People's Democratic Republic of Algeria

Ministry of Higher Education and Scientific Research

University of JIJEL

Practical Implementation of FOSS in Bioinformatics: A Focus on Biopython and GitHub

The students:

-Rekrouk Samah

-Hamidouche Chiraz

-Rekrouk Djihane

The teacher:

-Dr Bensalem Amel

2026-2027

Introduction:

The rapid advancement of high-throughput technologies in biological sciences has led to an explosion of genomic and proteomic data. To manage and analyze this information effectively, the field of bioinformatics relies heavily on Free and Open-Source Software (FOSS). These tools are essential not only for reducing the financial barriers to research but also for promoting the principles of 'Open Science,' which emphasize transparency, collaboration, and the global sharing of knowledge.

This report provides a comprehensive study on the practical application of FOSS in bioinformatics. It begins with a theoretical analysis of **Biopython**, one of the most significant Python libraries designed to simplify biological computation by providing tools for sequence manipulation and database access. Furthermore, the report details a practical implementation using **GitHub**, the leading platform for version control and collaborative development. By integrating these tools with archiving services like **Zenodo**, we demonstrate a modern workflow that ensures scientific data is not only accessible but also FAIR (Findable, Accessible, Interoperable, and Reusable). Through this work, we aim to highlight how open-source ecosystems empower researchers to conduct reproducible and high-impact scientific studies.

Part 1 - Theoretical Study of the Tool: Biopython

►Introduction

In the field of modern bioinformatics, the processing and analysis of biological data require efficient and flexible computational tools. Several libraries have been developed to facilitate these tasks, among which **Biopython** occupies an important position. This study aims to present **Biopython**, its main functionalities, technical aspects, as well as its strengths and limitations.

1. General Presentation of the Tool

Biopython is a long-standing, **mature open-source project** developed by an international collaboration of volunteer developers. Founded in **1999**, it was created to provide a comprehensive set of libraries for the **Python** programming language to address diverse problems in **bioinformatics and computational molecular biology**. The project is hosted and supported by the **Open Bioinformatics Foundation (OBF)**, which also supports related efforts like BioPerl, BioJava, and BioRuby. It is freely available under the Biopython license and is compatible with all major operating systems. **(Cock et al., 2009)**

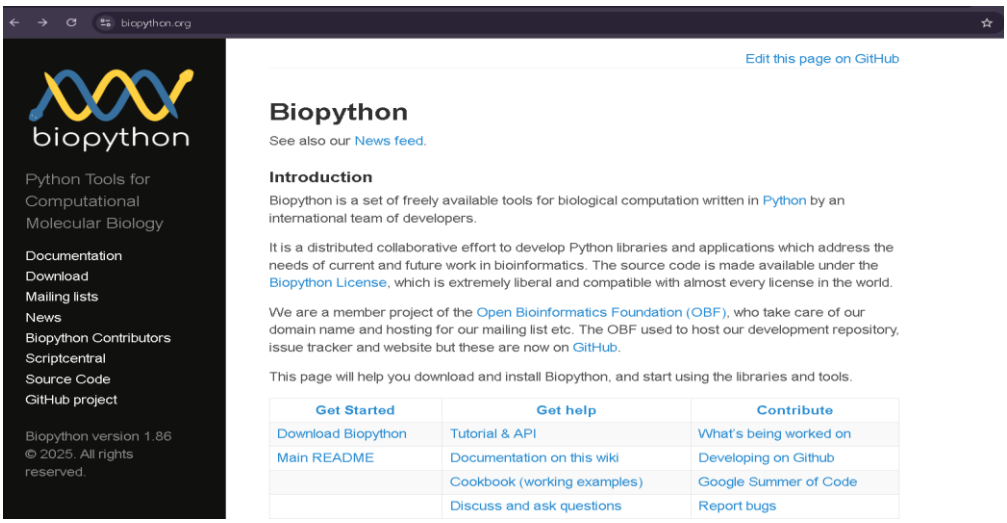


Figure : Biopython platform.

2. Main Functionalities

Biopython acts as a massive collection of modules that researchers can use in their own scripts or software. Its core functionalities include:

• **Sequence Manipulation:** The **Seq object** is the central representation of biological sequences, supporting biological methods like **transcription and translation**.

