



HAL
open science

How did you annotate your new viral genome without it ? VAZyMolO-2

Vincent Wilde, Bruno Canard, Francois Ferron

► **To cite this version:**

Vincent Wilde, Bruno Canard, Francois Ferron. How did you annotate your new viral genome without it ? VAZyMolO-2. ViBioM2024, May 2024, Luven, Belgium. hal-04616074

HAL Id: hal-04616074

<https://hal.science/hal-04616074>

Submitted on 18 Jun 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

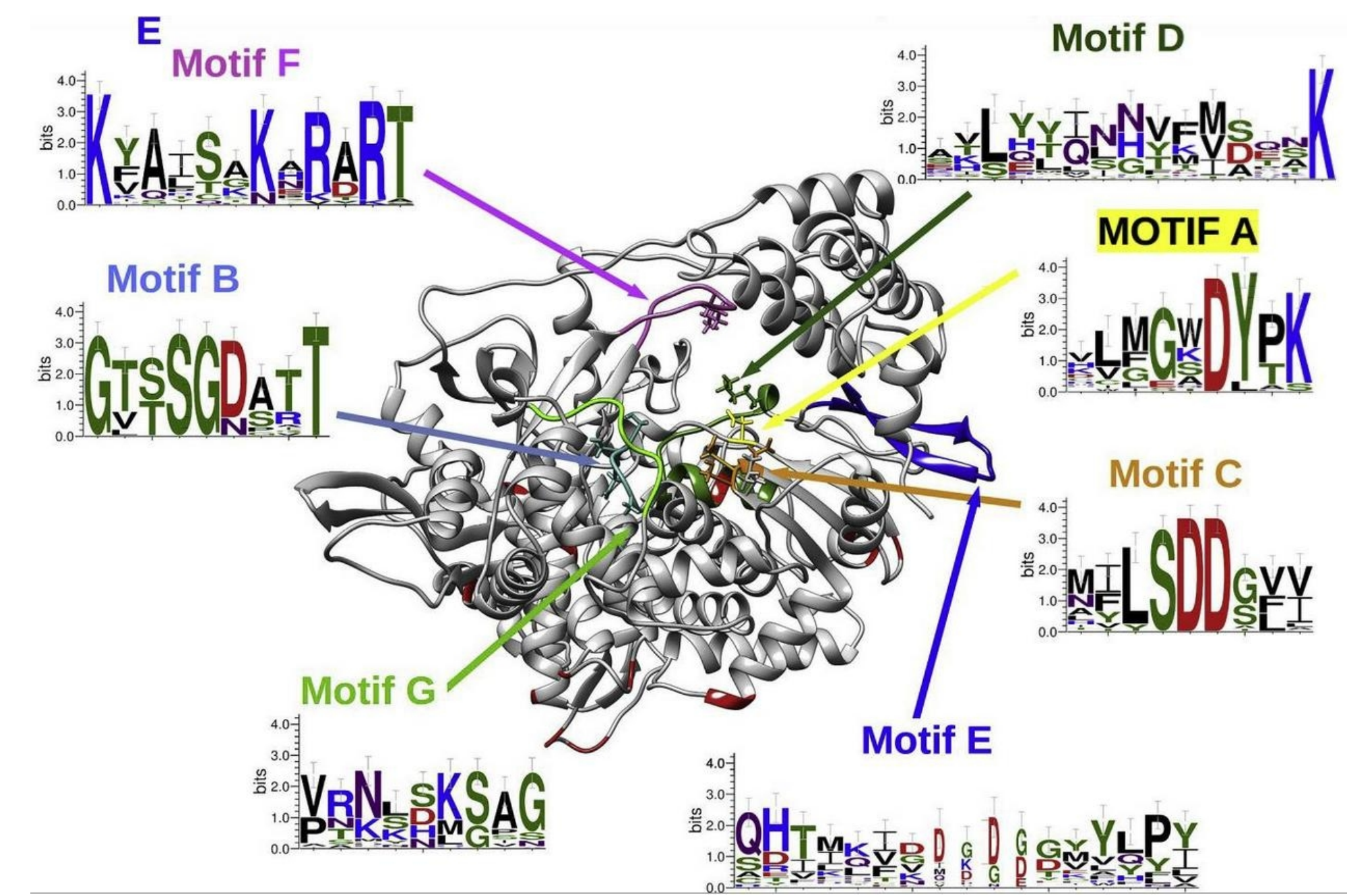
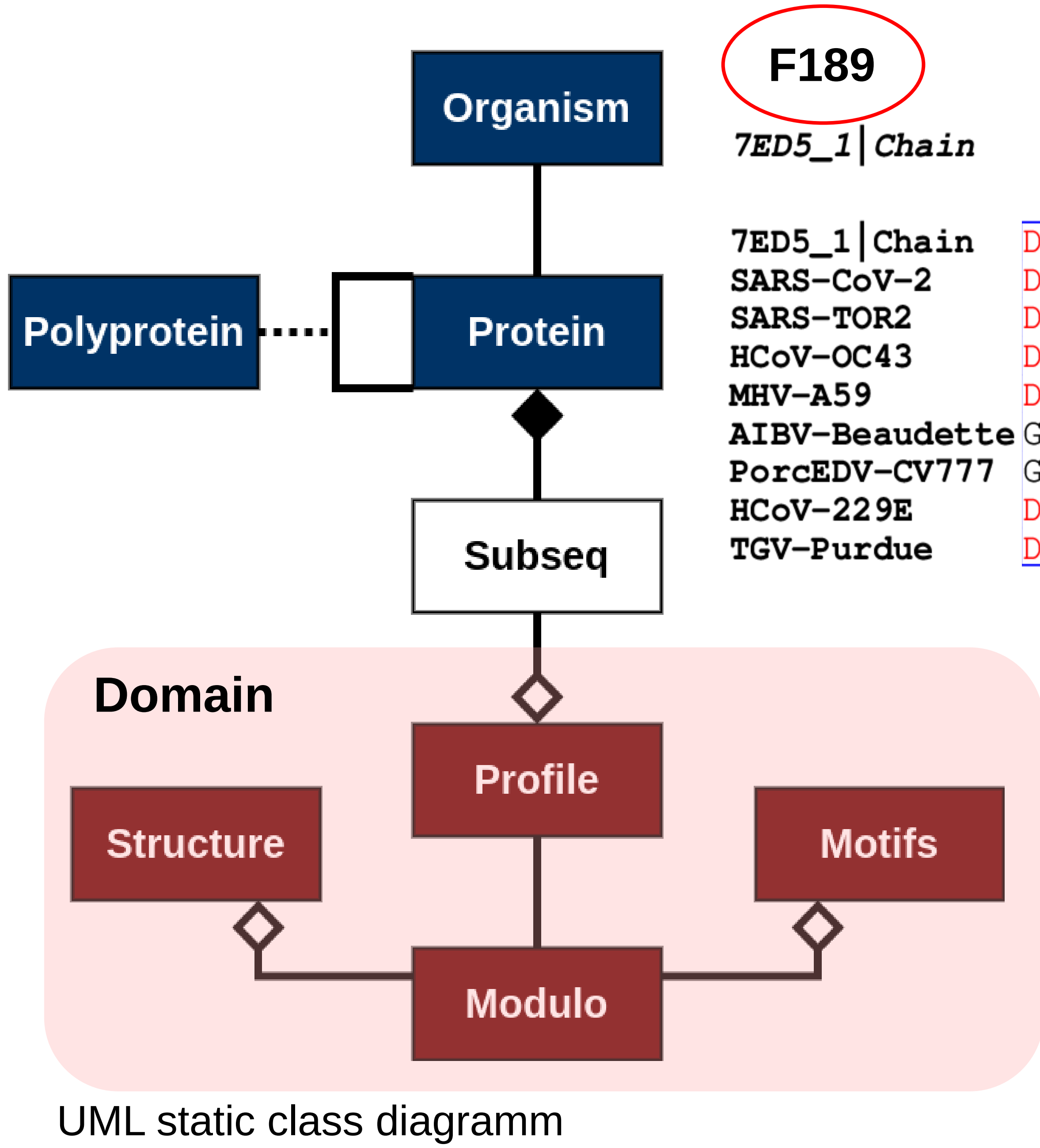
L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Public Domain

