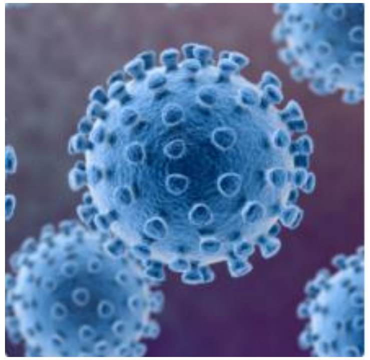


Viral Genomics and Bioinformatics (Latin America and the Caribbean)

10–14 October 2022 - Virtual course

wellcome connecting science

Introduction to multiple sequence alignments

Carolina Torres

Universidad de Buenos Aires, Facultad de Farmacia y Bioquímica, Instituto de Investigaciones en Bacteriología y Virología Molecular (IBaViM), Buenos Aires, Argentina.

Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires, Argentina.

caro.torr@gmail.com | ctorres@ffyb.uba.ar

@CaroTorr

Goals for sampling and sequencing

- **Genomic surveillance and molecular epidemiology:** frecuency of lineages, variants or mutations.

- **Origin and evolution:** introductions, diversification pattern.

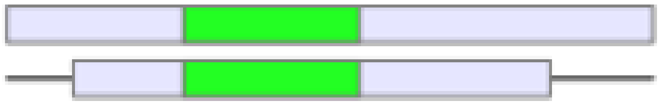
- **Outbreaks and transmission chains:** common sources of infection.

- **Evolutionary dynamics:** ancestral ages, rates of evolution, viral demography, dispersion rates and patterns, ancestral locations, predictors.

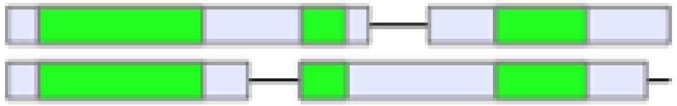
Alignments

✓ **An arrangement of DNA, RNA or amino acid sequences in which homologous sites are in the same position (they are “aligned”).**

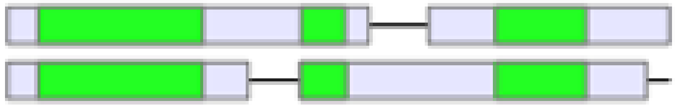
- **Local:** Align one or more stretches of similarity



