الجمهوريـــــــة الجزائريـــة الديمقراطيـــة الشعبيـــة
REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
وزارة التعليـم العالـــي والبحـث العلمـي
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

Université Mohammed Seddik Benyahia –Jijel

Faculté : des sciences de la nature et de la vie

Spécialité : biologie meleculaire et cellulaire

Présenté Par :

Anfal Yahiaoui

TP

Logiciels Libres et Open Source

TABLE DES MATIÈRES

PARTIE I : Étude théorique de Biopython

1. Présentation générale

Biopython est un projet international open-source qui fournit un ensemble de bibliothèques Python dédiées à la biologie moléculaire computationnelle et à la bio-informatique. Il a été développé de manière collaborative par des chercheurs et programmeurs volontaires à travers le monde. Contrairement à un logiciel unique, Biopython propose des modules réutilisables permettant d'intégrer les analyses biologiques directement dans l'environnement Python. [1] [2]

L'objectif principal de Biopython est de faciliter la réalisation des tâches courantes en bio-informatique en évitant aux chercheurs de tout programmer à partir de zéro. Il permet notamment l'analyse de séquences biologiques, la gestion de formats de fichiers standards, l'accès à des bases de données biologiques et l'intégration d'outils bio-informatiques largement utilisés. [1] [2]

Le projet a été fondé en 1999, inspiré du projet BioPerl, et fait partie de la famille des outils Bio soutenus par l'Open Bioinformatics Foundation (OBF), aux côtés de BioPerl, BioJava et BioRuby.

2. Fonctionnalités principales

Biopython offre un large éventail de fonctionnalités couvrant plusieurs domaines de la bio-informatique :

- Gestion des fichiers biologiques (SeqIO) :Lecture, écriture et conversion de nombreux formats de séquences tels que FASTA, GenBank et les formats d'alignement multiple.
- Analyse de séquences : Manipulation, traduction et analyse de séquences d'ADN, d'ARN et de protéines, ainsi que la gestion des alignements multiples.
- Accès aux bases de données en ligne : Accès programmatique aux bases de données biologiques majeures telles que NCBI (GenBank, PubMed), KEGG et ExPASy via les services Entrez.
- Bio-informatique structurale (Bio.PDB) : Analyse et manipulation de structures 3D de macromolécules à partir de fichiers PDB, incluant la gestion de modèles multiples et des résidus désordonnés.
- Intégration d'outils externes : Interfaces pour l'exécution et l'analyse des résultats d'outils comme BLAST, ClustalW et EMBOSS directement dans des scripts Python. [3] [4]

3. Aspects techniques

Biopython est entièrement implémenté en Python, ce qui lui permet de bénéficier de la lisibilité, de la portabilité et de l'intégration avec les bibliothèques scientifiques telles que NumPy et SciPy.

L'installation est simple et se fait généralement via pip, ce qui facilite son intégration dans différents environnements de recherche.

Le projet est distribué sous une licence libre de type BSD, garantissant une utilisation gratuite aussi bien pour la recherche académique que pour des applications commerciales. [3] [4]

4. Points forts

- Communauté et documentation : Projet mature avec une large communauté internationale, une documentation complète et de nombreux exemples d'utilisation.
- Open source et gratuit : Disponible gratuitement sur tous les systèmes d'exploitation majeurs, sans restrictions d'usage.
- Modularité : Organisation en modules spécialisés (SeqIO, PDB, Blast, Phylo…), permettant une utilisation ciblée et une forte réutilisabilité du code.
- Large adoption : Utilisé largement dans l'enseignement et les pipelines de recherche en bio-informatique. [4] [5]

5. Limites et points faibles

- Performance : Étant principalement écrit en Python, Biopython peut être moins performant que des bibliothèques optimisées en C/C++ pour l'analyse de très grands jeux de données.
- Courbe d'apprentissage : Son utilisation efficace nécessite une maîtrise de base du langage Python et des notions de programmation. [5] [2]

6. Conclusion

Biopython constitue un outil fondamental en bio-informatique moderne. Il permet d'automatiser des tâches répétitives telles que la gestion des séquences, les requêtes vers les bases de données et les analyses standards, améliorant ainsi la reproductibilité et la clarté des travaux scientifiques.

Grâce à son intégration dans des workflows Python et des environnements comme Jupyter Notebook, Biopython aide les chercheurs à se concentrer davantage sur l'interprétation des résultats et la rédaction de rapports scientifiques structurés.

