# BLASTApp User Manual

Version 1.0

BT 3172 - Special Topics in Bioinformatics

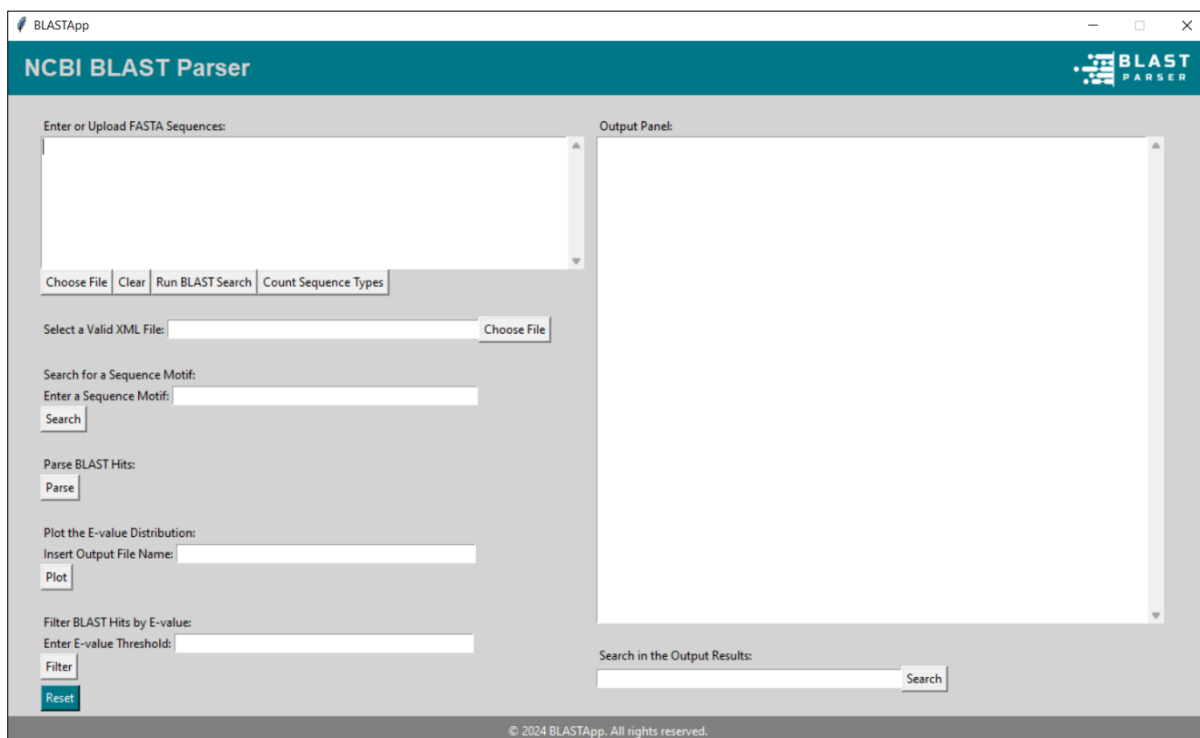15246 – T.H. Nishen Malinda

**Table of Contents**

# 1. Introduction

NCBI BLAST parser (BLASTApp) is a python application designed to perform NCBI BLAST operations on provided input sequences, offering various methods to examine and interpret its output. Through the use of Biopython and additional third-party packages, this application can handle multiple DNA, RNA, or amino acid sequences contained within a FASTA file. Its design integrates a suite of distinct methods, each finely tuned to address specific functionalities, guaranteeing robustness and effectiveness in use.

# 2. Features and Functionalities

1. Sequence Type Identification
   - Identify the type of sequences within a given FASTA file, whether DNA, RNA, or amino acid sequences

2. Sequence Counting
   - Count the number of sequences for each type (DNA, RNA, amino acid) in a FASTA file

3. BLAST Search Execution
   - Perform BLAST searches on multiple sequences within a single FASTA file
   - Automatically select appropriate BLAST algorithm types based on the sequence type detected
   - Save BLAST outputs for each alignment in XML format

4. Motif Search in BLAST Hits
   - Search for a user-defined sequence motif within BLAST hit sequences obtained from an XML file generated from previous BLAST searches

5. Parsing BLAST Hits
   - Parse through BLAST hits stored in an XML file and extract sequence details, length, E-value, and alignment information
   - Save the parsed information in a text file

6. E-value Distribution Visualization
   - Generate a histogram to visualize the distribution of E-values obtained from BLAST hits

7. Filtering BLAST Hits by E-value Threshold
   - Filter out BLAST hits with E-values below the user-defined threshold
   - Save the filtered information in a text file

8. Searching in Output Results
   - Perform a search within the output text widget to find specific information.

## 3. BLASTApp GUI



## 4. Credits

- ▪ Biopython: Used for handling biological sequence data and performing BLAST searches.

- ▪ Matplotlib: Utilized for plotting the E-value distribution.

- ▪ NCBIBlastParser: Custom module for parsing BLAST results.

## 5. Contact Information

For any questions, feedback, or support, please contact the project maintainer:

Nishen Malinda (nishenmalinda@gmail.com)

## 6. Installation and Setup

<u>Prerequisites</u>

- Ensure you have Python installed on your system. BLASTApp requires Python version 3.x or above.
- Make sure you have PyCharm installed on your system. If not, download and install it from the official website.

<u>Steps</u>

### 6.1 Install Required Python Libraries

- Open PyCharm.
- Open the terminal within PyCharm by clicking on "View" > "Tool Windows" > "Terminal".
- In the terminal, run the following command to install the required Python libraries:

```
pip install biopython pillow
```

### 6.2 Download BLASTApp Source Code

- Download the BLASTApp source code from the provided source.
- Save the downloaded source code in a directory of your choice.

### 6.3 Open BLASTApp in PyCharm

- Open PyCharm IDE.
- Click on "File" > "Open" and navigate to the directory where you saved the BLASTApp source code.
- Select the "BLASTApp.py" file and click "Open" to open the project in PyCharm.

### 6.4 Configure Python Interpreter

- PyCharm will automatically detect the Python interpreter installed on your system.
- If necessary, you can configure the Python interpreter by going to "File" > "Settings" > "Project: BLASTApp" > "Python Interpreter".
- Make sure the correct Python interpreter (version 3.x or above) is selected.

**6.5 Run BLASTApp**

- After configuring the Python interpreter, you can run BLASTApp directly from PyCharm.
- Navigate to the "BLASTApp.py" file in the project explorer.
- Right-click on the file and select "Run BLASTApp.py" from the context menu.
- Alternatively, you can open the file and click on the green "Run" button in the top-right corner of the editor window.