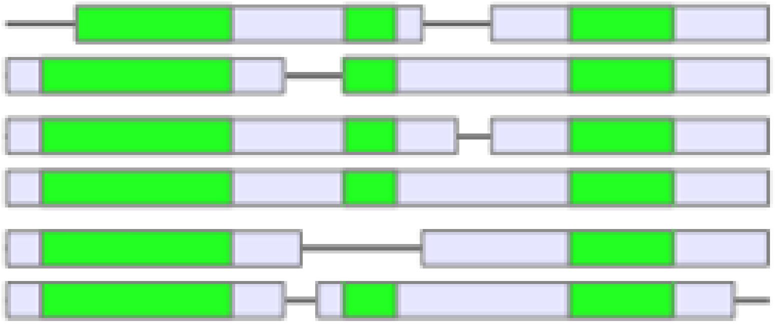
- **Pairwise:** Align two sequences



- **Global:** Align sequences end-to-end

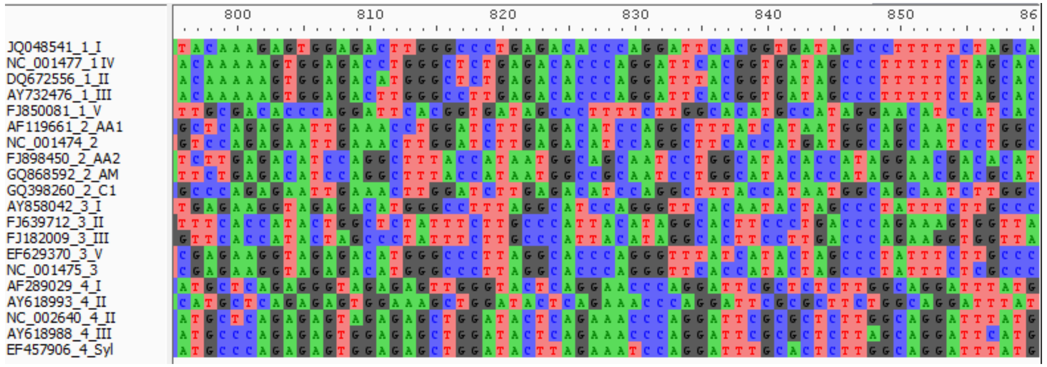


- **Multiple:** Align more than two sequences

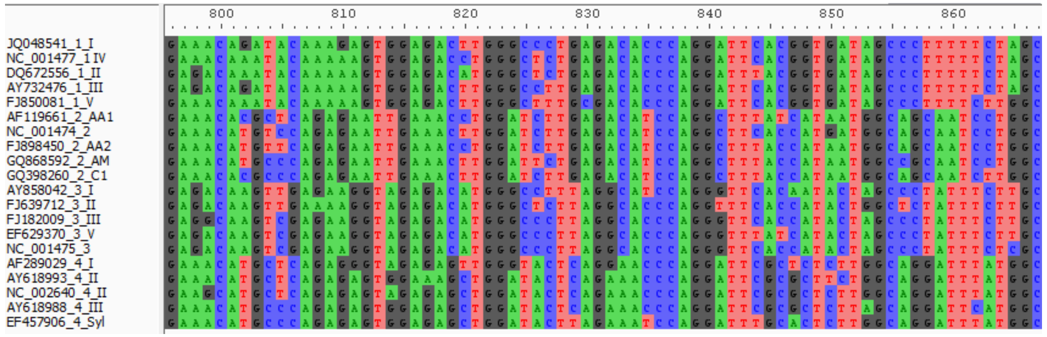


Multiple sequence alignments (MSA)

✓ **Positional homology**



Non-aligned sequences



Aligned sequences

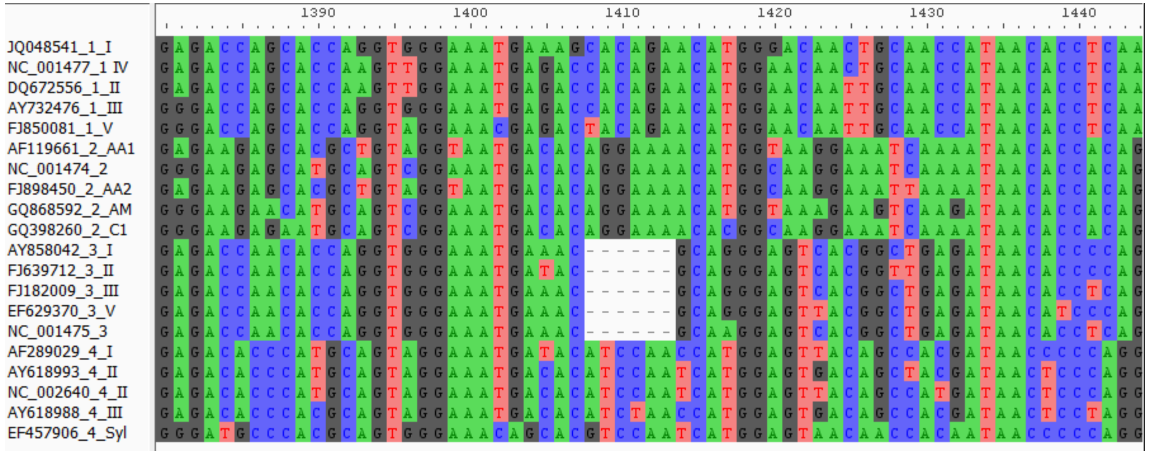
- The nts or AAs found at a position (column) in the sequences are considered descendants from a common ancestral site (homologous).

- All methods attempt to maximize matches/similarities and minimize mismatches/differences between sequences.

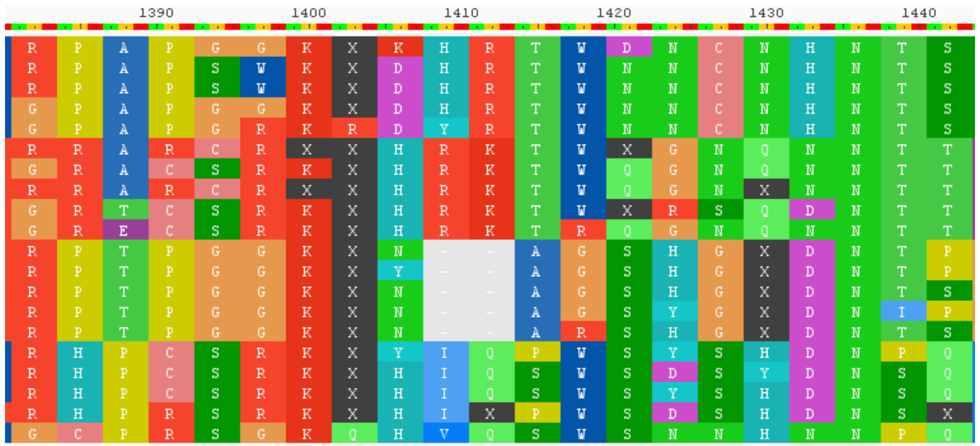
- Different programs search for the “best” alignment through different methods.

