USER GUIDE

CODON

Software to manual curation of prokaryotic genomes

Postgraduate Program in Applied Computing
Federal University of Pará (UFPA-CAMTUC)

What is CODON?

The CODON software is a tool for manual curation of genomic data, capable of performing the prediction and annotation process. This software makes use of a finite state machine in the prediction process and automatically annotates products based on information obtained from the Uniprot database.

1. Running CODON

1.1 Requirements

Java Runtime Environment 1.8.0 or superior.

1.2 To Running

All files needed to run CODON are in the dist folder. For Windows users, double-click the **codon.bat** file and for Linux users, give the necessary permissions to execute (for instance, chmod -R 777 dist) then execute the **codon.sh** file.

1.3 Input

The CODON accepts a FASTA file as a minimum required input, and users can customize parameters to the prediction and annotation process. In the case of reannotation, CODON can also take as input an EMBL file generated by another prediction and annotation tool.

1.4 STEP 1: Creating a workspace directory:

The user can create a specific directory (like Figure below) for his workspace or use the tool's default. The workspace will store all files used during the analysis (mainly all data retrieved from UniProt during the annotation process, rRNA e tRNA recognized, and some configuration data). The directory may growth significantly during the use time.

Figure 1. Default directory

1.5 STEP 2: The tasks

CODON software has two different ways to start a new project. The first one initiates from a FASTA file in order to perform both prediction and annotation processes.

The second one initiates from an EMBL file so as to re-annotate the predicted ORFs marked into the file. At the end of both task types, the annotation result can be saved as a CODON project for a further manual curation or exported as EMBL to be cured in another tool.

An interrupted task will continue while restarting it.

The initial Dialog box enables to select/change the workspace, to initiates a project from a FASTA or EMBL file, or to reload a CODON project.

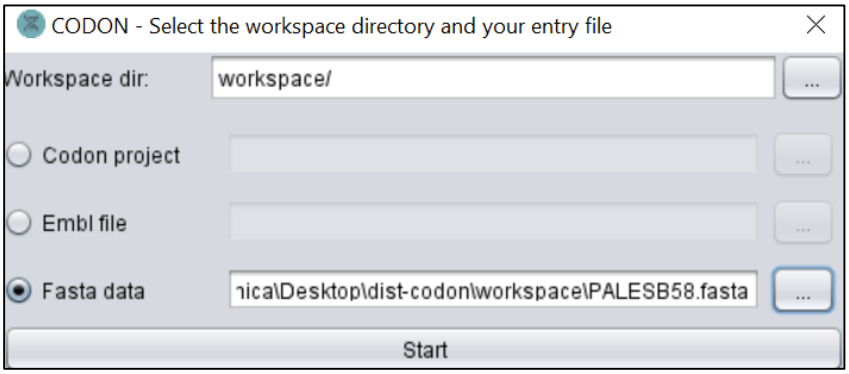

Figure 2. Input file

2. Basic usage scenarios

Scenario 1. Annotation from a FASTA file

1. Select a FASTA (section 1.5) and Start the project. Every possible ORF will be predicted and displayed as black boxes (section 4.2).
2. Recognize tRNA and rRNA ORFs, menu “Action Match tRNA” (section 5.3).
3. Select the menu “Optimized Blast with Uniprot” (4.3) and input “15” in the field simultaneous request in the opened Blasting options dialog box.
4. Start blasting. The annotation may take several days. Depending on the organism, about 20% of the ORFs will be blasted. The other will be discarded by the optimization algorithm during the process. **Case the annotation process stop. In such a situation you will have to restart CODON and redo steps 1, 3 and 4 (no need to perform step 2 another time). The annotation will restart where it has been stopped.**
5. After the end of the blast, remove the ORF with low accuracy, menu “curation ☢ Remove low accuracy” (Section 5.4).
6. Remove the overlaps that can be resolved by the CODON decision process, menu “curation" Remove overlaps” (Section 5.4).
7. Reduce the intergenic regions, menu “curation → Select entries to fulfill intergenic regions” (Section 5.4).

8. Save the project as CODON project, menu “file → Save CODON project” (Section 5.1).

   The project is ready to be cured manually.