Abstract

- Sequencing and Metagenomics lead to mass production of bioinformatic data feeding the virus-vers.
- Need of knowledge-collection that can assimilate and interpret this data.
- Revival and enhancement of VAZyMoIo (2003-2011) (F.Ferron et al., 2005).
- Harnessing the intricate interplay between protein structure and biological function.
- Providing a dynamic and effective solution for annotating new viral genome.

Contact me



Objectives

- Curated database
 - starting point previous VAZyMoIo and **leader** of other families.
 - annotating protein into **modules**.
 - merging **Phylogeny** and **Functional Annotation**.
- Web-service
 - **Search** user through a protein **sequence** or **keywords** (GO / EC).
 - detecting the **closest well-known virus orthologue**.
 - transferring the orthologue annotation on the query's sequence.

Methods

Routine analytic pipeline

- Motif Search**
Gathered from literature or Structure Analysis
Example: Motif A → "T-G-D-N"
→ Pattern search (regex)

- Sequence Alignment**
Using pair-wise alignment
"hit and run"
→ Clustering with MMseqs2

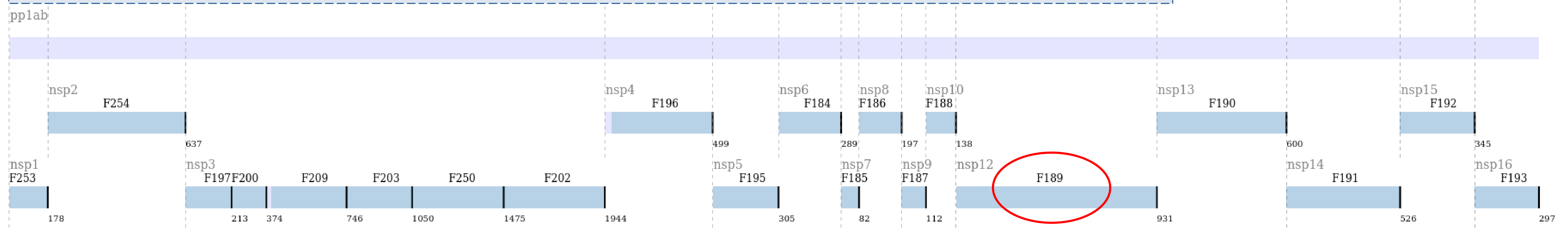
- Profile Alignment**
Depending on two previous output
Against the profile of 'each' domain
→ Aligning with Muscle

