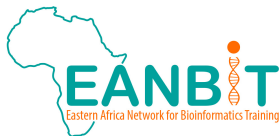


A short introduction to multiple sequence alignments

Jean-Baka Domelevo Entfellner

BecA-ILRI Hub, Nairobi, Kenya

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



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: Multiple Sequence Alignment**
- **character**: the atomic component of a sequence (nucleotide, 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

QSFGGFWRPTVSACNSVYP-TNVII-----PLKAYQLVCSICGGQQESFFKPI SNQTS
QSFGGFLRQPVSSYNSFYPSNNVVY-----SPKNFQLGSSFYNGQQETFEEPLECHSP
QSAGAYLRCPGSNCNSFYPTNNAVY-----AQRPQQLGSSFFGGQQESFSDPTDFETS
QSFGGFLRQPVSTYSSFYPTNNAVY-----SPKNFQLGSSFY-GQQETFEEPLEGYSP
CCFGSYLRYPVSTYNSFYPTNNAVY-----SPNTCQLDSSLYNGCQETTYCEPTSCQTS
RSFGGYLRYPSSSCGSSHPSNLVYRTDVCSPSTCQVGSSLHSGCQETCCEPTSCQTS
RSFGNYLGNSVSTCD SFYPTNVVY-----SPRTYQVGSSLQGTCCQETTFSEPTGFQTS
QSFGGFLRPTVSAYNSVYP-TNVII-----PLKTYQLGSSIYSGQQESFCEPIGNQTS
QSFGGFLRQPVSTYNSFYPIGNVVY-----SPKNFQQGSAFYNGQQETTFNEPLEGHL
QSFGGFLRQPVSTYNSFYPTSNVVY-----SPKNFQLGSSFYNGQQETTFSEPLEGHL
CSLGGYLG YQVPTYNAFYPTNNVVY-----SPRTFQVGSSNYNLSQENFCELP SFQRP
RSFGNGLGNSVSTCD SFYPTNNVVY-----SPGTYQVGSSPQGNCCQETFAEPTGFQAP

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.

- 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>

```
sel=0                                43                                104
UniRef90_A0A091DFM6  PLKAYQLVCSICGGQOESFFKPIISNQTSCGTGRSFQTS CFRPNNFI-SSPCQTNYTGS LGYG
UniRef90_A0A1A6HB96  SPKNFQLGSSFYNGQOETFEEPLECHSPCFGTRSFQASYFRPKQY-FSSPCHGGFTGSFGYG
UniRef90_A0A1S3G263  AQRPOQLGSSFFGGQOESFSDPTDFETSGVDA---TSCFRPKNFIFSRPCHTPYAGSFGCG
UniRef90_A0A1U7QK58  SPKNFQLGSSFY-GQQETFEEPLEGYSPCGGTRYFQTSSFRPKQY-FSSPCQGGFTGSFGYG
UniRef90_A0A2I2Z6F6  SPNTCQLDSSLYNGCQETICEPTSCQTSC TLRASYQTSCYCPKNSIFCSPRQTNYIRSLGCG
UniRef90_E2RS21      SPSTCQVGSSSLHSGCQETCCEPTSCQTSCVVSRPCQTSCYRPKSSIFFSPCQTNYTGS LGCG
UniRef90_G1TL07      SPRTYQVGSSSLQGTQOETFSEPTGFQTSFTVTRPCHTSFYHPKNSIFFSPCQTNHAGSFGHG
UniRef90_G5AZV2      PLKTYQLGSSSIYSGQOESFCEPIGNQTSCGTGRSFQTS CFHPSNSISFRPCQTNYTGS LGYG
UniRef90_M0RC38      SPKNFQOGSAFYNGQOETFNEPLEGHLPCVGTASFQTSCFRPKQY-FSSPCHGGFTGS LGYG
UniRef90_Q9QZU5      SPKNFQLGSSFYNGQOETFSEPLEGHLPCVGSASFHTSCFRPKQY-FSSPCQGGFTGSFGYG
UniRef90_UPI0002C472 SPRTFQVGSSNYNLSQENFCELP SFRPGFVGRSFQTSCYHPKNLILCSPCQTNFTESLGFG
UniRef90_UPI00032AF8 SPGTYQVGSSSPQGNCOETFAEPTGFQAPFAVTRPCHTSFYRPRNSIFSSSYQTNCAGSVGYG
UniRef90_UPI0003314A -----QMGPSPYGGCQESFFEPANFQPSYSGTRSFPA SCFRRKNSILYSPCQSN CAGPVAWG
UniRef90_UPI00033324 SPKSFHLGSSFYNGQOETFGEPI D CQETGAGFSYQSSCYRPKHFTFSRPCHANFTGSYGYG
UniRef90_UPI0003344D SPNNCQPGSSFYGGCQENFLEASGCQNPCAGTRTFPASCLRPKNMFMYN--QMNAV SQQCG
UniRef90_UPI00038C62 SPKNFQLGSSFY-GQQETFEEPLEDHSPCAGTSYFQTSCFRPKQY-FSRPCHGGFGGSFGYG
```

Visualizing and editing alignments: Jalview

Jalview's homepage: <http://www.jalview.org/>

