ContigPolishing
User Guide version 1.0

Dependencies

Before you can use ContigPolishing, you need to install the JDK on your machine. Installation varies according to the operation system used. You can download the JDK for the main operation system versions from Oracle's official website.

https://www.oracle.com/java/technologies/javase/jdk17-archive-downloads.html.

We recommend using version 17 or higher to ensure compatibility with ContigPolishing. Let's first check that you already have it installed, then follow the steps below:

How to check Java version?

First, open the terminal and run the command below to check that Java is installed and which version is active on your system.

```
allan@allan-Vostro-3470:~$ java --version
java 17.0.12 2024-07-16 LTS
Java(TM) SE Runtime Environment (build 17.0.12+8-LTS-286)
Java HotSpot(TM) 64-Bit Server VM (build 17.0.12+8-LTS-286, mixed mode, sharing)
allan@allan-Vostro-3470:~$
```

Warning: If it responds with something like this, you already have Java installed, skip to installing BLAST. Otherwise, follow the steps below.

How to install Java on Ubuntu Linux?

1- Step

Download the package version from website:

https://download.oracle.com/java/17/archive/jdk-17.0.12_linux-x64_bin.deb.

2 - Step

sudo apt install ./jdk-17.0.12_linux-x64_bin.deb

Now, to install BLAST+ (Basic Local Alignment Search Tool) from NCBI on Ubuntu, follow the steps:

1- Step
sudo apt update / sudo apt upgrade

2- Step
sudo apt -y install ncbi-blast+

The ContigPolishing

Let's run. See how below.

```
java -jar contigPolishing.jar
```

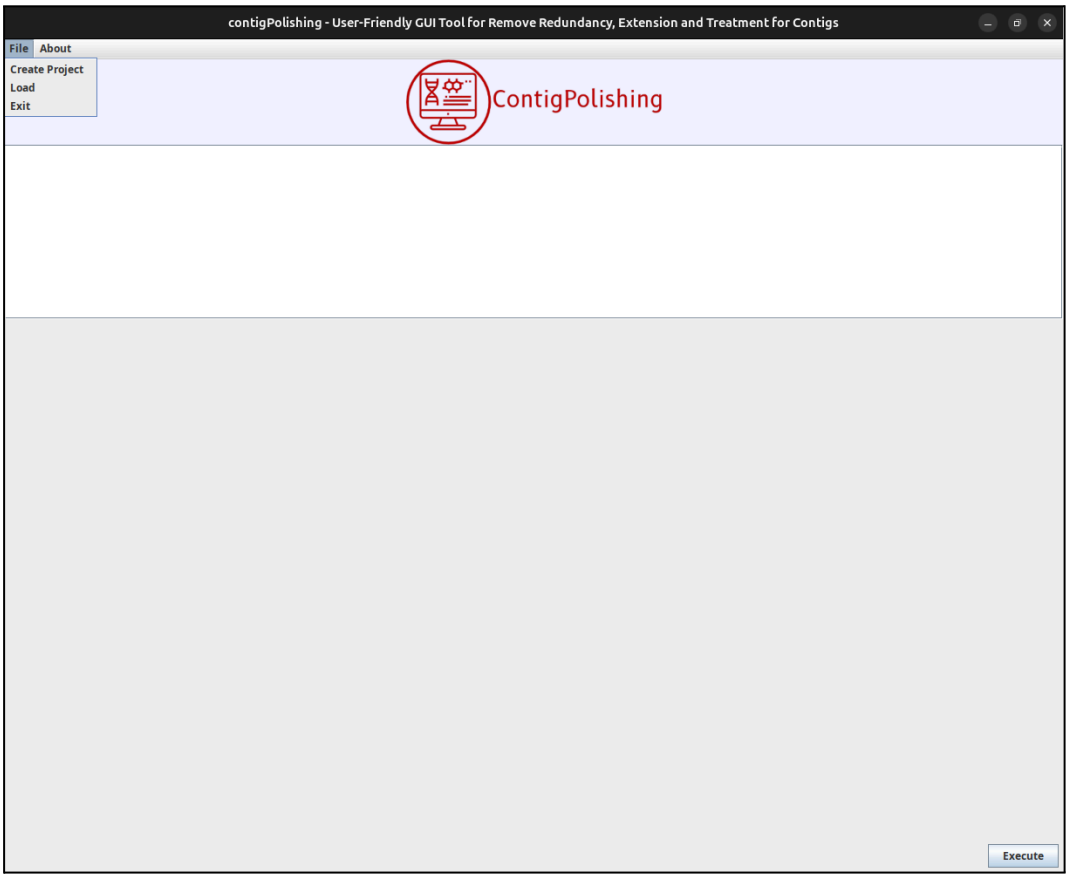
If you want to adjust the amount of memory available on your device, follow the example below

```
java -jar -Xmx16G contigPolishing.jar
```

Main Window

The Main Window is the starting point for interacting with the software. Here you can perform the following actions via the File menu:

1. Create a new project: This allows you to start a new project, setting it up according to your needs.
2. Load an existing project: Opens a previously saved project, allowing you to continue where you left off.
3. Stop running the software: Closes the software securely, ensuring that no information is lost.