**6.6 Using Terminal in PyCharm**

- You can also run BLASTApp using the terminal within PyCharm.
- Open the terminal by clicking on "View" > "Tool Windows" > "Terminal".
- Navigate to the directory containing the BLASTApp source code.
- Run the BLASTApp by executing the following command:

```
python BLASTApp.py
```

## 7. Using BLASTApp

### 7.1 Input FASTA Sequences

- You can either manually enter FASTA sequences into the text area or upload a FASTA file by clicking the "Choose File" button.
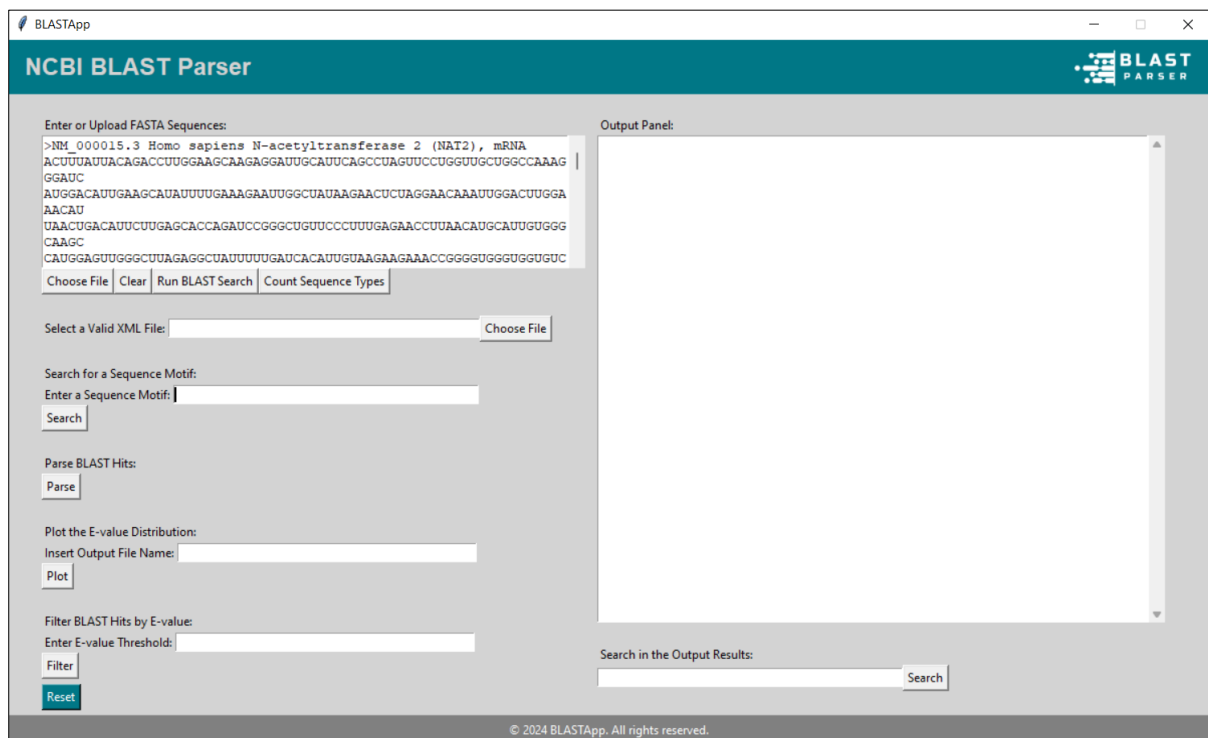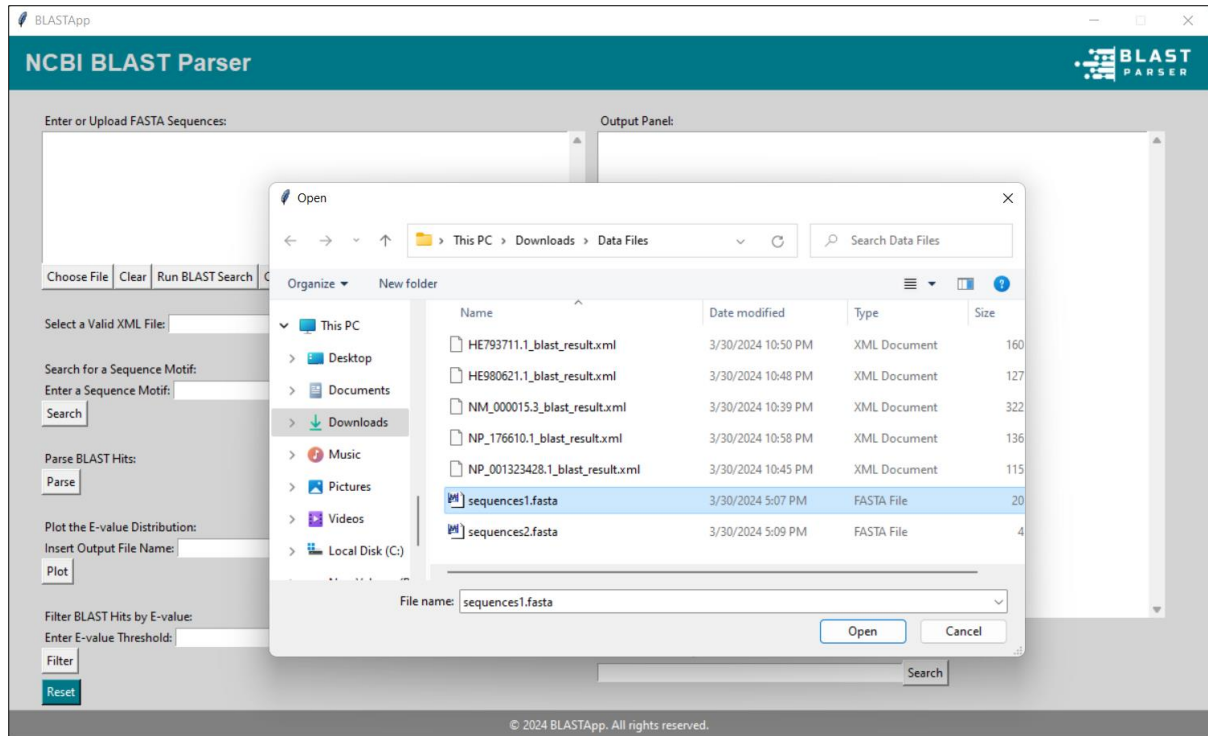
## 7.2 Running BLAST Search