PARTIE II

Présentation de Zenodo

A. Définition

Zenodo est un dépôt généraliste en libre accès hébergé au CERN qui préserve et diffuse des résultats de recherche variés issus de toutes les disciplines, y compris les sciences naturelles et les sciences de la vie. [6] [7]

B. Objectifs

Zenodo a été créé par le CERN et OpenAIRE (2013) pour soutenir la « longue traîne de la recherche » qui ne dispose pas de dépôts institutionnels ou de domaines spécifiques. Il vise à :

- Fournir gratuitement des services de stockage et de conservation à long terme pour les résultats de la recherche.
- Rendre les résultats FAIR (Faciles à trouver, Accessibles, Interopérables, Réutilisables) grâce aux DOI, aux métadonnées riches et aux options d'accès ouvert.
- Soutenez la science ouverte en permettant le partage des résultats positifs/négatifs, de tout format de fichier et des données non textuelles. [8]

C. Types de contenu hébergé

Zenodo accepte pratiquement tous les contenus liés à la recherche :

- Ensembles de données et logiciels (croissance rapide ; contenu majoritairement non textuel).
- Publications (articles de revues, actes de conférences, livres/chapitres, rapports, thèses).
- Objets multimédias et spécifiques à un domaine : images d'herbier, données 3D du patrimoine culturel, traitements taxonomiques, figures, etc…
- Il attribue des DOI à tous les téléchargements publics, améliorant la citabilité et le suivi via la bibliométrie et les altmetrics.

D. Intérêt de Zenodo

Zenodo fait progresser directement la science ouverte en :

- Garantir l'accès libre et les DOI, qui sont corrélés à une visibilité et à des altmetrics plus élevés pour certains types de documents 58 à la curation communautaire via les « communautés » et les flux de travail de domaine (par exemple, biodiversité, patrimoine culturel).
- Activation de l'ingestion automatisée, via API, à partir des pipelines de numérisation et des plateformes de domaine.

En sciences naturelles et en sciences de la vie, Zenodo sous-tend :

- Flux de travail sur la biodiversité (images d'herbier, littérature taxonomique, traitements, spécimens) qui deviennent conformes aux principes FAIR et sont liés à GBIF/ OpenBiodiv.

- De vastes ensembles de données partagés, tels que la collection de produits naturels COCONUT (> 400 000 composés).
- Données ouvertes de caractérisation des anticorps (YCharOS) et autres ressources biomédicales. [9]

E. Conclusion

Zenodo est un dépôt multidisciplinaire conforme aux principes FAIR dont les objectifs sont de préserver, partager et rendre citables un large éventail de résultats de recherche. Sa flexibilité, ses DOI et son intégration aux flux de travail du domaine en font un outil particulièrement important pour les pratiques de science ouverte et la recherche à forte intensité de données en sciences naturelles et en sciences de la vie.

Description des étapes réalisées

Formulation de la requête et exploration

La première étape a consisté à interroger la base de données de la plateforme **Zenodo** pour identifier des ressources pertinentes. La requête utilisée était **"tissue cancer tissue"**. Cette recherche ciblée visait à isoler des jeux de données (datasets) combinant à la fois des structures histologiques (tissus) et des pathologies oncologiques (cancer), afin de garantir une base de données riche pour une analyse biologique.

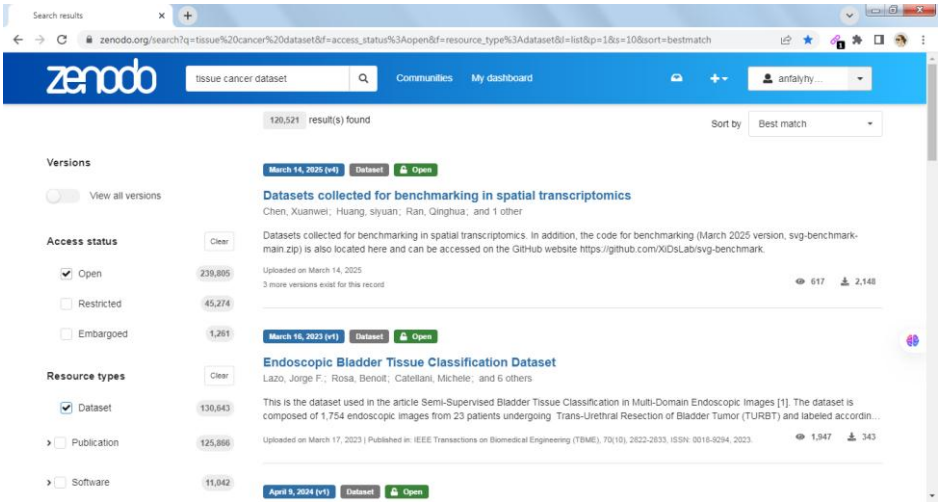

Figure 1 : Interface de recherche de Zenodo avec la requête "tissue

Critères de sélection du dataset

Parmi les résultats affichés, mon choix s'est porté sur le dataset intitulé *"Colorectal Cancer Histology Image Tiles for Tissue Multi-class Classification"*. Les critères de sélection retenus étaient :

- **La pertinence thématique :** L'adéquation parfaite avec les mots-clés de la recherche.
- **La fiabilité :** La présence d'auteurs affiliés à des institutions de recherche reconnues.
- **La qualité des données :** Un volume important d'images (plus de 15 000) permettant une classification multi-classe précise.