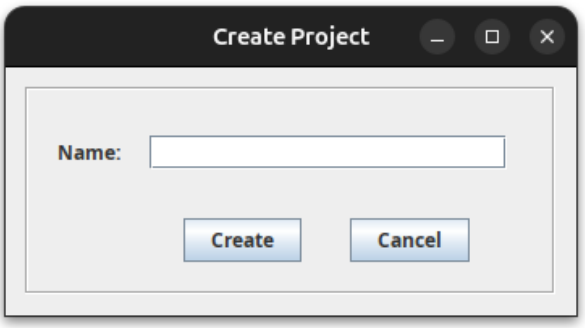


Creating your project

To create a new project, follow the steps below:

1. Click on the **File** menu on the main toolbar.
2. Select the **Create Project** option.
3. A new window will appear, allowing you to set up your project.
4. Enter the name of the project in the appropriate field.

After filling in the name, click on the **Create** button to finish creating the project.



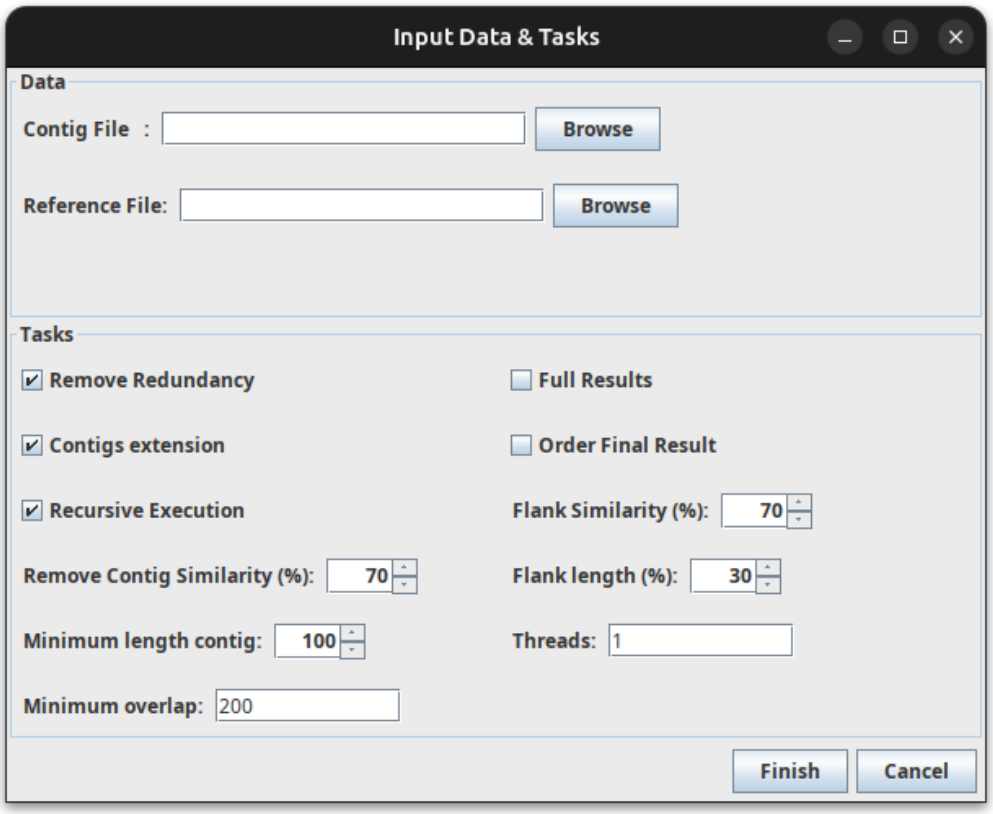
Input Data & Tasks window

After clicking the Create button, the user will be directed to the Input Data & Tasks window. At this stage, you need to provide the following information:

1. Input directory : Enter the path where the FASTA file(s) to be processed are located (the extension must be **.fasta**). Annotated reference file (optional).
2. If you wish to ordering the results, insert the reference file in GenBank format (extension .**gb**).

Important: If the ordering task is selected, the input folder must contain only the file you want to order. Otherwise, the software will process all the FASTA files found in the input folder using the same reference.