- Click the "Run BLAST Search" button to perform a BLAST search on the provided sequences.
- If a user clicks the button without entering a FASTA file, an error message will be shown.
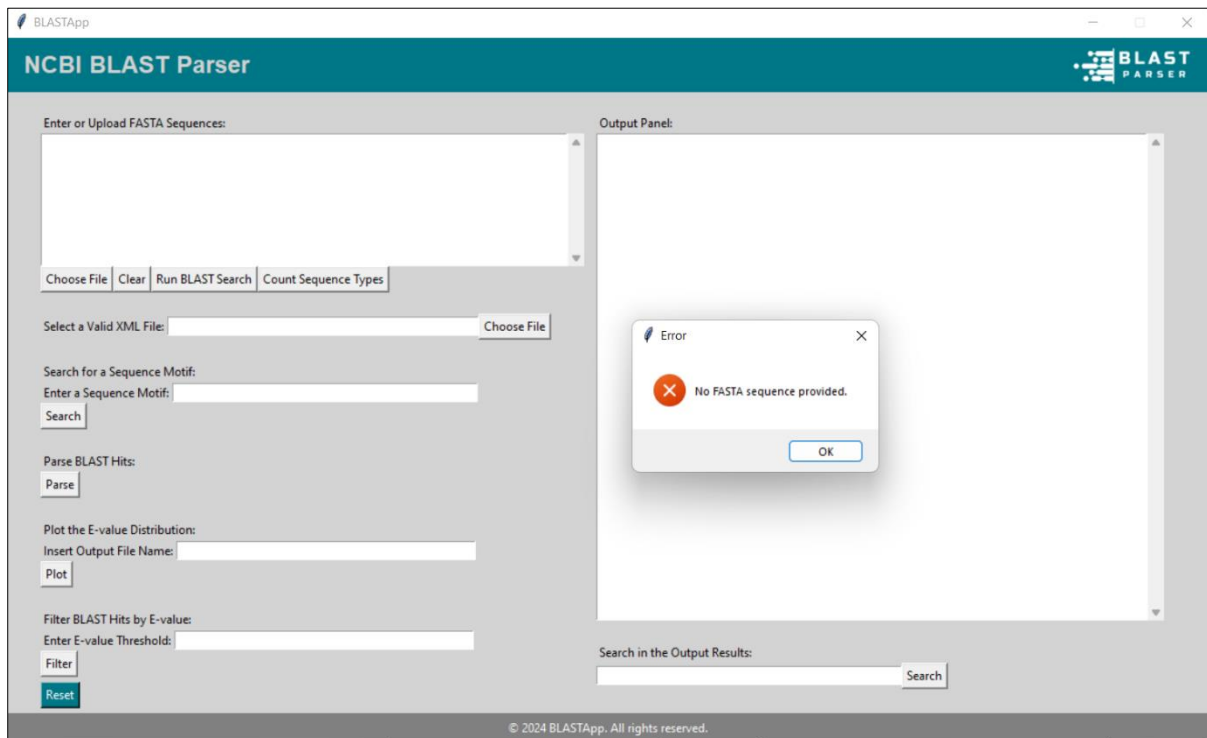


- BLAST results will be saved, and a success message will be displayed upon completion.
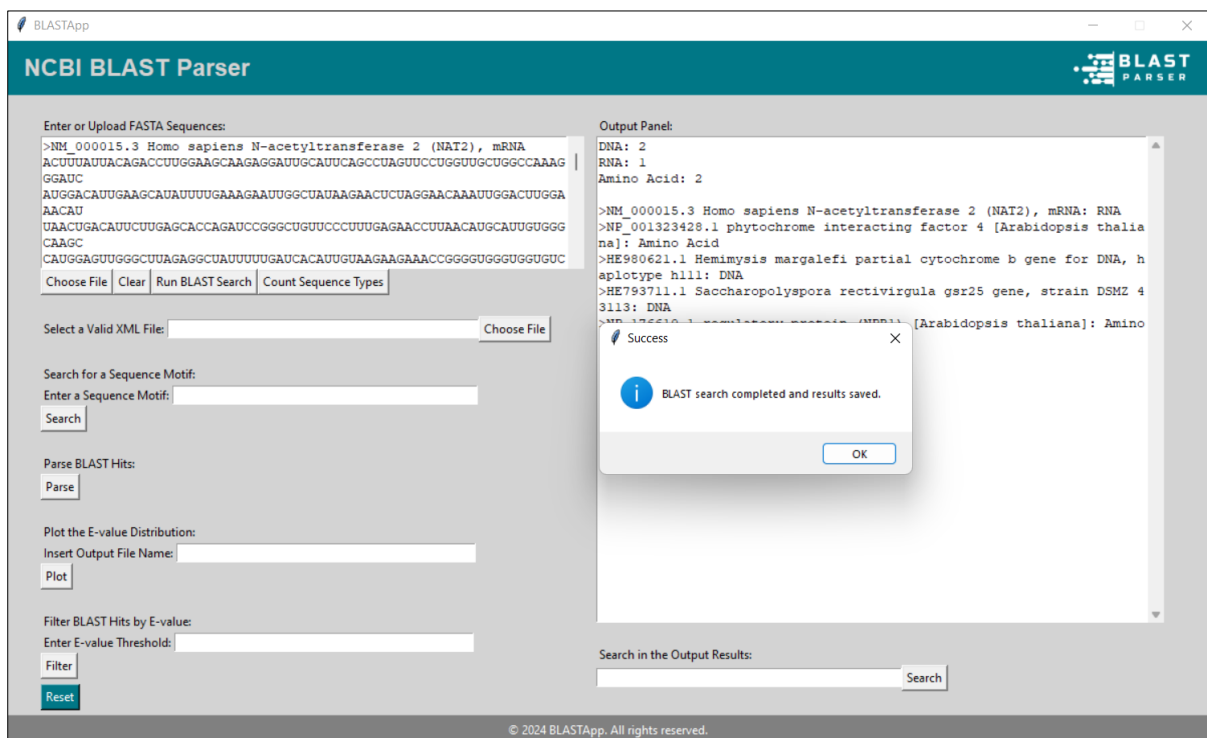


- If an error occurs during the BLAST search, an error message will be shown.

## 7.3 Counting Sequence Types

- Click the "Count Sequence Types" button to count the number of sequence types (DNA, RNA, amino acid) in the provided sequences.
- The counts will be displayed in the output panel.



- If a user clicks the button without entering a FASTA file, the counts will be displayed as zero in the output panel.



- Click the "Clear" button to clear the FASTA input field.

