BioT - Bioinformatics Tools Web Application

Help & Documentation

Kamil Pytlak

May 2022

Contents

1 Release history **3**

2 General information **4**
2.1 What is and what does the BioT cover? 4
2.2 Application functionality, included modules 4

3 Installation and usage **5**
3.1 Used technologies 5
3.2 Installation process 5
3.3 Use of individual functionalities 6
3.3.1 Homepage and sequence uploading/choosing 6
3.3.2 Sequence Summary 8
3.3.3 Two Sequence Alignment 10

4 License, future plans and contact **12**

1 Release history

Version	Release date	Description
1.0	27-05-2022	First release of the application containing two features: sequence summary and comparison of two sequences

2 General information

2.1 What is and what does the BioT cover?

Bioinformatics Tools (BioT) is a web application created using the Streamlit module in Python that allows for nucleotide/amino acid sequence analysis. It includes the tools necessary to perform descriptive statistics on sequences (substrate counts, verification of the hypothesis of sequence homogeneity), as well as the construction of a comparison matrix of the two nucleotide sequences (dotplot).

2.2 Application functionality, included modules

Version 1.0 of the application contains 3 root directories: apps, assets (divided into subdirectories: images and info) and data. All the application tools are located in the apps directory:

- home.py – homepage of the application;
- sequence_summary.py – the first tool which is descriptive analysis of sequence(s);
- sequence_upload.py – module providing functions for uploading and validating files;
- two_sequence_alignment.py – the second module which is the construction of a comparison matrix of two nucleotide sequences.

On the other hand, the assets directory contains image and description files (written in Markdown) necessary to properly format the application. Sample sequences written in the FASTA file format are included in the data directory. The full structure of the file tree is shown in fig. 1.

```
bioinformatics-tools/
├─ apps/
│   ├─ __init__.py
│   ├─ home.py
│   ├─ sequence_summary.py
│   ├─ sequence_upload.py
│   ├─ two_sequence_alignment.py
├─ assets/
│   ├─ images/
│   │   ├─ dna_strand.png
│   │   ├─ monitor.png
│   │   ├─ summary.png
│   ├─ info/
│   │   ├─ home_info.md
│   │   ├─ main_info.md
│   │   ├─ sequence_summary_info.md
├─ data/
│   ├─ homo-sapiens-beta-globin-region.fasta
│   ├─ homo-sapiens-tnf-after-alignments.txt
│   ├─ ls_orchid.fasta
│   ├─ sars-cov-2-replicase-after-alignments.fa
│   ├─ SARS-CoV-2_complete_genome.fasta
.gitignore
app.py
LICENSE
Pipfile
Pipfile.lock
README.md
```

Figure 1: Tree structure of BioT application files.

3 Installation and usage

3.1 Used technologies

The application was written in Python 3.9.12. Streamlit 1.9.1 was used to create the user interface. streamlit-option-menu 0.3.2 was used to divide the application into tabs (multi-page structure). The data analysis used pandas 1.4.2, SciPy 1.8.1, Biopython 1.79, plotly 5.8.0

3.2 Installation process

The project was uploaded to the web using Streamlit Cloud. You can use it online at the following link: `https://share.streamlit.io/kamilpytlak/bioinformatics-tools/app.py`. If you want to use this app on your local machine, install 'pipenv', copy the application to the target directory and run 'pipenv shell' and 'pipenv install' respectively:

1. Install the packages according to the configuration file 'Pipfile'.

   ```
   pipenv shell
   pipenv install
   ```

2. Ensure that the 'streamlit' package was installed successfully. To test it, run the following command:

   ```
   streamlit hello
   ```

3. If the example application was launched in the browser tab, everything went well. You can also specify a port if the default doesn't respond:

   ```
   streamlit hello --server.port port_number
   ```

   Where 'port_number' is a port number (8889, for example).

4. To start the app, type:

   ```
   streamlit run app.py
   ```

3.3 Use of individual functionalities

3.3.1 Homepage and sequence uploading/choosing

When you enter the application, the homepage is displayed. It contains information about the purpose of the application and its functionalities. Also included is a sidebar where you can select a tool, upload a FASTA format file with sequence(s), or use an example (fig. 2).

