Applied machine learning to transposable element identification

Project in Applied Bioinformatics

Professor Jan Komorowski
Department of Cell and Molecular Biology

The genomic content holds the recipe for the biomolecules that constitute each living organism we are aware of today. However, large part of the genome is plagued with what used to be called 'junk' DNA (Xing et al. 2009). Nowadays, we understand that this section of the genome is largely composed by a heterogenous group of entities that replicate themselves and invade host genomes, called transposable elements. However, transposable element composition varies greatly from species to species, and such effects are even more pronounce in some cases where genome expansions at the expense of transposable element radiation took place.

Diplomonads are a group of unicellular eukaryotes adapted to free-living and parasitic lifestyles. The group has members with compact and expanded genomes which makes them a suitable model system to work with on genome architecture. The genome of free-living *Hexamita inflata* is 34% composed of interspersed repeats in contrast to their parasitic counterparts where it ranges from 9-12%. Of these the repeat regions 90% in *H. inflata* remain uncharacterized.

The common way to classify transposable elements consists on comparing their sequences against known elements (Goerner-Potvin and Bourque 2018). However, this method falls short when numerous unknown elements are identified in a single species. To aid this problem, we propose using an unsupervised machine learning method for clustering transposable elements based on sequence features.

The results from this project will contribute to increase the repository of known transposable elements and their classification, which directly impacts some human genetic pathologies and it is essential for comparative genomics.

Candidates will require 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

Goerner-Potvin, Patricia, and Guillaume Bourque. 2018. 'Computational tools to unmask transposable elements', *Nature Reviews Genetics*, 19: 688-704.

Xing, J., Y. Zhang, K. Han, A. H. Salem, S. K. Sen, C. D. Huff, Q. Zhou, E. F. Kirkness, S. Levy, M. A. Batzer, and L. B. Jorde. 2009. 'Mobile elements create structural variation: Analysis of a complete human genome', *Genome Research*, 19: 1516-26.