## 7.4 Searching for Motif

- Select a valid XML file containing BLAST hits by clicking the "Choose File" button in the "Select a Valid XML File" section.
- Enter a motif in the "Enter a Sequence Motif" field.
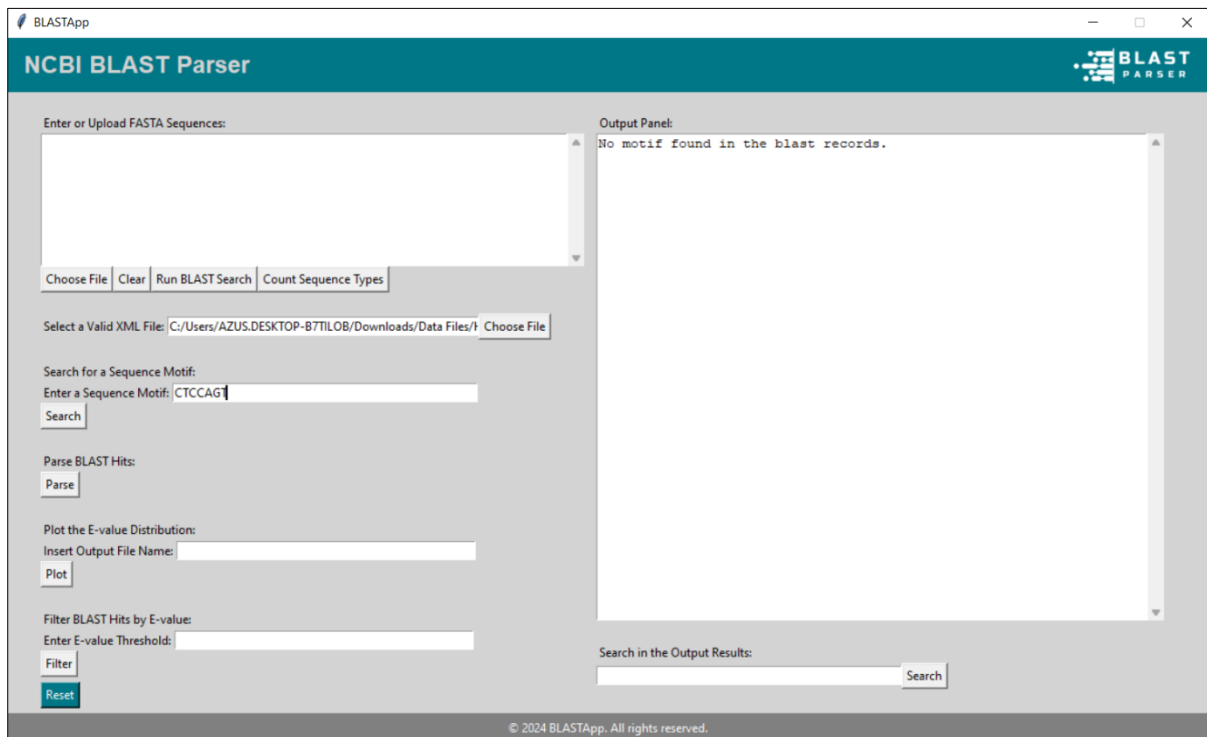- Click the "Search" button to search for the motif in the BLAST hits.
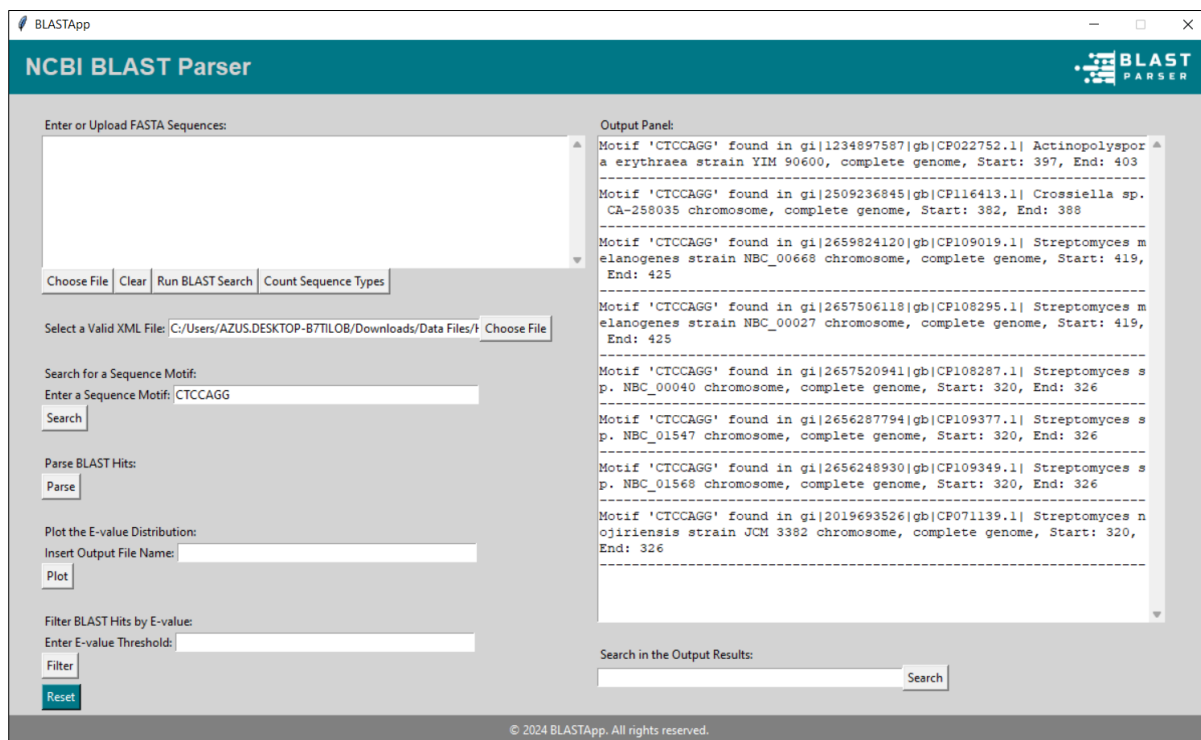


