Developing a standalone toolbox for processing FASTA files

Hugo López-Fernández[1,2,3,4], Miguel Reboiro-Jato[1,2], Noé Vázquez[1,2], Pedro Duque[3,4], Florentino Fdez-Riverola[1,2], Cristina P. Vieira[3,4], Jorge Vieira[3,4]

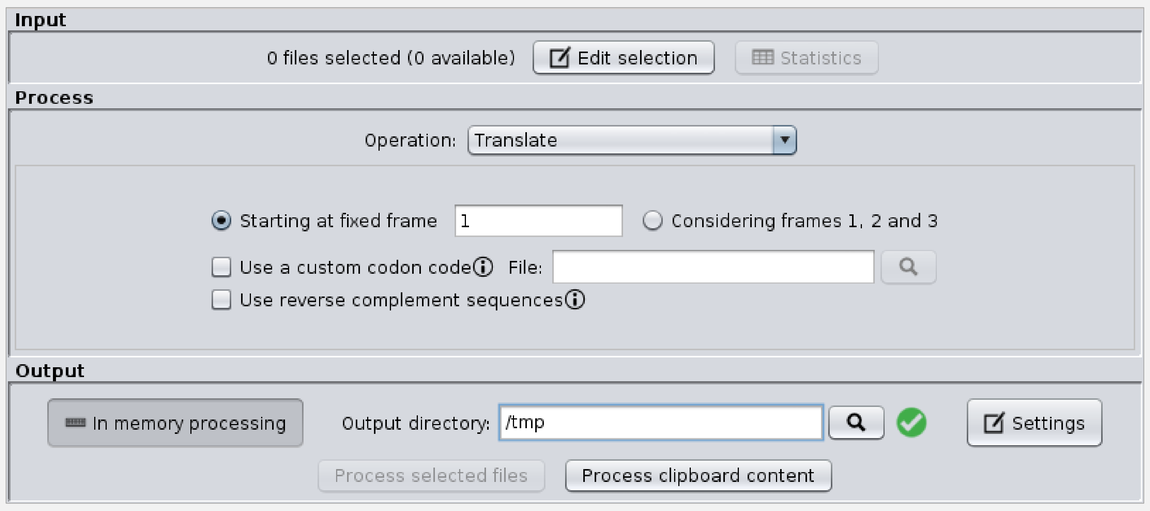
1 SING Research Group, Escuela Superior de Ingeniería Informática, University of Vigo, Edificio Politécnico, Campus Universitario As Lagoas s/n, 32004 Ourense, Spain.
2 CINBIO - Centro de Investigaciones Biomédicas, University of Vigo, Campus Universitario Lagoas-Marcosende, 36310, Vigo, Spain.
3 Instituto de Investigação e Inovação em Saúde (I3S), Universidade do Porto, Rua Alfredo Allen, 208, 4200-135 Porto, Portugal.
4 Instituto de Biologia Molecular e Celular (IBMC), Rua Alfredo Allen, 208, 4200-135 Porto, Portugal.

Motivation

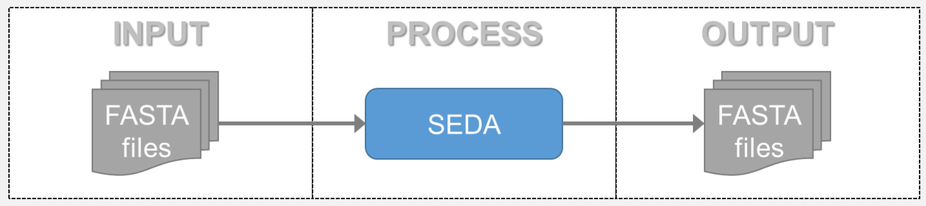
One of the most important types of data used in biological research is DNA or protein sequence data. They are usually stored in FASTA files, which can store one or more sequences. Public databases such as GenBank, NCBI or Ensembl provide huge collections of genomes, genome annotations, and so on, in FASTA format. Nevertheless, downloaded files usually must be preprocessed before subsequent analysis depending on each researcher needs. Despite the simplicity of these preprocessing operations (e.g. remove sequences without a minimum number of bases), processing of large batches of FASTA files is a complex task that usually requires advanced bioinformatics skills and the combination of different tools (including the bash command line) to achieve the desired result. In order to allow researchers to easily perform these operations we are developing the SEDA software application.

Results: SEDA

www.sing-group.org/seda



What is SEDA? *SEDA* is a desktop multiplatform application for ease the processing of FASTA files containing DNA and protein sequences.



SEDA Functions

- Sequence filtering.
- Sequence filtering based on text patterns.
- Remove redundant sequences.
- Sort sequences.
- Split FASTA files.
- Reallocate reference sequences.
- Rename sequence headers.
- Reformat FASTA files.
- Grow sequences.
- Rename NCBI accession numbers.
- Merge FASTA files.
- Concatenate sequences.
- Undo alignment.
- Sequence translation.
- Disambiguate sequence names.
- Consensus sequence creation.
- BLAST analysis.

SEDA for Programmers

Additionally:

- Programmers can take advantage of the *SEDA* core to develop new operations to process FASTA files.
- *SEDA* has a plugin-based architecture, new functions can be easily added to the *SEDA* GUI through the creation of new plugins.

Acknowledgements

This article is a result of the project Norte-01-0145-FEDER-000008 - Porto Neurosciences and Neurologic Disease Research Initiative at I3S, supported by Norte Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (FEDER). H. López-Fernández is supported by a post-doctoral fellowship from Xunta de Galicia (ED481B 2016/068-0). SING group thanks CITI (*Centro de Investigación, Transferencia e Innovación*) from University of Vigo for hosting its IT infrastructure. This work was partially funded by Consellería de Cultura, Educación e Ordenación Universitaria (Xunta de Galicia) and FEDER (European Union).

INSTITUTO DE INVESTIGAÇÃO E INOVAÇÃO EM SAÚDE
UNIVERSIDADE DO PORTO

Rua Alfredo Allen, 208
4200-135 Porto
Portugal

+351 220 408 800

www.i3s.up.pt