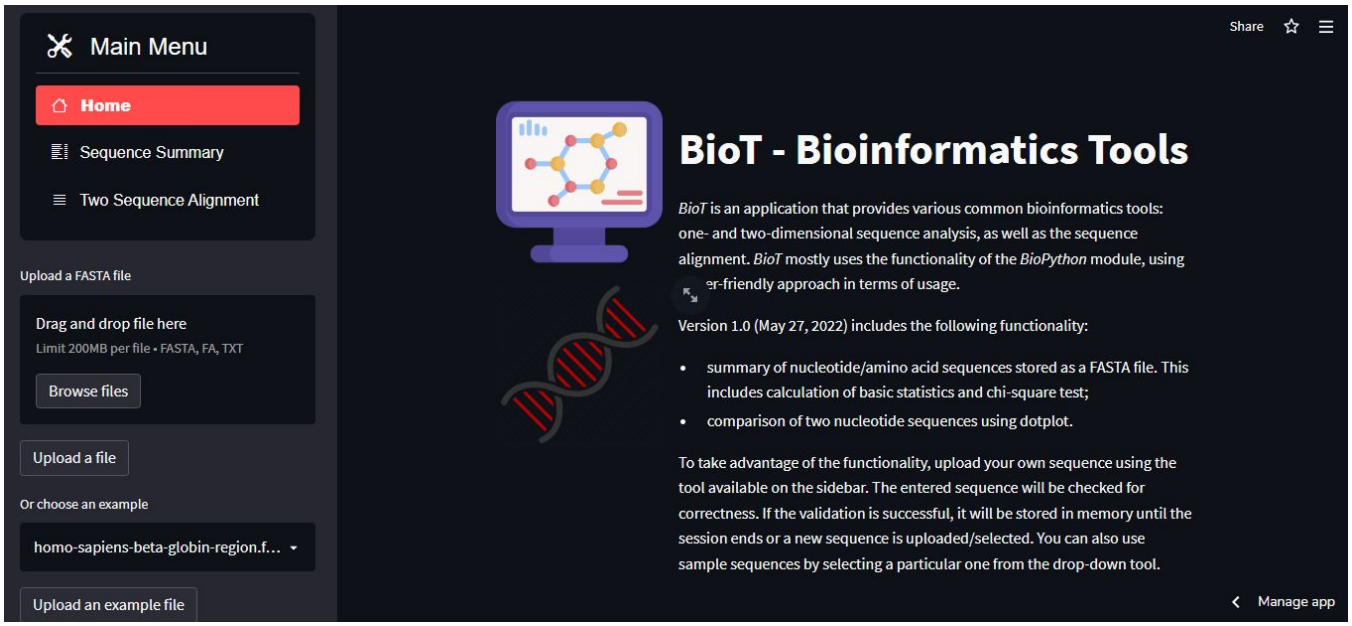

Figure 2: BioT home page with sidebar.

Let's say we want to upload a file with our own sequence. We select it by pressing "Browse files" and "Upload a file". After successful validation, a message will be displayed (fig. 3).

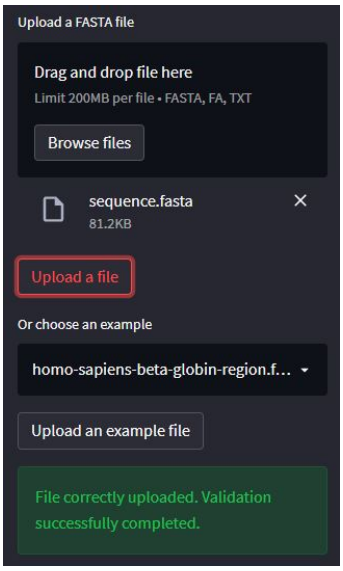

Figure 3: Tools to upload a sequence file or a selected sample sequence with a message that the sequence has been successfully uploaded.

When a sequence is uploaded, it is stored in the application cache. This way you don't have to reselect it when you select a new tool (4).

Figure 4: Message about capturing a sequence to the application cache.

If you select a file that is not saved in FASTA format, validation will fail and you will be notified (fig. 5).