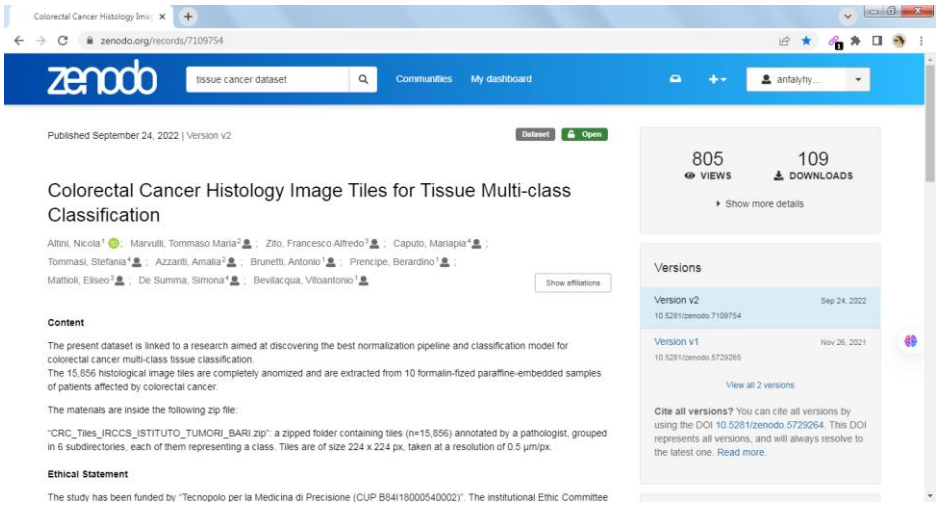

***Figure 2 : Page d'accueil du dataset sélectionné illustrant le titre et les* auteurs.**

Navigation et interaction sur la plateforme

Après avoir sélectionné la ressource, la navigation s'est poursuivie sur la page dédiée du dataset. J'ai examiné la description technique, les auteurs et la date de publication. Conformément aux objectifs pédagogiques, j'ai procédé au téléchargement de l'archive ZIP pour valider la disponibilité des fichiers et la structure de l'organisation des données.

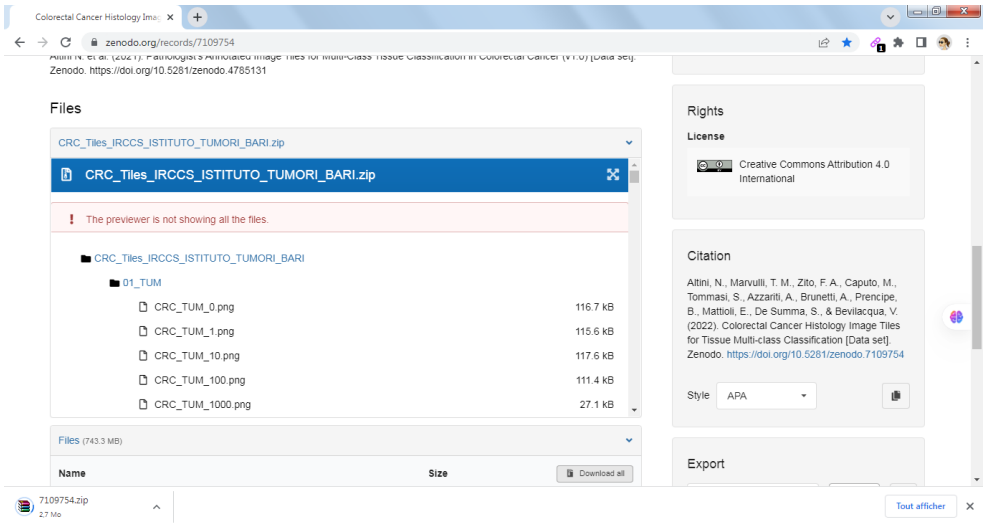

Figure 3 : Accès aux fichiers de données et initiation du téléchargement

Récupération des métadonnées (Dublin Core)

La phase finale a consisté à extraire l'identité numérique du dataset. Pour ce faire, j'ai utilisé l'outil d'exportation intégré à Zenodo, en choisissant la norme **Dublin Core**. Cette étape a permis de récupérer de manière structurée des informations essentielles comme le **DOI**, la licence d'utilisation (Creative Commons) et les identifiants uniques des chercheurs, garantissant ainsi la traçabilité et l'éthique de la science ouverte.