9. You can use the CODON interface to cure the annotation or export the file as EMBL to cure it in another tool. Note that the exportation as EMBL will export only the ORFs that have not been removed during the previous step, while a CODON project save the situation for every ORF, even the removed and unblasted ones that can be analyzed using the menu “View → Hide removed orfs”.
10. The graphic interface enable analyze the remaining overlap situation and bring you graphical information from UniProt database to decide what to do: resizing an ORF (menu “Selection → Move start to the left/right”, section 5.4), removing an ORF (menu “Selection → Force to remove”, section 5.4), selecting another entry for the ORF (Section 5.), etc.
11. It also enables exploration of the Intergenic Regions (section 6.2), comment ORF (Section 6.3) and edit the sequence (Section 7.)

Scenario 2. Reannotation from a FASTA file

1. Select an EMBL (section 1.5) and Start the project. Every possible ORF predicted into the EBML file will be displayed as black boxes (section 4.2).
2. Recognize tRNA and rRNA ORFs, menu “Action Match tRNA” (section 5.3).
3. Select the menu “Full Blast with Uniprot” (4.3) and input “15” in the field simultaneous request in the opened Blasting options dialog box.
4. Start blasting. The time to perform annotation process depend to internet services and amount of product into the organism. The annotation may take several days. **It is recommended to monitor (section 6.4) the process every 4h. In such a situation you will have to restart CODON and redo the steps 1, 3 and 4 (no need to perform the step 2 another time). The annotation will restart where it has been stopped.**
5. Perform an automatic resize of every ORF, menu “curation→Resize using alternative start” (Section 5.4).
6. After the end of the blast, remove the ORF with low accuracy, menu “curation→Remove low accuracy” (Section 5.4).
7. Remove the overlaps that can be resolved by the CODON decision process, menu “curation" Remove overlaps” (Section 5.4).
8. Reduce the intergenic regions, menu “curation →Select entries to fulfill intergenic regions” (Section 5.4).
9. Save the project as CODON project, menu “file" Save CODON project” (Section 5.1).

   The project is ready to be cured manually. (see 9/10/11 of the scenario 1) scenario 1)