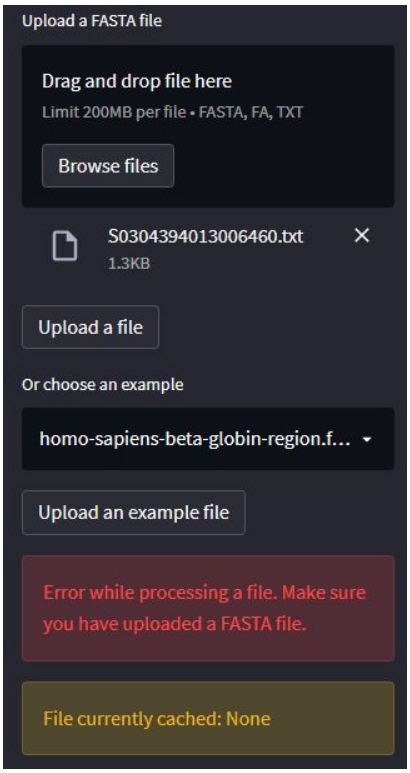

Figure 5: Message about unsuccessful validation of an unsaved file in FASTA format.

3.3.2 Sequence Summary

Once we have uploaded a good sequence, we can proceed to analyze it. Let's say we want to use an example: "sars-cov-2-replicase-after-alignments.fa". After clicking on the "Sequence Summary" button, you will be redirected to the tool page (fig. 6).