• **File Format Support:** Through the Bio.SeqIO and Bio.AlignIO modules, the tool can **read and write** various biological formats, including **FASTA, GenBank, EMBL, Swiss-Prot, and Clustal W**. For example, it is widely used to parse **FASTA** files in sequencing analysis pipelines.

• **Database Interaction:** It provides interfaces to access key online databases such as **NCBI Entrez, ExPASy, KEGG, and SCOP**.

• **Tool Integration:** Biopython includes **wrappers** for widely used command-line tools like **BLAST, ClustalW, and EMBOSS**, allowing users to run these tools directly from Python scripts.

• **Structural Biology and Motifs:** The Bio.PDB module parses macromolecular structures, while Bio.Motif supports sequence motif analysis, including searching and **de novo learning**.

• **Advanced Analytics:** It offers modules for **supervised and unsupervised statistical learning** (e.g., Bayesian methods, Markov models, and clustering) and tools for **population genetics**.

• **Visualization:** The inclusion of **GenomeDiagram** allows for the graphical representation of large-scale genomic data. (**Cock et al., 2009)**

3. Technical Aspects

From a technical perspective, Biopython is based on the Python programming language, which is well known for its easy-to-learn syntax and object-oriented programming capabilities.

• **Core Objects:** Unlike simple strings, the **Seq** object incorporates biological metadata (e.g., alphabets) and supports native biological operations like transcription. This is further extended by the **SeqRecord** and **SeqFeature** classes, which allow for the management of annotated sequences.

• **Performance:** Although Python is a high-level language, Biopython can interface with optimized code written in **C, C++, or FORTRAN**. It also integrates with the **NumPy** project for intensive **numerical computations**.

• **Data persistence:** Through **BioSQL**, **Biopython** allows storage and retrieval of annotated sequences in **SQL databases** using a standardized schema shared with other "Bio*" projects. **(Cock et al., 2009)**

4. Strengths

The strengths of Biopython mainly lie in its robustness and ecosystem:

• **Extensive Functional Coverage:** It covers almost all areas of bioinformatics, from basic file manipulation to complex population simulations.

• **Interoperability:** By following format naming standards used by **BioPerl and EMBOSS**, it facilitates transitions between different tools and platforms.

• **Documentation and community**: The project benefits from a comprehensive tutorial, detailed API documentation, and active mailing lists for user support.

• **Workflow Integration:** It acts as a central hub, bridging the gap between online biological databases (like NCBI) and local command-line tools.

• **Open-access and Cross-platform Availability:** Being free of charge and available on all operating systems, it reduces entry barriers for researchers and students, supported by comprehensive documentation and an active global community. **(Cock et al., 2009)**

5. Limitations and Weaknesses

Although sources highlight the success of **Biopython**, several limitations can be identified:

• **Programming dependency:** **Biopython** is an application programming interface (API) designed for programmers. It therefore requires Python scripting skills and does not provide a ready-to-use graphical user interface for non-programmers.

• **High computational demands**: For extremely demanding tasks such as molecular dynamics, the use of third-party libraries (such as **NumPy**) or interfaces with lower-level languages (**C/FORTRAN**) is necessary to compensate for Python's speed limitations.

• **Extensive Learning Curve:** Due to the vast scope of the library, the functionalities described in primary literature represent only a fraction of its total capabilities. Consequently, users must invest substantial time in exploring the **API documentation** to fully utilize the tool's potential. **(Cock et al., 2009)**

6. Conclusion:

Biopython has established itself as an essential tool for **the development of bioinformatics software** and **the creation of daily scripts for routine tasks**. Its ability to integrate external tools, query global databases, and manipulate complex biological structures makes it a cornerstone of modern computational biology. As **an open-source project** supported by an international community, it continues to evolve to address new challenges in biological data analysis and the growing needs of scientific research.

Part 2 – Practical Study: Exploration of Zenodo.

1. Presentation of Zenodo.

►General Overview

Zenodo is a global open-access repository designed to enable researchers, scientists, and scholars to share and preserve their research outputs. It was launched in **2013** through a collaboration between **CERN** (the European Organization for Nuclear Research) and the **OpenAIRE** project, funded by the European Commission **(Zenodo, n.d.)**. Unlike many repositories, **Zenodo** is "catch-all," meaning it is open to all fields of science and all types of digital artifacts.

