

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Mohamed Seddik Benyahia – Jijel

Faculté des Sciences de la Nature et de la Vie

Département de Biologie Moléculaire et Cellulaire



**Master 01 LMD Biologie Moléculaire
Module : Logiciels Libres et Open Source**

**Utilisation des Logiciels Libres et de l'Open
Science : Cas de Biopython et Zenodo**

Réalisé par :

**BELARBI Zakaria
KHAYAR Mohamed**

proposé par :

Dr: BENSALEM. A

-Année universitaire 2025/2026

PARTIE 1 – Étude théorique d'un outil : Biopython

1. Présentation générale de l'outil

Biopython est une bibliothèque open source écrite en langage Python, dédiée à la bioinformatique et aux sciences de la vie. Elle fournit un ensemble d'outils permettant la manipulation, l'analyse et le traitement de données biologiques telles que les séquences ADN/ARN, les protéines, les structures moléculaires et les données issues de bases de données biologiques [1].

Biopython est largement utilisée dans la recherche académique et industrielle grâce à sa simplicité, sa flexibilité et son intégration facile avec d'autres bibliothèques scientifiques Python [2].

2. Fonctionnalités principales

Lecture et écriture de fichiers biologiques (FASTA, GenBank, EMBL, PDB, etc.)

Manipulation et analyse de séquences (traduction, transcription, alignement)

Accès aux bases de données biologiques en ligne (NCBI, ExPASy)

Analyse phylogénétique

Manipulation de structures protéiques

Outils pour l'annotation et la comparaison de séquences [3] [4].

3. Aspects techniques

Sur le plan technique, Biopython s'appuie sur une architecture modulaire qui facilite son intégration dans divers environnements de développement.

Langage : Python

Licence : Biopython License (open source)

Compatibilité : Windows, Linux, macOS

Dépendances : Python $\geq 3.x$

Intégration possible avec NumPy, SciPy, Pandas et Matplotlib [5] [6].

4. Points forts

L'adoption massive de Biopython par la communauté scientifique s'explique par plusieurs atouts majeurs : Gratuit et open source

Facilité d'apprentissage

Documentation riche et communauté active

Forte intégration avec l'écosystème Python

Adapté à l'automatisation et aux analyses à grande échelle [7] [8].

5. Limites et points faibles

Bien que Biopython soit un outil puissant et incontournable en bioinformatique, il présente certaines limitations :

Courbe d'apprentissage : Pour les biologistes n'ayant aucune base en programmation, l'utilisation de Biopython peut être difficile car elle nécessite une maîtrise préalable du langage Python.

Performance pour les grands jeux de données : Étant écrit principalement en Python (un langage interprété), Biopython peut être plus lent que des outils écrits en C ou C++ lors du traitement de fichiers génomiques extrêmement volumineux (Big Data génomique).

Documentation inégale : Bien qu'une documentation officielle existe, certaines parties de la bibliothèque (les modules les moins utilisés) sont moins bien documentées ou contiennent des exemples obsolètes.

Dépendance aux bibliothèques tierces : Pour certaines fonctionnalités avancées, comme l'alignement multiple complexe, Biopython dépend de l'installation externe d'autres logiciels (comme ClustalW BLAST), ce qui peut compliquer l'installation pour l'utilisateur.

Visualisation limitée : Biopython n'est pas une bibliothèque dédiée à la visualisation graphique avancée ; il est souvent nécessaire de la coupler avec Matplotlib ou Seaborn pour générer des graphiques complexes. [9] [10].

6. Conclusion

Biopython constitue un outil puissant et flexible pour l'analyse bioinformatique. Son caractère open source et sa compatibilité avec Python en font une solution idéale pour la recherche et l'enseignement en sciences de la vie.

Références

- [1] Cock, P. J. A. et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11), 1422–1423.
- [2] Documentation officielle Biopython. <https://biopython.org/wiki/Documentation>
- [3] . Biopython Tutorial and Cookbook. <https://biopython.org/wiki/Documentation>
- [4] Chapman, B., Chang, J. (2000). Biopython: Python tools for computational biology. *ACM SIGBIO Newsletter*
- [5] Documentation officielle Biopython – Installation et dépendances.
- [6] Python Software Foundation. <https://www.python.org>
- [7] Cock, P. J. A. et al. (2009). *Bioinformatics*.
- [8] Articles de revue sur l'utilisation de Biopython en bioinformatique.
- [9] Retours utilisateurs – Documentation et forums Biopython.
- [10] Comparaisons d'outils bioinformatiques open source.