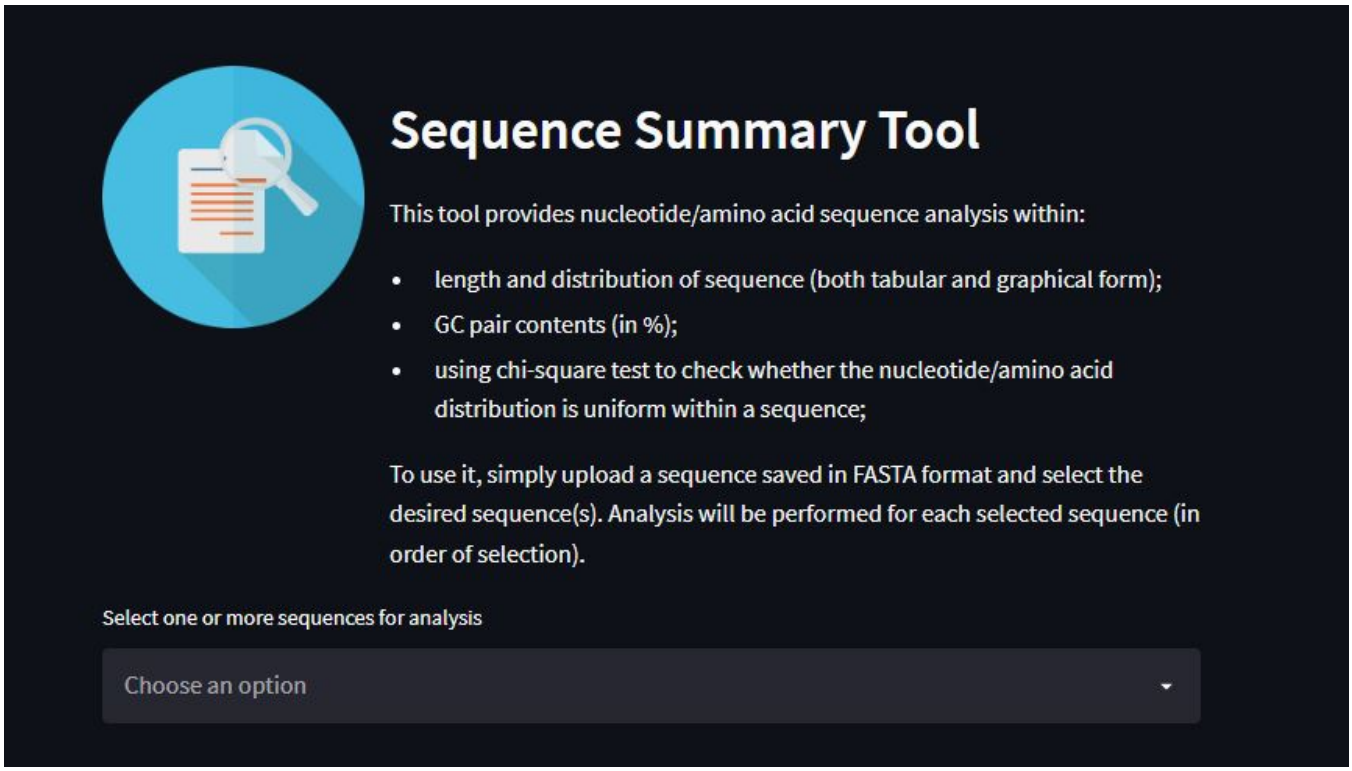

Figure 6: Main window of the "Sequence Summary" tool.

From the drop-down menu, we can select the sequences we want to analyze. For our purposes, we will select the first three: sp|P0DTD1|R1AB_SARS2, sp|P0C6X7|R1AB_SARS and tr|Q6UZF5|Q6UZF5_SARS (fig. 7).

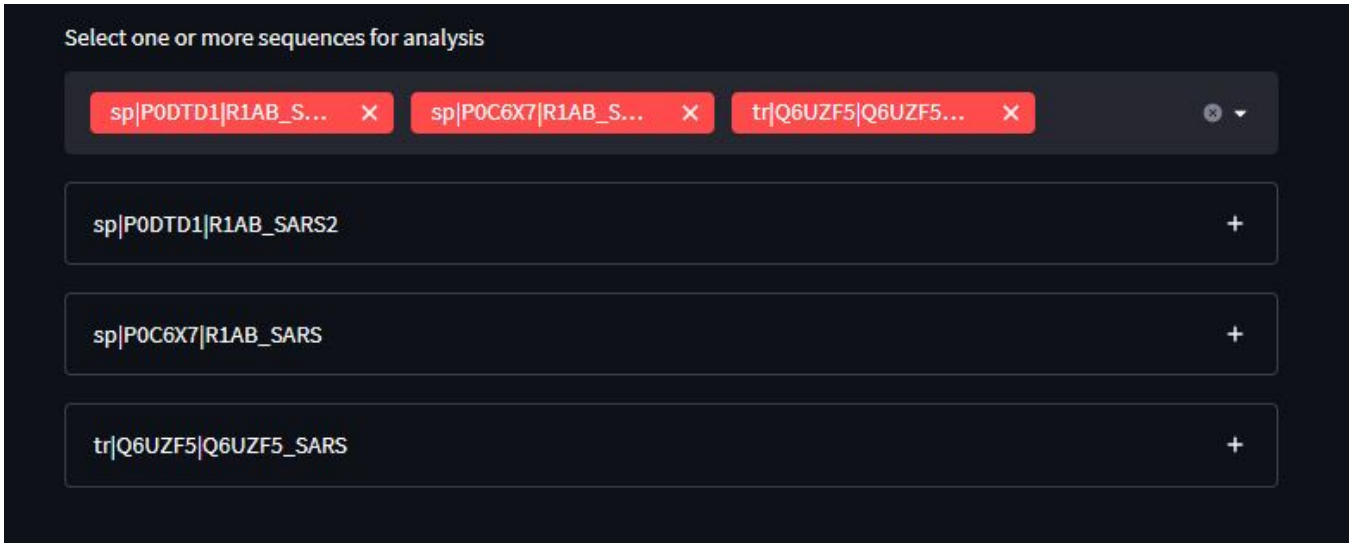

Figure 7: Selected sequences for analysis and three boxes – each box contains an analysis report.

In each of the three boxes (in our case) is an analytics report that we can review. We get it by expanding the respective box by clicking ”+”.

3.3.3 Two Sequence Alignment

”Two Sequence Alignment” is a tool that allows you to visually compare two sequences using the popular match matrix. You can read more about this methodology here: `https://omicstutorials.com/interpreting-dot-plot-bioinformatics-with-an-example/`. After successfully uploading/selecting a sequence and navigating to this tool, your eyes should see the main window of the program along with two drop-down tabs for selecting a particular sequence (fig. 8).