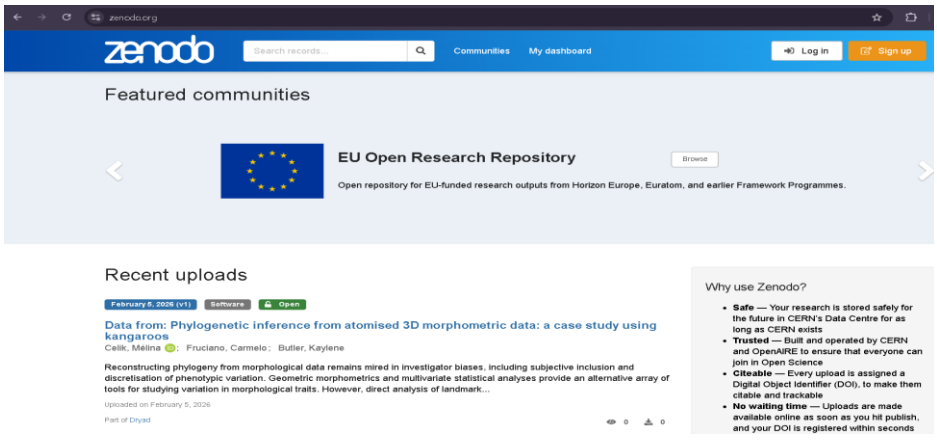


Figure : zenodo platform.

►Platform Objectives

The primary mission of **Zenodo** is to support the transition toward **Open Science** by providing:

- **Persistent Identification:** Assigning a Digital Object Identifier (DOI) to every upload to make the work citable and trackable **(CERN, 2023)**.
- **Accessibility:** Ensuring research data is free to access for anyone, anywhere, immediately upon publication.
- **Sustainability:** Leveraging CERN's high-performance IT infrastructure to guarantee the long-term storage of data for at least 20 years **(European Commission, 2021)**.

►Types of Hosted Content

Zenodo accepts a wide variety of research-related digital objects, including:

- **Datasets:** Raw or processed data from experiments.
- **Publications:** Peer-reviewed articles, preprints, and conference papers.
- **Software:** Code and scripts (often integrated with GitHub).
- **Multimedia:** Images, videos, and posters related to scientific dissemination.

►Importance for Open Science and NLS (Natural Life Sciences)

In the field of **Natural Life Sciences (NLS)**, **Zenodo** plays a critical role in the "FAIR" data principles (Findable, Accessible, Interoperable, and Reusable). For instance, sharing genomic sequences or cellular imaging data on **Zenodo** allows for:

1. **Transparency:** Allowing other researchers to verify findings.

2. **Collaboration:** Enabling scientists to build upon existing datasets without repeating expensive experiments **(OpenAIRE, 2022)**.