Multiple sequence alignments (MSA)

✓ **Positional homology**



```
                       1390      1400      1410      1420      1430      1440
JQ048541_1_I    GAGACCAGCACCAGGTGGGAAATGAAAGCACAGAACATGGGACAACTGCAACCATAACACCTCAA
NC_001477_1 IV  GAGACCAGCACCAAGTTGGAAATGAGACCACAGAACATGGAACAACTGCAACCATAACACCTCAA
DQ672556_1_II   GAGACCAGCACCAAGTTGGAAATGAGACCACAGAACATGGAACAATTGCAACCATAACACCTCAA
AY732476_1_III  GGGACCAGCACCAGGTGGGAAATGAGACCACAGAACATGGAACAATTGCAACCATAACACCTCAA
FJ850081_1_V    GGGACCAGCACCAGGTAGGAAACGAGACTACAGAACATGGAACAATTGCAACCATAACACCTCAA
AF119661_2_AA1  GAGAAGAGCACGCTGTAGGTAATGACACAGGAAAACATGGTAAGGAAATCAAAATAACACCACAG
NC_001474_2     GGGAAGAGCATGCAGTCGGAAATGACACAGGAAAACATGGCAAGGAAATCAAAATAACACCACAG
FJ898450_2_AA2  GAGAAGAGCACGCTGTAGGTAATGACACAGGAAAACATGGCAAGGAAATTAAAATAACACCACAG
GQ868592_2_AM   GGGAAGAACATGCAGTCGGAAATGACACAGGAAAACATGGTAAAGAAGTCAAGATAACACCACAG
GQ398260_2_C1   GGGAAGAGAATGCAGTCGGAAATGACACAGGAAAACACGGCAAGGAAATCAAAATAACACCACAG
AY858042_3_I    GAGACCAACACCAGGTGGGAAATGAAAC------GCAGGGAGTCACGGCTGAGATAACACCCCAG
FJ639712_3_II   GAGACCAACACCAGGTGGGAAATGATAC------GCAGGGAGTCACGGTTGAGATAACACCCCAG
FJ182009_3_III  GAGACCAACACCAGGTGGGAAATGAAAC------GCAGGGAGTCACGGCTGAGATAACACCTCAG
EF629370_3_V    GAGACCAACACCAGGTGGGAAATGAAAC------GCAGGGAGTTACGGCTGAGATAACATCCCAG
NC_001475_3     GAGACCAACACCAGGTGGGAAATGAAAC------GCAAGGAGTCACGGCTGAGATAACACCTCAG
AF289029_4_I    GAGACACCCATGCAGTAGGAAATGATACATCCAACCATGGAGTTACAGCCACGATAACCCCCAGG
AY618993_4_II   GAGACACCCATGCAGTAGGAAATGACACATCCAATCATGGAGTGACAGCTACGATAACTCCCAGG
NC_002640_4_II  GAGACACCCATGCAGTAGGAAATGACACATCCAATCATGGAGTTACAGCCATGATAACTCCCAGG
AY618988_4_III  GAGACACCCACGCAGTAGGAAATGACACATCTAACCATGGAGTGACAGCCACGATAACTCCTAGG
EF457906_4_Syl  GGGATGCCCACGCAGTGGGAAACAGCACGTCCAATCATGGAGTAACAACCACAATAACCCCCAGG
```



```
       1390      1400      1410      1420      1430      1440
 R  P  A  P  G  G  K  X  K  H  R  T  W  D  N  C  N  H  N  T  S
 R  P  A  P  S  W  K  X  D  H  R  T  W  N  N  C  N  H  N  T  S
 R  P  A  P  S  W  K  X  D  H  R  T  W  N  N  C  N  H  N  T  S
 G  P  A  P  G  G  K  X  D  H  R  T  W  N  N  C  N  H  N  T  S
 G  P  A  P  G  R  K  R  D  Y  R  T  W  N  N  C  N  H  N  T  S
 R  R  A  R  C  R  X  X  H  R  K  T  W  X  G  N  Q  N  N  T  T
 G  R  A  C  S  R  K  X  H  R  K  T  W  Q  G  N  Q  N  N  T  T
 R  R  A  R  C  R  X  X  H  R  K  T  W  Q  G  N  X  N  N  T  T
 G  R  T  C  S  R  K  X  H  R  K  T  W  X  R  S  Q  D  N  T  T
 G  R  E  C  S  R  K  X  H  R  K  T  R  Q  G  N  Q  N  N  T  T
 R  P  T  P  G  G  K  X  N  -  -  A  G  S  H  G  X  D  N  T  P
 R  P  T  P  G  G  K  X  Y  -  -  A  G  S  H  G  X  D  N  T  P
 R  P  T  P  G  G  K  X  N  -  -  A  G  S  H  G  X  D  N  T  S
 R  P  T  P  G  G  K  X  N  -  -  A  G  S  Y  G  X  D  N  I  P
 R  P  T  P  G  G  K  X  N  -  -  A  R  S  H  G  X  D  N  T  S
 R  H  P  C  S  R  K  X  Y  I  Q  P  W  S  Y  S  H  D  N  P  Q
 R  H  P  C  S  R  K  X  H  I  Q  S  W  S  D  S  Y  D  N  S  Q
 R  H  P  C  S  R  K  X  H  I  Q  S  W  S  Y  S  H  D  N  S  Q
 R  H  P  R  S  R  K  X  H  I  X  P  W  S  D  S  H  D  N  S  X
 G  C  P  R  S  G  K  Q  H  V  Q  S  W  S  N  N  H  N  N  P  Q
```