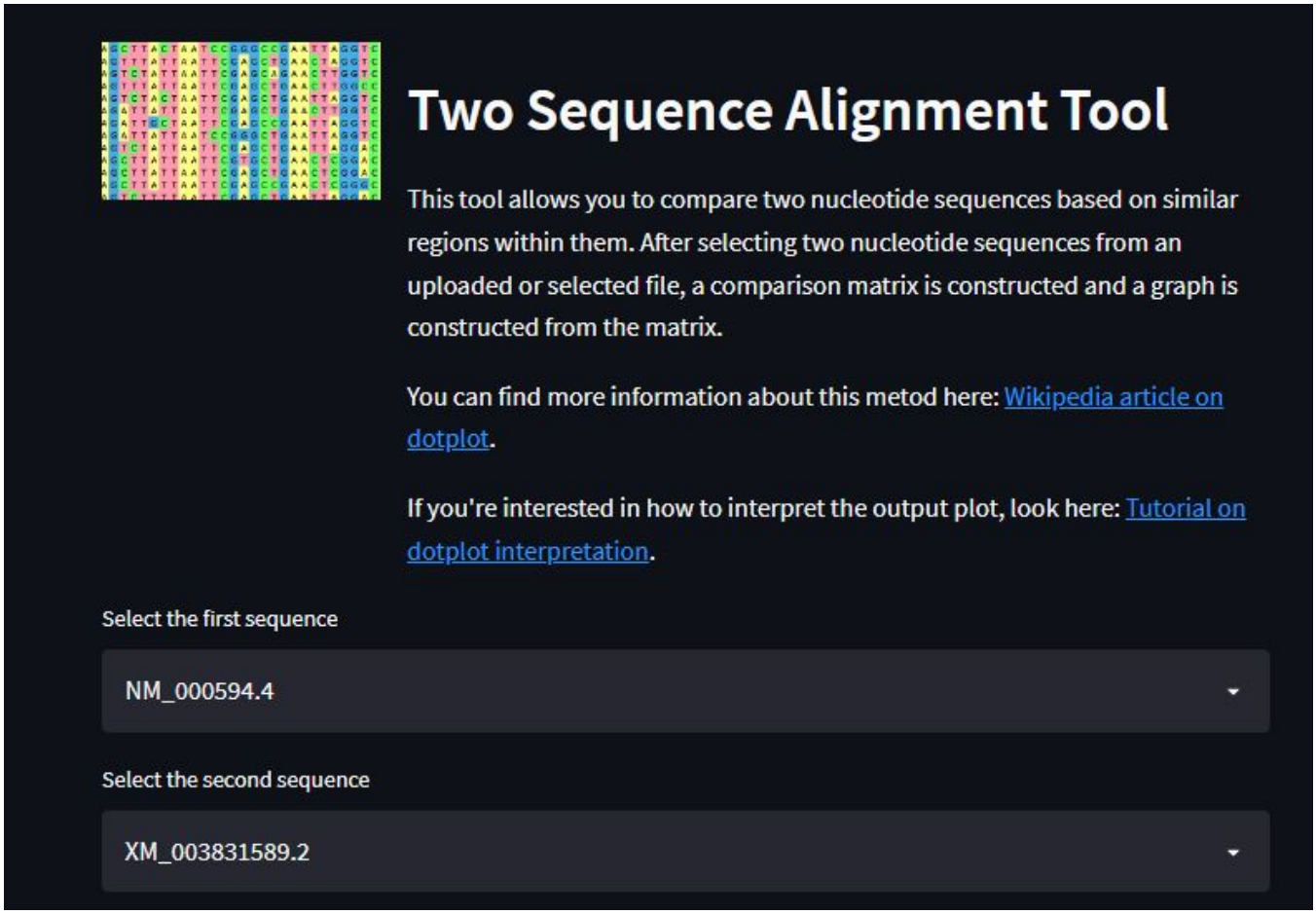

Figure 8: Main window of the ”Two Sequence Alignment” tool.

In this example, I selected two homologous sequences of the homo sapiens TNF gene (from a sample set of files). I left the window size at the default of 10. Knowing that the sequences are strongly similar, the eye should see a graph with the diagonal highlighted (fig. 9).

