Genome Analysis

CNCA aligns small annotated genomes

Jean-Noël Lorenzi[1,2,3,*], François Graner[1,4], Virginie Courtier-Orgogozo[1,2] and Guillaume Achaz[1,3]

1: Université Paris Cité, F-75006 Paris, France.

2: CNRS, Institut Jacques Monod, F-75013 Paris, France.

3: SMILE group, Center for Interdisciplinary Research in Biology (CIRB), Collège de France, F-75006 Paris, France.

4: CNRS, Matière et Systèmes Complexes, F-75013 Paris, France.

*To whom correspondence should be addressed.

Abstract

Motivation: We present CNCA, an online tool to align annotated genomes from GenBank files. It generates a nucleotide alignment that is then updated based on the protein sequence alignment. The output final nucleotide alignment matches the protein alignment and guarantees no frameshift. CNCA was designed to aligned closely related small genome sequences up to 50 kb (typically viruses) for which the gene order is conserved.

Availability: The interface is available at https://cnca.ijm.fr/ and the code is available at https://github.com/jnlorenzi/CNCA_standalone

Contact: jean-noel.lorenzi@ijm.fr

A naive nucleotide alignment of complete genomes usually result in many frameshifts and other oddities that do not exist in protein alignments.

Several methods have been developed to perform nucleotide alignments taking protein alignment into account. One approach is "back-translation", where coding nucleotide sequences are translated into amino acids, that are aligned. Corresponding codons are then aligned in a final nucleotide alignment. The web-based tool web-prank (https://www.ebi.ac.uk/goldman-srv/webprank/; (Löytynoja and Goldman, 2010)) is such an example. Other tools based on back-translation proposes specific options like the choice of genetic codes (PAL2NAL (Suyama *et al.*, 2006), transAlign (Bininda-Emonds, 2005), RevTans (Wernersson and Pedersen, 2003)). Some are designed to consider cases in which frameshifts or stop codons can occur (MACSE (Ranwez *et al.*, 2011, 2018), PAL2NAL (Suyama *et al.*, 2006), transAlign (Bininda-Emonds, 2005)). TranslatorX (Abascal *et al.*, 2010) checks the relevance of the amino acid alignment by finding regions of uncertainties in the amino acid alignment (masked by Gblocks (Castresana, 2000)) and reports them in the nucleotide alignment. Others are optimized for virus gene sequences (NucAmino (Tzou *et al.*, 2017), VIRULIGN (Libin *et al.*, 2019)). To the best of our knowledge, none of these methods processes genome alignment with both coding and non-coding regions. We have thus developed CNCA (Coding / Non-Coding Aligner), a genome-wide solution that returns a complete genome alignment compatible with the protein sequence alignment. The method was designed for small (up to 50 kb) homologous annotated syntenic genomes devoid of introns, such as virus genomes. It will ease the subsequent evolutionary analysis of these complete genomes.

CNCA takes as input two or more GenBank files of annotated genomes (Fig. 1). It first MAFFT-aligns (Nakamura *et al.*, 2018) the nucleotide (nt) sequences of all genomes and produces a Multiple Sequence Aligmnent (MSAnt). It then generates MSAaa, the MAFFT-alignment of the concatenate of all protein sequences. As the concatenate takes protein sequence on the order of gene annotations, synteny must be conserved. The MSAnt is then updated using MSAaa for all coding regions where both alignments are not concordant. A final MSAcnca is returned that contains no contradiction with MSAaa and thus no frameshift (figure 1). We choose to implement a graphical web version of the pipeline to widen the potential users to non-experts. Results (logs and the three alignments MSAcnca, MSAnt, MSAaa in both nexus and fasta formats) are stored locally for a week. An email with a link to access the results is sent to the user at the end of the procedure.

As an illustration, the whole CNCA pipeline runs in 45 min for a dataset of 12 genomes closely related to SARS-CoV-2.

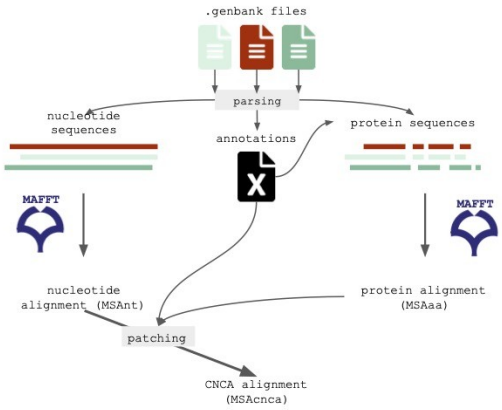

Figure 1. Schematic view of the CNCA pipeline.

Funding

This work was supported by the Labex "Who Am I?", ANR-11-LABX- 0071 and the Université Paris Cité, Idex ANR-18-IDEX-0001, funded by the French Government through its "Investments for the Future" program.

Conflict of Interest: none declared.

References

Abascal,F. *et al.* (2010) TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res.*, **38**, W7–W13.

Bininda-Emonds,O.R. (2005) transAlign: using amino acids to facilitate the multiple alignment of protein-coding DNA sequences. *BMC Bioinformatics*, **6**, 156.

Castresana,J. (2000) Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Mol. Biol. Evol.*, **17**, 540–552.

Libin,P.J.K. *et al.* (2019) VIRULIGN: fast codon-correct alignment and annotation of viral genomes. *Bioinformatics*, **35**, 1763–1765.

Löytynoja,A. and Goldman,N. (2010) webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. *BMC Bioinformatics*, **11**, 579.

Nakamura,T. *et al.* (2018) Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics*, **34**, 2490–2492.

Ranwez,V. *et al.* (2011) MACSE: Multiple Alignment of Coding SEquences Accounting for Frameshifts and Stop Codons. *PLOS ONE*, **6**, e22594.

Ranwez,V. *et al.* (2018) MACSE v2: Toolkit for the Alignment of Coding Sequences Accounting for Frameshifts and Stop Codons. *Mol. Biol. Evol.*, **35**, 2582–2584.

Suyama,M. *et al.* (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.*, **34**, W609–W612.

Tzou,P.L. *et al.* (2017) NucAmino: a nucleotide to amino acid alignment optimized for virus gene sequences. *BMC Bioinformatics*, **18**, 138.

Wernersson,R. and Pedersen,A.G. (2003) RevTrans: multiple alignment of coding DNA from aligned amino acid sequences. *Nucleic Acids Res.*, **31**, 3537–3539.