- The goal of alignment methods is to maximise a score based on:
  - rewarding matches (+ score).
  - penalising mismatches/rare substitutions (- score).
  - penalising gaps or indels (- score), which requires a gap opening and extension penalty scheme (these values are chosen arbitrarily).

- Gaps are introduced in the sequences or at the ends of the alignment.
- In coding regions, gaps are usually introduced in triplets.

Programs to build alignments

- **Clustal W/X: Progressive alignment.**
- **Muscle: Iterative method.**
- **MAFFT: multiple methods.**
- **T-Coffee:** consistency-based alignment (capable of combining a collection of multiple/pairwise, global/local alignments into one).
- **ProbCons:** combination of probabilistic modeling and consistency-based alignment techniques (protein sequences).
- **Probalign:** combines amino acid posterior probability estimation using partition function methods and computation of maximal expected accuracy alignment.

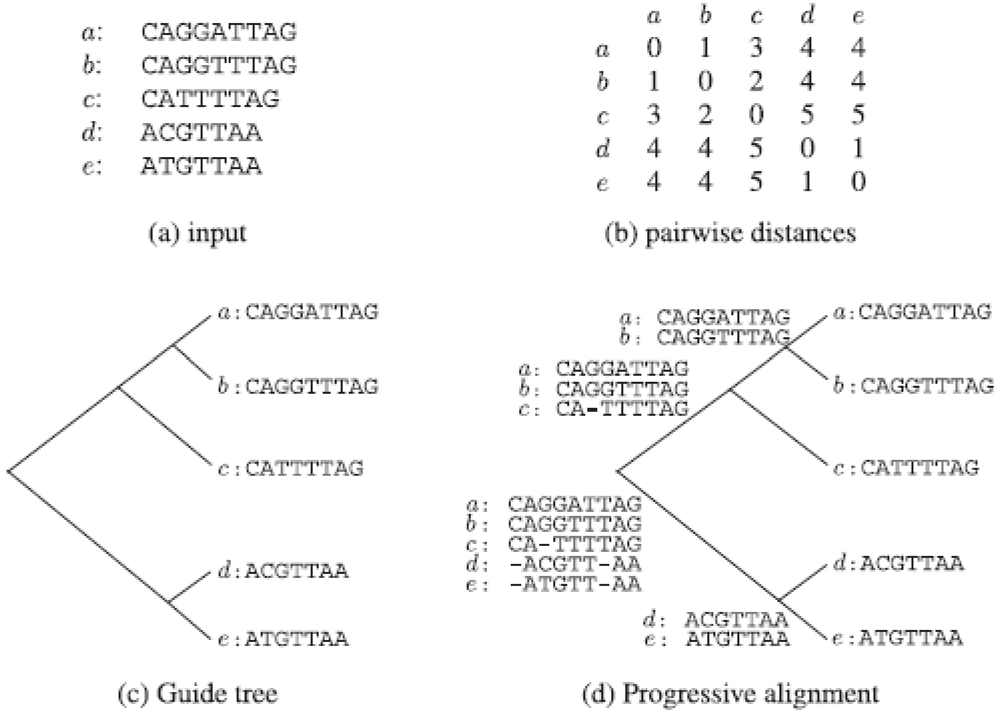
Methods to build alignments

- The **progressive algorithm** consists of three main stages:

  (i) All pairs of sequences are aligned separately (**pairwise alignments**) in order to calculate a "distance" matrix (this is done using dynamic programming);

  (ii) A guide tree is built from the distance matrix (using a clustering algorithm, such as Neighbor Joining);

  (iii) The guide tree is used to cluster and align the sequences progressively according to the branching order (starting with the two closest sequences and ending with the most distant).



