XoveTIC: Impulsando el talento científico

Editorial Office

Servizo de Publicacións (SPU)
Universidade da Coruña
Maestranza 9, 15001 A Coruña, Spain

Handle: 2183/40661
DOI: 10.17979/spudc.9788497498913

Cover image courtesy of CITIC—Research Center of Information and Communication Technologies, University of A Coruna, Spain
CITIC, as a center accredited for excellence within the Galician University System and a member of the CIGUS Network, receives subsidies from the Department of Education, Science, Universities, and Vocational Training of the Xunta de Galicia. Additionally, it is co-financed by the EU through the FEDER Galicia 2021-27 operational program (Ref. ED431G 2023/01)

Proceedings XoveTIC 2024. Impulsando el talento científico.

This volume was typeset in LaTeX by the editors using *varianTeX* — a reusable template for journals in the Humanities, developed by Wout Dillen. *varianTeX* is open source, available on GitHub, and deposited in the Zenodo Open Science Repository. DOI: 10.5281/zenodo.3484651.

New Biopython Library to Support Molecular Biology

Patrícia Nogueira, and Vítor J. Sá

LabRP/CIR, ESS, Polytechnic of Porto, Porto, Portugal
Correspondence: vitor.sa@ess.ipp.pt

DOI: *https://doi.org/10.17979/spudc.9788497498913.1*

Abstract: Biopython is a library that facilitates the development of applications for Bioinformatics, using the Python programming language. Maintained by an international association of programmers - the Open Bioinformatics Foundation - since 1999, it provides tools for analyzing biological sequences and accessing online databases like NCBI. It features modules for sequence alignment, protein structures, population genetics, and more. Since the library is open-source with the collaboration of several developers, the project aimed to create a new library that used to existing API services such as Expasy, Blast, Uniprot and DrugBank. This enabled a single-call module to reference multiple services, obtain results, and generate a final report for a searched sequence. The first step to start the project was a study of existing Biopython libraries to assess their alignment with the proposed objectives. Access to the DrugBank service required a formal request. It was necessary to justify the request, explaining how and why the data would be used within the scope of the project. After the preceding steps, the entire architecture and design of the solution were conceptualized and, subsequently, we started the development. As a result of this project, there was an endpoint that, when invoked by any software or platform, returns all information found for a genetic sequence in JSON format. A small user interface was also developed to display the search results as an alternative to using only the API. Because BioPython is a free package and science is the motivation, a new library will be available for several services such as Expasy, Blast, Uniprot, and DrugBank. It will be free for all schools, labs, researchers, and developers who want to use it.
Keywords: BioPython; ExPASy; Blast; UniProt; DrugBank

1 Introduction

This project aims to develop a new library within BioPython, an open-source bioinformatics toolkit written in Python and maintained by the Open Bioinformatics Foundation since 1999. BioPython provides tools for biological sequence analysis, database access (e.g., NCBI), and methods for processing files from various formats such as BLAST, FASTA, GenBank, and others. This project aims to create the API FoundSeq 1.0, integrating online tools via public APIs to extract data related to searched sequences.

2 Theoretical Reference

Before detailing the methodology, it will be necessary to provide a theoretical explanation about the tools that the application will work with by invoking their APIs or information files:

- **ExPASy Translate Tool**: A tool designed to translate nucleotide sequences into protein sequences and identify open reading frames (ORFs). ORFs are DNA regions containing codons that serve as templates for protein synthesis; larger ORFs are more likely to produce proteins.

- **BLAST** (**Basic Local Alignment Search Tool**): This tool is developed for searching and comparing biological sequences against extensive databases. It returns sequences that are most similar to the queried sequence.
- **UniProt** (**Universal Protein**): A free database providing detailed information on protein sequences, including their biological functions, genetic variations, and related pathologies. The data is sourced from scientific literature and includes databases like Swiss-Prot and TrEMBL, which are part of UniProtKB.
- **DrugBank**: An online, freely accessible database that contains comprehensive information about medications, including chemical properties, targets, sequences, and pathways. It is widely used in the pharmaceutical industry and by researchers. While the database is free to access, its content requires a license for use and redistribution. Academic users can request a free license for specific cases, whereas others must obtain a paid license.

