



Identification of biotechnological potential on genomic nonfunctionalized orthologs elements

Objectives: Integrated use of biological systems for unveiling biotechnological potential on putative encoding genomic data.

Methodology: The biotechnological potential of unexploited ecosystems is of paramount importance of disclose new solutions for the current societal challenges. Scaling law predicts Earth is home to upward of 1 trillion (10^{12}) microbial species. Following present classification, there are a little less than 9,300 known species of prokaryotes. The advent of 3rd Next Generation Sequencing has enabled the exponential retrieval of genomic information in real time at single molecule level, bringing IoT to disclose massive biotechnological potential. Currently it produces >60 gigabytes of data per day with the potential to increase dramatically to terabase-per-day in the coming years. However, more than 40% of the orthologs elements identified from genomic data do not have assigned function by standard methods of classification, new approaches for data analytics are required.

The fellow will be in charge to produce and analyse different type of experimental data from the programme's partners (Genomic & Phenotypic @ M&B; Gene Expression @ GER; Metabolomic @ Carlos Cordeiro's Lab). The student will also test Rosetta stone method, develop and compare new integrative approaches on data reduction, data mining for creation of a novel Knowledge database. This data will be linked and semantically enriched in collaboration with F. Couto's Lab. During visits to Christopher's group (ANL, Illinois - US), the fellow will test automated reconstruction of genomescale metabolic models based at the previous knowledge database. New data validation and reconciliation methods will be used to development of new tools for functional imputation. The student will test, develop and compare new approaches for real time identification of biotechnological potential as well as development of cutting-edge tools.

Locey KJ, Lennon JT. Scaling laws predict global microbial diversity. *Proc Natl Acad Sci U S A*. 2016; 113(21):5970-5. doi: 10.1073/pnas.1521291113.

Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. *Nat Rev Genet*. 2015;16(6):321-32. doi: 10.1038/nrg3920.

Riley R, Haridas S, Wolfe KH, et al. Comparative genomics of biotechnologically important yeasts. *Proc Natl Acad Sci U S A*. 2016;113(35):9882-7. doi: 10.1073/pnas.1603941113.

Supervisor: Ricardo Dias **Co-Supervisor:** Christopher Henry

Type of fellowship:

Mixed (Portugal and abroad: Argonne National Laboratory, US)