a: CAGGATTAG
b: CAGGTTTAG
c: CATTTTAG
d: ACGTTAA
e: ATGTTAA

(a) input

	a	*b*	*c*	*d*	*e*
a	0	1	3	4	4
b	1	0	2	4	4
c	3	2	0	5	5
d	4	4	5	0	1
e	4	4	5	1	0

(b) pairwise distances

(c) Guide tree

```
a: CAGGATTAG
b: CAGGTTTAG

a: CAGGATTAG
b: CAGGTTTAG
c: CA-TTTTAG

a: CAGGATTAG
b: CAGGTTTAG
c: CA-TTTTAG
d: -ACGTT-AA
e: -ATGTT-AA

d: ACGTTAA
e: ATGTTAA
```

(d) Progressive alignment

https://doi.org/10.1017/CBO9780511974076.004

✓ Gaps are inserted at identical positions in all sequences of a cluster and are preserved.

- **Programs:** Clustal W/X

Methods to build alignments

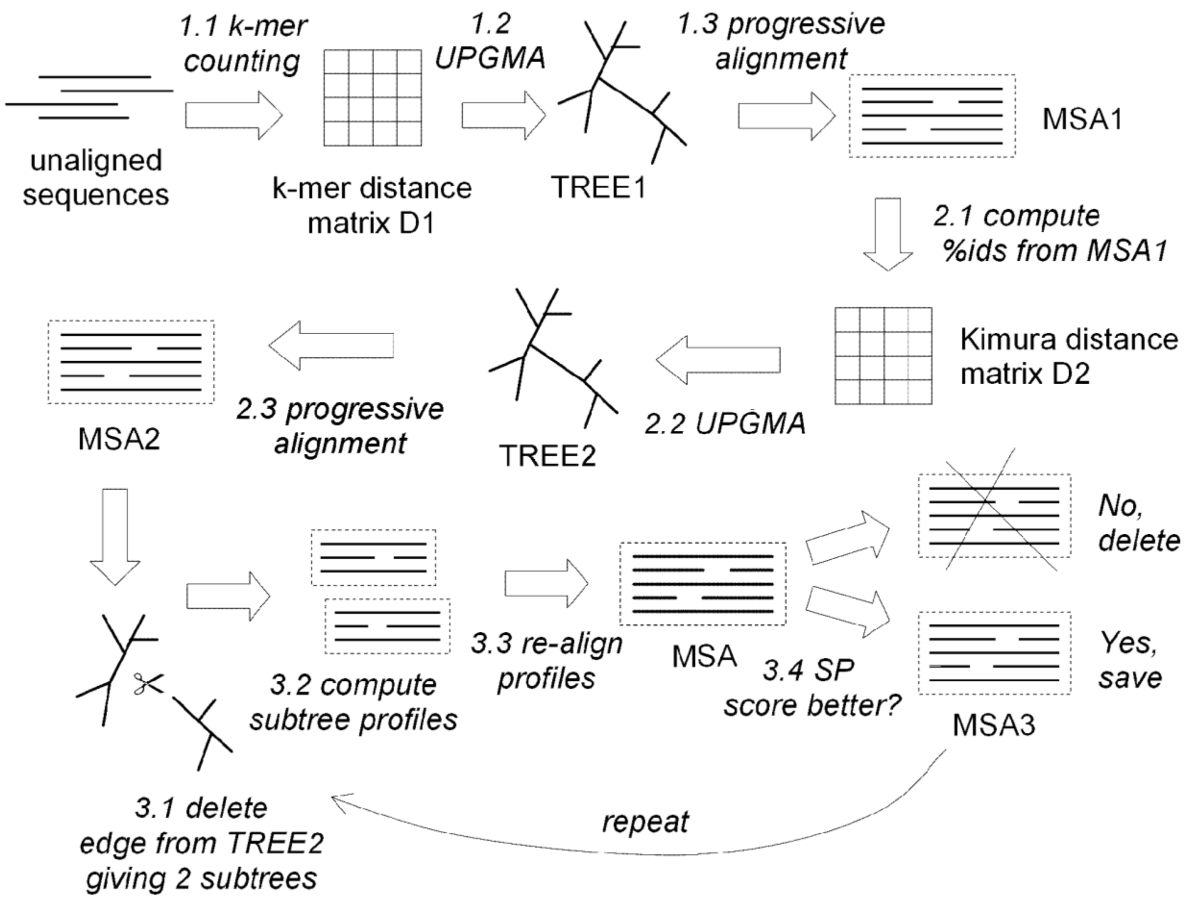
- The **iterative methods** work similarly to progressive methods, but they repeatedly realign the initial sequences and add new sequences to the MSA.

