

Table 1: Revision History

<b>Date</b>	<b>Developer(s)</b>	<b>Change</b>
April 7th, 2025	Uriel Garcilazo Cruz	Initial version

# User Guide for SubLiMat Substitution Matrix Benchmarking Tool

Uriel Garcilazo Cruz

April 10, 2025

## 1 Introduction

SubLiMat is a bioinformatics tool for benchmarking substitution matrices using the Needleman-Wunsch global alignment algorithm. This guide provides instructions for installing and using the software.

## 2 System Requirements

- **Operating Systems:** Windows 10+, macOS 13+, or Linux
- **Python:** Version 3.8 or higher
- **Dependencies:** pandas, numpy

## 3 Installation

### 3.1 From Source

1. Clone the repository:

```
1 git clone https://github.com/UGarCil/UGarCil_capstone.git
```

2. Install dependencies:

```
1 pip install -r requirements.txt
```

## 4 Quick Start

### 4.1 Running the Program

Execute the main script from the command line:

```

1 from sublimat import file_manager
2 from sublimat.main import main
3
4 # To run the program:
5 # define the absolute path to:
6 # - substitution_matrices.txt
7 benchmark = "<path_to_your_file>/substitution_matrices.txt"
8 # - input_sequences.fasta
9 sequence_data = "<path_to_your_file>/input_sequences.fasta"
10 # - output directory
11 output_path = "<path_to_your_file>/data/"
12 # Execute the function composition export(main()) to export the
    results to the output directory
13 file_manager.export(main(sequence_data, benchmark), output_path)

```

Then execute the script using the Python interpreter. If your file is called main.py:

```

1 python main.py

```

## 4.2 File locations

You need to specify the locations of your input files. The suggested locations are as follows:

- Input sequences: data/input\_sequences.fasta
- Substitution matrices: data/substitution\_matrices.txt
- Results output: data/benchmark\_results.csv

## 5 Input File Formats

### 5.1 Sequence File (FASTA Format)

```

1 >Sequence1
2 ATGCGTACGT
3 >Sequence2
4 TGC GTACGTA

```

Requirements:

- Exactly two sequences
- Only characters A, T, C, G allowed
- Minimum length: 1 base pair

## 5.2 Substitution Matrix File

```
1 >Matrix1
2 1.0,-0.33,-0.33,-0.33
3 -0.33,1.0,-0.33,-0.33
4 -0.33,-0.33,1.0,-0.33
5 -0.33,-0.33,-0.33,1.0
6 >Matrix2
7 1.0,-1.0,-0.5,-1.0
8 -1.0,1.0,-1.0,-0.5
9 -0.5,-1.0,1.0,-1.0
10 -1.0,-0.5,-1.0,1.0
```

Requirements:

- Each matrix must be 4×4 (for A,T,G,C)
- Numeric values only
- One matrix per block (header + 4 lines)

## 6 Output

The program generates a CSV file with alignment results:

```
1 matrix,score
2 Matrix1,0.5
3 Matrix2,-12
4 ...
```

Columns:

- **matrix**: Name of substitution matrix
- **score**: Alignment score (lower is better)

## 7 Troubleshooting

Table 2: Common Issues and Solutions

Issue	Solution
"Invalid FASTA File Error"	Ensure file has exactly two sequences with headers starting with ">"
"Invalid Substitution Matrix"	Verify all matrices are 4×4 and contain only numbers
"Zero-length sequences"	Check that sequences contain at least one base pair

## 8 Examples

### 8.1 Basic Example

You can find a basic example of execution in the following demo: [HERE](#).

## 9 Contact

For support or questions, contact:

- Uriel Garcilazo Cruz
- Email: [garcilau@mcmaster.ca](mailto:garcilau@mcmaster.ca)
- GitHub: <https://github.com/UGarCil>