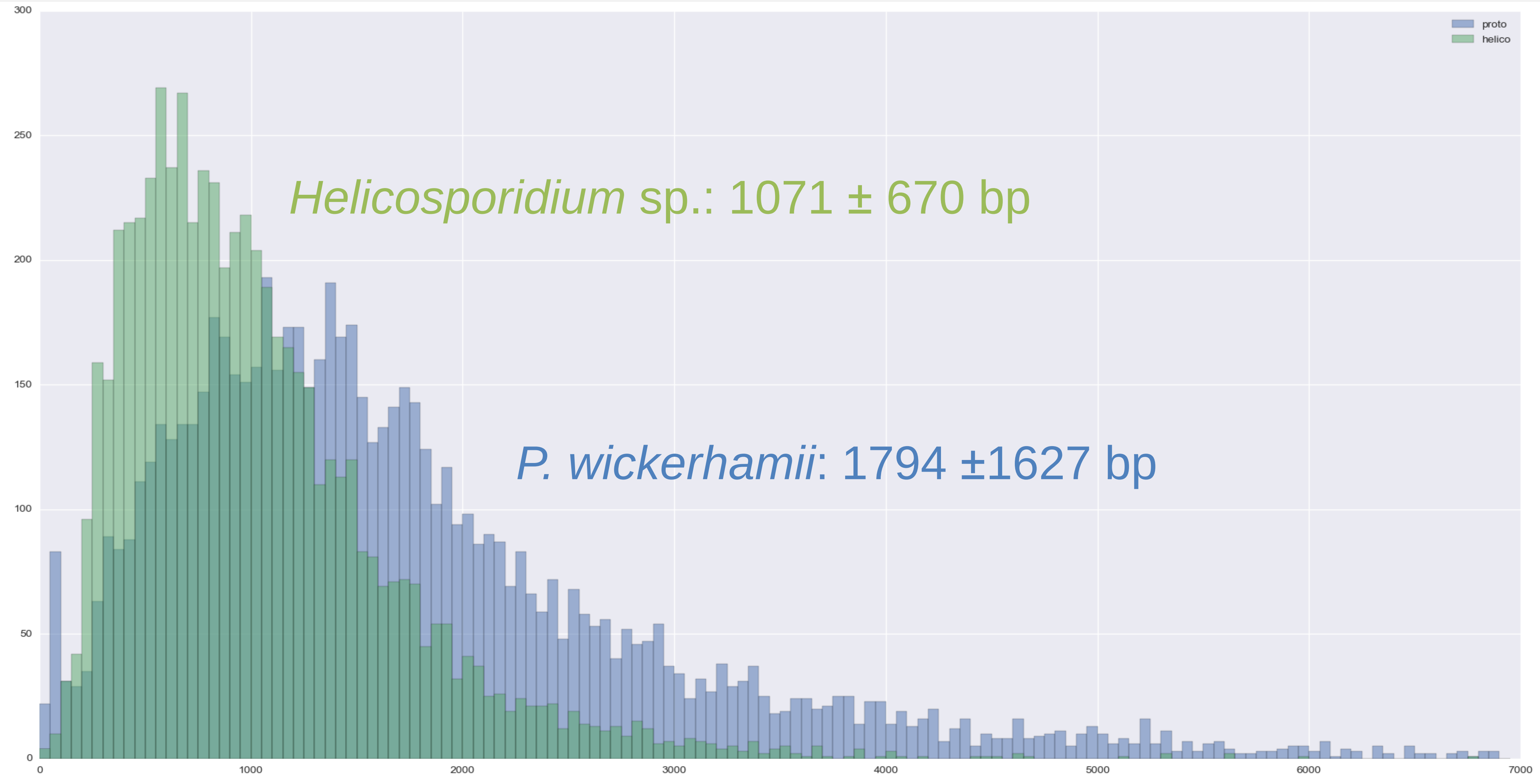
Results

- Genome sequencing yielded 2.860 scaffolds. The size of the genome was estimated at *ca*. 20.7 Mbps.
- Functional annotation resulted in Gene Ontology categories for ¼ genes. Additionally, we predicted InterProDomains for almost ¾ genes.
- *P. wickerhamii* seems to be characterized by a short average gene length and by a low abundance of repeat sequences.

Tab. 1. Comparison of *P. wickerhami*i annotation statistics with same/similar order organisms.

	P. wickerhamii	*C. protothecoides*	*Helicosporidium sp.*	*C. variabilis*	*C. reinhardtii*
Assembly length (Mb)	20.7	22.9	17	46.2	121
GC content (%)	61	63	62	67	64
Number of gene	7287	7039	6035	9791	15.143
Average gene length (bp)	**1794**	2863	1071	2928	4312
Average no of exons per gene	**3.98**	5.72	2.3	7.3	8.33
Average exon length (bp)	326	207	366	170	190
Average intron length (bp)	355	246	168	209	373
Coding sequence ratio (%)	2.8	3.2	no data	4.7	8.0
Repeat sequences (%)	**0.77**	6.1	no data	8.9	16.7

Fig. 1. Gene length distribution in *P. wickerhamii* and *Helicosporidium* sp.



Conclusion

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 grant «PRELUDIUM» (Contract no.: 2013/09/N/NZ2/00248)

E-mail: zofiabakula@biol.uw.edu.pl