

CNCA Manual

The CNCA (Coding Non Coding Aligner) program aligns multiple closely related annotated genomes of a size up to 50kb (typically viruses). It extracts protein sequences from the annotations and corrects frameshifts in the nucleotide alignment using protein alignments.

CNCA is accessible at <https://cnca.ijm.fr/>.

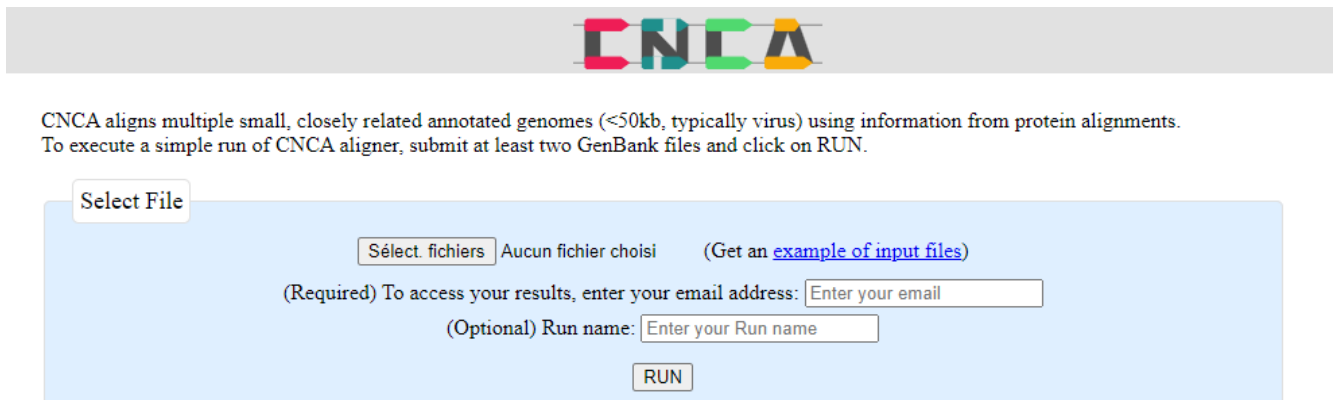
Comments and questions should be sent to: jean-noel.lorenzi@ijm.fr.

To run CNCA, simply:

1. Provide at least two GenBank files.
2. Enter a valid email address.
3. Click on the "RUN" button (**Figure 1**).

The pipeline is then launched with default parameters, and you shall receive an email notification after its completion, with a link to download the result output files: a "patched" nucleotide alignment as well as raw nucleotide and protein alignments (in nexus and fasta formats).

Figure 1. CNCA webtool homepage.



The screenshot shows the CNCA webtool homepage. At the top, there is a logo for CNCA with the letters in different colors (C: red, N: blue, C: green, A: yellow). Below the logo, there is a text box with the following content:

CNCA aligns multiple small, closely related annotated genomes (<50kb, typically virus) using information from protein alignments.
To execute a simple run of CNCA aligner, submit at least two GenBank files and click on RUN.

Below the text, there is a form with the following elements:

- A "Select File" button.
- A text box labeled "Sélect. fichiers" with the text "Aucun fichier choisi" and a link "(Get an [example of input files](#))".
- A text box labeled "(Required) To access your results, enter your email address:" with a text input field "Enter your email".
- A text box labeled "(Optional) Run name:" with a text input field "Enter your Run name".
- A "RUN" button.