3 Methodology

3.1 Biopython Modules Research

To start this project, was necessary to evaluate existing Biopython libraries and determine their relevance to the project goals. The **Bio.Blast** module, which includes sub-modules like **Bio.Blast.NCBIWWW** and **Bio.Blast.NCBIXML**, provided code for the web version of BLAST offered by NCBI. However, this module has been deprecated and was updated to**ElasticBLAST**, a cloud-native distributed system running on GCP and AWS, designed for scalability and user-friendliness. ElasticBLAST uses a Platform as a Service (PaaS) model for resource management, leveraging cloud technologies, batch computing (AWS Batch), serverless computing (AWS Lambda), and object storage.

The impossibility of using ElasticBLAST is related to its requirement for cloud services, which have costs. After further research, the **Blast Tool** provided by **EMBL-EBL** was identified as an acceptable alternative. This tool does the same work as NCBIWWW. Another module studied was **Bio.ExPASy**, but it was concluded that its sub-modules do not fulfill the requirements for translating nucleotide sequences into their corresponding proteins, specifically returning all 5'-3' and 3'-5' frames, as the ExPASy Translate Tool does. The mentioned service provides a URL for obtaining the translation by passing the necessary parameters, such as a FASTA file, to retrieve the desired results.

Regarding **UniProt**, the sub-modules provided by BioPython also do not meet the project's needs. Alternatives such as **PyDPI**, which focuses on protein-drug interaction analysis, were considered but they are too specific and not enough for project purposes. Further research was discovered The UniProt API that retrieves protein data using a unique identifier(ex.P42898).

At last, access to **DrugBank** was studied. Since BioPython lacks a specific module for this, an academic license was necessary to change the approach.

3.2 APIs Research

The services ExPASy, UniProt and Blast offer URLs to their public APIs with an exception for DrugBank, which license has costs involved. However, DrugBank has a kind of student license, and to get it was necessary to explain the reason for the request and the scope in which the data is being used. At the end of the project, it will also be necessary to submit a brief document with the outcome. After obtaining this license, it was possible to download a .xml file containing all the available information. To make reading and searching through the file easier, an additional tool was implemented in this project, which reads the information and migrates it to an SQLite3 database also developed as part of the project.

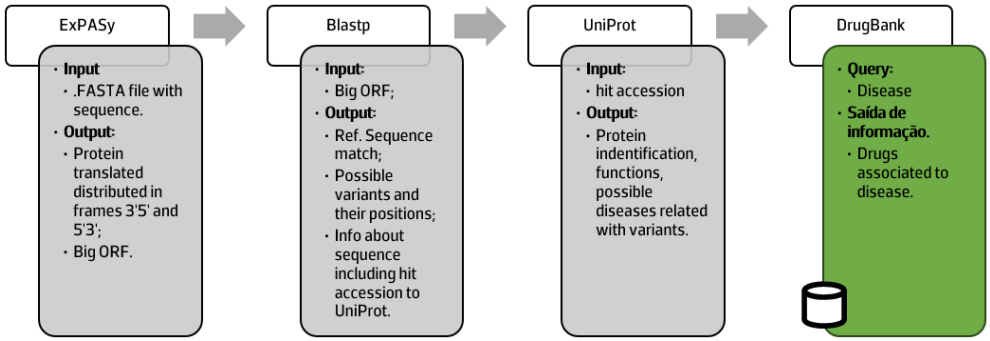

Figure 1: APIs interaction with input and output data

3.3 Architecture

The application is divided into two components:

- API – which makes calls to services and retrieves information. This component can be used individually by calling each service's endpoints and obtaining information in text format (ExPASy, DrugBank) and in .json format (Blast, UniProt);
- APP – which provides a user interface and allows the information to be displayed in a more organized and user-friendly way.

The user will interact with the application or, if the user only wants to access to endpoint and retrieve information for its own application this is possible either. An internet connection is necessary to work with the application and/or API once external endpoints for ExPASy, Blast, and UniProt are called.

3.4 Database

As previously mentioned, there is a database that will store the most relevant information about drugs provided by DrugBank. To build this database, the content of the .xml file was first analyzed, and all the fields that would be important for finding a drug were identified. After this analysis, a relational model was designed for the database, duly standardised, starting with the usual entity-relationship diagram, then the logical model and finally the physical model. The following tables were created: *drug*, *classification*, *groups*, *products*, and *pathway*. The table drugs has a one-to-many relation to the other tables.

