ANNOTATING GENES IN NOISY READS THROUGH DEEP LEARNING

Student background:** Master CS/DIAM

The goal of this project is to implement gene finding in noisy DNA reads as a deep-learning model

Background
Deep learning has shown remarkable performance in a variety of learning tasks, particularly when presented with large amounts of (noisy) data. One task in bioinformatics that we like to solve is to identify the boundaries of genes directly in long DNA sequence reads. Long read technologies have revolutionized the way we can measure DNA, but they come with a major drawback: they contain up to 10% errors.

Current gene finding algorithms do not work with error containing sequences

We could reformulate the classical gene finding algorithms in a deep learning architecture. This would allow us to directly identify genes in noisy reads, which would then allow studying structural variation and population dynamics within microbial communities.

Aims

- Reformulate classical gene finding as a deep learning problem: How can we translate gene annotation to a deep learning problem? What feature encodings? What is the most suitable deep-learning architecture for gene annotation?
- Training and benchmarking: How do we acquire sufficient training data for our model? How do we evaluate such a model?
- Assess real-world impact: How well does this model perform compared to baseline alternatives in realistic real-world scenarios.

It is expected that at the end of the project we have an annotation model that can provide gene boundary predictions in Nanopore reads with sufficient accuracy for downstream synteny analysis.

Responsible supervisor: Thomas Abeel – t.abeel@tudelft.nl (EEMCS – INSY - Pattern Recognition and Bioinformatics)

*Types of project: Bachelor seminar (TI3706): 5 ECTS literature review course // Research Assignment (IN5010): 15 ECTS bioinformatics literature review // BEP: 10-15 ECTS Bachelor End/Honors Project // MEP: 30-60 ECTS Master End (thesis) Project // internship: 3 month no credit project. The type of project you are completing will impact the scope and depth that you will be expected to accomplish.

** Student background: The Delft Bioinformatics Lab serves a broad student community with a variety of projects. The background mentioned here is a suggestion and not a restriction. We will adapt the scope and focus of the project to connect well with your expertise and program requirements.