- If a user clicks the button without entering an XML file, an error message will be shown.
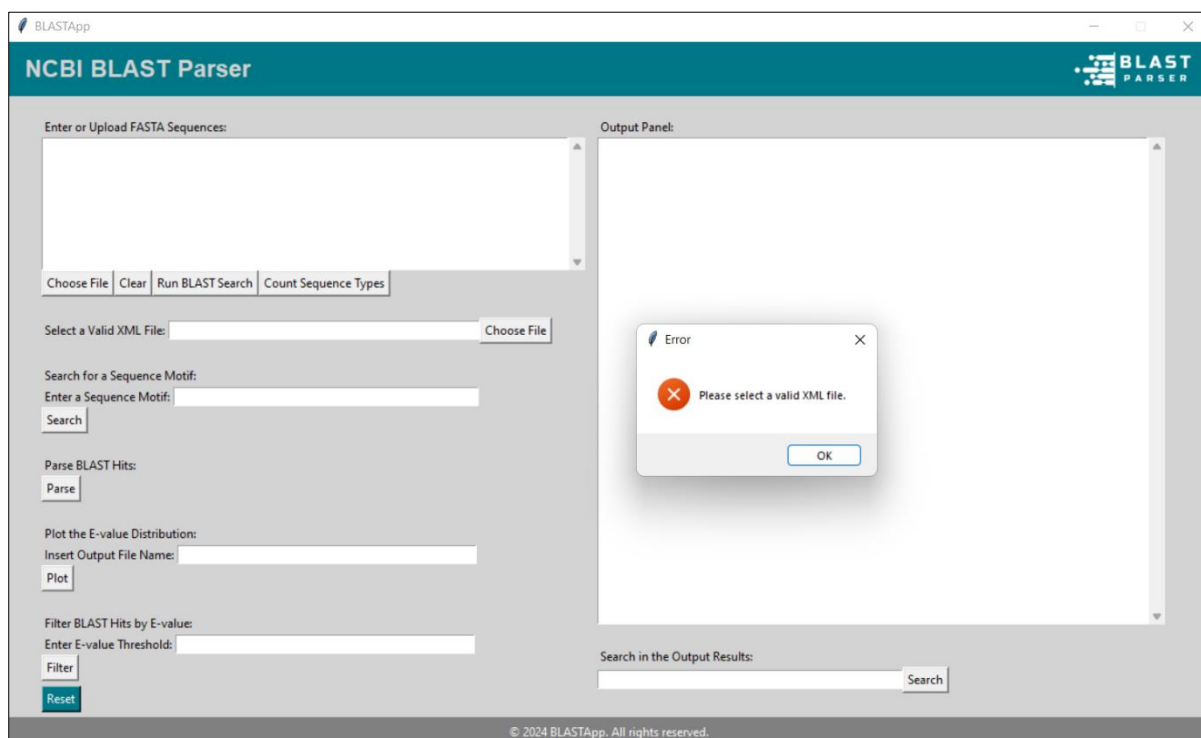
- If a user clicks the button without entering a sequence motif, an error message will be shown.
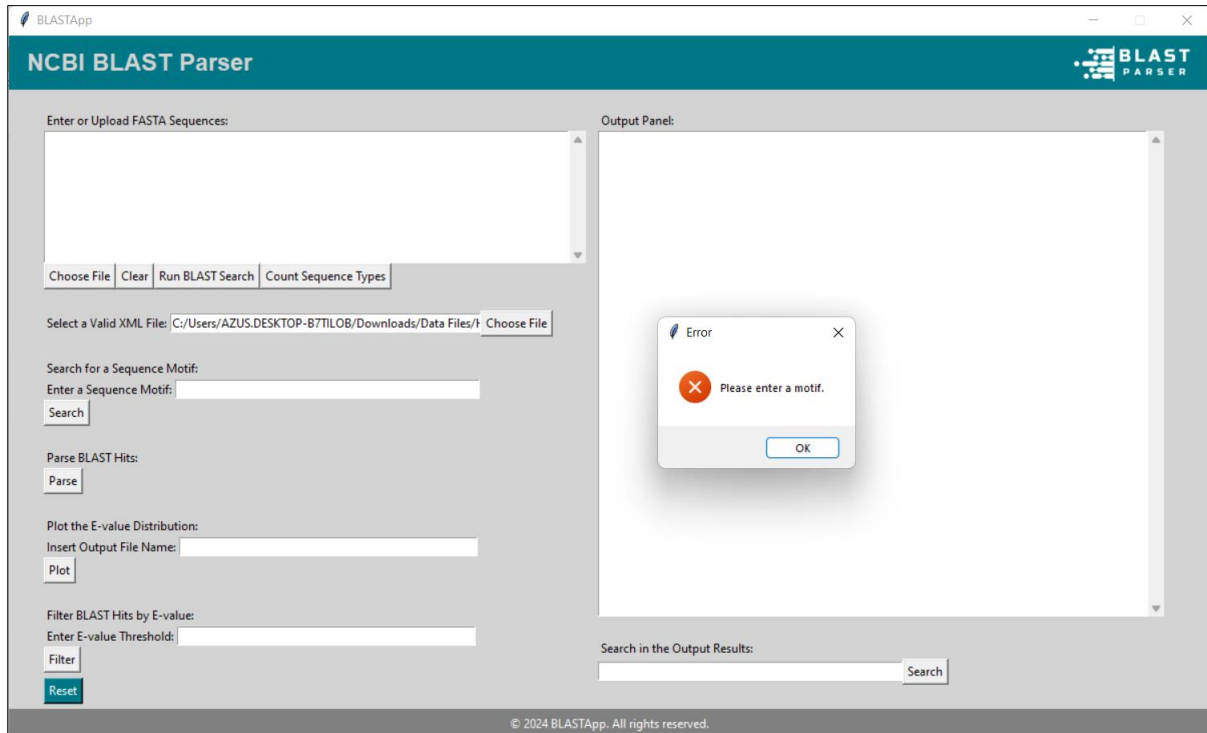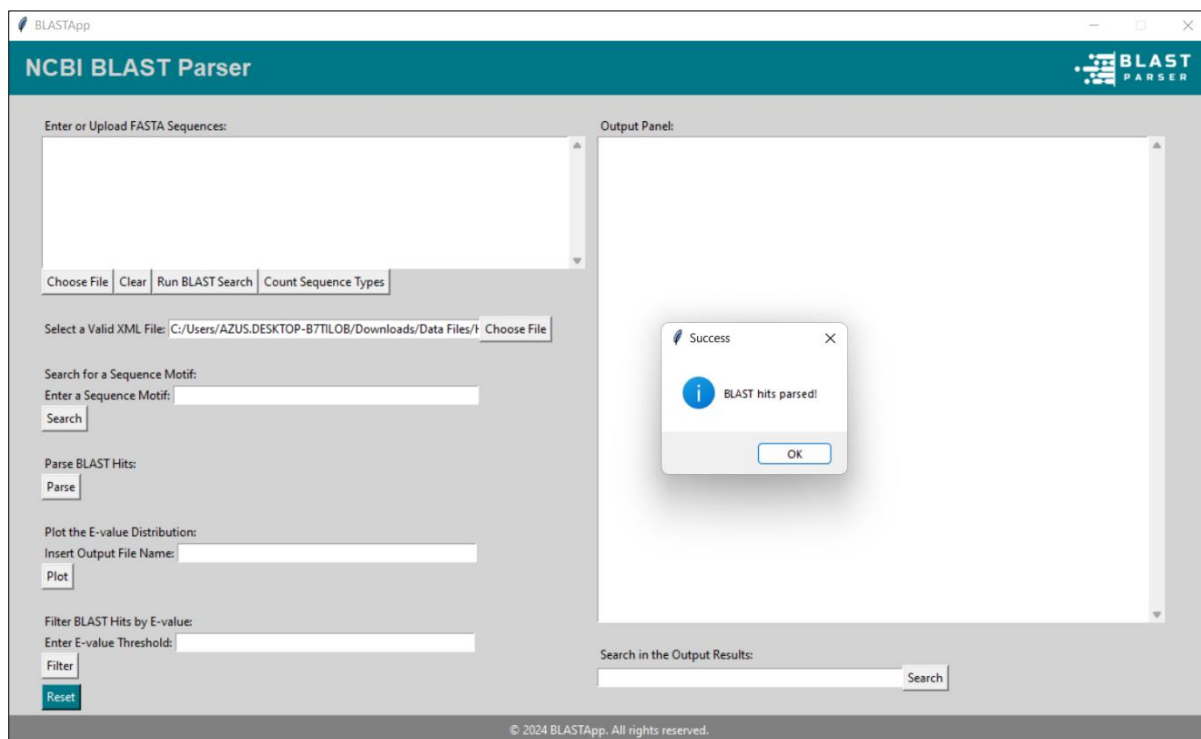