3.5 Solution Development

For development we used the Visual Studio Code editor, the SQLite3 database management system and the Python v3.10 programming language for the entire application/API. The solution is divided into two projects:

- **FoundSeq App v1.0** – This represents the user interface (front office), where a file in .FASTA format will be uploaded, and the result will be returned, enabling the creation of a report in .pdf format from a .html page.
- **FoundSeq API v1.0** – This represents the entire back-office layer, where there is external access to APIs, use of the Biopython module, and access to databases. Here, the response to the user will be built and returned to the front office.

Flowchart

In general terms, the solution flows as follows:

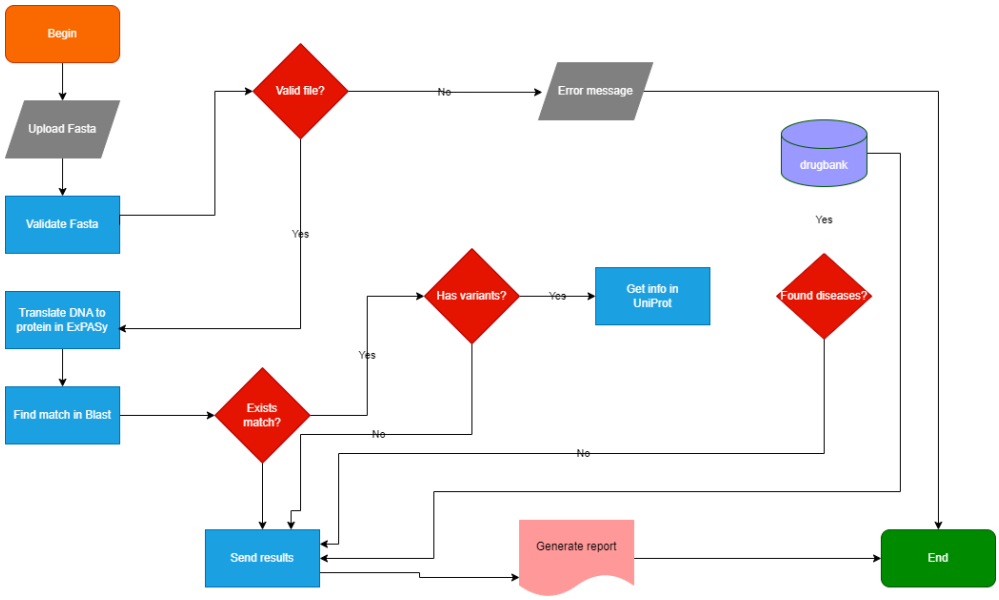

Figure 2: Solution flowchart.

4 Results

According to the sequence provided by the software, we can have numerous possible results:

Table 1: Possible results

Result type	Service	Inputs	Outputs
Complete	ExPASY	Valid FASTA file	Translated protein and Big ORF
	Blast	Big ORF	Found a match, variants, their positions, and other info as hit accession
	UniProt	hit accession	Protein info, structure, and disease associated to variants
	DrugBank	disease name	one or more drugs to help, prevent or delay the disease
No drugs to disease	ExPASY	Valid FASTA file	Translated protein and Big ORF
	Blast	Big ORF	Found a match, variants, their positions, and other info as hit accession
	UniProt	hit accession	Protein info, structure, and disease associated to variants
	DrugBank	disease name	not found any drugs
No disease for variants	ExPASY	Valid FASTA file	Translated protein and Big ORF
	Blast	Big ORF	Found a match, variants, their positions, and other info as hit accession
	UniProt	hit accession	Protein info, structure, but no disease found
No variants	ExPASY	Valid FASTA file	Translated protein and Big ORF
	Blast	Big ORF	Found a match, but no variants found, i.e the sequence is similar to the reference sequence
No match	ExPASY	Valid FASTA file	Translated protein and Big ORF
	Blast	Big ORF	No match found for sequence
Error results	ExPASY	Invalid FASTA file	Empty file, bad formatted file, invalid nucleotides

5 Discussion

An analysis of this project, along with feedback collected (notably on May 3rd during the presentation at SBBAH 2024 - Symposium on Biostatistics and Bioinformatics Applied to Health), showed that the developed library can be particularly useful for students in genomics. The application met its objective of providing access to services like ExPASy, Blast, UniProt, and DrugBank in one place. Various result scenarios were considered and tested, ensuring the application responds efficiently, even when errors occur.