3. Overview

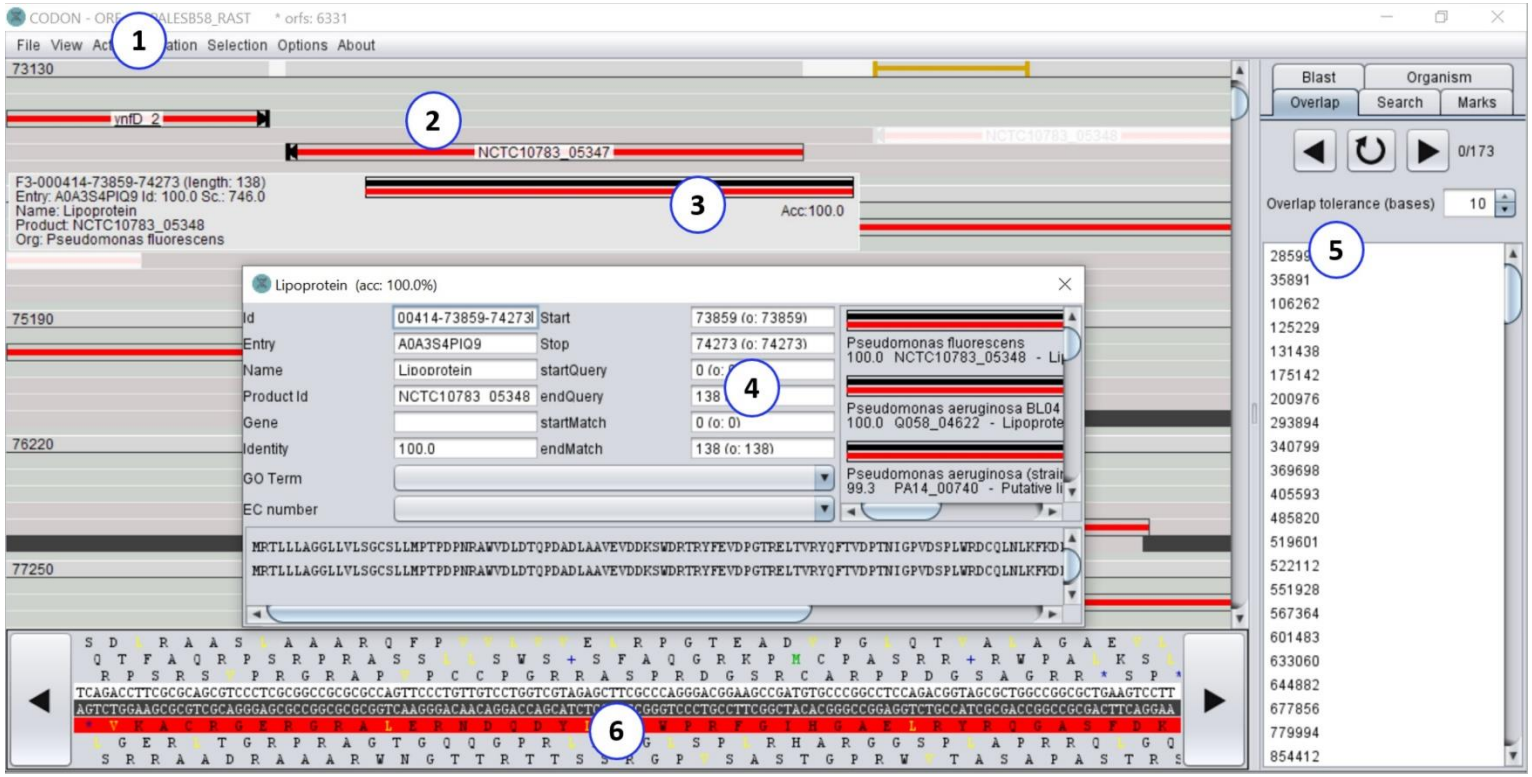

Figure 3. CODON Overview

1: **Main Menu**
Shows all options available in the tool.

2: **Main view**
Displays a compact result of the prediction and annotation process. Every predicted ORF can be analyzed.

3: **ORF Annotation Information**
Displays basic information about the ORF (product name, acronym gene, organism entry math, and percentage of identity) when the ORF is selected by a click.

4: **ORF Details**
Displays more details about the ORF and enables to explore deeper the annotation result. To view this option, double click on the ORF.

5: **Side bar**
Enables to monitor the annotation task progression and to explore the results.

6: **Sequence details view**
Sequence details view enables to analyze accurately specifics subsequences and to edit the sequence.

4. Main View

The main view displays the predicted ORF (or the ORF loaded from an EMBL file), the user can follow the annotation process in each ORF and later perform the manual curation process.

Figure 4. ORFs

In the figure below, the white areas represent the intergenic regions, the overlapping areas are represented by a darker gray tone, however, the user has a list with all ORF's in the overlapping region. The blue markings represent the identified tRNA and rRNA.

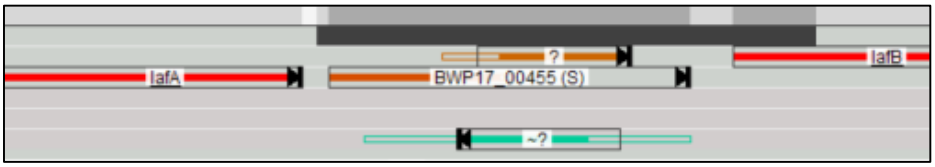

Figure 5. Frame strips

GUI color pallete

Initially, after the process of identification of the ORF's they have a rectangular shape in dark gray color. However, as the annotation process takes place, ORF's are represented using the same color palette used in the Uniprot database. It is a gradient between RED (100%) and GREEN (50%) when the identity is up to 50%, and between GREEN (50%) and BLUE (0%) when the identity is down to 50%. The part of the colored strip filled materializes the query and subject subsequences that are matching together. The part of the colored strip unfilled materializes the rest of the subject (the subject may be higher than the query).

Note: Unblasted ORF are displayed by a simple filled black box.

