nf-core/multiplesequencealign

A pipeline to benchmark and deploy Multiple Seqeunce Aligners (MSAs)

Luisa Santus[1], Jose Espinosa-Carrasco[1], Leon Rauschning[1,2], Cameron Glichrist[3], Adam Gudys[4], Sebastian Deorowicz[4], nf-core community, Martin Steinegger[3], Cedric Notredame[1]

1 Centre for Genomic Regulation (CRG), Barcelona, Spain
2 Ludwig-Maximilians-Universität, Munich, Germany
3 Artificial Intelligence Institute, Seoul, South Korea
4 Silesian University of Technology, Gliwice, Poland

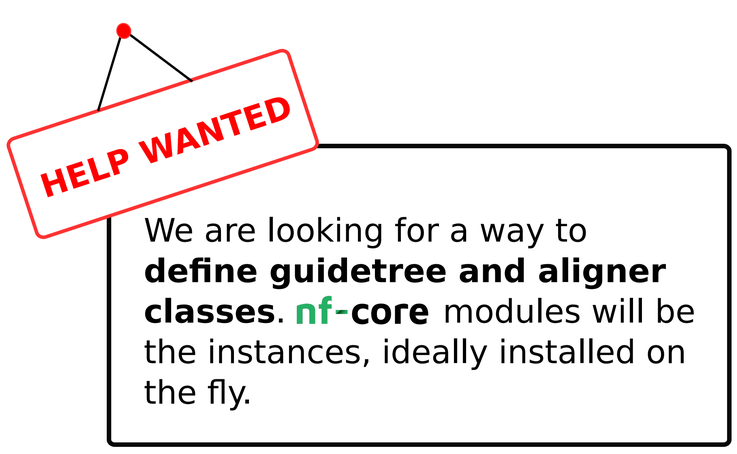
Why do we need this pipeline?

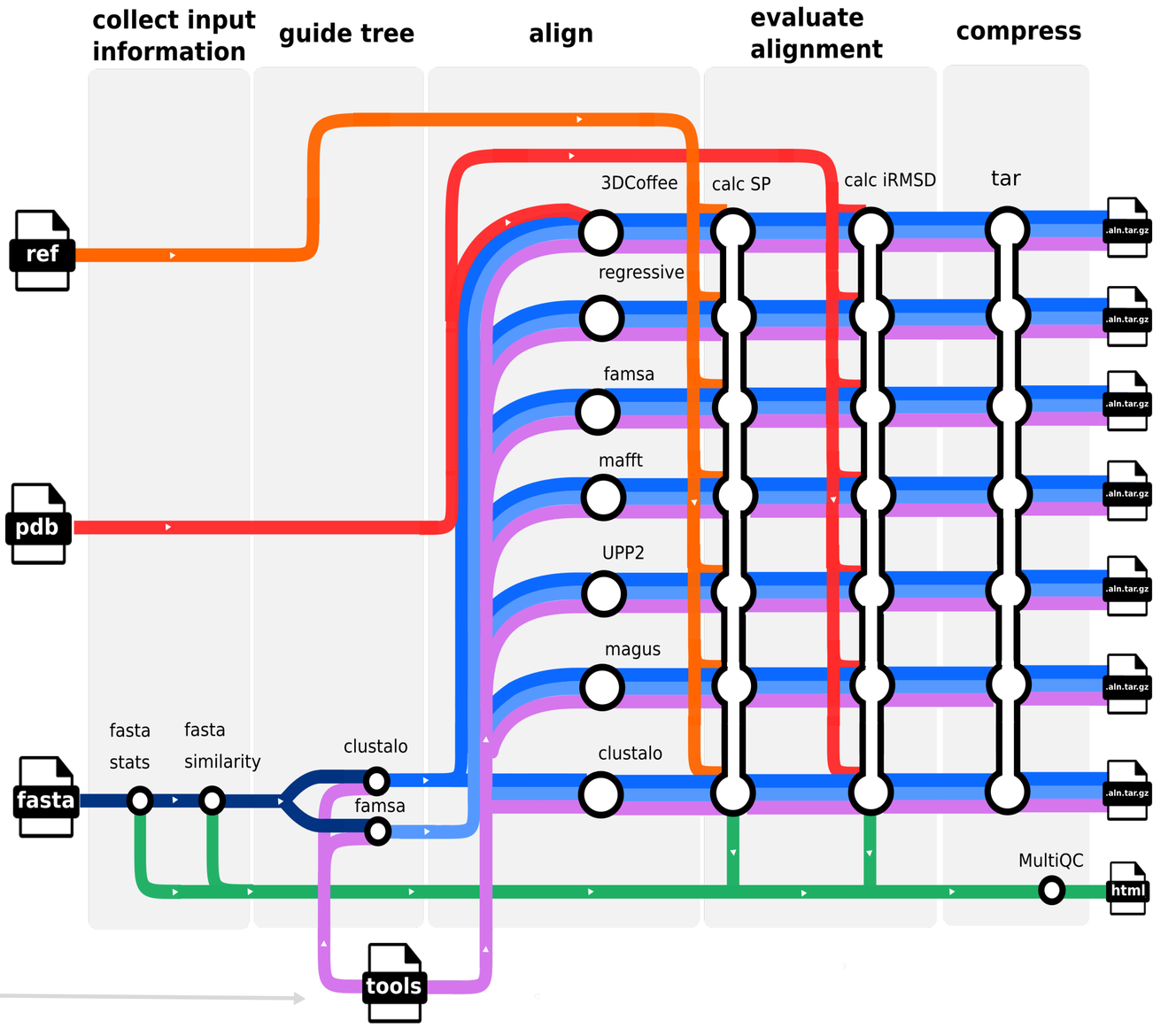
Multiple sequence alignment tools are **highly popular modelling methods** in bioinformatics, used in various downstream applications such as protein structure prediction and phylogenetic reconstruction.