- **MUSCLE:** there are three main stages:

  (i) Draft progressive to build a MSA (from a distance matrix with approximate values)**;**

  (ii) Improved progressive: a new distance matrix is created from the first MSA and sequences realigned to reflect new guide tree;

  (iii) Refinement: The tree is split into 2 subtrees, profiles are built and aligned, different bipartitions are tried until convergence is reached.

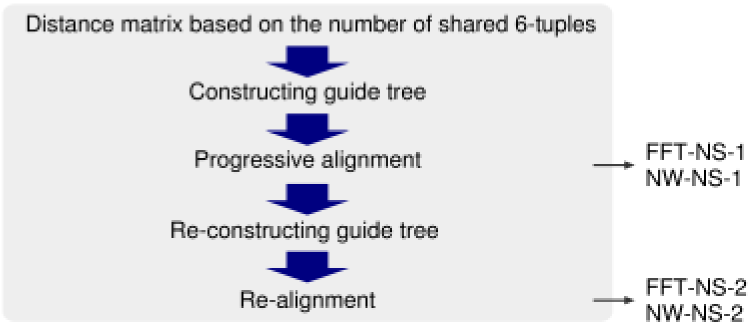


unaligned sequences
1.1 k-mer counting
k-mer distance matrix D1
1.2 UPGMA
TREE1
1.3 progressive alignment
MSA1
2.1 compute %ids from MSA1
Kimura distance matrix D2
2.2 UPGMA
TREE2
2.3 progressive alignment
MSA2
3.1 delete edge from TREE2 giving 2 subtrees
3.2 compute subtree profiles
3.3 re-align profiles
MSA
3.4 SP score better?
No, delete
Yes, save
MSA3
repeat

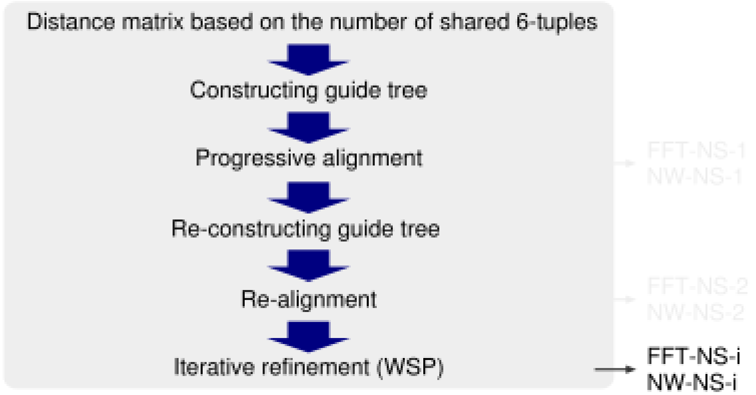
Methods to build alignments

- **MAFFT** offers a range of multiple alignment strategies:

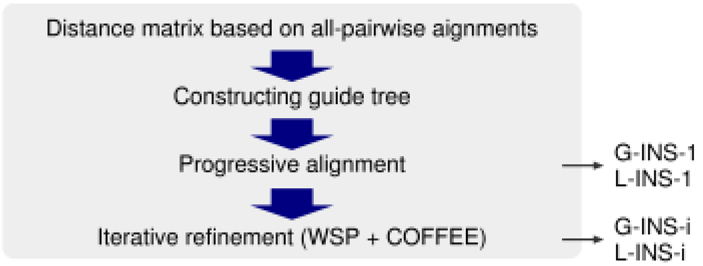
(a) FFT-NS-1, FFT-NS-2 — Progressive methods



(b) FFT-NS-i, NW-NS-i — Iterative refinement method


Distance matrix based on the number of shared 6-tuples
Constructing guide tree
Progressive alignment
Re-constructing guide tree
Re-alignment
Iterative refinement (WSP)
FFT-NS-1
NW-NS-1
FFT-NS-2
NW-NS-2
FFT-NS-i
NW-NS-i

(c) L-INS-i, E-INS-i, G-INS-i — Iterative refinement methods using WSP and consistency scores