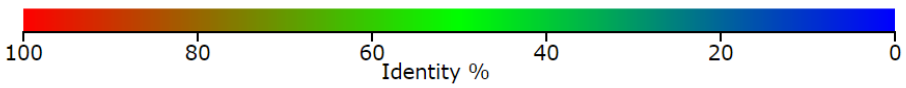

Figure 6. Color scale

5. Main Menu

5.1 File menu

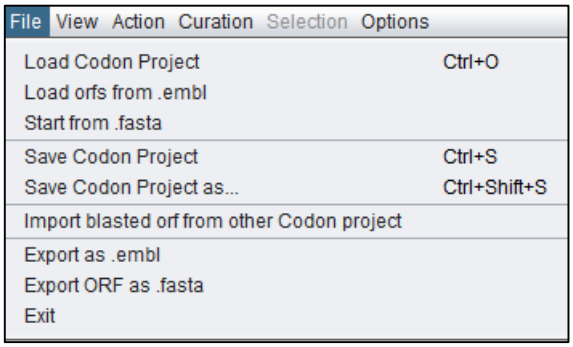

Figure 7. File menu

- **Load Codon Project (Ctrl+O):** Load a CODON project created previously.
- **Load From ORFS:** Load the ORF coordinates from the EMBL file.
- **Start from. FASTA*:*** Start a new project from a FASTA file.
- **Save Codon Project (Ctrl+S):** Save the CODON project.
- **Save Codon Project as (Ctrl+Shift+S):** Save as Project CODON with a new name.
- **Import blasted orf from other Codon project:** Load information about annotation from another CODON project.
- **Export as .embl :** Export as EMBL file.
- **Export ORF as .fasta :** Export the ORF on FASTA file.
- **Exit:** Leave the application.

5.2 View menu

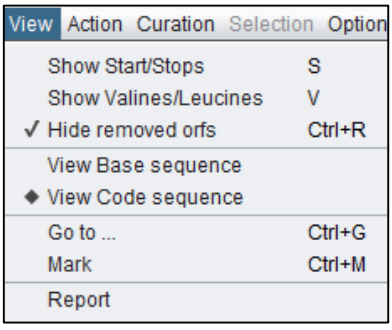

Figure 8. View menu

- **Show Start/Stops (S):** Displays all start (Methionine) and stop codons into the Main view. Green and Blue dashes represent the start and stop codon respectively.

- **Show Valines (V):** Displays all alternative start (valines and leucines) as yellow dashes into the Main View.
- **Hide removed orfs (Ctrl+R):** Hide/Show all ORF discarded during the automatic annotation process and during the manual curation.
- **View Bases sequence:** Displays the subject and query sequences in nucleotide format into the dialog box showing the ORF details.
- **View Codes sequence:** Displays the subject and query sequences in amino acid format into the dialog box showing the ORF details.
- **Go to (Ctrl+G) :** Navigate to a specific coordinate.
- **Mark (Ctrl+M) :** Add a comment to the ORF.
- **Report:** Show a statistical report for the annotation result.

5.3 Action menu

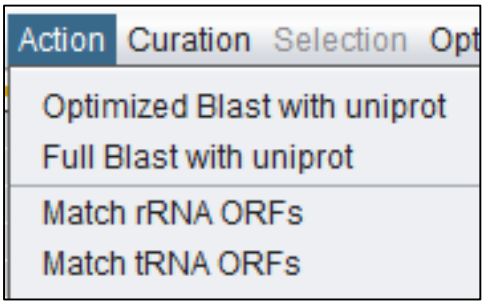

Figure 9. Action menu

- **Optimized Blast with uniprot:** Starts the search for similarity (blast) in the Uniprot database. In this option, BLAST will be executed following the strategy adopted in the tool to optimize the process. Thus, the algorithm will select the most relevant ORFs excluding the least according to the tool's metrics.**This option may be preferred when starting a project from a FASTA File.**
- **Full Blast with uniprot:** Starts the search for similarity (blast) in the Uniprot database for all the predicted ORFs. **This option may be preferred when starting a project from an EMBL File to reannotation process.**
- **Match/Re-match rRNA ORFs:** Annotate/re-annotate the rRNA feature.
- **Match/Re-macth tRNA ORFs:** Annotate/re-annotate the tRNA feature.