This feature has been developed to make it easier to process multiple datasets of the same genre efficiently.



In the same window (Input Data & Tasks), the user can configure the tasks to be performed and set the corresponding parameter values.

Available tasks:

- **Redundancy removal**: Eliminates redundant sequences to optimize results.
- **Contig extension**: Extends contiguous sequences based on the data provided.
- **Recursive execution**: Allows tasks to be executed iteratively until the defined criteria are reached.
- **Sorting the results**: Sorts the results using the GenBank reference file (if provided).
- **Complete or intermediate results**:
    - The user can choose to save only the final results.
    - Otherwise, the intermediate results will be discarded.

Configurable parameters:

- **Percentage of flank length:** Defines the minimum proportion of the flank that will be considered for extension or sorting.
- **Similarity percentage between flanks**: Specifies the minimum identity ratio for combining flanks.
- **Minimum contig length**: Determines the minimum size a contig must have to be considered.
- **Minimum overlap length:** Establishes the minimum number of bases that two sequences must share in order to be joined.

- **Number of threads (CPUs)**: Sets the number of threads the software can use to optimize performance.

Warning: If the user does not change any of the values, the default value for each parameter will be used.

Starting Processing

Once you have finished configuring the input data and setting the parameters, follow the steps below to start processing:

- Click the Finish button to confirm the settings in the Input Data & Tasks window.
- You will be redirected to the Main Window.
- In the Main Window, click the Execute button to start processing.

The software will start executing the selected tasks, applying the parameters set. Make sure all the settings are correct before starting the process.

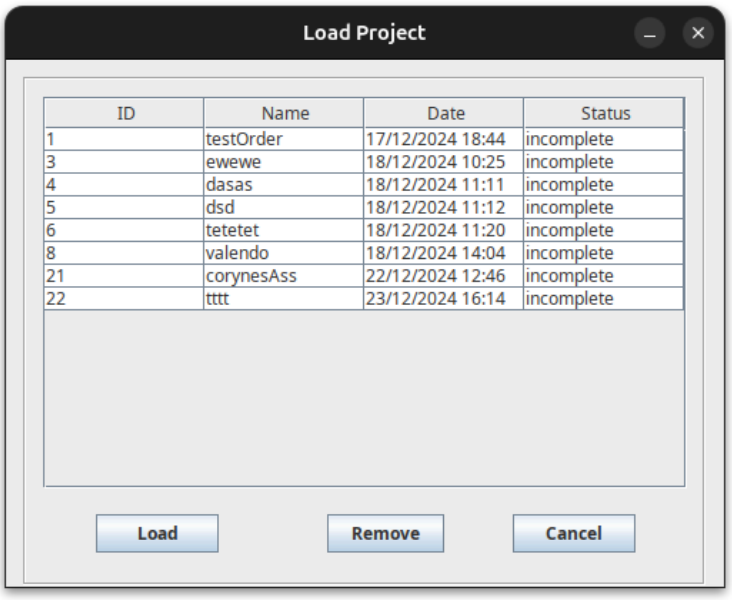
Load Window

The Load window allows the user to continue processing a previously created project or resume a project that has been interrupted.

How to load an existing project?

1. In the Main Window, click on the File menu.
2. Select the Load option.
3. A new window will appear containing a list of available incomplete projects.
4. Choose the desired project from the list and click Load.

The project will open, allowing the user to continue processing where they left off.



Loading or Removing Projects

To manage projects in the Load Window, follow the instructions below:

Load a project: Select the desired project from the list displayed. Click the Load button to open the project and continue processing.

Remove a project: Select the project you want to delete from the list. Click the Remove button to remove it permanently.

Warning: Make sure you really want to delete a project before confirming the removal, as this action is irreversible.