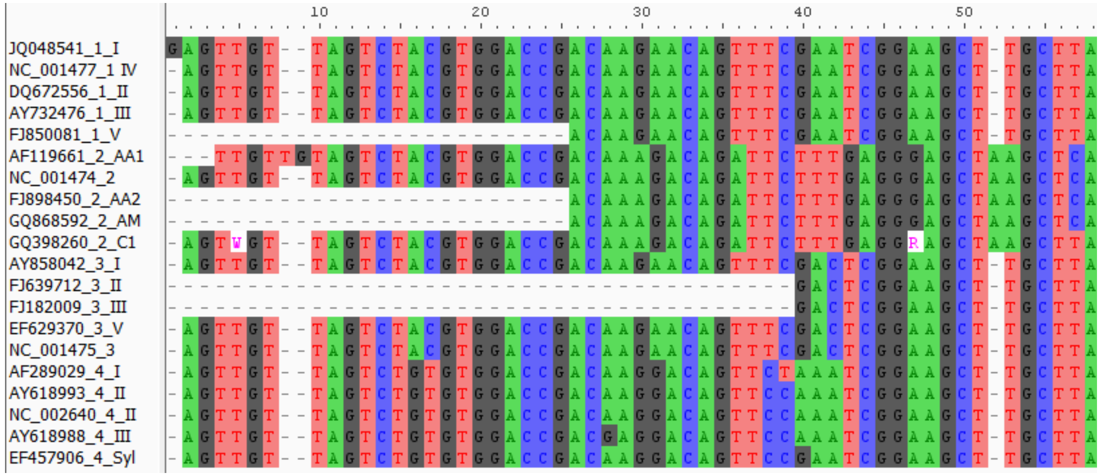

Distance matrix based on all-pairwise alignments
Constructing guide tree
Progressive alignment
Iterative refinement (WSP + COFFEE)
G-INS-1
L-INS-1
G-INS-i
L-INS-i

In general, there is a tradeoff between speed and accuracy. The order of speed is a > b > c, whereas the order of accuracy is a < b < c.

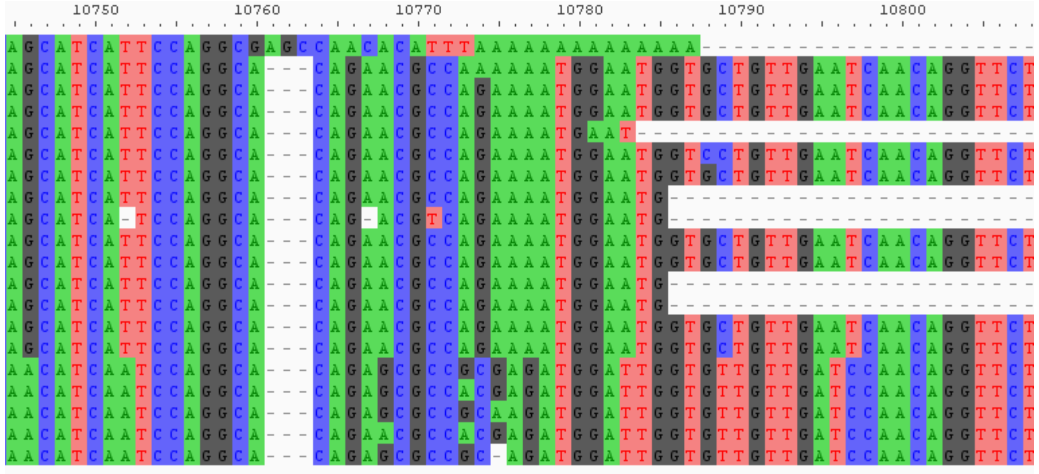
https://mafft.cbrc.jp/alignment/software/algorithms/algorithms.html

Editing alignments

- **It is a good practice to always visually inspect the alignment** to check the position of gaps, outlier sequences and poorly-aligned regions.



...