Two main challenges emerged during development. First, DrugBank does not provide a free API, so an academic license had to be requested by explaining the use case. Second, late in the project, the Biopython Blast module was discontinued, threatening the project's viability. This was resolved by switching to the EMBL-EBL Blast Tool.

The new library can be used via the user interface or by invoking API endpoints to retrieve results from individual services or all at once. The latter option is suitable for integrating the library into other desktop, web, or mobile applications. The user interface was designed from scratch, starting with wireframes. Microsoft Copilot (free version) was used to generate the logo, and icons were sourced from Flaticon.

All the source code (frontend and API) is hosted on GitHub, a platform for version control and collaboration on private or open-source projects globally.

6 Conclusion

In conclusion, Biopython is an essential tool for developing bioinformatics applications. With its support, along with services like ExPASy, Blast, UniProt, and DrugBank, the proposed project was completed, leading to the release of FoundSeq version 1.0. This application enables users to obtain results for a searched sequence and generate a corresponding report.

Future improvements and features already envisioned for this version include:

- Retrieving non-sequential information from each service, allowing for the upload of different files (or the same one) for each service independently;
- 3D modeling of proteins, with and without mutations, if applicable;
- Bulk upload of up to three sequences to obtain sequential results.

As this is an open-source library that invites collaboration from other developers, it was a rewarding challenge to contribute something new that may be useful not only academically but also potentially in laboratories and for other developers. FoundSeq version 2.0 is expected to be released soon.

Bibliography

G. Alder. *draw.io*, 2024. URL *https://app.diagrams.net/*.

Biopython. *Biopython · Biopython*, 1999. URL *https://biopython.org/*.

D.-S. Cao, Q.-S. Xu, Q.-N. Hu, and Y.-Z. Liang. ChemoPy: freely available python package for computational biology and chemoinformatics. *Bioinformatics*, 2013.

B. Chapman and J. Chang. Biopython: Python tools for computational biology. *ACM SIGBIO Newsletter*, 2000.

DrugBank. *DrugBank | Clinical Drug Data API*, 2006. URL *https://www.drugbank.com/clinical*.

DrugBank. *API Reference | DrugBank Help Center*, 2020. URL *https://docs.drugbank.com/v1/*.

Flaticon. *Free Icons and Stickers - Millions of images to download*, 2010. URL *https://www.flaticon.com/*.

N. C. for Biotechnology Information. *BLAST: Basic Local Alignment Search Tool*, 1988. URL *https://blast.ncbi.nlm.nih.gov/Blast.cgi*.

E. B. Institute. *Data resources and tools*, 2023. URL *https://www.ebi.ac.uk/services/data-resources-and-tools/*.

kushinauwu. *OBF » Projects » Projects*, 2012. URL *https://www.open-bio.org/projects/*.

K. Laboratories. *KEGG: Kyoto Encyclopedia of Genes and Genomes*, 1995. URL *https://www.genome.jp/kegg/*.

T. d. P. Lehugeur and H. C. S. Melo. BIOINFORMÁTICA APLICADA NO DESENVOLVIMENTO DE NOVOS FÁRMACOS, 2018.

D. A. H. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine. *OMIM API Help - OMIM*, 1966. URL *https://www.omim.org/help/api*.

Microsoft. *Microsoft Copilot: o seu complemento de IA para o dia a dia*, 2023. URL *https://ceto.westus2.binguxlivesite.net/*.

S. National Institutes of Health, EMBL-EBI. *UniProt*, 2002. URL *https://www.uniprot.org/*.

S. S. I. of Bioinformatics. *Expasy - Translate tool*, 1993. URL *https://web.expasy.org/translate/*.

OpenAI. *ChatGPT*, 2022. URL *https://chatgpt.com*.

C. Pitassi and A. A. Gonçalves. Fatores que influenciam a adoção de ferramentas de TIC nos experimentos de bioinformática de organizações biofarmacêuticas: um estudo de caso no Instituto Nacional do Câncer, 2014.

SUBASH. Basic Local Alignment Search Tool(BLAST) of isolated Sars-Cov-2 genome using Biopython scripts at windows os, 2023.