- If the motif is found, the details will be displayed in the output panel.

## 7.5 Parsing BLAST Hits

- Click the "Parse" button to parse BLAST hits from the selected XML file.
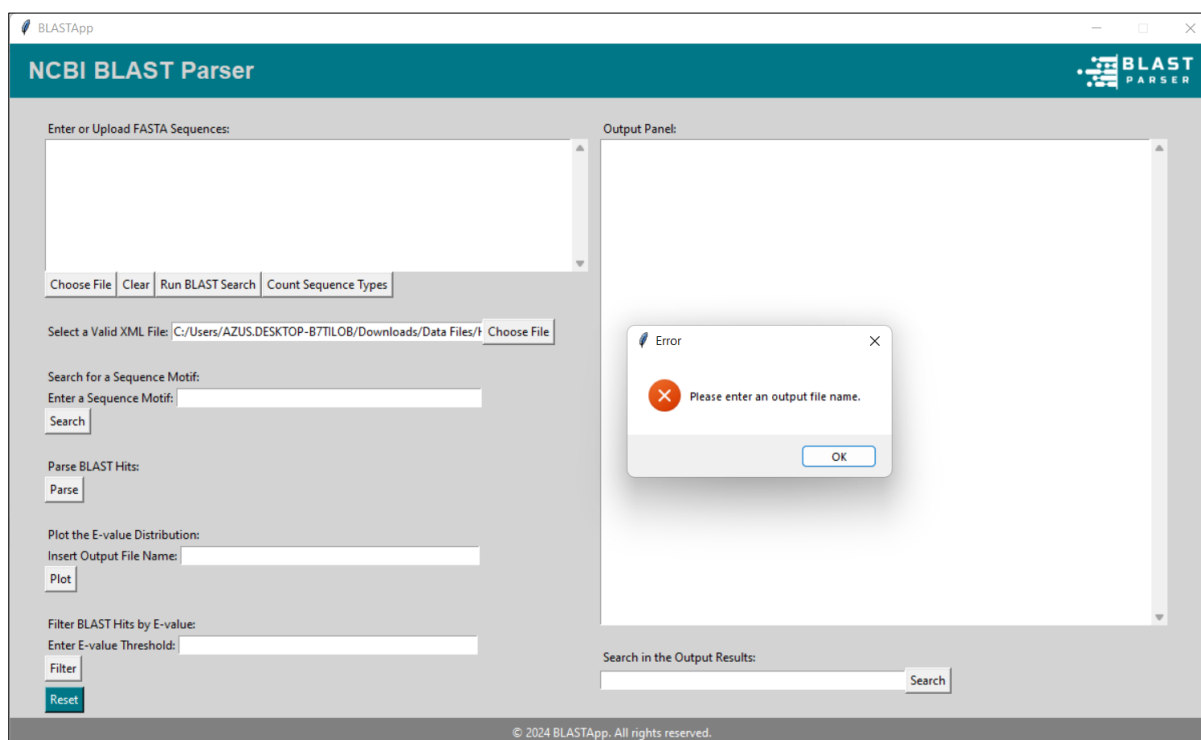- If successful, a success message will be displayed.



- The output text file will be saved in the same directory as the input XML file.

## 7.6 Plotting E-value Distribution

- Select a valid XML file containing BLAST hits by clicking the "Choose File" button in the "Plot the E-value Distribution" section.
- Enter an output file name in the "Insert Output File Name" field.
- Click the "Plot" button to plot the E-value distribution.
- If a user clicks the button without entering an output file name, an error message will be shown.



- If successful, a success message will be displayed.
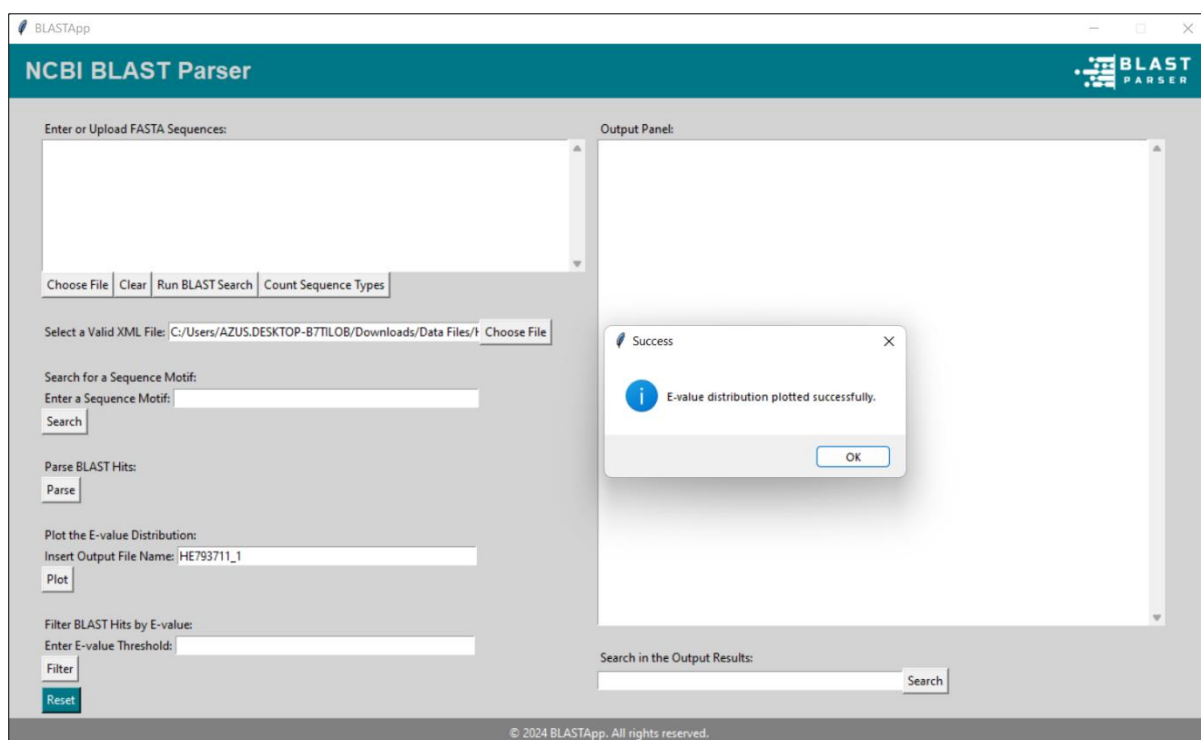
- The output file (.png) will be saved in the same directory as the input XML file.