5.3.1 Blasting options dialog box

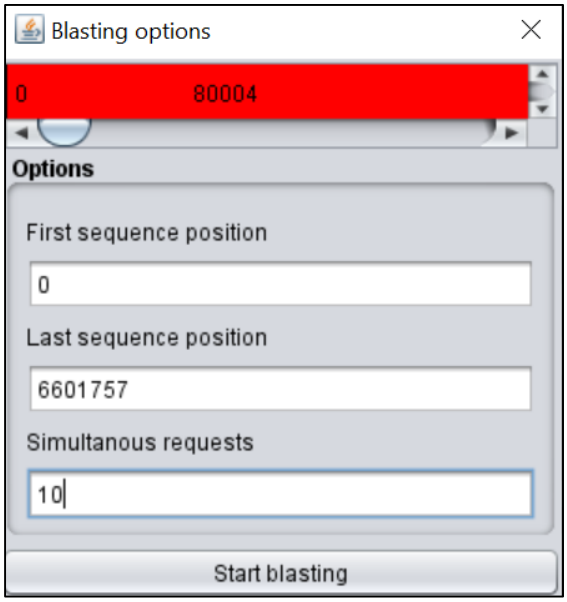

Figure 10. Blasting options

When selecting the options Optimized Blast with UniProt or Full Blast with UniProt. The user has the option of performing BLAST in the entire genomic sequence or can distribute this process between more than one computer. A dialog box enables to select the initial and final coordinates that will be processed in this CODON instance. On the same screen, the user can choose the number of strings sent to BLAST simultaneously. **The maximal simultaneous blast that can be performed depends on the user's Internet bandwidth and on the Uniprot workload. The performance tends to increase while the simultaneous BLAST increases until reaching a top and then decreases while the simultaneous BLAST continues increasing. It also may change over time.**

5.4 Curation menu

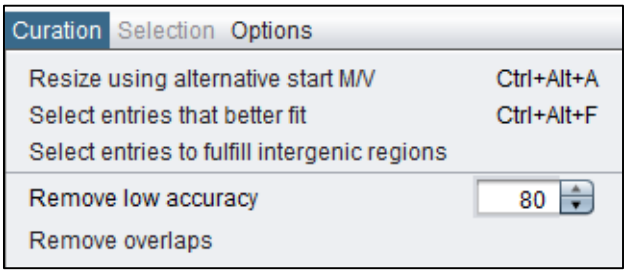

Figure 11. Curation menu

This menu process filters/algorithms for every unremoved ORF.

- **Resize using alternative start M/V (Ctrl+Alt+A):** When possible, it adjusts the length of the query by selecting an alternative start to make the query better matching with the subject obtained from the Uniprot database.
- **Select entries that better fit (Ctrl+Alt+F):** The blast result for an ORF may suggest several different alternatives (entries). It selects the most appropriate entry for an ORF without allowing resizing the ORF.
- **Select entries to fulfill intergenic regions:** The result for an ORF may suggest several different alternatives (entries). It selects the best entry allowing resizing the ORF and trying to maximize the use of intergenic areas.
- **Remove low accuracy:** Removes the ORFs that have a low accuracy according to the tool's metrics.
- **Remove overlaps**: Makes adjustments to minimize the cases of overlapping areas by:
    - Selecting an alternative entry when possible for one of the both ORFs creating the overlap region;
    - Removing one of the ORF when the accuracy or pertinence for the discarded one is lower than for the other.

5.5 Selection menu

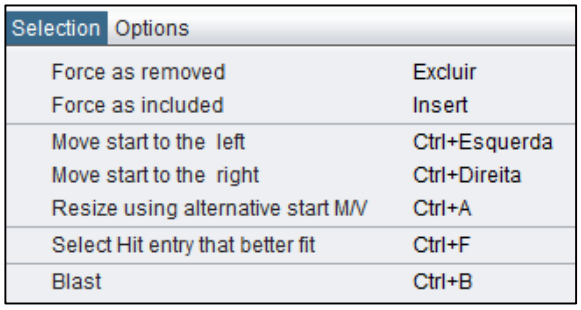

Figure 12: Selection menu

This menu is activated when an ORF is selected in the Main view. Every action of this menu will be applied exclusively to the selected ORF.

