Multiple Sequence Alignment Project

Damien Garcia – Florian Echelard

January 2023

Contents

List of Figures

List of Tables

1 Compilation and execution

In order to facilitate usage of the program, a Makefile is provided in the projects archive allowing for easy compilation of the diverse programs:

- **`make all`** : Compiles the complete project.
- **`make clean`** : Removes all the compiled files and potential precompiled headers.
- **`make rebuild`** : Combination of **`make clean`** and **`make all`**
- **`make data_generation`** : Compiles the traces generation program.
- **`make alignment`** : Compiles the alignment program.
- **`make quality_analysis`** : Compiles the scoring program.

Six example are at disposal to test the program and check the execution process. The program runs syntaxic and semantic checks on the expression to avoid errors and unpredictable behavior during traces generation. Working examples and basic syntaxic/semantic error case can be executed with the commands:

- **`make test_simple`** : Working trace generation using simple generative regions.
- **`make test_complex`** : Working trace generation using complex generative regions.
- **`make test_semantic[1-4]`** : Execution returns syntaxic and semantic errors handled by the data_generation program[1].

Each section of the project can be executed independently with the same command line structure: **`./main [inputFile] [outputFile]`**. Input file is either the parameters file for the data_generation program, the traces to align for the alignment program, or the aligned traces for the quality_analysis program. Output file is the path where the results should be written.

A Python3 script is also at disposal for easy and efficient complete execution of the project. Use the command: **`python3 pipeline.py [pathToDatasetDirectory]`**. The pipeline will handle any necessary compilation, execute the program accordingly to previously explained process and create the results directories if necessary.

2 Traces generation

The first section of the project aims to generate sequences called traces which are composed of tops and events. Events can be anchors, meaning they are shared by all generated traces and only exist once for each trace, or simple events that depending on the generation parameters, either exist in a trace or not, and could in some cases have repetitions. For easier further analysis of the traces, tops are represented by a dot (“.”), events always start by a full case “E” for anchors and lower case “e” for other events.

In order to generate traces, the program relies on parameters files providing a RegEx-like[2] sequence, the number of traces to generate and the maximal size of the traces to generate.

[1] `make test_semantic4` is currently not working. In theory it should return an error, yet it generates traces.
[2] Regular Expression

Listing 1: Parameter file structure

```
# GENERATION PARAMETERS
expression=(2-3)E1<(4-8)e2X10|E2%25>E2(2-4)E3<(2-5)e1%40>
number_of_traces=20
maximum_length=100
```

The expression is separated in 3 sections types :

- Simple generative regions : Containing only tops, delimited by parenthesis.
- Complex generative regions : Containing tops and events, delimited by a less-than and greater-than signs.
- Anchors : Events outside of generative regions.

Table 1: Generative regions format

Expression	<	(	min	-	max	)	+	event	X	val	|	...	>
Required for simple generation		×	×	×	×	×							
Required for complex generation	×	×	×	×	×	×		×					×

Traces are generated and written in a single file, with on expression per line as seen below.

Listing 2: Traces generated from expression (2-3)E1<(4-8)e2X10|E2%25>E2(2-4)E3<(2-5)e1%40>

```
. . . . E1 . e2 . e2 E2 . . . . . E3 . . . . .
. . E1 . . e2 e2 . e2 . . E2 . . E3 e1 . e1 e1 e1
. . E1 . . . . . . . E2 . . . . . E3 . e1 e1
. . . . E1 e2 . e2 . . . . E2 . . E3 . . . . . .
. . . . E1 e2 . e2 . E2 . . . . . E3 . . .
. . . E1 . . . . . . . . . . . . E2 . . . . . E3 . .
. . . E1 . . . e2 . . e2 . . E2 . . . . E3 . . . . .
. . . . E1 . . e2 e2 . . . . . E2 . . . E3 . e1
. . E1 . e2 . . . e2 E2 . . E3 . . . e1
. . E1 . . . . . . . . . E2 . . E3 . e1 .
. . . . E1 e2 . . . . . . . e2 . . E2 . . E3 . . . . .
. . E1 . . . . . E2 . . . . . E3 . e1
. . . E1 . . . . . . . . . E2 . . . . . E3 . . .
. . E1 e2 e2 . . . . . E2 . . E3 . .
. . E1 . . . . . . . E2 . . . . E3 . e1 . e1 e1 e1
. . . . E1 . . . . . . . . E2 . . . . . E3 . e1 e1
. . . . E1 e2 . . . . . E2 . . . . . E3 . . .
. . . E1 . . . . . . . e2 E2 . . E3 . . . .
. . . E1 e2 . e2 e2 E2 . . . . . E3 . . .
. . . E1 . . e2 e2 e2 . . . E2 . . E3 . .
```