2. Description of the steps carried out.

1. Visit the **Zenodo** platform (https://zenodo.org/)

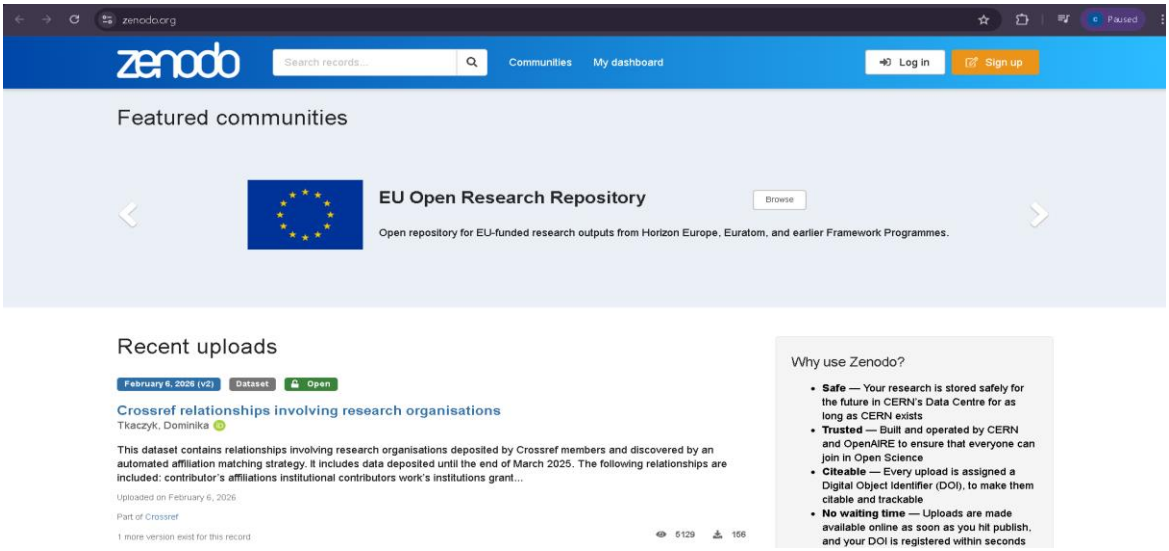


Figure 1 : Zenodo Platform.

2.Search for a dataset using a query contain keyword: **genome**

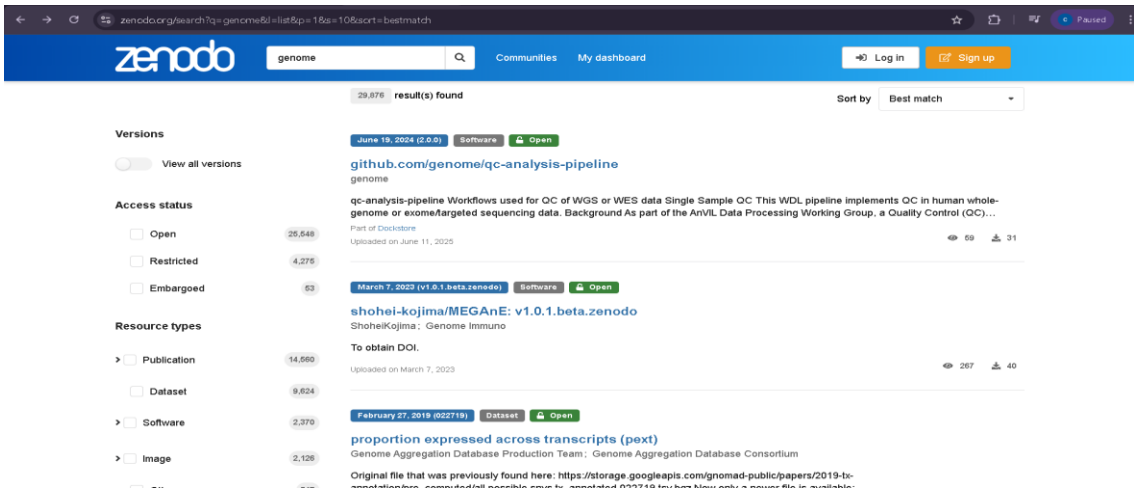


Figure 2: Search results for the query 'Genome' on **Zenodo**.

3. Select a relevant dataset.

Figure 3: Detailed view of the selected dataset. The record displays the title, authors, and the unique DOI assigned to the research on antibiotic resistance in *K. pneumoniae*.

4. Download the selected dataset.

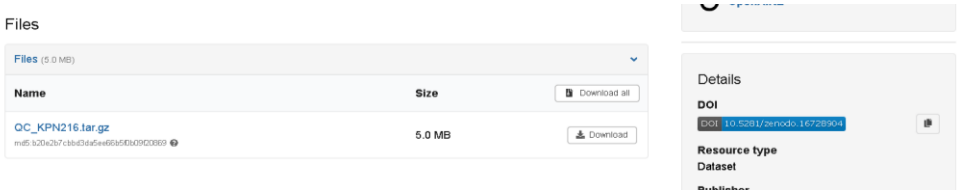
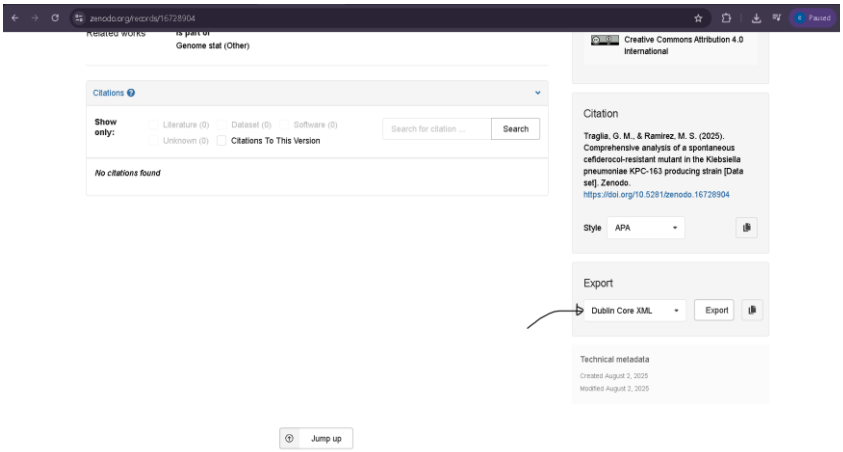


