Universitat Pompeu Fabra Barcelona | Master in Bioinformatics for Health Sciences

Master project 2020-2021

Personal Information

Supervisor	Rosa Trobajo & David Mann
Email	rosa.trobajo@irta.cat
Institution	IRTA
Website	http://www.irta.cat/ca/grup/aigues-marines-i-continentals/
Group	Aigües Marines i Continentals

Project

Computational genomics

Project Title:

Functional evaluation of HTS reads using protein sequence data

Keywords:

HTS, protein sequence, environmental DNA, microalgae

Summary:

High Throughput Sequencing (HTS) is currently being developed as a substitute for traditional methods in biomonitoring aquatic ecosystems (e.g. for the European Union Water Framework Directive). For example, methods involving counting cells of microscopic algae are being replaced by metabarcoding: a short region of the gene (rbcL) coding for the large subunit of RuBisCO (ribulose-1,5-bisphosphate carboxylase/oxygenase, the key enzyme of photosynthesis) is amplified from environmental samples and sequenced by Illumina; the reads are processed through a bioinformatics pipeline, where they are filtered to remove sequences containing errors, and then identified to species by reference to a database of known 'barcodes' from Sanger sequencing. Although these pipelines work adequately, some of the methods used to reject 'faulty' sequences are crude. For example, sequences that occur only rarely in a dataset are often rejected because of the error rate in Illumina sequencing, which leads to incorrect nucleotides occurring ±anywhere along a sequence. A threshold is therefore set, that a particular sequence must be observed twice or more times before it is accepted as real, on the basis that any particular error during sequencing is unlikely to occur many times repeatedly. The proportion of reads rejected on this basis can be of the order of 50% and probably involves many type II errors in the quantification of diversity. The suggested project would involve developing a method that is able to assess HTS reads, even when the sequences have never been encountered before (and are therefore not in the reference database), by taking account of the fact that RuBisCO (like all proteins) has a function and that function can only be performed if the protein folds in the correct way. Hence certain changes in the DNA coding for RuBisCO are evolutionarily 'easy' (because they have no effect on protein function, e.g. many codon 3rd position changes), whereas others are strongly or fully constrained. The project aims to determine the likelihood of different changes at a particular DNA site by evaluating variation among known rbcL gene sequences in the group of organisms being studied and also considering RuBisCO structure (which is well-known), and to use this information to develop an 'intelligent filter' for metabarcoding pipelines. In this, reads would be evaluated on the basis of whether they code for biologically plausible peptides, rather than solely on the basis of their frequency. Though applied to rbcL, the approach developed could be applicable with appropriate modification to any coding sequence used for metabarcoding (e.g. the CO1 gene used to barcode animals).

Expected skills::

Bioinformatics pipeline development, protein structure prediction, programming, sequence alignment

Possibility of funding::

No

Possible continuity with PhD: :

To be discussed

Comments:

This project would require a combination of skills from different areas of specialization in the syllabus, mainly from computational genomics and structural bioinformatics, and include the need for some programming.