**SeReNA**

The software has been designed to retrieve, align and produce phylogenetic trees using CDS and protein sequences from an internal database of publicly available records. The database can be updated with new sequences imported locally, generated from an assembled genome or from a paired end Illumina dataset.

**Installation**

To install the software navigate inside the main folder and double click on the file install.command . A terminal will open, and all the required dependences will be installed. The following message will inform you that the installation is complete:

Text

Description automatically generated

Graphical user interface

Description automatically generated

You can now double-click on the SeReNA file to run the program. After few seconds the main window open:

Graphical user interface, waterfall chart

Description automatically generated

**Data structure**

Each entry in the database is expected to have 4 data available, that are (a) sample type, (b) sample number, (c) sample date and (d) sample compartment. If the sampling date is not available a generic string XXXXXX can be used, otherwise the date is expected to be provided in the format YYMMDD. Similarly, the letter X can be used if the compartment is not known. No information about sampling date and compartment are available for the entries reported in the database at the time of writing so XXXXXX and X are reported instead. The sample type can be any string. In this first release, only the samples downloaded from NCBI are present and for all of them I assigned a sample type code of “NCBI”. Finally, the sample number can be any string you want (not even necessarily a number), and for the NCBI samples reported in this database it corresponds to the NCB accession number.

**Retrieving CDS/Protein sequences**

To retrieve CDS or protein sequences, first select the sample type, followed by the sample number, the date and the compartment. All these fields accept multiple selections. Clicking will select/deselect the entry while clicking and dragging will select adjacent entries. Finally, click on the gene you want to retrieve the sequences for (again, multiple selections can be done at this stage). To import the sequences into the Preview box press either the add CDS or the add proteins buttons. The added sequences can be saved in fasta format using the menu File -> Save sequences (fasta).

The preview window is editable. This means that reported sequences can be modified/deleted while additional ones can be added. Operations performed in the Preview windows will not have any effect on the database.

**Adding new samples to the database**

Each sample is reported in a folder named after the sample name and placed in the DB folder of the distribution. Several methods can be used to add new samples to the database. After a new entry is added to the database please press File -> Update database to see the changes.

**Creating a blank entry.**

This can be done by clicking the menu item Insert -> New entry -> Blank entry. The following window will open:

Graphical user interface, text, application

Description automatically generated

Once the required data are added, a new empty entry is created upon clicking the “Create new blank entry”. A blank entry is just an empty folder named as the sample name in the DB folder, but it is necessary to add single CDS or protein sequences to the database.

**Add a single CDS or protein sequence**

A single CDS or protein sequence can be added by clicking the menu item Insert -> New sequence -> Nucleotide or Insert -> New sequence -> Protein respectively. This will result in the following window to open:

Graphical user interface, application

Description automatically generated

First insert the data relative to the sample (type/name/date/compartment) you want to add the CDS or the protein to, then select the name of the gene you are adding the sequence for from the Gene drop-down menu. Finally add the sequence (just the sequence, not fasta format header) and click the “Add sequence to DB” button. If the sequence is already present in the database (e.g. you are replacing a sequence with an updated version), you will be warned that the old sequence will be overwritten and will be given the possibility to cancel the operation.

An error message is returned if the sample with the chosen name/type/date/compartment is not present in the database.

**Add new CDS and proteins starting from a genomic sequence**

You can use this option If you have a new genomic sequence and want to automatically annotate and add both CDS and proteins to the database. Using a complete genome is not mandatory as the software will only annotate what it finds. This module can be accessed by clicking the menu item Insert -> New Entry -> From genome (fasta) which will open the following window:

Graphical user interface, text, application

Description automatically generated

Again, all the data about the sample name/type/date/compartment must be provided. In this case, you don’t need to have a blank entry ready as this module will create the entry once the CDS/proteins have been annotated. If an entry with the same sample data is present in the database it will be overwritten (no warning message in this release).

After selecting the genome file (in fasta format) using the “open file” button, press the “Retrieve sequences” button to start the annotation. The annotation steps will be shown in the Log window. This module uses the same algorithm GRACy uses to annotate a new genome and can take between 30 and 60 minutes depending on the machine used to run it.

**Add a new entry directly from paired end fastq files.**

This module can be used to annotate CDS and proteins directly from a paired end Illumina dataset without prior availability of an assembled genome. In this case, for each gene, the reads are aligned to a collection of deposited sequences and those mapping are extracted from the initial dataset. Such extracted reads are used to perform a de novo assembly of the single genes, while using the deposited sequences as a guide. Given that this process involves a de novo assembly of 170 genes, the user can expect the annotation to take a while to complete (even several hours).

This module can be accessed by clicking on the menu Insert -> New entry -> From reads (fastq) which will open the following window:

Chart, waterfall chart

Description automatically generated

As usual the sample data needs to be provided (any sample in the database with the same data will be overwritten). The reads1 and reads2 need to be provided using the corresponding “Open” button and the number of threads to be used can be specified in the box. After clicking the button “Run” the annotation will start and the process progress will be reported in the Log window. The “Found” box will report the genes for which a CDS has been successfully assembled, “Not found” will report those missing and “Pseudo” will report the genes for which a CDS was found but featuring some disrupting mutation or missing start or stop codons (incomplete sequence).

**Add a new entry from an NCBI accession number**

If a new sequence becomes available on NCBI the corresponding CDS and proteins can be downloaded directly on the database. This can be done by clicking on the menu Insert -> New Entry -> From NCBI accession number which will open the following window:

Graphical user interface, application

Description automatically generated

Just fill-up the required fields and press the “Get record” button to retrieve the sequences and add them to the database.

Graphical user interface, application

Description automatically generated

**Modify an entry**

If you want to modify the metadata of a sample such as the sample name, date or compartment you can do so by clicking the menu Edit -> Modify entry which will open the window on the right

Just insert the old sample data (top) and the new (bottom) and click on the “Update record” button to modify an entry metadata.

**Remove an entry**

Click the Edit -> Remove entry to permanently delete an entry from the database. More than one entry can be selected at this stage.

**Sequence alignment**

After adding CDS or proteins to the preview box in the main window, the sequences can be aligned by clicking the “Align” button. This will perform a clustal omega alignment and reports the result in the preview box. This alignment can be saved by clicking on the menu File -> Save alignment (fasta) or can be copied and pasted in your favourite alignment viewer software.

**Homology tree**

Once the CDS or protein sequences have been added to the Preview window, a phylogenetic tree can be built by clicking on the “Tree” button. When doing this, a clustal omega alignment will be performed on the sequences, followed by a tree built with the tool fasttree. A window will open reporting the tree which can be navigated using the tool bar on the top. The graphic of this representation is not great. However, after constructing the tree, the newick formatted tree will appear in the Preview window and this can be copied in your favourite software (e.g. Mega). The newick file can also be saved into a file using the menu File -> Save -> Save tree.