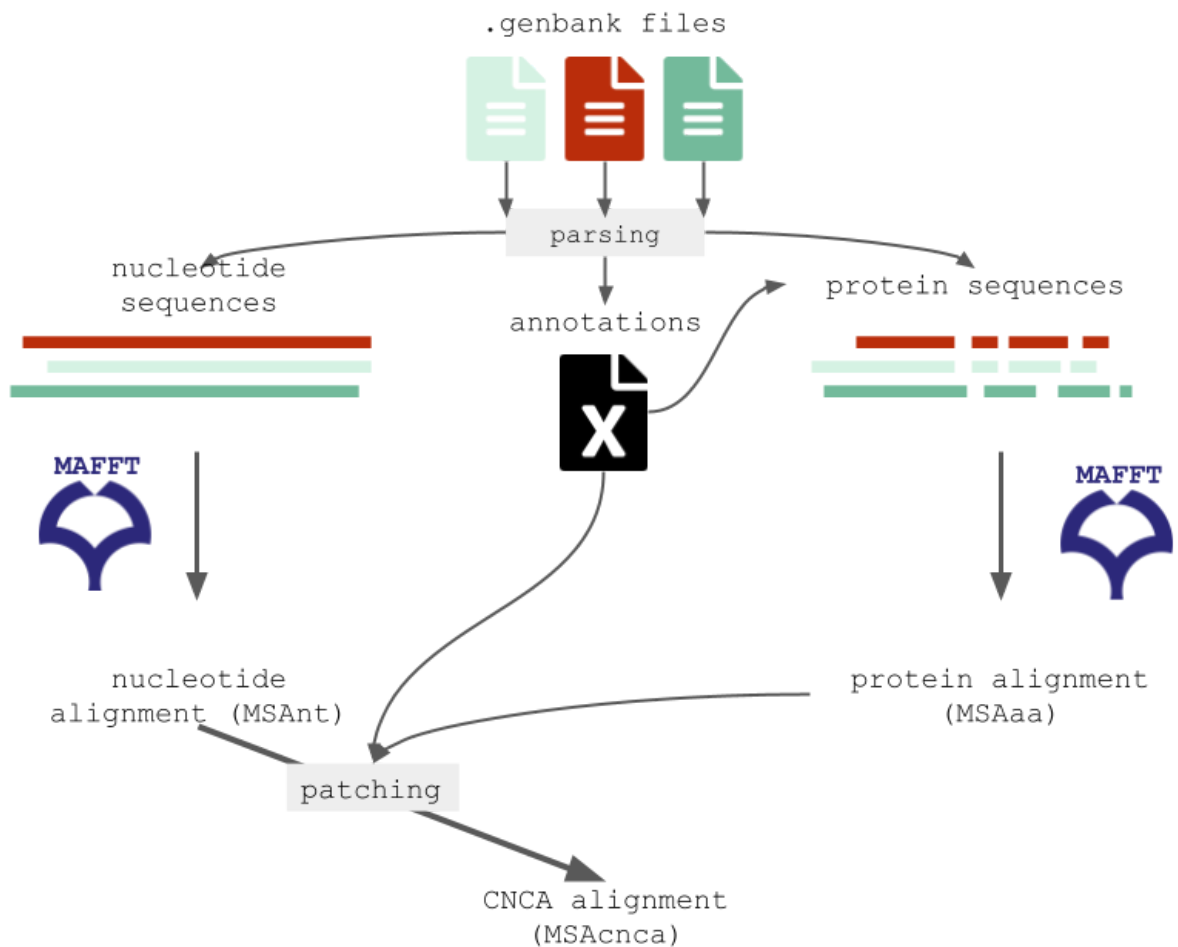
Table of content

OVERVIEW	2
INPUT	3
PROCESS	3
Preparation.....	3
Multiple Sequence Alignment of the nucleotide sequences (MSAnt).....	3
Multiple Sequence Alignment of the amino acid sequences (MSAaa).....	4
Patching.....	4
1- Indexing.....	4
2- Detection of misaligned regions.....	6
3- Correction of misaligned regions.....	7
4- Alignment correction.....	8
5- Proofreading.....	8
OUTPUT	8
DATA AVAILABILITY	8
REFERENCES	8

OVERVIEW

CNCA is a bioinformatics tool designed to generate a nucleotide alignment from closely related genome sequences based on a mixture of nucleotide and protein alignments. Unlike traditional retro-translation aligners which operate on coding regions only, CNCA aligns both coding regions (using amino acid sequence) and non coding regions. First, it performs a complete genome nucleotide alignment and then locally adjusts it using protein alignments. When there are inconsistencies between the nucleotide and the protein alignments, the protein alignment takes precedence in correcting the overall alignment (**Figure 2**).

Figure 2. A schematic view of the CNCA pipeline.