| Name | Date modified | Type | Size |
|---|---|---|---|
| HE793711.1_blast_result.xml | 3/30/2024 10:50 PM | XML Document | 160 KB |
| HE793711.1_blast_result_filtered_hits.txt | 3/30/2024 11:06 PM | Text Document | 43 KB |
| HE793711.1_blast_result_parsed_output.txt | 3/30/2024 11:05 PM | Text Document | 40 KB |
| HE793711_1.png | 3/30/2024 11:06 PM | PNG File | 11 KB |
| HE980621.1_blast_result.xml | 3/30/2024 10:48 PM | XML Document | 127 KB |
| NM_000015.3_blast_result.xml | 3/30/2024 10:39 PM | XML Document | 322 KB |
| NP_176610.1_blast_result.xml | 3/30/2024 10:58 PM | XML Document | 136 KB |
| NP_001323428.1_blast_result.xml | 3/30/2024 10:45 PM | XML Document | 115 KB |
| sequences1.fasta | 3/30/2024 5:07 PM | FASTA File | 20 KB |
| sequences2.fasta | 3/30/2024 5:09 PM | FASTA File | 4 KB |

This PC > Documents > Data Files

## 7.7 Filtering BLAST Hits by E-value

- Select a valid XML file containing BLAST hits by clicking the "Choose File" button in the "Filter BLAST Hits by E-value" section.
- Enter an E-value threshold in the "Enter E-value Threshold" field.
- Click the "Filter" button to filter BLAST hits based on the E-value threshold.
- If a user clicks the button without entering an E-value threshold, an error message will be shown.



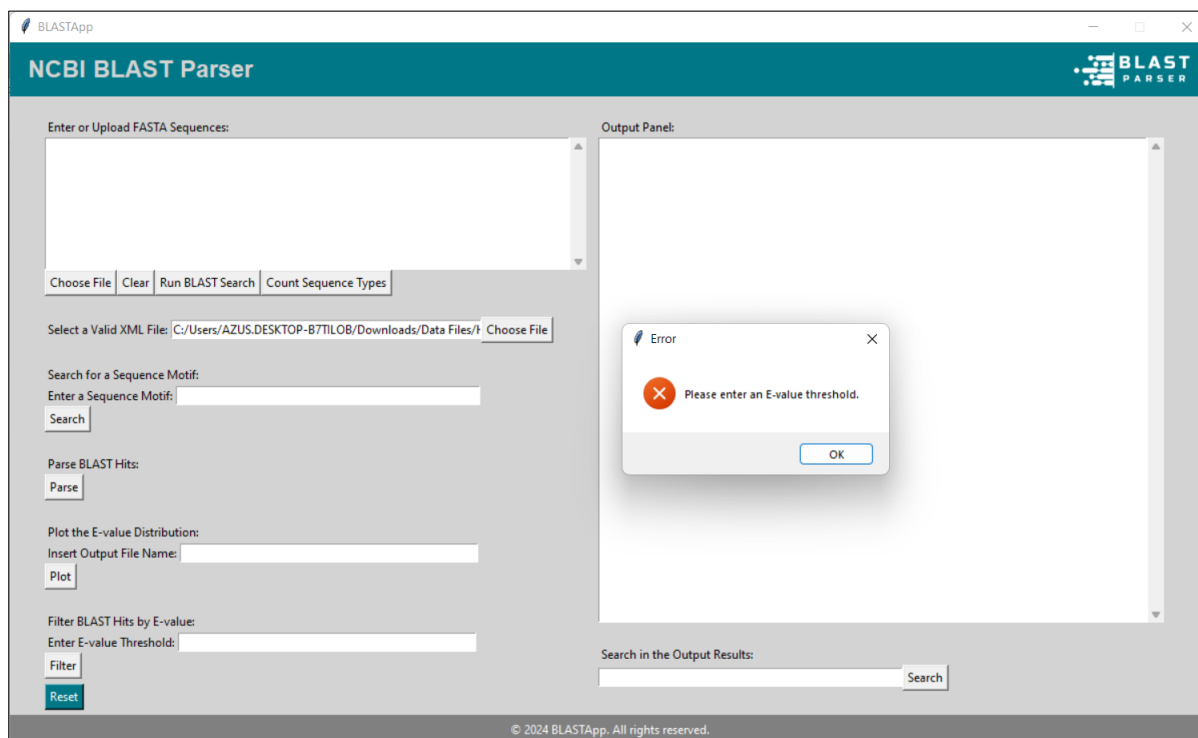- If successful, a success message will be displayed.

