Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Roderic
Email	roderic.guigo@crg.cat
Institution	Center for Genomic Regulation
Website	https://genome.crg.cat/
Group	Bioinformatics and Genomics

Project

Computational genomics

Project Title:

: Efficient gene annotation across the entire phylogenetic spectrum

Keywords:

Bioinformatics, gene finding, transcriptomics,

Summary:

Understanding Earth's biodiversity and responsibly administrating its resources is among the top scientific and social challenges of this century. The Earth BioGenome Project (EBP) aims to sequence, catalog and characterize the genomes of all of Earth's eukaryotic biodiversity over a period of 10 years (http://www.pnas.org/content/115/17/4325). The outcomes of the EBP will inform a broad range of major issues facing humankind, such as the impact of climate change on biodiversity, the conservation of endangered species and ecosystems, and the preservation and enhancement of ecosystem services. It will contribute to our understanding of biology, ecology and evolution, and will facilitate advances in agriculture, medicine and in the industries based on life: it will, among others, help to discover new medicinal resources for human health, enhance control of pandemics, to identify new genetic variants for improving agriculture, and to discover novel biomaterials and new energy sources, among others. The value of the genome sequence depends largely on the precised identification genes. The aim of the research project is to develop a gene annotation pipeline that produces high quality gene annotations that can be efficiently scaled to more than one million species. Our group has a long-standing interest in gene annotation. Roderic Guigo developed one of the first computational methods to predict genes in genomic sequences (geneid, Guigó et al, 1992), which has been widely used to annotate genomes during the past years. On the other hand, we are part of GENCODE, which aims to produce the reference annotation of the human genome. Within GENCODE we have developed experimental protocols to efficiently produced full-lengh RNA sequences. Our pipeline will be based on identifying the genes that can be precisely predicted computationally in a given species, subtract them from RNA samples, and produced high quality RNA sequences for the genes that are more difficult to annotate. The master student will work specifically on the identification of selenoprotein genes

Expected skills::

Good programming skills python, C, or similar. Good unerstandgin of molecular biology concets

Possibility of funding::

To be discussed

Possible continuity with PhD: :

To be discussed