INPUT

The input of CNCA is a collection of annotated GenBank files (<https://www.ncbi.nlm.nih.gov/genbank/>). It is mandatory that the GenBank files contain closely related sequences, with an Average Nucleotide Identity (ANI) (Richter and Rosselló-Móra 2009) of 90% or more. You can verify the pairwise ANI values for a set of multiple sequences using the *jspeciesws* tool at <https://jspecies.ribohost.com/jspeciesws/> (Richter et al. 2015).

This sequence identity requirement is central as the CNCA tool relies on a comparable organization of genomes. More specifically, the order and the orientation of the coding sequences (gene synteny) must be conserved.

For the sake of server speed, sequences longer than 50 kb are not accepted as input.

All the nucleotide sequences should be in the same, relative orientation. Furthermore, no intron nor alternative splicing annotation should be present. All the coding sequences should be in the same orientation (between genomes). Nucleotides U and T are considered to be equivalent. Sequence files can contain both. Furthermore, it is essential that a single piece of nucleotide sequence is annotated for only one protein sequence. In other words, each segment of the nucleotide sequence should be associated with a unique protein coding sequence to avoid any conflicts or ambiguities during the CNCA process.

PROCESS

Preparation

When a user submits its data to CNCA, a temporary directory is automatically created. This directory serves as a storage location for all the GenBank files that are submitted by the user.

```
./temp/cnc_align_[unique id]/genome/file1.gbk
                               /file2.gbk
                               ...
```

CNCA extracts the coding sequences (CDS) of each genome from the annotations of the GenBank files. The GenBank files are parsed into temporary fasta (for sequences) and json files (for annotations), to ease their access.

When a piece of nucleotide sequence is annotated to be more than one protein sequence, such as joined CDS with overlapping frames, the process is stopped and an error is returned.

Multiple Sequence Alignment of the nucleotide sequences (MSAnt)

The nucleotide aligner of CNCA is MAFFT (Rozewicki et al. 2019). It aligns the complete nucleotide sequence of all input genomes. Default MAFFT parameters are used unless specified by the user.

On the CNCA main page the user can tune the available options of MAFFT (<https://mafft.cbrc.jp/alignment/server/>), to customize the alignment process.

Multiple Sequence Alignment of the amino acid sequences (MSAaa)

The protein sequences are concatenated based on the order of appearance of the coding sequences (CDS) in the genomes, from 5' to 3', and from N-terminal to C-terminal amino acid. It is therefore crucial that the order of the CDSs is identical among all input genomes. This ensures that the artificial concatenated protein sequences accurately reflect the organization of the coding regions in each genome.

The amino acid sequence concatenate is then aligned using MAFFT. Default MAFFT parameters are used unless specified by the user. Similar to the nucleotide alignment, users can adjust MAFFT parameters to optimize alignment quality and accuracy based on the desired alignment output and the nature of the protein sequences.

Several distance matrices can be used for the sequence alignment. The most widely used substitution matrix is BLOSUM62. Pearson (2013) describes in depth the factors one should consider when choosing a substitution matrix.

Patching

In CNCA, the protein alignment information (when available) can result in changes in the nucleotide alignment. When there is a conflict between the nucleotide alignment (MSAnt) and the protein alignment (MSAaa), the information from the protein alignment (MSAaa) takes precedence. This preference for protein information may result in insertions of additional gaps in the nucleotide sequence alignment when necessary.

The patching process of CNCA generates a new “patched” alignment that is a nucleotide alignment (MSAnt) corrected by the protein alignment (MSAaa). The patching method in CNCA involves several steps, from data indexing to final verification checks. The sequence of steps ensures the accuracy of the merging between the nucleotide and protein alignments.

1- Indexing

The indexing step involves the creation of a correspondence table between MSAnt and MSAaa progressively by examining each alignment site (for both nucleotide and corresponding amino acid sites) from left to right.