- **Force as removed:** Option if the user wants to permanently remove an ORF, so the filters/algorithms will ignore it later, not allowing it to return into the analysis without a reverse action performed by the user.
- **Force as included:** An included ORF will not be removed by an automatic filter/algorithm and may be removed by a remove action performed by the user.
- **Move start to the left:** Change the start of the ORF by selecting the next alternative start (M, V ou L) at the right of the actual start.
- **Move start to the right:** Change the start of the ORF by selecting the next alternative start (M, V ou L) at the left of the actual start.
- **Resize using alternative start M/V (Ctrl+A):** When possible, it adjusts the length of the query by selecting an alternative start to make the query better matching with the subject obtained from the Uniprot database.
- **Select Hit entry that better fit (Ctrl+F):** The blast result for an ORF may suggest several different alternatives (entries). It selects the most appropriate entry for an ORF without allowing to resize the ORF.
- **Blast (Ctrl+B):** Running BLAST for one selected ORF.

5.6 Options -→ Parameters menu

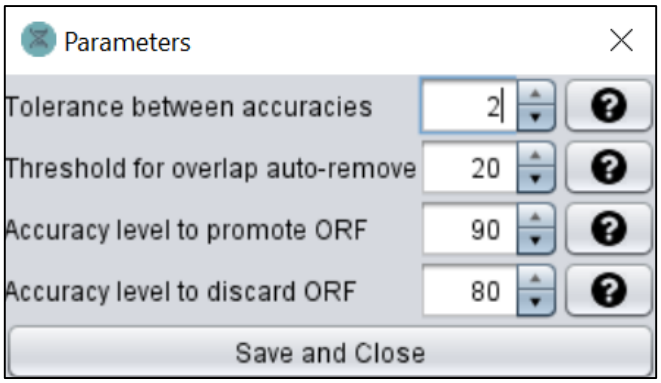

Figura 13. Parameters

- **Tolerance between accuracies:** The parameter is used to compare entries. When the accuracy difference is lower than this threshold, the difference is not considered significant. For instance, an entry with a lower accuracy than another (with accuracy difference down to this threshold), but that matches with a gene when the others do not, will be considered better. If the difference is up to this threshold, the entry with the higher accuracy will be considered as better independently of other comparison criterion.
- **Threshold for overlap auto-remove:** If Overlap percentage is lower than this threshold, the Overlap suppressing filter will try to resize the ORFs using alternative entry, but will not remove completely any ORF.
- **Accuracy level to promote ORF:** During the blast, if the accuracy of a blasted ORF is higher than this threshold and if the ORF is characterized, the algorithm will try to remove the overlaps by resizing or removing the other unblasted ORF.
- **Accuracy level to discard ORF:** During the blast, if the accuracy of a blasted ORF is higher than the Promote threshold and if the ORF is characterized, the algorithm will try to remove the ORF than provoke an Overlap and that have an accuracy lower than his threshold.

For the examples contained in this manual, the accepted parameters were in accordance with the standards of the tool, but if the user wishes to adjust them, this action must be performed before the execution of the Blast step.