Monitoring processing

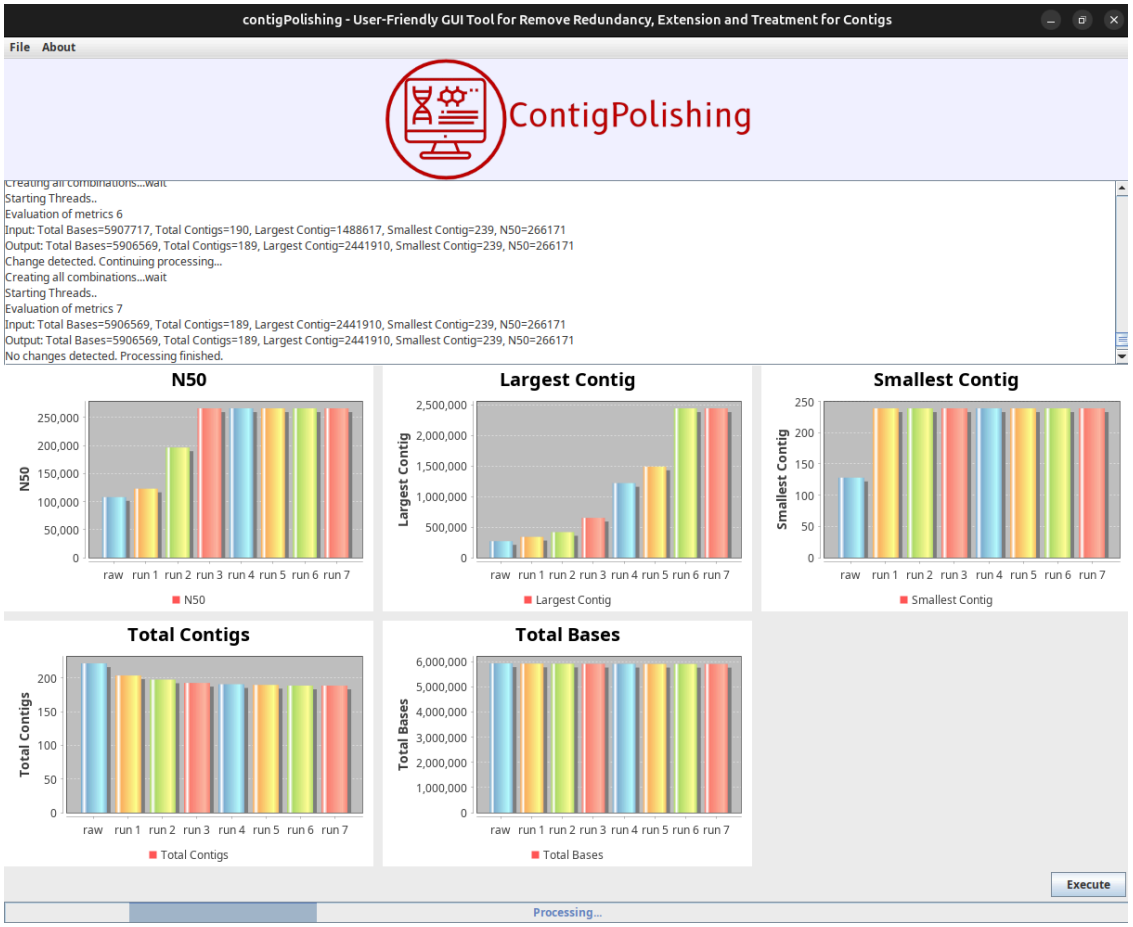
When you start processing the project, you can follow the progress in real time via the Log Area.

Log area: Displays detailed information on the progress of processing. It shows messages about each stage, allowing the user to identify possible errors or bottlenecks in the flow.

Graph display: After the FASTA file has finished processing, graphs with metric evaluations will be displayed automatically.

For multiple runs, the graphs will show a comparison between the raw dataset and the results processed in each run.

This visualization makes it easier to analyze the impact of each run on the metrics evaluated.



Expected file structure

- File with redundancy removed (ending with TratedCuckoo.fasta):

Example: 5275.2_ASM1659527v2_genomic_TratedCuckoo.fasta

- Final processing FASTA file (with the executions):

Exemple: GCA_016595275.2_ASM1659527v2_genomic_TratedCuckoo_run1_run2.fasta

- If it has been ordered, the extension will contain the name ordered.fasta:

Exemple: GCA_016595275.2_ASM1659527v2_genomic_TratedCuckoo_run1_run2_ordered.fasta

- PDF file as total report:

  A PDF file generating a final report of the entire processing.

PDF example:

ContigPolishing

Metrics Report

Analyzed organism GCA_022869005.1_ASM2286900v1_genomic

Metric analyzed raw data:

Smallest Contig = 1800, Total Bases = 3023073, N50 = 530434, Total Contigs = 17, Largest Contig = 713109

Metric analyzed by run 1:

Smallest Contig = 1800, Total Bases = 3018710, N50 = 662755, Total Contigs = 12, Largest Contig = 1271277

Metric analyzed by run 2:

Smallest Contig = 1800, Total Bases = 3013366, N50 = 662755, Total Contigs = 10, Largest Contig = 1280340

Metric analyzed by run 3:

Smallest Contig = 1800, Total Bases = 3012602, N50 = 699606, Total Contigs = 9, Largest Contig = 1280340

Metric analyzed by run 4:

Smallest Contig = 1800, Total Bases = 3012602, N50 = 699606, Total Contigs = 9, Largest Contig = 1280340