AMBASSADE
DE FRANCE
AU ROYAUME-UNI
Liberté
Égalité
Fraternité

EMBL-EBI

Internship report
4th March-30th August

Master Bioinformatique

Lees Group
Supervisors: Dr. John Lees, Dr.Samuel Horsfield

Lale MAOULOUD
Normandie Univ., Univ. Rouen
Normandie, MSc in Bioinformatics

Development and application of a metagenome annotation workflow with the graph-based annotation tool, ggCaller

This internship is hosted at the EMBL-EBI, the home of big data in biology. It is facilitated through a collaboration between the EBI and the French Embassy in London, which offers paid internships to students enrolled in master's programs in France. Driven by my curiosity for the invisible, I am a first-year master's degree student in Bioinformatics, Modeling, and Statistics. I believe that this program will provide a valuable and enriching experience. With a keen interest in bacterial genomics and software, I am grateful for the opportunity to spend six months with Dr. John Lees's Pathogen Informatics and Modeling group at the EMBL-EBI.

The Lees group, also known as the Pathogen Informatics and Modeling team, is renowned for its innovative contributions to microbial genomics and bioinformatics. The team has continually advanced the field through the development of cutting-edge software tools like CELEBRIMBOR (Core ELEment Bias Removal In Metagenome Binned ORthologs) and ggCaller (Graph-Gene-Caller). ggCaller predict genes collectively within a network-like structure constructed from bacterial pangenomes, known as a ‘graph,’ to identify putative gene sequences called open reading frames (ORFs). When constructing pangenomes from metagenomic data, researchers aim to obtain a comprehensive overview of all genes present in a microbial community, encompassing both core and accessory genes. The primary challenge in metagenomic analysis lies in handling incomplete and fragmented genomic data, which complicates gene annotation and pangenome construction. Traditional methods often fall short in providing accurate and reliable results due to the variability and complexity of metagenome-assembled genomes (MAGs). This problematic landscape necessitated the development of robust pipelines capable of efficiently imputing missing genes and improving the accuracy of gene predictions.

The objective of my internship was to develop a workflow that not only handles incomplete genomic data with high accuracy but also improves the overall quality of MAG annotations. The specific aims included simulating MAGs with varying completeness and exploring different imputation methods to identify the most effective strategies for metagenomic data analysis. I first adapted, improved, and secured the pipeline scripts to make them easier to use, especially on a large scale, by reorganizing the inputs and outputs. The comprehensive evaluation demonstrated that no single imputation method outperformed the others. This workflow, which I named MAGGImpute (Metagenome-Assembled Genomes Gene-Graph Caller Imputation-based Workflow), is efficient for handling incomplete genomic data with high reliability. If pursued as a PhD project, the focus could extend to integrating new methods with real-time data analysis, exploring co-assembly techniques, and developing novel algorithms tailored to the unique challenges of accessory genes. The work I have done for my internship represents a next step in the evolution of metagenomic analysis tools, building on the foundation laid by tools like CELEBRIMBOR. While CELEBRIMBOR has contributed significantly to the field, MAGGImpute aims to advance this further by addressing specific challenges in gene imputation and prediction within MAGs. The insights gained from this comprehensive evaluation pave the way for future improvements, shaping ggCaller, CELEBRIMBOR, and MAGGImpute into pioneering tools in metagenomic annotation methodologies, ensuring robust analyses in microbial genomics. In terms of future work over the next one and a half months, I have started a new project. I aim to predict the expression of short ORFs using ggCaller, and I plan to utilize a machine learning model such as Random Forest or Neural Network.

I extend my heartfelt appreciation to the Embassy of France in London for their generous financial support, which has been instrumental in enabling me to undertake this exceptional internship experience in Cambridge. This opportunity has been a pivotal moment in my academic journey, and I am truly grateful for the embassy's unwavering commitment to fostering academic excellence. I am hopeful that this collaborative partnership will continue to empower and inspire many other aspiring researchers in their pursuit of knowledge. Your support is not just a financial contribution, but a significant investment in the future of research and innovation.

I am profoundly indebted to my supervisors, John Lees and Samuel Horsfield, for providing me with an enriching and supportive environment throughout my internship. Their mentorship and expertise have not only broadened my understanding across various domains but have also instilled in me the importance of patience and perseverance as inherent parts of the research journey.I am especially grateful to Samuel Horsfield for his unwavering attention, invaluable guidance, and for always being there to assist me with any challenges I encountered. Thank you once again to all who have made this experience possible for believing in and supporting the next generation of scholars.. Your generosity and support have left an indelible mark on my academic journey, and I am committed to carrying forward the knowledge and skills I have gained during this exceptional experience.