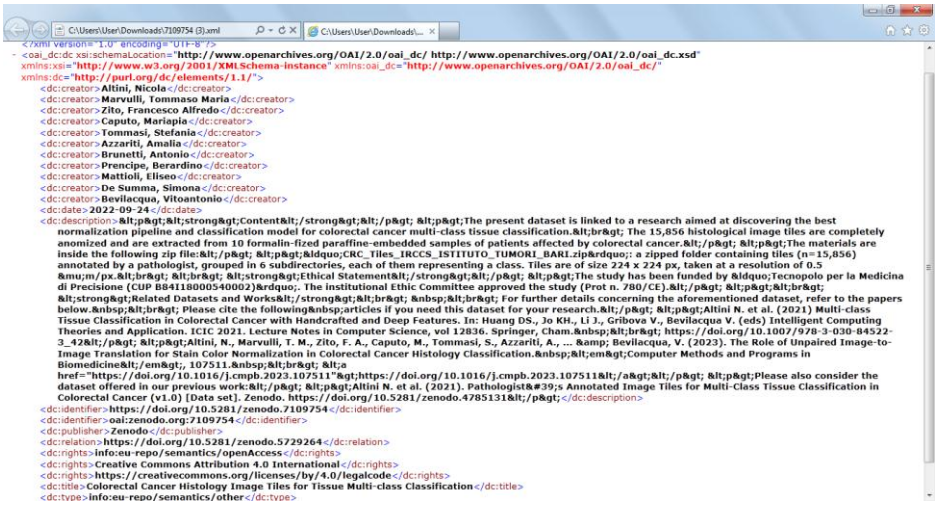

Figure 4 : Extraction des métadonnées structurées selon la norme Dublin Core

Métadonnées du dataset

Élément Dublin Core	**Données extraites**
Title	Colorectal Cancer Histology Image Tiles for Tissue Multi-class Classification
Creator	Altini, Nicola ; Marvulli, Tommaso Maria ; Zito, Francesco Alfredo ; Caputo, Mariapia ; Tommasi, Stefania ; Azzariti, Amalia ; Brunetti, Antonio ; Prencipe, Berardino ; Mattioli, Eliseo ; De Summa, Simona ; Bevilacqua, Vitoantonio
Date	2022-09-23
DOI	10.5281/zenodo.7109754
Publisher	Zenodo
Type	Dataset
Language	Eng
Rights	Creative Commons Attribution 4.0 International (CC BY 4.0)
Description	Le dataset contient 15 856 tuiles d'images histologiques (224 x 224 px) extraites de 10 échantillons de patients atteints de cancer colorectal, annotées par un pathologiste et classées en 6 catégories.
Subject	Colorectal Cancer; Tissue Classification; Histology; Deep Learning; Multi-class classification

Références Bibliographiques :

- [1] Chapman, B. A., & Chang, J. (2000). Biopython: Python tools for computational biology. ACM SIGBIO Newsletter, 20(2), 15–19.
- [2] Cock, P. J. A., Antão, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., … de Hoon, M. J. L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics, 25(11), 1422–1423.
- [3] Hamelryck, T., & Manderick, B. (2003). PDB file parser and structure class implemented in Python. Bioinformatics, 19(17), 2308–2310.
- [4] Talevich, E., Invergo, B. M., Cock, P. J. A., & Chapman, B. A. (2012). Bio.Phylo: A unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython. BMC Bioinformatics, 13, 209.
- [5] Yeoh, J. W. Y., Swainston, N., Vegh, P., Mulholland, A. J., & Turner, N. J. (2021). SynBiopython: An open-source software library for synthetic biology. ACS Synthetic Biology, 10(1), 40–46.
- [6] Nielsen, L. H., Gonzalez Lopez, J. B., Smith, T., et al. (2022). Zenodo data repository: Providing practical solutions for data storage and data sharing. Regulatory Affairs Watch.
- [7] Nowak, K., & Nielsen, L. H. (2016).Zenodo, a free and open platform for preserving and sharing research output.
- [8] Peters, I., Kraker, P., Lex, E., et al. (2017). Zenodo in the spotlight of traditional and new metrics.
- [9] Ioannidis-Pantopikos, A., & Agosti, D. (2021).Biodiversity Literature Repository: Building the customized FAIR repository by using custom metadata.