- The output text file will be saved in the same directory as the input XML file.
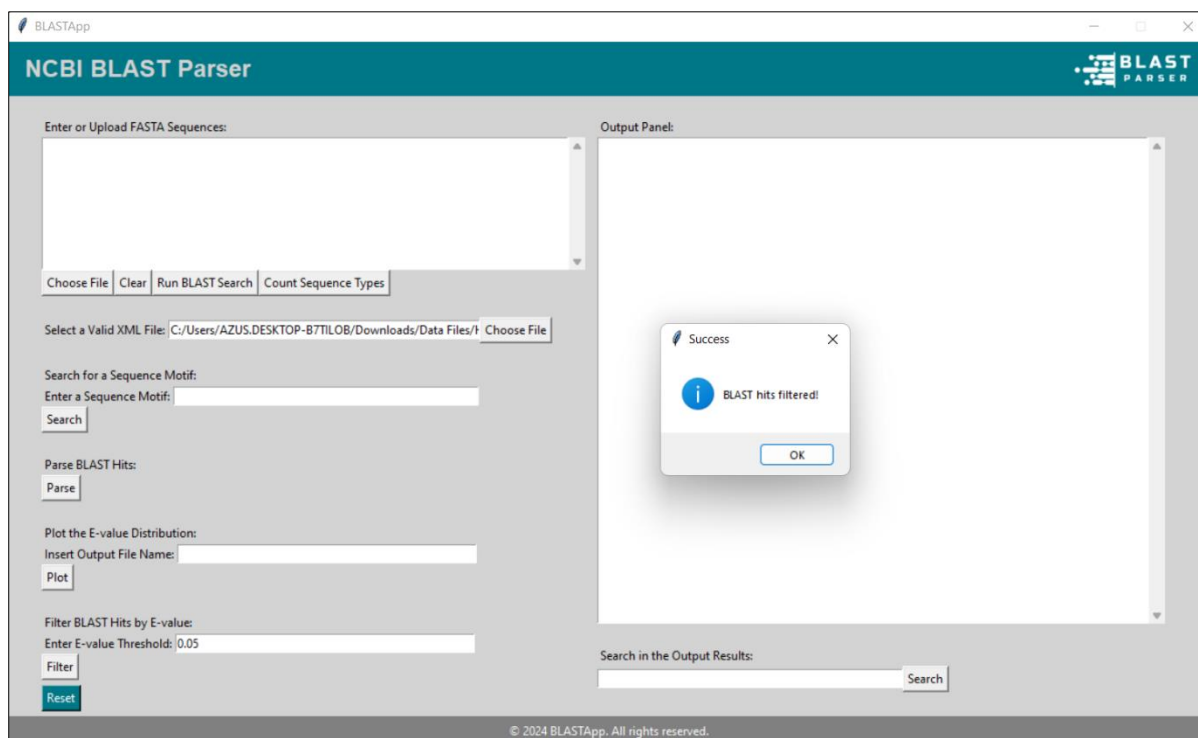
> This PC > Documents > Data Files

| Name ^ | Date modified | Type | Size |
|---|---|---|---|
| HE793711.1_blast_result.xml | 3/30/2024 10:50 PM | XML Document | 160 KB |
| HE793711.1_blast_result_filtered_hits.txt | 3/30/2024 11:06 PM | Text Document | 43 KB |
| HE793711.1_blast_result_parsed_output.txt | 3/30/2024 11:05 PM | Text Document | 40 KB |
| HE793711_1.png | 3/30/2024 11:06 PM | PNG File | 11 KB |
| HE980621.1_blast_result.xml | 3/30/2024 10:48 PM | XML Document | 127 KB |
| NM_000015.3_blast_result.xml | 3/30/2024 10:39 PM | XML Document | 322 KB |
| NP_176610.1_blast_result.xml | 3/30/2024 10:58 PM | XML Document | 136 KB |
| NP_001323428.1_blast_result.xml | 3/30/2024 10:45 PM | XML Document | 115 KB |
| sequences1.fasta | 3/30/2024 5:07 PM | FASTA File | 20 KB |
| sequences2.fasta | 3/30/2024 5:09 PM | FASTA File | 4 KB |

## 7.8 Searching in Output Results

- Enter the text you want to search for in the "Search in the Output Results" field.
- Click the "Search" button to highlight all occurrences of the entered text.



- If no matches are found, a message will be displayed indicating that no matches were found.

## 7.9 Resetting Inputs

- Click the "Reset" button to clear all input fields and the output panel.

## 8. Frequently Asked Questions (FAQ)

**Q1: How do I troubleshoot errors encountered during the BLAST search process?**
A: Troubleshooting errors during the BLAST search involves several steps. Ensuring that sequences are correctly formatted and compliant with required standards is crucial. Additionally, verifying a stable internet connection is essential as BLAST searches rely on accessing online databases. If issues persist, reviewing input parameters, error messages, and seeking assistance from project maintainers are recommended steps.

**Q2: What factors should I consider when interpreting the E-value distribution plot?**
A: Interpreting the E-value distribution plot involves understanding the statistical significance of BLAST results. Lower E-values signify higher significance, indicating that observed sequence similarities are less likely to occur by chance. Analyzing the distribution's shape and spread can help assess the reliability and significance of the matches identified.

**Q3: Can I adjust the filtering criteria for BLAST hits in BLASTApp?**
A: Yes, BLASTApp offers customization of filtering criteria, primarily through adjusting the E-value threshold. Users can tailor filtering parameters to their research objectives, focusing on matches meeting desired levels of confidence. Other criteria such as alignment length and sequence identity can also be customized.

**Q4: What steps can I take if my BLAST search results seem inaccurate or unexpected?**
A: Inaccurate or unexpected BLAST results may require reviewing input sequences, search parameters, and understanding the characteristics of the chosen BLAST algorithm. Verifying sequence quality, database selection, and considering alternative search strategies are advisable.

**Q5: How can I optimize BLAST search performance for large sequence datasets?**
A: Optimizing BLAST search performance involves several strategies, including preprocessing sequences, selecting appropriate BLAST algorithms, and leveraging parallel processing capabilities. Breaking down large datasets into smaller batches and utilizing optimized search parameters can also enhance efficiency.

**Q6: What are some best practices for preparing input sequences for BLAST searches?**
A: Best practices for preparing input sequences include ensuring correct formatting, providing informative sequence headers, and verifying sequence integrity. Cleaning sequences of artifacts or non-biological characters and validating sequence metadata enhance search accuracy.

**Q7: Can BLASTApp be used for analyzing sequences beyond nucleic acids and proteins?**
A: Yes, BLASTApp supports analysis of various biological sequences beyond nucleic acids and proteins, including RNA sequences, genomic DNA, and user-defined sequence data. Flexibility in input formats and customizable search parameters enable diverse sequence analysis applications.

## 9. Future Improvements

1. Parallel Processing
   Implement parallel processing techniques to improve the speed and efficiency of BLAST searches, especially when dealing with large datasets.

2. Integration with Cloud Services
   Allow users to use cloud computing resources for performing BLAST searches, which can handle larger datasets and complex analyses more efficiently.

3. Advanced Filtering Options
   Include additional filtering options for BLAST hits, such as filtering by percent identity, alignment length, or other relevant parameters.

4. Integration with Other Bioinformatics Tools
   Integrate the BLAST parser with other bioinformatics tools and databases to provide more comprehensive analyses and insights into the biological data.

5. Customizable Output Formats
   Allow users to customize the output formats for the parsed BLAST hits and other analysis results to suit their specific needs.

6. Optimization for Different Computing Environments
   Optimize the program to run efficiently on various computing environments, including different operating systems and hardware configurations.

7. Community Contributions and Feedback
   Encourage community contributions and feedback to continually improve the program, address bugs, and add new features based on user requirements and suggestions.