3 Multiple Sequence Alignment (MSA)

To perform the MSA, the program relies on two main methods being (i) pairwise alignment algorithm and (ii) Unweighted Pair Group Method with Arithmetic mean (UPGMA) algorithm. The decision guiding the order in which traces will be aggregated by UPGMA algorithm is by measuring the dissimilarity between each pair of traces. This measure is inherently linked to the pairwise alignment process.

Step 1 At first, all the traces are stored into a data type that can either contain one or more traces. This way, when a pair of traces will be aggregated, the algorithm won't have to rely on different methods and functions depending on the progress of the MSA. All functions and methods are programmed to process two single traces as well as batches of traces.

Step 2 Next, every pair is aligned which allows to get the dissimilarity score. The pair with the lowest dissimilarity score is then selected and modified with considerations to the alignment results. At this point the traces might include gaps represented by hyphens ("-"). The updated traces are then aggregated into one batch of traces.

Step 3 The dissimilarity matrix[3] is updated by removing the pairs and adding the new batch of traces. The dissimilarity score for the new batch is then calculated with every other traces and/or batches. This method allows the algorithm to avoid recalculating every dissimilarity score by keeping unmodified data and only calculating scores for new batches.

Steps 2 and **3** are repeated until all traces are aggregated into one batch.

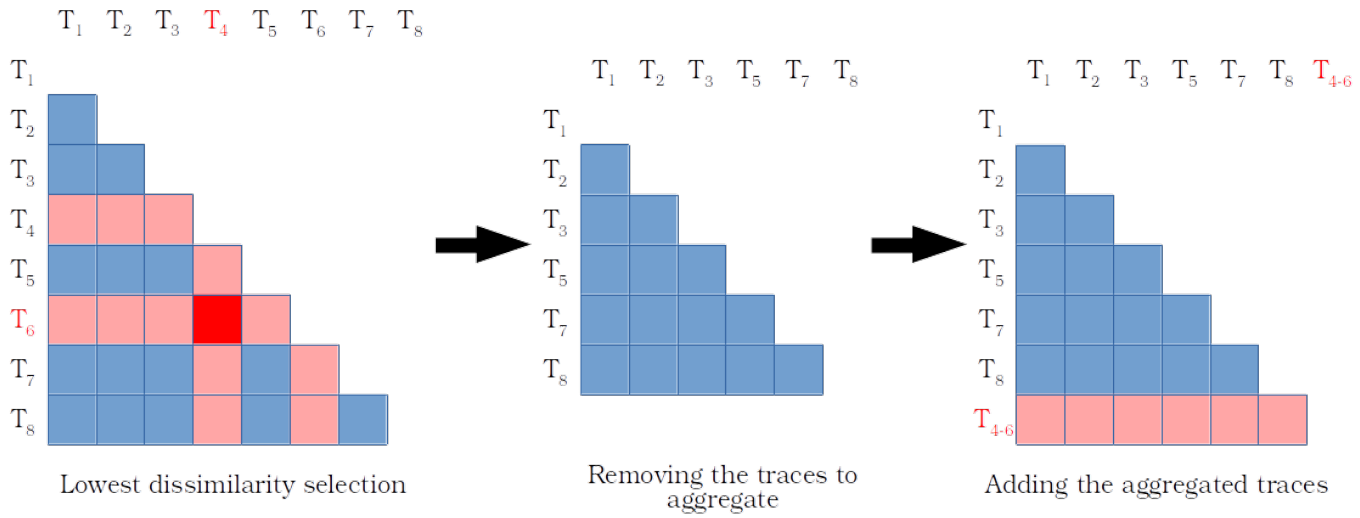

Figure 1: Blabla

4 MSA quality assessment

4.1 Scoring functions

4.2 Quality of the method

5 Prospects

Supplementary figures

[3]Represented by a lower triangular matrix for optimization purpose. Dissim(T1, T2) ≡ Dissim(T2, T1)