Currently, **MSAs** are challenged to align an ever-increasing number of sequences and their **deployment has become increasingly challenging**. Furthermore, the lack of a proper and rigorous benchmarking framework remains a significant challenge for MSA development.

nf-core/multiplesequencealign is a comprehensive pipeline that **facilitates the seamless computation of MSAs while offering rigorous performance evaluation**.

Workflow





samplesheet.csv

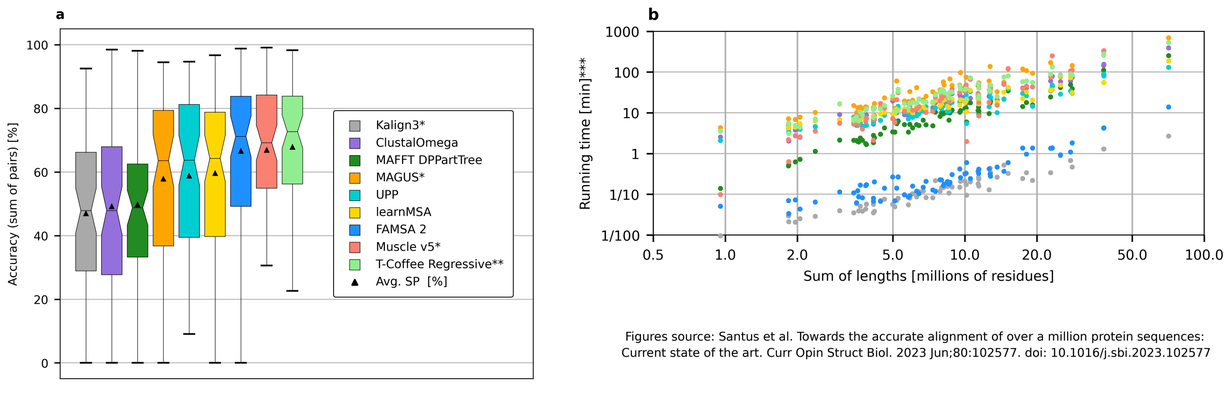
id	fasta	ref	structures
fam1	fam1.fa	fam1.ref	dir_fam1_str
fam2	fam2.fa	fam2.ref	dir_fam2_str
fam3	fam3.fa	fam3.ref	dir_fam3_str

toolsheet.csv

tree	args_tree	aligner	args_aligner
famsa	-gt upgma -parttree	clustalo	
		mafft	--anysymbol --dpparttree

Summary input/output

The pipeline generates informative plots that summarize main aspects of the input dataset as well as the benchmarking main results.



Figures source: Santus et al. Towards the accurate alignment of over a million protein sequences: Current state of the art. Curr Opin Struct Biol. 2023 Jun;80:102577. doi: 10.1016/j.sbi.2023.102577

On the TODO list

- Add all tools to nf-core modules
- Add proper visualization tools
- Add the subworkflows to nf-core
- Make the output more informative
- First release!

Acknowledgments: Many thanks to the contributors of nf-core/taxprofiler: its structure was very helpful in designing the main workflow and subworkflows of nf-core/multiplesequencealign.