Figure 4: Downloading the research files for local analysis. The platform provides direct access to open-source data according to Open Science principles.

5. Retrieve the dataset metadata using the **"Dublin Core"** Standard.



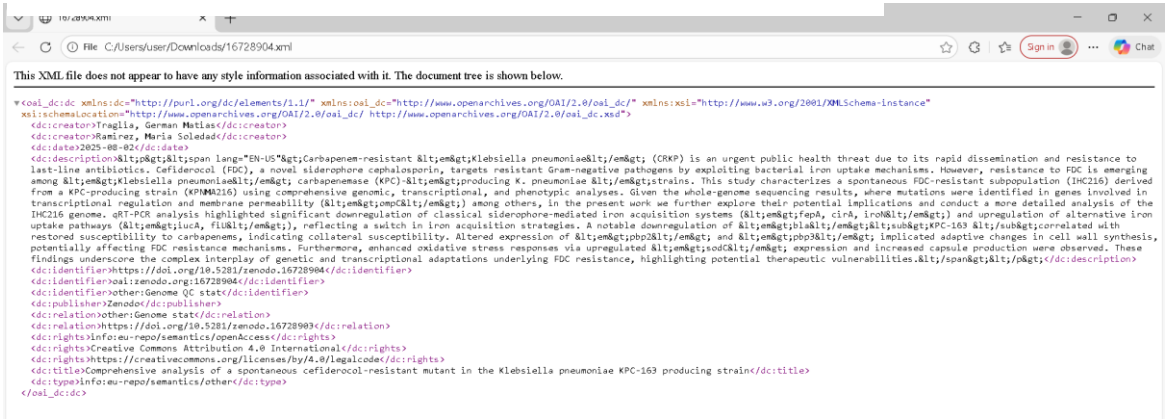


Figure 5: Exporting the dataset metadata in Dublin Core (XML format). This step demonstrates the interoperability of the platform and the standardized description of the research components.

6. Extract and present as much information as possible.

Dublin Core Era	Information
Title	● Comprehensive analysis of a spontaneous cefiderocol-resistant mutant in the Klebsiella pneumoniae KPC-163 producing strain
Creator	1. Traglia, German Matias 2. Ramirez, Maria Soledad
Date	● 2025-08-02
Langue	● EN-US
Description	● Carbapenem-resistant *Klebsiella pneumoniae* (CRKP) is an urgent public health threat due to its rapid dissemination and resistance to last-line antibiotics. Cefiderocol (FDC), a novel siderophore cephalosporin, targets resistant Gram-negative pathogens by exploiting bacterial iron uptake mechanisms. However, resistance to FDC is emerging among *Klebsiella pneumoniae* carbapenemase (KPC)-producing *K. pneumoniae* strains. This study characterizes a spontaneous FDC-resistant subpopulation (IHC216) derived from a KPC-producing strain (KPNMA216) using comprehensive genomic, transcriptional, and phenotypic analyses. Given the whole-genome sequencing results, where mutations were identified in genes involved in transcriptional regulation and membrane permeability (*ompC*) among others, in the present work we further explore their potential implications and conduct a more detailed analysis of the IHC216 genome. qRT-PCR analysis highlighted significant downregulation of classical siderophore-mediated iron acquisition systems (*fepA, cirA, iroN*) and upregulation of alternative iron uptake pathways (*iucA, fiu*), reflecting a switch in iron acquisition strategies. A notable downregulation of *bla*KPC-163 correlated with restored susceptibility to

	carbapenems, indicating collateral susceptibility. Altered expression of *pbp2* and *pbp3* implicated adaptive changes in cell wall synthesis, potentially affecting FDC resistance mechanisms. Furthermore, enhanced oxidative stress responses via upregulated *sodC* expression and increased capsule production were observed. These findings underscore the complex interplay of genetic and transcriptional adaptations underlying FDC resistance, highlighting potential therapeutic vulnerabilities.
Identifier	● https://doi.org/10.5281/zenodo.16728904 ● oai:zenodo.org:16728904 ● other:Genome QC stat
Publisher	● Zenodo
Relation	● other:Genome stat ● https://doi.org/10.5281/zenodo.16728903
Rights	● info:eu-repo/semantics/openAccess ● Creative Commons Attribution 4.0 International ● https://creativecommons.org/licenses/by/4.0/legalcode
Type	● info:eu-repo/semantics/other