PARTIE 2 – Étude pratique : Exploration de Zenodo

1. Présentation de Zenodo

Zenodo est une plateforme de dépôt de données scientifiques développée par le CERN et soutenue par la Commission européenne. Elle permet aux chercheurs de partager, archiver et citer leurs données de recherche.

Objectifs de la plateforme : Zenodo est un entrepôt de données de la recherche en accès ouvert (Open Access) développé par le CERN. Il vise à permettre aux chercheurs de partager, conserver et citer leurs données.

Types de contenus : La plateforme héberge divers types de documents tels que des jeux de données (datasets), des logiciels, des publications et des présentations.

Intérêt pour les SVN : Elle favorise la science ouverte en facilitant la réutilisation des données biologiques et génomiques, garantissant ainsi la transparence de la recherche[1] [2].

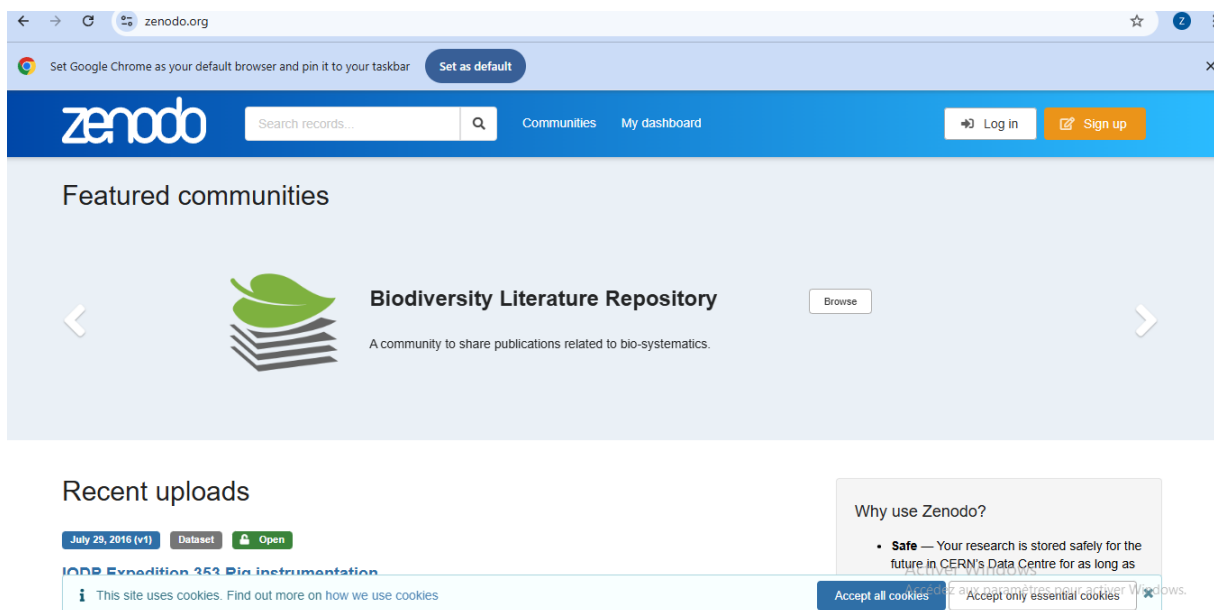


Figure 1 : Page d'accueil de la plateforme Zenodo

Cette capture d'écran montre l'interface principale du site Zenodo (<https://zenodo.org>). C'est à partir de cette page que nous avons initié notre recherche en utilisant la barre de recherche centrale pour trouver des données spécifiques au domaine de la biologie cellulaire et génomique

