



Abstract

Many bioinformatics problems, are computationally complex, leading to diverse heuristic solutions with no clear best option. This complicates tool selection for users and evaluation for developers, especially as data volumes grow. Multiple sequence alignment, essential for common tasks like protein structure prediction and phylogeny reconstruction, exemplify this challenge due to its NP-hard complexity, making optimal solutions impractical. We present a pilot nf-core framework designed to streamline MSA tool deployment and performance evaluation. By integrating popular MSA tools in a modular, extensible architecture, this framework aims to support deployment, evaluation, and algorithm development for the MSA community, while serving as a model for broader bioinformatics applications.

Configuration

Toolsheet

- Each line defines one *procedure*
- Specifies aligner/guidetree tool, along with parameters
- => reproducible, stored in one place
- Each sample is run with each procedure
- => useful for evaluation
- To swap aligner, just modify the file

tree	args_tree	aligner	args_aligner
FAMSA	-gt upgma -parttree	FAMSA	
FAMSA	-gt nj -parttree	FAMSA	
FAMSA	-gt upgma -parttree	CLUSTALO	
FAMSA	-gt nj -parttree	CLUSTALO	
		LEARNMSA	
		3DCOFFEE	-method TAlign_pair

Nextflow config

- Enable/disable subworkflows
- configure input/output
- Documented, follows nf-core standards

```
pipeline.conf
params {
  input          = './samplesheet.csv'
  tools          = './toolsheet.csv'
  skip_stats     = false
  calc_seq_stats = true
  skip_eval      = false
  calc_sp        = true
  calc_tc        = true
}
```