For each position (column) in MSAnt of each genome (row), several pieces of information are retrieved:

1. Position in MSAnt (integer)
2. Position in the raw sequence (integer)
3. Name of the raw sequence (text)
4. Nucleotide coding or not (TRUE, FALSE, NA)
5. Nucleotide in the raw sequence (IUPAC nucleotide code, “NA” if it corresponds to a gap in the nucleotide or protein sequence)

6. Nucleotide in MSAnt (IUPAC nucleotide code, "-" if gap in the nucleotide sequence, "NA " if it corresponds to a gap in the protein sequence)
7. Coded amino acid in the raw sequence (IUPAC amino acid code, "NA" if non coding or gap)
8. Phase of the nucleotide (0, 1, 2 or "NA" if non coding or gap)
9. Amino acid in MSAaa (one-letter code for amino acids, "-" if gap, "NA" if non coding)
10. Position in MSAaa (integer)

Table 1. Example of part of a correspondence table.

1	2	3	4	5	6	7	8	9	10
3379	3316	EPI_ISL_412977	TRUE	T	T	I	2	I	1023
NA	NA	EPI_ISL_412977	NA	NA	NA	NA	NA	-	1024
NA	NA	EPI_ISL_412977	NA	NA	NA	NA	NA	-	1025
3380	3317	EPI_ISL_412977	TRUE	G	G	E	0	E	1026
3381	3318	EPI_ISL_412977	TRUE	A	A	E	1	E	1026
3382	NA	EPI_ISL_412977	NA	-	NA	NA	NA	NA	NA
3383	NA	EPI_ISL_412977	NA	-	NA	NA	NA	NA	NA
3384	NA	EPI_ISL_412977	NA	-	NA	NA	NA	NA	NA
3385	3319	EPI_ISL_412977	TRUE	A	A	E	2	E	1026

For the first line: "For the EPI_ISL_412977 genome [3], the position 3379 in MSAnt [1] corresponds to position 3316 in the raw sequence [2]. This nucleotide is coding [4] and phased 2 in the codon [8]. This site corresponds to a T in the raw sequence [5] and in the MSAnt [6]. The corresponding position in MSAaa is 1023 [10]. It codes for an I in the raw sequence [7] and in MSAaa [9]." Column [5] and [7] serve as validation for the data retrieved in column [6] and [9], respectively.

The second and third lines correspond to two gaps (for genome EPI_ISL_412977) at position 1024 and 1025 in the MSAaa alignment. The last four lines indicate that there is a gap of three consecutive nucleotides at positions 3382-3384 in the MSAnt for the EPI_ISL_412977 genome.

Each row corresponds to a position in MSAnt and/or MSAaa for a given sequence. If both MSAnt and MSAaa alignments match and involve an amino acid/codon, only one row corresponds to both MSAnt and MSAaa sites. Otherwise, two rows are present: one for MSAnt and one for MSAaa. In the case of a gap, multiple lines are generated for each position within the gap, with distinct lines for MSAnt and MSAaa.

The correspondence table is created as follows:

The MSAnt is scanned genome by genome (line by line) from top to bottom and for each genome position by position (column by column) from left to right. A given position in MSAnt can be either a nucleotide or a gap.