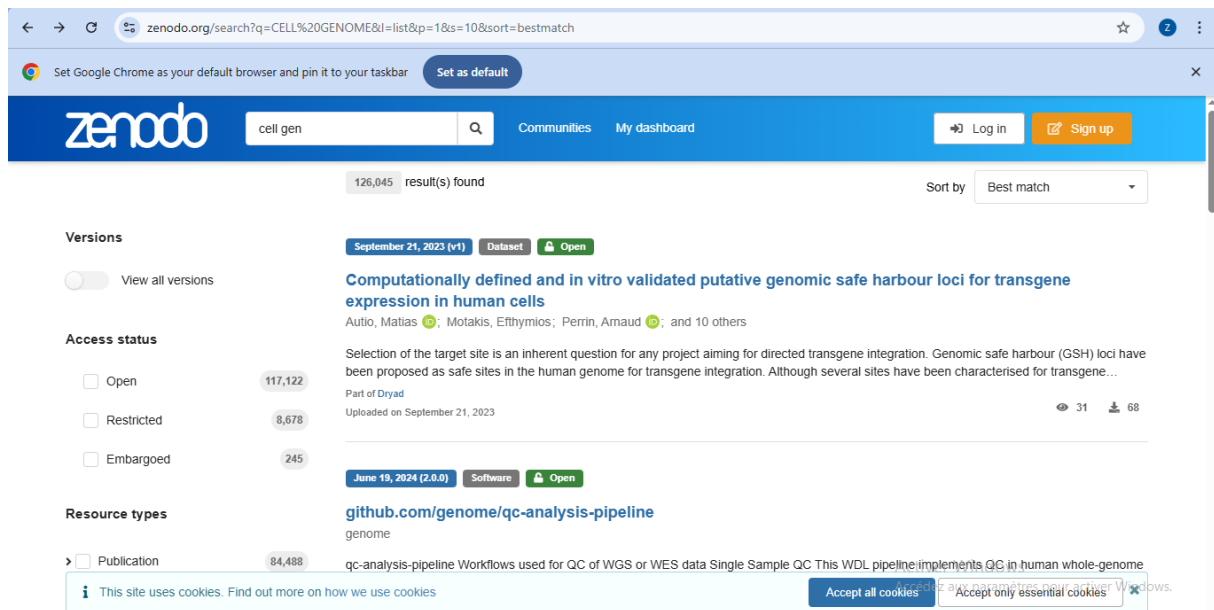


Figure 2 : Résultats de recherche sur Zenodo avec la requête “cell genome”

Après avoir accédé à la plateforme, nous avons effectué une recherche en utilisant les mots-clés 'cell genome'. Comme le montre la Figure 2, la plateforme a généré une liste de résultats variés (plus de 126 000 résultats). Pour affiner notre recherche, nous avons observé les filtres disponibles à gauche, tels que le type d'accès (Access status) et le type de ressource (Resource types).

Critères de sélection du dataset

Pertinence : Le dataset choisi correspond directement à notre thématique de recherche (génomique et cellules humaines).

Type de contenu : Nous avons sélectionné un fichier de type 'Dataset' pour répondre aux exigences du devoir.

Disponibilité : Le document est en 'Open Access', ce qui permet son téléchargement et l'extraction de ses métadonnées.

