Sequence Alignments

Gepoliano Chaves, Ph. D.

March 20th, 2024

Overview

After completing today's session, students should:

- Be able to define computational algorithms via the Arab influence in Europe
- Define bioinformatics
- Define pseudo-code
- Practice global Needleman Wunsch algorithm in their notebook
- BLAST a sequence
- Download R
- Think about downloading and aligning sequences from the State of Bahia

Sequence alingnments: Global Alignment

1) Human language

- Describe what needs to be done in human language
- What the problem is
  - For global alignment: to align two whole sequences against each other to evaluate how much similar they are
  - Compare nucleotide by nucleotide, what the identity match is or is not
  - Account for nucleotides that have the same or different identities
- Needlemand-Wunsch algorithm
  - Alignment of the entire sequence
  - Match, mismatch and gap penalty score
  - https://www.youtube.com/watch?v=18vt6k-2Jbs
  - https://www.youtube.com/watch?v=FIxYGV7WPA8
  - https://www.slideshare.net/HarshitaBhawsar/needlemanwunch-algorithm-harshita

Sequence alingnments: Global Alignment

1) Human language: Goals of Sequence Alignments

- Goals of the alignment
  - Measure similarity
  - Observe patterns of sequence conservation between related biological species and variability of sequences over time and geographic location
  - Infer evolutionary relationships

Sequence alingnments: Global Alignment

ALGORITHM

Sequence alingnments: Global Alignment

1) Human language: Goals of Sequence Alignments

- Steps:
    - Initialization
    - Matrix fill or scoring
    - Traceback and alignment

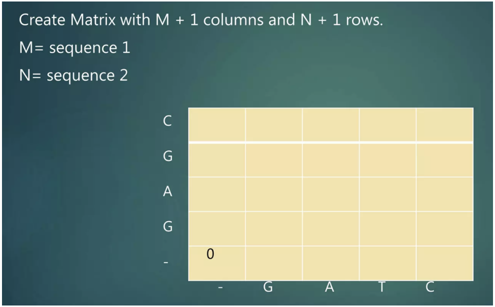
2) Pseudocode: Use equations to describe the calculations to be made in the algorithm

- Rules:
    - Fill the first column and the last row with gap values
    - Value of box beside + Gap value
    - Value of box bottom + Gap value
    - Diagonal value + {match/mismatch}

Sequence alingnments: Global Alignment

Initialization

- In your notebook, please create columns to align two sequences



Create Matrix with M + 1 columns and N + 1 rows.

M= sequence 1

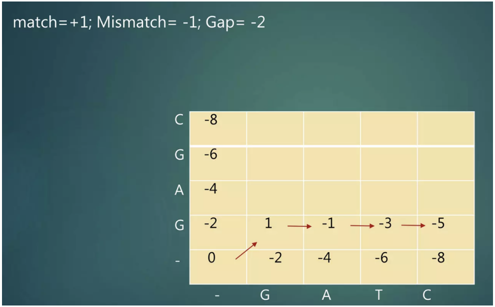
N= sequence 2

C					
G					
A					
G					
-	0				
	-	G	A	T	C

Figure 1: Scoring Matrix. Figure from Bhawsar (2016)

Sequence alingnments: Global Alignment

Scoring: Filling the matrix



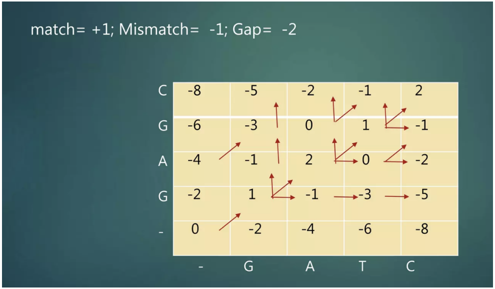
match=+1; Mismatch= -1; Gap= -2

C	-8				
G	-6				
A	-4				
G	-2	1	-1	-3	-5
-	0	-2	-4	-6	-8
	-	G	A	T	C

Figure 2: Scoring (filling) the matrix. Figure from Bhawsar (2016)

Sequence alingnments: Global Alignment

2) Pseudocode: Continuing the procedure



match= +1; Mismatch= -1; Gap= -2

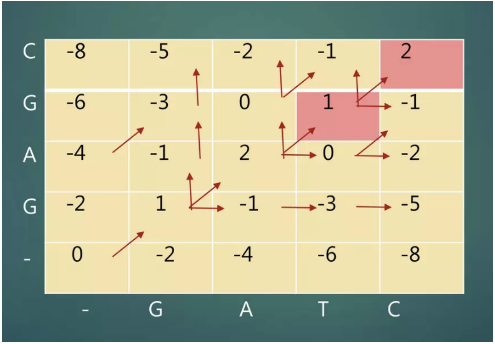
C	-8	-5	-2	-1	2
G	-6	-3	0	1	-1
A	-4	-1	2	0	-2
G	-2	1	-1	-3	-5
-	0	-2	-4	-6	-8
	-	G	A	T	C

Figure 3: Scoring (filling) the matrix. Figure from Bhawsar (2016)

Sequence alingnments: Global Alignment

2) Pseudocode: Implementation (not using code) of the Needlemand-Wunsch Scoring Matrix

- We could have used code to fill in this matrix
- For the Traceback step, we follow the pointers (the arrows)



