A short introduction to multiple sequence alignments

Jean-Baka Domelevo Entfellner

BecA-ILRI Hub, Nairobi, Kenya

2nd EANBiT residential training
Kilifi, Kenya, 11 July 2019

biosciences eastern and central africa

EANBiT Eastern Africa Network for Bioinformatics Training

ILRI INTERNATIONAL LIVESTOCK RESEARCH INSTITUTE

Why do we align sequences?

The rationale for sequence alignments:

1. biological sequences have evolved in many different ways (insertions, deletions, inversions, translocations, etc)
2. yet, *homologous* sequences derive from a common *ancestral sequence*
3. to investigate evolution, we first have to *align* (i.e. put in the same column) characters that (are believed to) descend from a common ancestor

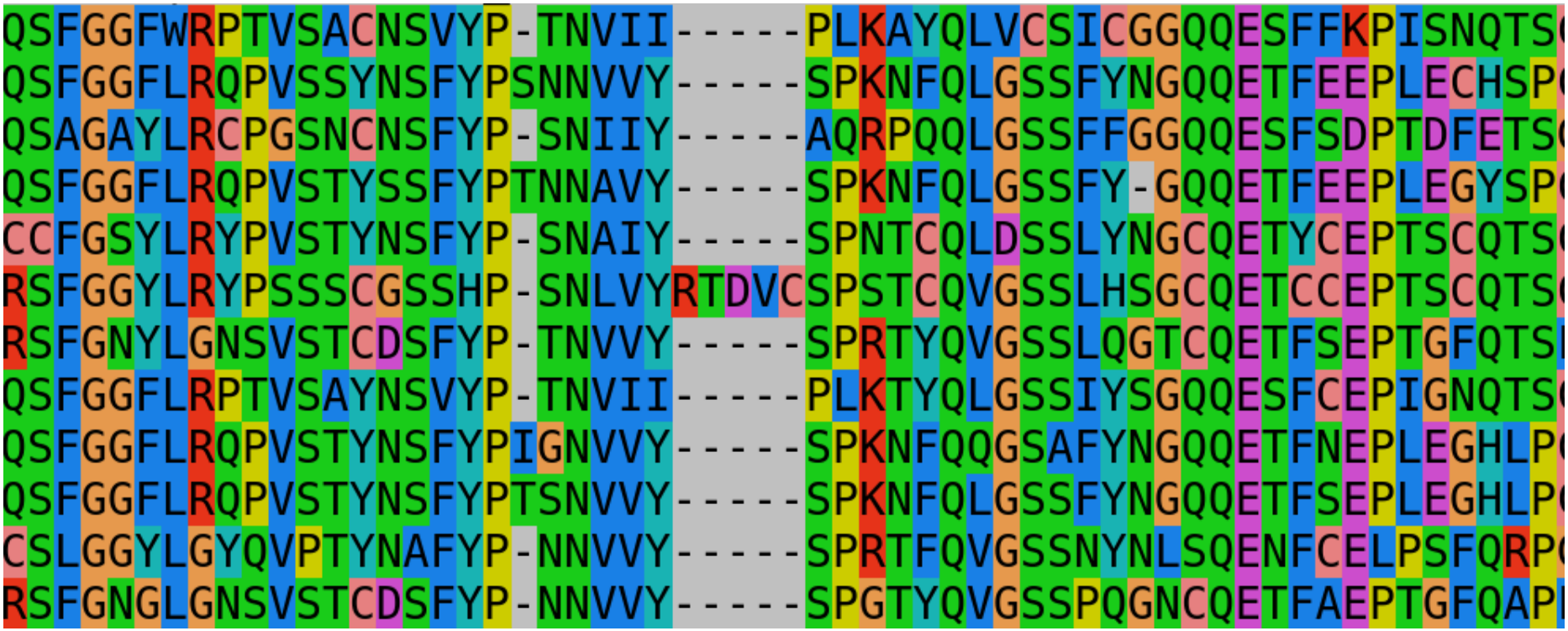
Caveat #1

Characters are often aligned based on **sequence similarity**, but the alignment with highest conservation scores is not necessarily the one depicting the true **evolutionary history**.

Some vocabulary

- **MSA**: **M**ultiple **S**equence **A**lignment
- **character**: the atomic component of a sequence (nucelotide, amino acid or quantitative trait)
- **homologous** sequences have evolved from a common ancestral sequence. Often, they still share the same function.
- **site**: a column in a multiple sequence alignment
- **gap**: a placeholder, *virtual* character ('-') used when displaying an MSA, to preserve the alignment within previous or subsequent sites (no "empty space" in an alignment)
- **indels** are insertions or deletions of one or several consecutive characters in a subset of the sequences. Whether they are called "insertions" or "deletions" only depends on the choice of a reference sequence.
- **conserved** sites are those where all taxa share the same character

Example of a proteic alignment



How do computers build MSA?

To align n sequences together, most automatic aligners use a **progressive alignment** greedy heuristic:

1. determine a **guide tree** using a quick, distance-based tree inference
   1. calculate all $n(n-1)/2$ pairwise alignments (easy: Needleman-Wunsch)
   2. determine the half-matrix of pairwise distances
   3. use some quick inference method (e.g. Neighbour-Joining) to get the guide tree
2. use a modified N-W algorithm to **perform successive** seq-to-seq, seq-to-group and group-to-group **"pairwise" alignments**
   1. pick the closest 2 sequences in the tree and align them
   2. go up into the guide tree, iteratively aggregating sequences or groups of sequences (consensus) to the alignment. Partial alignments are never un-done.
3. once no sequence remains unaligned, return the resulting alignment

Scoring schemes for pairwise alignments

Total score of a pairwise alignment made of:

1. match score: a positive number (reward) when aligning identical or similar characters
2. mismatch penalty: penalty when aligning discordant characters. Either fixed cost, or higher penalty for more dissimilar chars
3. gap opening penalty: a negative number (penalty) when opening a gap
4. gap extending penalty: a negative number (penalty) when extending a gap (usually penalized less than opening a new one)

Caveat # 2

Most aligners will prefer to "align garbage" than open several gaps.

Exercises: pairwise alignments

Let us use the following scoring scheme:

- match: +8
- mismatch: -3
- gapopen: -5
- gapextend: -3

Exercise 1

Calculate the alignment score for:

```
ATGGTT-TGA
A-TGTTTGA-
```

Can one find a better alignment (yielding a better score)?

Exercise 2

Align CATTGTGGA and CTTGTGA, and give (best) score.

Popular aligners

- Clustal, ClustalW, **ClustalΩ** (Des Higgins, Julie Thompson et al., 1988, 1994, 2011): the historical assembler. Thompson et al. 1994 (NAR) ranking #10 in *Nature*'s Top100
- **Muscle** (RC Edgar, 2004): fast and efficient
- **Mafft** (Katoh et al., 2002): very efficient, especially for protein sequences. Based on quick Fast Fourier Transforms on recoded data.
- **T-Coffee** (Notredame et al., 2010): interesting approach using aligned pairs from pairwise alignments
- **Prank** (Löytynoja & Goldman, 2008): attempt to take better care of inferred evolutionary events. Tends to insert more gaps.

Visualizing alignments: Seaview

Seaview's homepage:
`http://doua.prabi.fr/software/seaview`

Visualizing and editing alignments: Jalview

Jalview's homepage: `http://www.jalview.org/`