6. ORF Details dialog

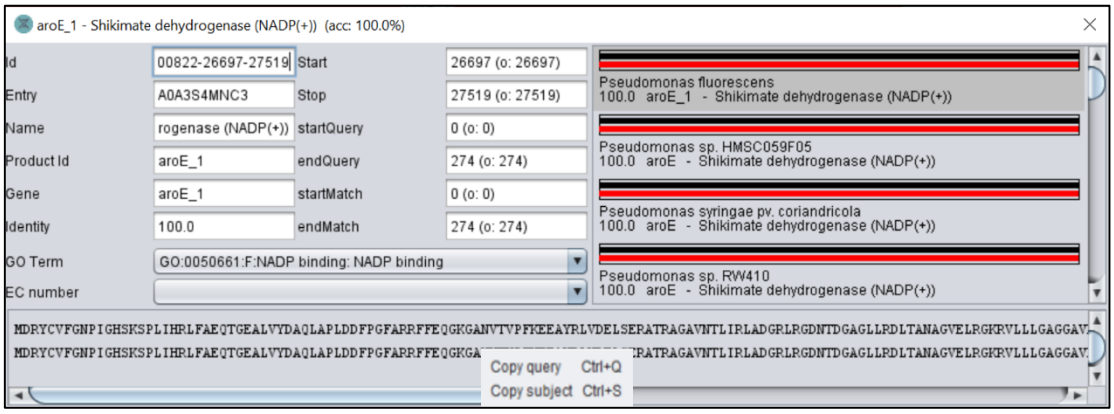

Figure 14. ORF details

This dialog box displays the ORF details resulting from the blast for the current selected entry. The right panel displays every hit entry from the blast result. Another entry can be manually selected in this panel updating. The selection updates the dialog box data.

The dialog box also enables to copy the query and the subject of the current selected entry into the clipboard