Part 3 – Bonus: GitHub Repository

In this part, a GitHub account was created in order to share and store the final report of the practical study.

GitHub is a collaborative platform that allows researchers and students to manage versions of their projects and make their work publicly accessible.

First, a GitHub account was created on the official platform (https://github.com).

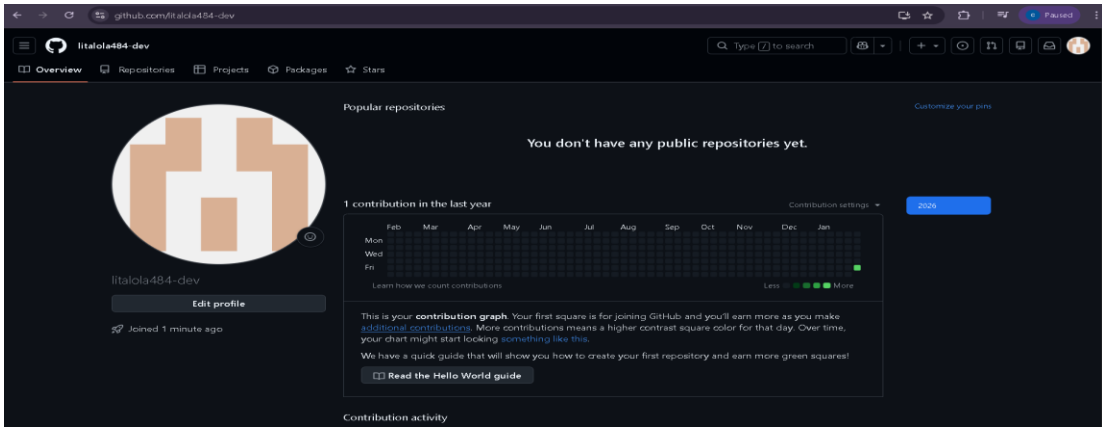


Figure 1: Creation of GitHub account.

Then, a new public repository entitled "**Practical-Work-Free-and-Open-Source-Software**" was created. This repository was dedicated to storing the final report of Part I and Part II.

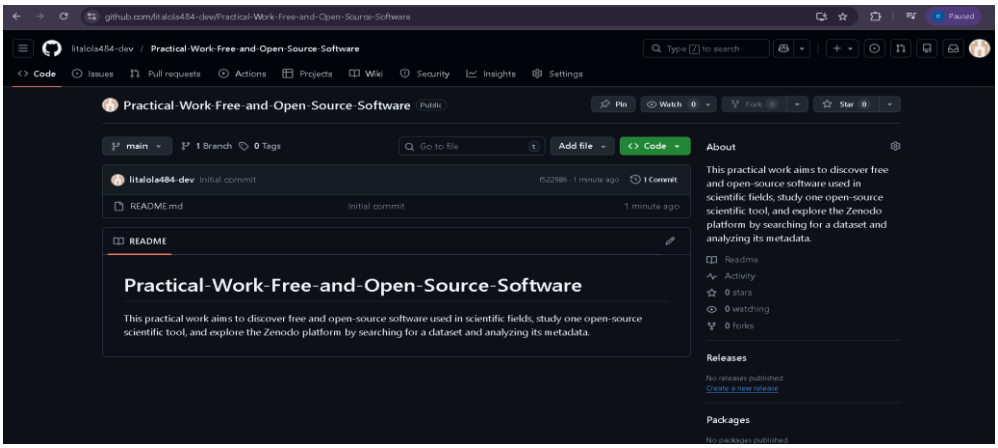


Figure 2: Creation of the repository.

Finally, the report was uploaded in PDF format to the created repository. This step allows easy access to the work, ensures transparency, and facilitates sharing with teachers and other users.