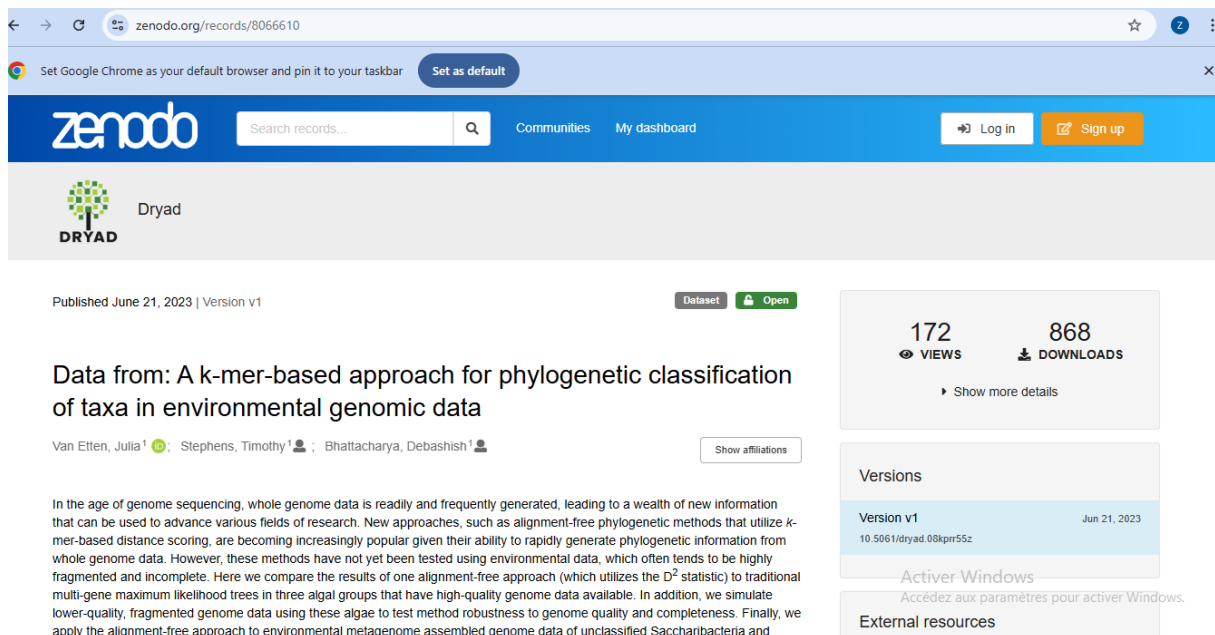


Figure 3 : Détails du dataset sélectionné sur Zenodo.

Après l'analyse des résultats, nous avons sélectionné le dataset intitulé : 'Data from: A k-mer-based approach for phylogenetic classification of taxa in environmental genomic data'. Ce choix est justifié par la pertinence des données génomiques présentées et la présence d'un résumé détaillé ainsi que d'un identifiant DOI permanent.

Section 3 : Métadonnées du dataset

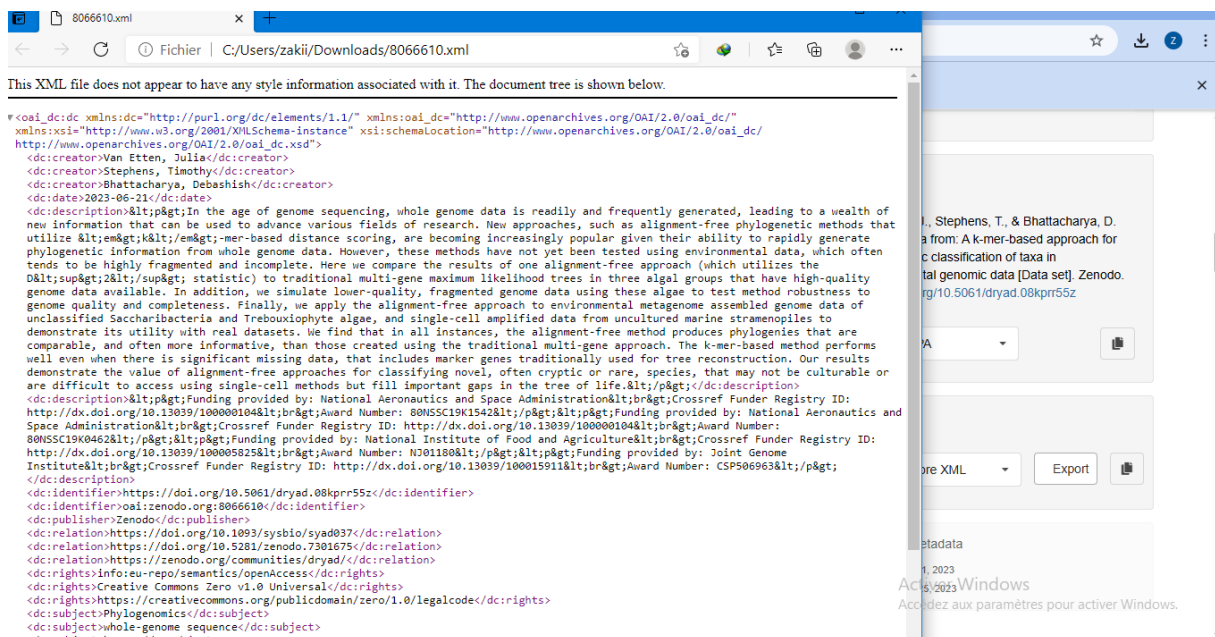


Figure 4 : Extraction des métadonnées au format Dublin Core (XML).

Conformément aux exigences du travail pratique, nous avons récupéré les métadonnées associées au dataset choisi en utilisant la norme Dublin Core. La Figure 4 montre le fichier XML généré par Zenodo, contenant les balises sémantiques essentielles pour l'identification et l'indexation de la recherche.

Présentation structurée des métadonnées

Élément Dublin Core	Valeur extraite (d'après le fichier XML)
Title	Data from: A k-mer-based approach for phylogenetic classification of taxa in environmental genomic data
Creator	Van Etten, Julia; Stephens, Timothy; Bhattacharya, Debashish
Date	2023-06-21
Description	Dataset related to a k-mer-based approach for phylogenetic classification using environmental genomic data.
Type	Dataset
Publisher	Zenodo
Identifier	DOI 10.5061/dryad.08kpr55z
Rights	Creative Commons Zero v1.0 (Open Access)
Keywords and subjects	Phylogenomics whole-genome sequence k-mer alignment-free phylogeny reconstruction Biodiversity cryptic species Metagenomics environmental data
Format	XLSX, RTF,TXT,CSV
Language	English

Tableau 1 : Métadonnées du dataset extraites selon la norme Dublin Core.