Internal analytic pipeline

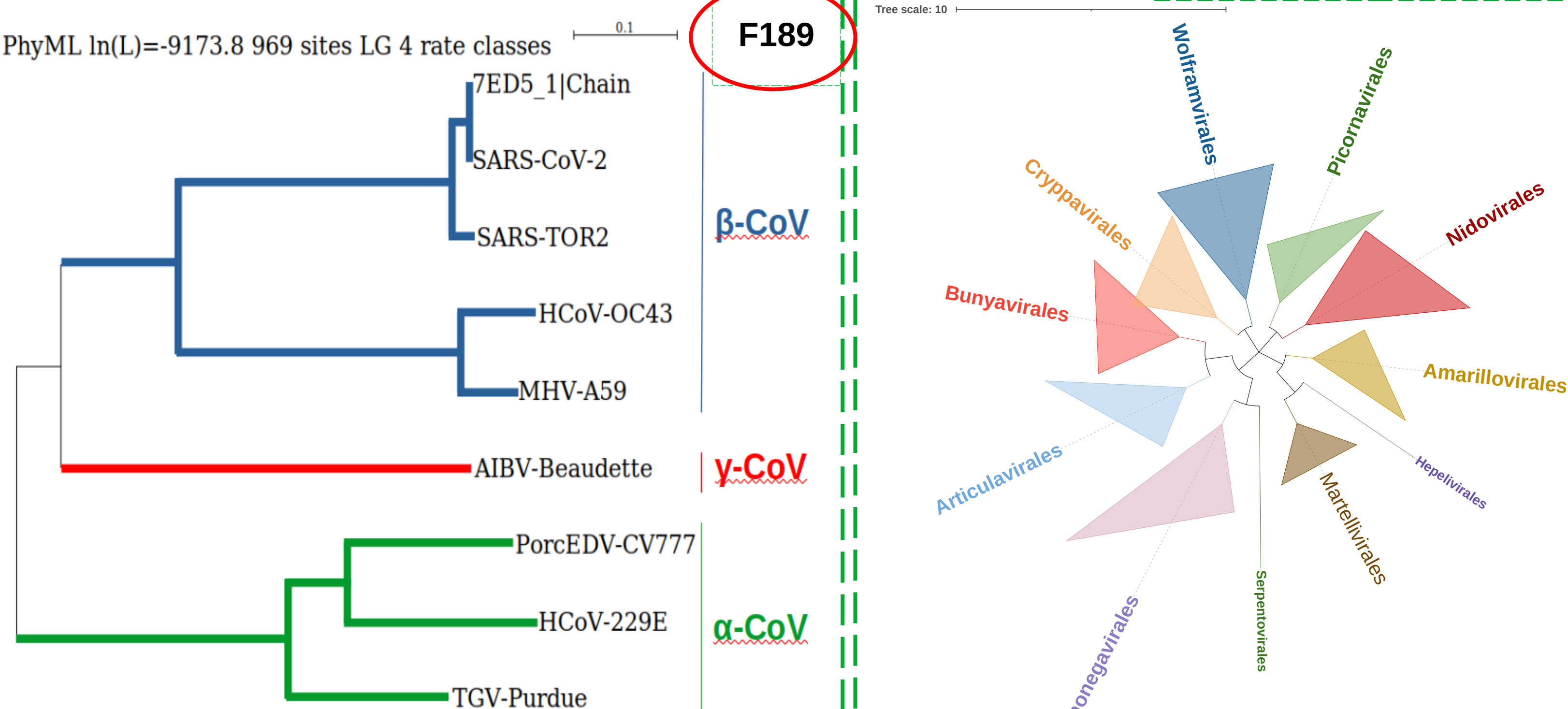
- ✓ **Structure Analysis**
Internal usage to characterize :
→ structure determination
→ association with a biological function
→ Domain validation

Specificity Sensitivity

Result -> Processing of pp1ab Severe acute respiratory syndrome coronavirus (SARS-CoV-2 - 2697049)



Phylogeny



Conclusion and Perspectives

- Project still in development
- Focusing on **Coronaviruses** towards all **Nidovirales** by :
 - **updating** previous project database VAZyMoIo.
 - following **current team research** :
replicase of emergent viruses.
- Tracing usage to prioritize the processing effort.
- When enough data is collected, usage of AI assistance.

Acknowledgements



C A R E
CORONA ACCELERATED R&D IN EUROPE

