Project in Applied Bioinformatics

Applied machine learning to comparative genomics

Professor Jan Komorowski
Department of Cell and Molecular Biology

The genomic revolution, initiated by the sequencing of the human genome about twenty years ago, has facilitated the development of high-throughput sequencing and improved assembling methods. With this, enormous steps have been taken in increasing our understanding of genome architecture and gene content of different organisms. However, we soon realized that genomes are not simple a collection of gene, but rather harbor complex dynamic relationships and numerous simultaneous effects on each other. Therefore, it becomes complex to evaluate their relationships by standard methods and in isolation.

Diplomonads are one of the most common human parasites affecting 200 million people worldwide. With a broad zoonotic potential, they can cause disease in mammals, fish, and birds. Increasing the knowledge about diplomonads will help us understand the relationship between host-parasite and their evolutionary mechanisms. Diplomonads are flagellated unicellular eukaryotes that are adapted to low oxygen environments. They are usually parasitic in the intestine of the vertebrates, but some species are free-living in pond sediment. This paraphyletic structure in the group makes them special to work as a model organism. The transition from parasitic to free-living lifestyle is considered rare in evolution since the parasitic genome reduces and becomes more compact than the free-living species.

However, in this group, we can also find free-living species like *Hexamita inflata*. This species is potentially a secondary free-living organism. Previous studies shows that *Trepomonas Sp.*, free-living species in the diplomonad group, escaped from parasitic lifestyle by acquiring genes missing in the parasitic ancestor from bacteria via horizontal gene transfer (Xu et al. 2016).

Therefore, we focus on an evolutionary transition between free-living and parasitic organisms within eukaryotes using free-living diplomonads and their close parasitic relatives as the main model system using genomics, transcriptomics, and comparative genomics methods. The comparison between parasitic and free-living diplomonads leads us to understand better how species escaped from a parasitic lifestyle. Furthermore, everything we learn about this process can be extrapolated to understand how a parasitic species evolved in the first place

To address this problem, we propose to use interpretable machine learning techniques, that is methods where relationship among the features and the outcomes can be identified, to analyze genome metabolic pathways from different sister species adapted to different environment.

Importantly, the results from these analyses could potentially be applied to other systems and organisms to unravel complex relationships, for example emergent pathogens, such as *Influenzavirus*, *Coronavirus*, or multidrug resistant bacterial strains.

Candidates shall posses strong programing skills in languages commonly used for data science, R, python or Julia, as well as interest in machine learning algorithms. Experts in computation, genomics and machine learning will support and supervised the proposed project.

Contact:
Daniel Rivas: daniel.rivas@icm.uu.se
Jan Komorowski: jan.komorowski@icm.uu.se

Xu, Feifei, Jon Jerlström-Hultqvist, Martin Kolisko, Alastair G. B. Simpson, Andrew J. Roger, Staffan G. Svärd, and Jan O. Andersson. 2016. 'On the reversibility of parasitism: adaptation to a free-living lifestyle via gene acquisitions in the diplomonad Trepomonas sp. PC1', *BMC Biology*, 14.