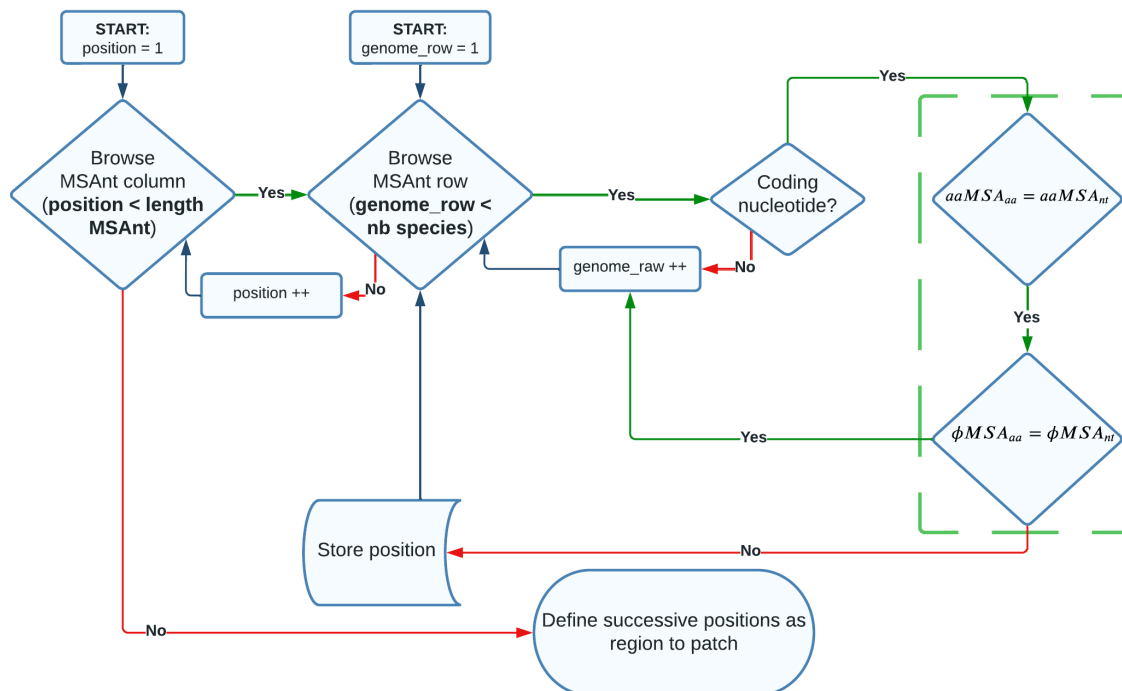
- A gap is recorded as such if there is no correspondence to the protein alignment nor the raw nucleotide sequence.
- For non-coding nucleotides, CNCA includes information from the MSAnt into the correspondence table.
- For coding nucleotides, in addition to the information from MSAnt, the encoded amino acid and the nucleotide phase are also recorded. The MSAaa is then scanned, and information from MSAaa is collected for the corresponding position. If a gap is encountered in MSAaa, it generates a new line in the correspondence table.

2- Detection of misaligned regions

CNCA leverages the correspondence table to detect misaligned regions. All nucleotide positions are individually checked according to the criteria shown in **Figure 3**. The key steps in this process are highlighted within the green dotted rectangle. The criteria for determining misaligned regions are as follows:

1. Different matching positions in MSAaa: if, for a given MSAnt position (column), the retrieved matching positions in MSAaa differ among genomes (row), it indicates a misalignment.
2. Different coding phases (ϕ) in MSAaa: Another factor that indicates a potential misalignment is when different coding phases are obtained from MSAaa for the tested MSAnt position. The coding phase refers to the position of the nucleotide within a codon (0, 1, or 2). If different coding phases are retrieved from MSAaa, it suggests a potential misalignment.
- 3.

Figure 3. Protocol for detecting misaligned regions. ++ means that the protocol increments to the next nucleotide or genome.



In pseudo-code:

```

for each position p of MSAnt:
  for each genome g of MSAnt:
    if p is coding in g:
      if coded amino acid (aaMSAnt) != amino acid in MSAaa
      (aaMSAaa) or phase in MSAnt ( $\phi$ MSAnt) != phase in MSAaa ( $\phi$ MSAaa):
        mark position

return marked regions of consecutive positions to patch

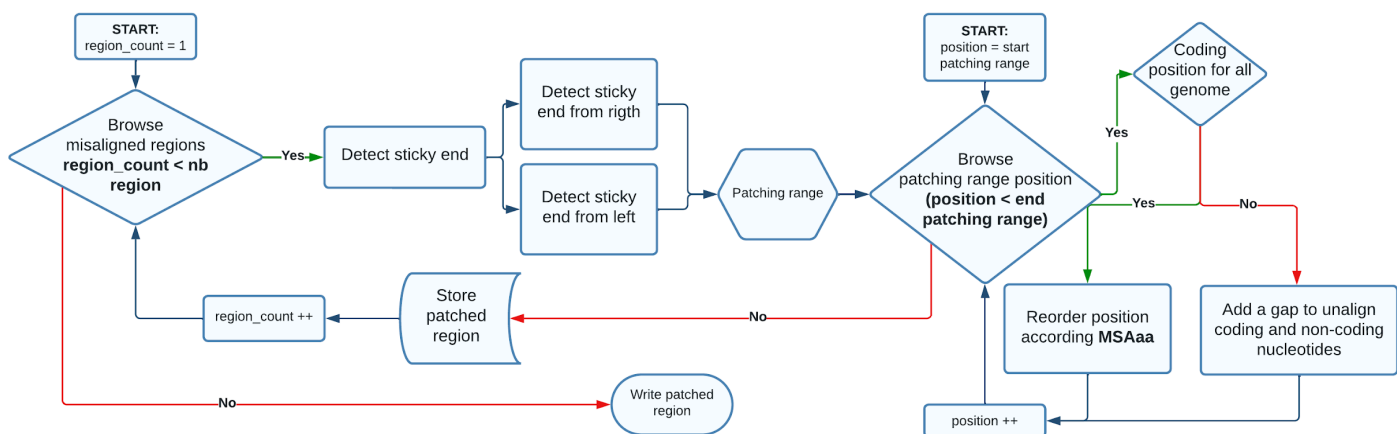
```