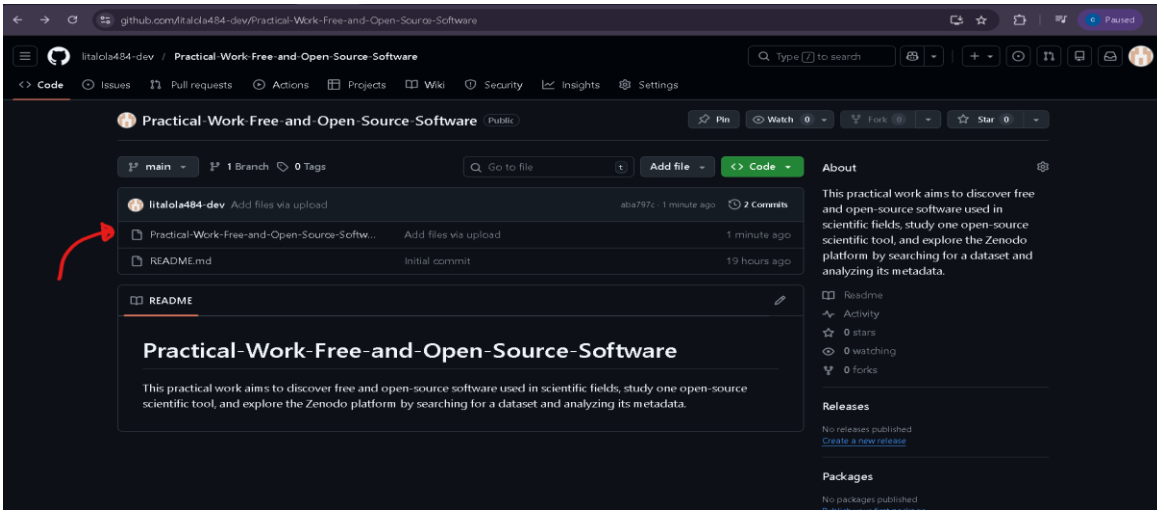


Figure 3: Successful upload of the final report (PDF) combining Part I and Part II to the GitHub repository.

●Final Submission Link

The complete project, including all documentation and the final report, is hosted and publicly accessible via the following GitHub repository link:

URL:

https://github.com/litalola484-dev/Practical-Work-Free-and-Open-Source-Software

This platform is widely used in scientific research for open science practices and reproducibility.

Conclusion:

In conclusion, this report has demonstrated that the integration of Free and Open-Source Software (FOSS) is not merely an alternative, but a cornerstone of modern bioinformatics. Through our exploration of **Biopython**, we have seen how standardized libraries can streamline complex biological computations, from sequence analysis to database interaction, with high efficiency and precision. Furthermore, the practical application of **GitHub** and **Zenodo** illustrates the vital importance of the 'Open Science' movement. By adopting version control and open-access archiving, researchers can ensure their work is transparent, traceable, and ready for global collaboration. Ultimately, mastering these tools is an essential skill for any bioinformatician aiming to contribute to a more open, accessible, and reproducible scientific future. This practical work serves as a foundational step in our journey to utilize digital innovation for biological discovery.

References :

● **CERN. 2023**. ***Zenodo*** – *Research data repository*. European Organization for Nuclear Research.

● **Cock**, P.J.A., **Antao**, T., **Chang**, J.T., **Chapman**, B.A., **Dalke**, A., **Friedberg**, I., **Hamelryck**, T., **Kauff**, F., **Wilczynski**, B., and **de Hoon**, M.J.L. **2009**. ***Biopython****: freely available Python tools for computational molecular biology and bioinformatics*. Bioinformatics, 25(11): 1422–1423.

● **European Commission. 2021.** *Open science and data sustainability*. European Union.

● **OpenAIRE. 2022**. *FAIR data principles and open science*. OpenAIRE Initiative.

● **Traglia**, G.M. and **Ramirez**, M.S. 2025. *Comprehensive analysis of a spontaneous cefiderocol-resistant mutant in the Klebsiella pneumoniae KPC-163 producing strain*. **Zenodo**. DOI: 10.5281/zenodo.16728904.

● **Zenodo. n.d**. *Zenodo: Research. Shared*. European Organization for Nuclear Research (**CERN**).