7. Side bar

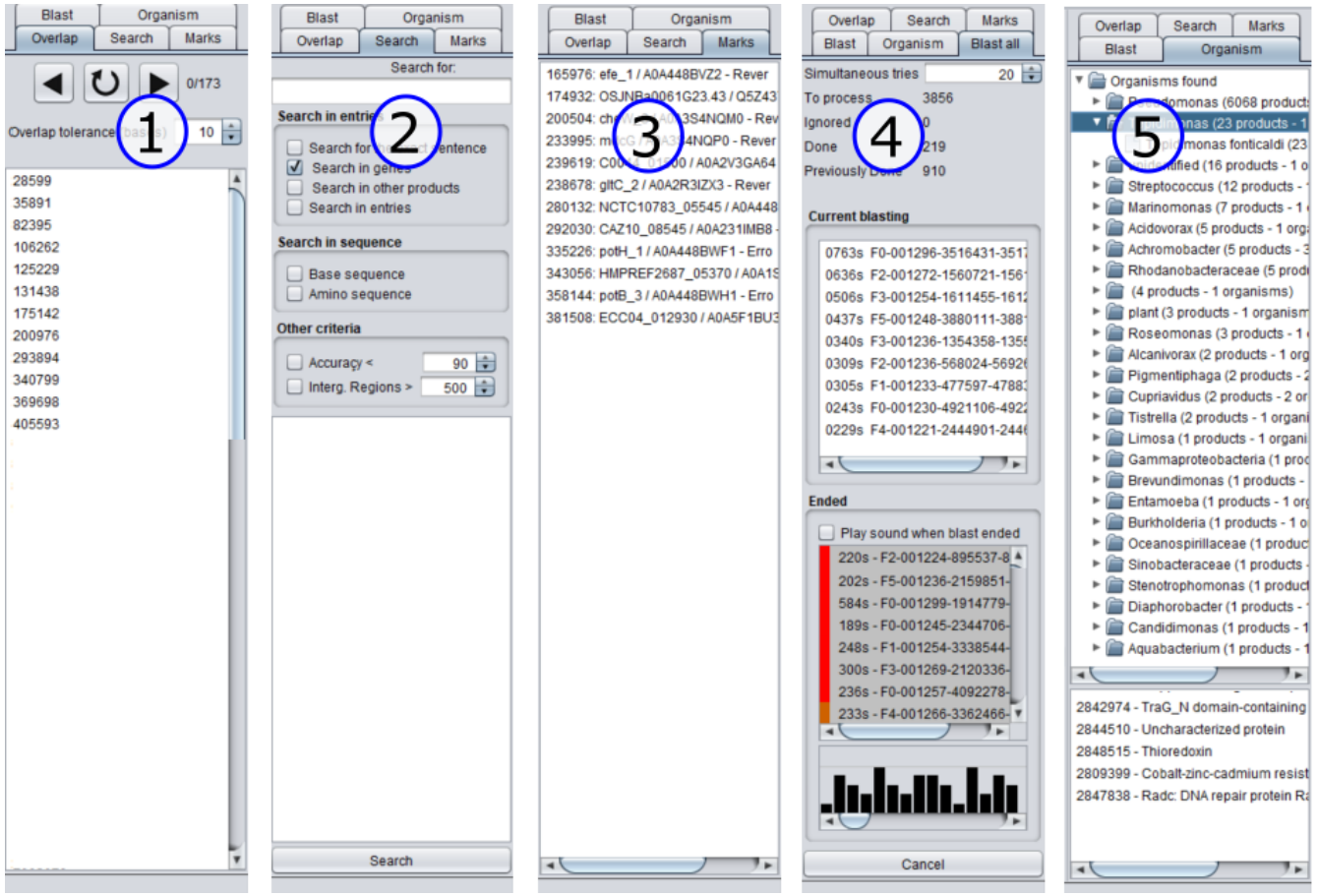

Figure 15. Side bar

1: **Overlap navigator**
Enables to navigate into the overlap areas filtered by a minimum size.

2: **Search panel**
Tools to perform searches in annotation result or into the nucleotides/amino sequence

3: **Mark's navigator**
Enables to navigate into the ORF marked by the user

4: **Blast/BlastAll monitors**
Monitor the annotation process.

5: **Organism navigator**
Displays all organisms and species identified during the annotation.

7.1 Overlap navigator

The overlap navigator provides a shortcut to access the nucleotide sequence regions where at least two ORF are overlapped. To be considered into this filter, the overlapped spans must be higher than a threshold specified by the user (default value: overlap of 10 bases).

7.2 Search panel

The search panel offers searching options enabling to localize:

- ORF with characteristics such as:
    - gene/entry/name having containing a specific string
    - Accuracy lower than a user specified threshold
- Sub-sequences matching with a nucleotide string or amino acid string
- Intergenic areas higher than a specified threshold

7.3 Mark's navigator

The mark's navigator provides shortcuts enabling access to previous marked/commented ORF.

7.4 Blast/BlastAll monitors

This panel enables the blast processes evolution.

The first parameter enables to alter the number of simultaneous blasts performed (see 4.3.1).

The following data illustrates:

- How many blast remains to be blasted (depending on the blasting strategy, many ORF may be ignored afterward. See 4.3).
- How many ORFs have been ignored/discarded during the whole blasting process.
- How many ORFs were processed during this session.
- How many ORFs had ever been blasted by a previous interrupted session.

A first list monitors what are the ORFs currently blasted. A double click into an item select the corresponding ORF into the main view.

A second list monitors the ORFs that have ever been blasted. A double click on an item selects the blasted ORF into the main view. The header of the item, shown by a colored box, illustrates if the accuracy of the better blast entry is high or not (According to the color scale presented in 4.3). The header also characterized the nature of the product identified: if the product is an uncharacterized protein, the box display a "?"; if the product is characterized, whereas is not an identified gene, the box displays by a "P" (Products); instead, the box is empty but shorter, and highlighted because of it, identifying a gene.

At last, a bar graphics illustrates how many blasts were successfully performed (one bar per minute) to monitor eventual internet troubleshooting or Uniprot overload needing to reconsider the number of simultaneous blasts performed.

7.5 Organism navigator

The tab lists which organisms were encountered into for the annotation results.

8. Sequence details view

Figure 16. Sequence details view

The sequence details view displays the nucleotide sequences in forward and reverse frames, and their amino acid transcriptions in the 6 frames. The ORFs are also displayed there by a filled box. The colors of the box depends on the ORF accuracy according to the color scale define in section 4.

The sequence may be edited. To select the subsequence to edit. Perform a CLICK followed by a SHIFT+CLICK on the nucleotide bases that will begin and end the subsequence.

Following a sequence edition, the ORF are relocated into their new frames and the new ORFs generated by the edition are displayed.

ACKNOWLEDGMENTS

CNPq

UNIVERSIDADE FEDERAL DO PARÁ

PPCA

Collaboration

Ufra
UNIVERSIDADE FEDERAL RURAL DA AMAZÔNIA

UEPA
UNIVERSIDADE DO ESTADO DO PARÁ