By applying the described process, positions are identified where a contradiction is present between the MSAnt and the MSAaa. The marked adjacent positions are grouped into regions that display inconsistencies between the two alignments. The resulting output is a list of regions to be realigned.

3- Correction of misaligned regions

The misaligned regions are then corrected using the process outlined in **Figure 4**. The first step of the process involves identifying sticky ends within the region to be patched. A sticky end is defined as a sequence of at least 3 consecutive nucleotides involved in the same codon that are correctly aligned (*i.e.*, matching between MSAaa and MSAnt) for all genomes. By identifying sticky ends, it is ensured that rearranging the columns between these sticky ends will not affect the alignment outside of the patched region. This allows for safe modifications within the misaligned regions without impacting the overall alignment.

Figure 4. Patching process. ++ means that the protocol increments to the next misaligned region or to the next position.



In pseudo-code:

```

for each marked region mr
  scan edges of mr to find sticky ends (se_l left and se_r right)
  for each position of MSAnt between se_l and se_r:
    if coding in all genomes:

```



```

        align MSAnt(p,g=1 to last) as in MSAAA
    else:
        insert gaps to prevent aligning coding with non-coding
        record aligned nucleotides within mr

return the collection of all aligned mr

```

4- Alignment correction

The raw nucleotide alignment (**MSAnt**) is modified by incorporating the patched regions obtained from the alignment process. Between the identified sticky ends, the original alignment is replaced with the corresponding patched region. This operation results in the creation of a new alignment called **MSAcnca**.

5- Proofreading

To validate the **MSAcnca** alignment, each genome in **MSAcnca** is browsed and compared to the corresponding raw sequence. The identity between the **MSAcnca** sequence and the raw sequence is checked to ensure identity between genome sequences of **MSAcnca** and GenBank input sequences.

If there is any inconsistency or mismatch between the **MSAcnca** sequences and the original sequences, an error is returned.

OUTPUT

If the CNCA pipeline runs successfully, it returns the nucleotide and protein alignments (**MSAnt** and **MSAAA**) along with the new patched alignment (**MSAcnca**). The three alignments are provided in both fasta and nexus formats. A link to access the results is sent to the user via email.

If an error occurs during the pipeline execution, an email is also sent to the user. This email contains troubleshooting tips and guidance based on the specific error encountered.

DATA AVAILABILITY

The data provided by the user and the generated results are stored on the data server for a period of one week. After one week, the data is deleted.

REFERENCES

- Pearson, William R. 2013. "Selecting the Right Similarity-Scoring Matrix." *Current Protocols in Bioinformatics* 43(1). <https://onlinelibrary.wiley.com/doi/10.1002/0471250953.bi0305s43> (June 27, 2023).
- Richter, Michael, and Ramon Rosselló-Móra. 2009. "Shifting the Genomic Gold Standard for the

Prokaryotic Species Definition.” *Proceedings of the National Academy of Sciences* 106(45): 19126–31.

Richter, Michael, Ramon Rosselló-Móra, Frank Oliver Glöckner, and Jörg Peplies. 2015.

“JSpeciesWS: A Web Server for Prokaryotic Species Circumscription Based on Pairwise Genome Comparison.” *Bioinformatics* 32(6): 929–31.

Rozewicki, John et al. 2019. “MAFFT-DASH: Integrated Protein Sequence and Structural Alignment.” *Nucleic Acids Research* 47(W1): W5–10.