Functional annotation of a divergent genome using sequence and structure-based similarity

Dennis Svedberg[1,2†], Rahel R. Winiger[1†], Alexandra Berg[1,2†], Himanshu Sharma[1,2], Christian Tellgren-Roth[3], Bettina A. Debrunner-Vossbrinck[4], Charles R. Vossbrinck[5], Jonas Barandun[1]

† These authors contributed equally to this work.

[1] Department of Molecular Biology, The Laboratory for Molecular Infection Medicine Sweden (MIMS), Umeå Centre for Microbial Research (UCMR), Science for Life Laboratory, Umeå University, 90187 Umeå, Sweden,

[2] Department of Medical Biochemistry and Biophysics, Umeå University, 90736 Umeå, Sweden.

[3] Science for Life Laboratory, Department of Immunology, Genetics and Pathology, Uppsala University, Uppsala, Sweden.

[4] Department of Math/Science, Gateway Community College, 20 Church Street, New Haven, CT 06510, USA.

[5] Department of Environmental Science, Connecticut Agricultural Experiment Station, New Haven, CT 06504, USA.

Microsporidia are obligate intracellular parasites with extremely compacted genomes and an unusually high sequence divergence. This degree of divergence limits functional genome annotation using traditional, sequence-based methods as they result in numerous genes of unknown function. Compared to primary sequence, protein structure is generally more conserved as it is usually tightly linked to protein function. Therefore, with the current software for fast and accurate protein structure prediction and comparison, structure-based similarity searches can serve as a valuable, complementary approach to traditional functional annotation.

In this study, we combined traditional, sequence-based, and structure-based functional annotation and visualize the results in a ChimeraX plugin called ANNOTEX (Annotation Extension for ChimeraX). We applied this approach on our newly sequenced, high-quality genome of the microsporidian *Vairimorpha necatrix*, a parasite of Lepidoptera. First, we predicted protein-coding DNA sequences and confirmed 89% by RNA sequencing data. For the structural data, we folded the *V. necatrix* proteome using ColabFold and performed structural searches with FoldSeek against the PDB and AlphaFold databases. Both the structural and the sequence-based hits are then manually inspected and curated in ANNOTEX. The curation and the addition of structural matching enhanced the quality and accuracy of the *V. necatrix* genome annotation compared to when blindly relying on sequence similarity. With this approach we can present a comprehensive annotation of the *V. necatrix* genome and highlight the most prevalent protein folds in this understudied organism.

Taken together, we established a workflow for functional genome annotation of divergent, non-model organisms and used it to elucidate the underlying biology of a divergent lepidopteran parasite. We believe that the complementation of functional annotation through structural similarity is a valuable addition for gene function prediction in microsporidia and other divergent organisms.