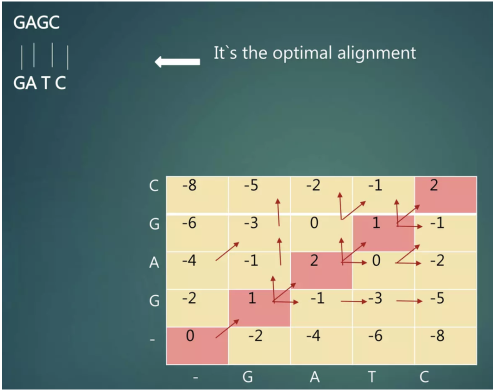
C	-8	-5	-2	-1	2
G	-6	-3	0	1	-1
A	-4	-1	2	0	-2
G	-2	1	-1	-3	-5
-	0	-2	-4	-6	-8
	-	G	A	T	C

Figure 4: Scoring Matrix. Figure from Bhawsar (2016)

Sequence alignments: Global Alignment

3) Traceback: alignment of the Needlemand-Wunsch Matrix

- We could have used code to fill in this matrix
- For the Traceback step, we follow the pointers (the arrows)



GAGC
| | | |
GA T C

It`s the optimal alignment

C	-8	-5	-2	-1	2
G	-6	-3	0	1	-1
A	-4	-1	2	0	-2
G	-2	1	-1	-3	-5
-	0	-2	-4	-6	-8
	-	G	A	T	C

Figure 5: Complete traceback and alignment. Figure from Bhawsar (2016)

Sequence alingnments: Local Alignment

- Smith-Waterman
  - Match +1, Mismatch -1, GAP penalty -2
  - https://www.youtube.com/watch?v=bFDRny7T3_s&t=3s
  - Query sequence vs. database sequence on a character to character level
  - Dinamic programming: divide problems into sub-problems for optimal solution
  - initalization, matrix filling and trace back
- BLAST
  - http://www.ncbi.nlm.nih.gov/BLAST/
  - Fragment of SARS-CoV-2 sequence to blast:
    - ACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAAC

Sequence alignements: the scoring system

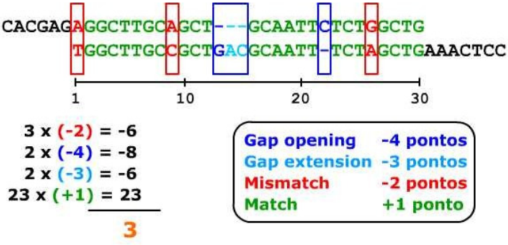


Figure 6: Sequence alignment: the scoring system.

Algorithm or Pipeline

The algorithm (also called a pipeline) needs to objectively explain how we go about answering our question or solving a problem

- Align to reference sequence (FASTA)
- Compare alignment to reference (SAM)
- Annotate differences (mutations) (VCF)
- Extract mutations from VCF (Frequency Table)

Bioinformatics Software Development

- Software development considers the analytical steps in human language
  - What are the exact steps that are necessary for execution of the analysis?
- Then, the software product considers the steps the machine will execute
- How files are produced and what are the processing steps?
- Where in the computational infra-structure are the files stored?

Bioinformatics Software Development

Conclusions

- We can develop our own computational methods to understand biology and propose solutions
- In order to do that we need to follow these three steps for developing a computational algorithm that will solve a problem:

Bioinformatics Software Development

Conclusions

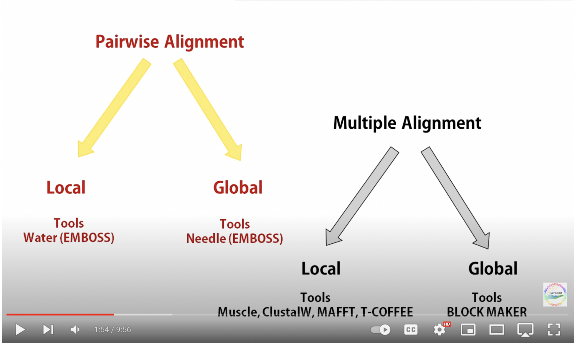
- 1) Describe the problem in human language and propose solutions in ways that are inteligible to human collaborators
- 2) Start using mathematical equations and figure out a computational language to write code to process data related to a problem
  - For example: the genetics of racial groups in Brazil, a population of mixed descent
- 3) Write a script or code to run computational experiments that demonstrate possibilities to solve or address the problem

Bioinformatics Software Development

- Let's move on to describe the DNA sequencing methods

Multiple Alignments

- In multiple sequences the alignment is much more significant than just two sequences
- Score higher when multiple sequences align
- The similarities refer to functional equivalence and evolutionary relationships between the two proteins



Pairwise Alignment

Local

Tools
Water (EMBOSS)

Global

Tools
Needle (EMBOSS)

Multiple Alignment

Local

Tools
Muscle, ClustalW, MAFFT, T-COFFEE

Global

Tools
BLOCK MAKER

Figure 7: Multiple Sequence Alignment.

Multiple Alignments

- Load the msa library

```
library(msa)
```

- Read FASTA file and create DNAStringSet object

```
dna_sarsCov2_start_30000 <- readDNAStringSet(file="~/Deskto
```

- Visualize DNAStringSet object

```
dna_sarsCov2_start_30000
```

HMM

- https://www.youtube.com/watch?v=vO_6xfLwGao
- https://www.youtube.com/watch?v=i3AkTO9HLXo
- Classifying proteins with Markov Chains
    - https://www.youtube.com/watch?v=HbA0odlLuZs

Bioinformatics Software Development (continued)

- How can these files be accessed?
- What information do the files containe?
- The present program is about how a scientific question is answered, not what the final answer is
- If how the question is answered is not addressed, opportunity is lost in terms of information that is embedded in the process of data analysis
- This is an important notion to have when developing computational tools that answer a scientific question

References

Bhawsar, Harshita. 2016. *Needleman-Wunch Algorithm*. https://www.slideshare.net/HarshitaBhawsar/needlemanwunch-algorithm-harshita.