

Day 2 - Project

Thomas Junier, Robin Engler

22-23 May 2024

FASTA to TSV converter script project

The Problem with Fasta

Here are some operations that one might wish to perform on a FASTA file:

- ▶ Count the number of records
- ▶ Count the number of species
- ▶ Extract a record by ID
- ▶ Find the longest sequence
- ▶ Reject sequences with ambiguous nucleotides (N, Y, etc.)
- ▶ Discard aligned sequences with too many gaps
- ▶ Partition records by species

With shell tools, the first two are trivial, but **the last four are next to impossible.**

Why ?

- ▶ Unix shell tools (sed, awk, grep, etc.) are predominantly *line-oriented*.
- ▶ Some bioinformatics formats are line-oriented (*e.g.* **GFF**, **VCF**)... but Fasta is not (neither are GenBank, UniProt, ...).
- ▶ Converting FASTA to some line-oriented format (*e.g.* TSV) would solve the problem.¹

¹The *format* problem, that is - the rest can be left to grep and the like.

The project

To be able to perform more operations easily on FASTA file content, we are going to write a **FASTA → TSV converter script**.

WARNING

Didactical Script!

The script is meant to *illustrate* concepts, **not** to be efficient.
⇒ We'll write it in pure style. A real-world script would be in delegation style and very different (but mostly useless for teaching).

WARNING

Didactical Script!

The script is meant to *illustrate* concepts, **not** to be efficient.
⇒ We'll write it in pure style. A real-world script would be in delegation style and very different (but mostly useless for teaching).

Indeed, a FASTA → TSV converter could be written in one line of sed:

```
sed -n '1{h;b};${H;bo};/^>/!{H;b};:o;x;s/^>//;s/\n/\t/;s/\n//g;p'
```