Input

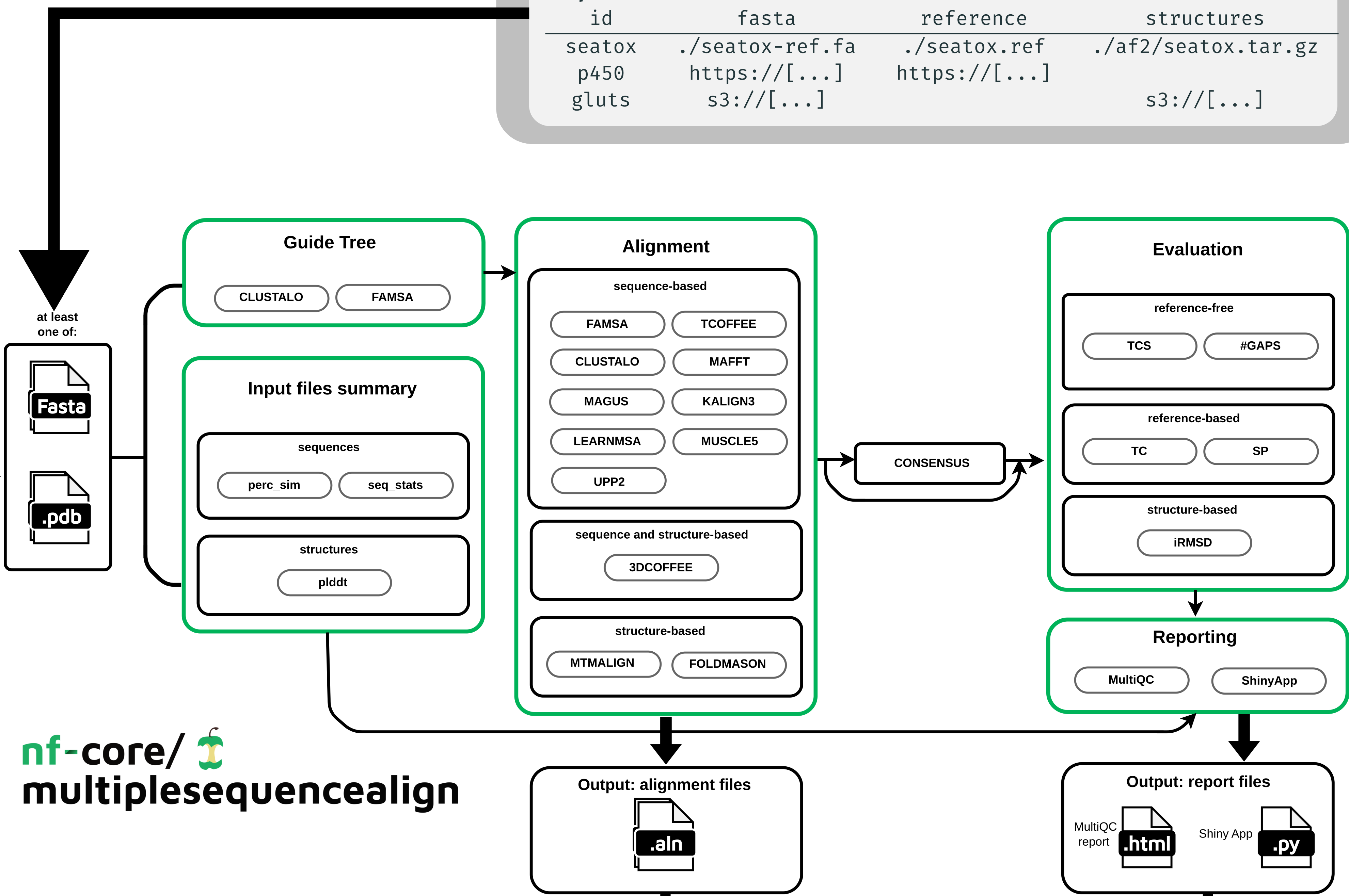
>p1
MSQA
>p2
MS-A
>p3
MSRA

- AA sequences
- multi-seq FASTA
- **required input**
- reference aln
- gapped FASTA
- optional => evaluation
- protein structures
- gzipped .PDB files
- optional => structural aligners

Samplesheet

- CSV collecting all input data (local files, web, s3, ...)
- One line per sample, processed in parallel
- => reproducible & shareable

id	fasta	reference	structures
seatox	./seatox-ref.fa	./seatox.ref	./af2/seatox.tar.gz
p450	https://[...]	https://[...]	
gluts	s3://[...]		s3://[...]



Features

- Standardized, straightforward deployment across local/HPC/cloud
- Reproducible, recorded runs
- Configurable, modular design => subworkflows, tool modules
- Extensible
- => Documented
- Integrated benchmarking/QC

Our Learnings

- Standardized deployment is possible, even in an old & fragmented ecosystem => modularity and configurability is key!
- Parameters are as important as tools
- Decomposing algorithms into their steps provides benefits for users & developers => opportunities for new tools
- Tools don't perfectly fall into »classes«, but do have common interfaces

Evaluation Reports

MultiQC Summary

id	fasta	tree	aligner	n_sequences	seqlength_meanperc_sim	sp	tc	iRMSD	plddt
1.0	seatox-ref	CLUSTALO	REGRESSIVE	5.0	47.0	40.20%	81.0	46.9	0.9
2.0	seatox-ref	FAMSA	MAGUS	5.0	47.0	40.20%	85.4	50.1	0.9
3.0	seatox-ref	MAGUS	5.0	47.0	40.20%	85.4	50.1	0.9	
4.0	seatox-ref	TCOFFEE	5.0	47.0	40.20%	81.9	51.0	1.0	
5.0	seatox-ref	FAMSA	5.0	47.0	40.20%	81.0	46.9	0.9	
6.0	seatox-ref	MAFFT	5.0	47.0	40.20%	86.3	50.2	0.9	
7.0	seatox-ref	REGRESSIVE	5.0	47.0	40.20%	81.0	46.9	0.9	
8.0	seatox-ref	REGRESSIVE	5.0	47.0	40.20%	81.7	46.9	0.9	
9.0	seatox-ref	MAFFT	5.0	47.0	40.20%	85.4	51.1	1.1	
10.0	seatox-ref	MUSCLES	5.0	47.0	40.20%	83.6	50.1	1.0	
11.0	seatox-ref	CLUSTALO	5.0	47.0	40.20%	81.9	51.0	0.9	
12.0	seatox-ref	KALIGN	5.0	47.0	40.20%	82.6	51.0	0.9	
13.0	seatox-ref	FAMSA	5.0	47.0	40.20%	81.0	46.9	0.9	
14.0	seatox-ref	MTALIGN	5.0	47.0	40.20%	80.6	50.1	0.6	
15.0	seatox-ref	3DCOFFEE	5.0	47.0	40.20%	85.6	52.3	0.8	
16.0	seatox-ref	LEARNMSA	5.0	47.0	40.20%	87.8	10.3	1.2	

- MultiQC-based report in HTML/PDF format
- Collects evaluation, input and runtime stats
- For each sample & procedure
- => useful for tool selection & benchmarks

Evaluation Explorer

nf-core/multiplesequencealign Stats & Evaluation Explorer

Mappings: X axis: Number of Sequences, Y axis: Total Column Score (1), Color: Assembly, Show linear model (intercept):

Style: General

- Interactive Shiny Web-App
- How do different factors affect performance? => useful for parameter tweaking & tool development

Output Files

>p1
MSQA--
>p2
MS-A
>p3
MSRA

Alignments
FASTA (gzipped)
=> by each procedure & consensus

Guide trees
Newick

Evaluation metrics
raw output/report/aggregated csv

Input statistics
CSV/report

CPU, mem use
CSV/report

Tool versions
YML by modules

The release version of the pipeline is available on the nf-core website and GitHub.

We'd love to hear from you! Feedback or suggestions are welcome on GitHub or slack!

Acknowledgements
Thanks to Igor Trujnara and Leila Mansouri for contributing to nf-core/msa, and for Adam Gudys, Sebastian Deorowicz, Martin Steinegger and the nf-core community for feedback, discussions and support!