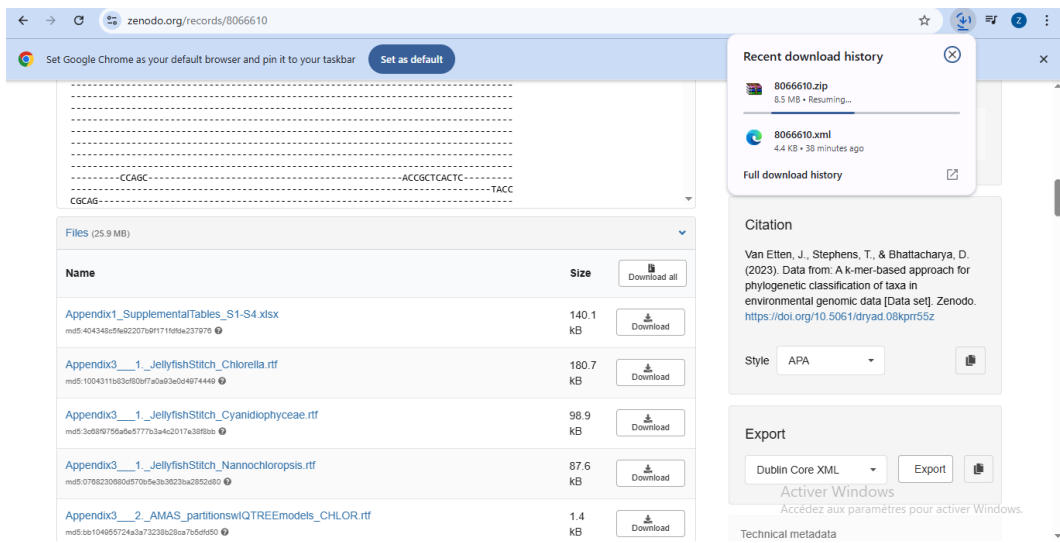


Figure 5 : Téléchargement des fichiers du dataset.

La dernière étape de la partie pratique consiste à télécharger les données pour une éventuelle exploitation. Comme illustré dans la Figure 5, nous avons procédé au téléchargement des fichiers sources. Cette étape confirme l'accessibilité des données partagées sous licence libre sur la plateforme Zenodo.

Conclusion

L'exploration de la plateforme Zenodo nous a permis de comprendre l'importance cruciale de l'indexation des données via des métadonnées standardisées, telles que le Dublin Core, pour garantir la visibilité et la réutilisabilité des travaux scientifiques. Cette expérience pratique souligne la complémentarité entre les outils d'analyse, comme la bibliothèque Biopython, et les infrastructures de partage en accès ouvert. En somme, la maîtrise de ces outils libres et open source est devenue une compétence indispensable pour tout chercheur en sciences de la nature et de la vie, permettant de contribuer efficacement à une science plus transparente, collaborative et accessible à tous.

Référence

- [1] European Organization for Nuclear Research (CERN). (2013). Zenodo: A general-purpose open-access repository. Récupéré de : <https://about.zenodo.org/>
- [2] Peters, S., & Siciliano, F. (2017). Zenodo: An innovative service for sharing and preserving research data. CERN Document Server.

PARTIE 3 : BONUS – DÉPÔT DU RAPPORT SUR GITHUB

1. Description de la démarche :

Conformément aux directives du devoir, nous avons créé un dépôt (Repository) public sur la plateforme GitHub afin d'héberger ce rapport. Cette étape permet de s'initier aux outils de gestion de version et au partage de documents dans un environnement collaboratif et open source.

2. Informations sur l'entrepôt :

Nom de l'entrepôt : TP-Logiciels-Libres-2025

Visibilité : Public

Contenu : Rapport final (Partie I & Partie II) au format PDF.

3. Lien d'accès (URL) :

Le rapport est consultable et téléchargeable via le lien suivant :"

Lien GitHub : [<https://github.com/ZakariaBelarbi-alt/TP-Logiciels-Libres-2025>]