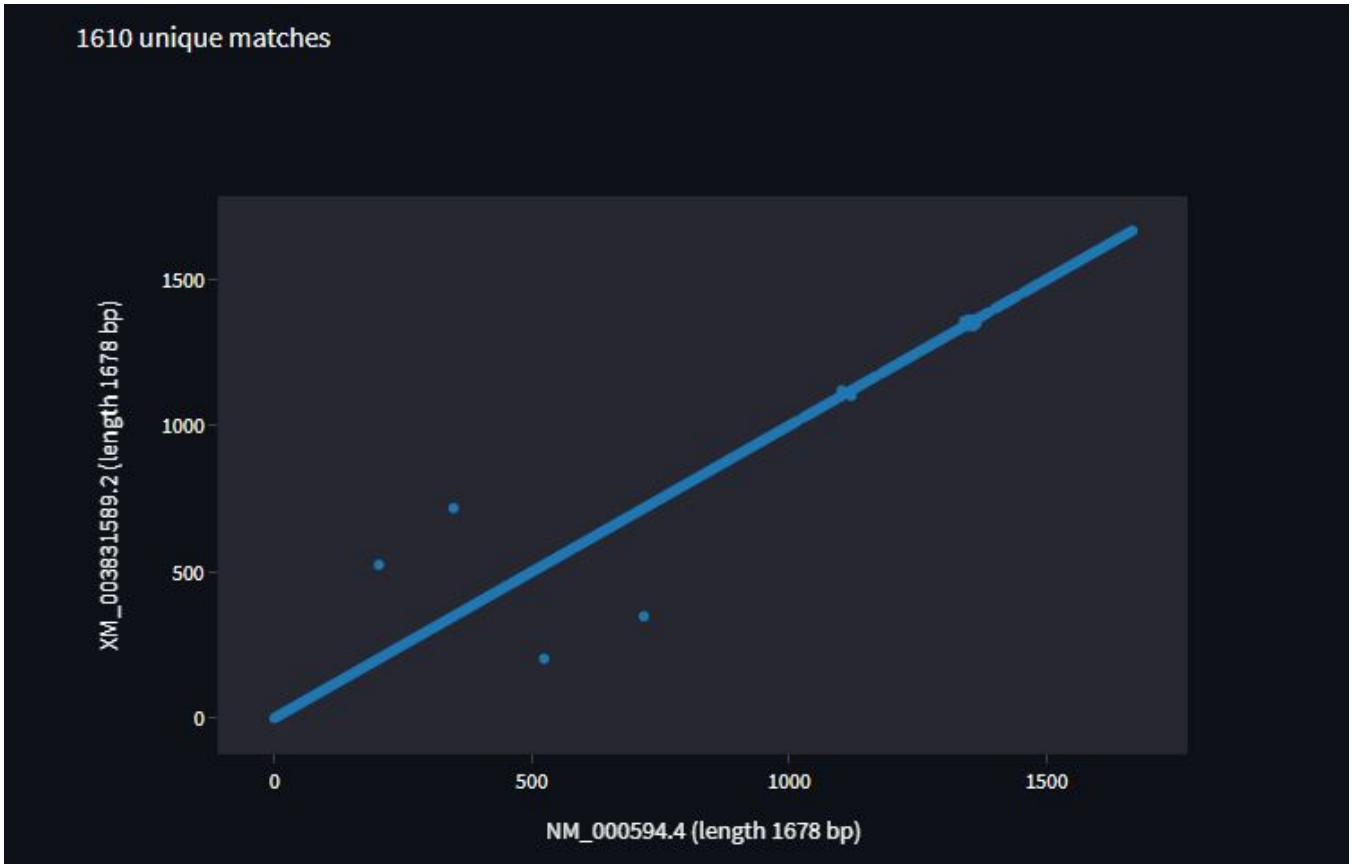

Figure 9: Main window of the ”Two Sequence Alignment” tool.

4 License, future plans and contact

The application has been released under an MIT license, which allows full distribution of the program's source code by any user. In case of distribution, the author (source) must be cited. In the future, it is planned to further extend the functionality of the application with functionalities offered by the Biopython module, as well as with popular tools used in the work of a bioinformatician, such as blast. If you have any ideas for its development or have noticed a bug and want to report it, please contact me via e-mail kam.pytlak@gmail.com. I also invite you to check out my GitHub for other projects: `https://github.com/kamilpytlak`.