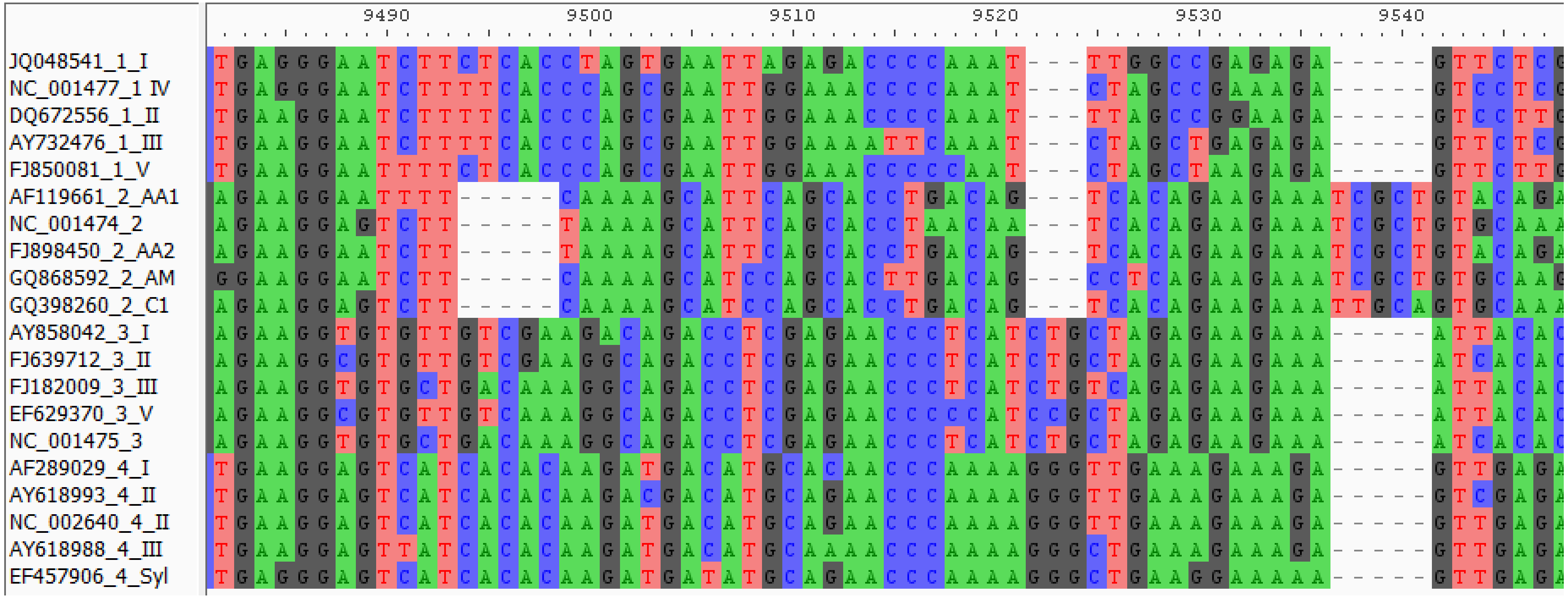


- **Manual edition may be necessary:**
  - Trim the ends or specific positions that cannot be aligned unambiguously (Gblocks may be useful).
  - Realign blocks.
  - Use biological knowledge to improve the alignment.

- **Program:** Aliview, BioEdit, Seaview, MEGA, UGENE, others.

Editing alignments

- **It is a good practice to always visually inspect the alignment** to check the position of gaps, outlier sequences and poorly-aligned regions.



```
                       9490      9500      9510      9520      9530      9540
JQ048541_1_I     TGAGGGAATCTTCTCACCTAGTGAATTAGAGACCCCAAAT---TTGGCCGAGAGA-----GTTCTCG
NC_001477_1 IV   TGAGGGAATCTTTTCACCCAGCGAATTGGAAACCCCAAAT---CTAGCCGAAAGA-----GTCCTCG
DQ672556_1_II    TGAAGGAATCTTTTCACCCAGCGAATTGGAAACCCCAAAT---TTAGCCGGAAGA-----GTCCTTG
AY732476_1_III   TGAAGGAATCTTTTCACCCAGCGAATTGGAAAATTCAAAT---CTAGCTGAGAGA-----GTTCTCG
FJ850081_1_V     TGAAGGAATTTTCTCACCCAGCGAATTGGAAACCCCCAAT---CTAGCTAAGAGA-----GTTCTTG
AF119661_2_AA1   AGAAGGAATTTT-----CAAAAGCATTCAGCACCTGACAG---TCACAGAAGAAATCGCTGTACAGA
NC_001474_2      AGAAGGAGTCTT-----TAAAAGCATTCAGCACCTAACAA---TCACAGAAGAAATCGCTGTGCAAA
FJ898450_2_AA2   AGAAGGAATCTT-----TAAAAGCATTCAGCACCTGACAG---TCACAGAAGAAATCGCTGTACAGA
GQ868592_2_AM    GGAAGGAATCTT-----CAAAAGCATCCAGCACTTGACAG---CCTCAGAAGAAATCGCTGTGCAAG
GQ398260_2_C1    AGAAGGAGTCTT-----CAAAAGCATCCAGCACCTGACAG---TCACAGAAGAAATTGCAGTGCAAA
AY858042_3_I     AGAAGGTGTGTTGTCGAAGACAGACCTCGAGAACCCTCATCTGCTAGAGAAGAAA-----ATTACAC
FJ639712_3_II    AGAAGGCGTGTTGTCGAAGGCAGACCTCGAGAACCCTCATCTGCTAGAGAAGAAA-----ATCACAC
FJ182009_3_III   AGAAGGTGTGCTGACAAAGGCAGACCTCGAGAACCCTCATCTGTCAGAGAAGAAA-----ATTACAC
EF629370_3_V     AGAAGGCGTGTTGTCAAAGGCAGACCTCGAGAACCCCCATCCGCTAGAGAAGAAA-----ATTACAC
NC_001475_3      AGAAGGTGTGCTGACAAAGGCAGACCTCGAGAACCCTCATCTGCTAGAGAAGAAA-----ATCACAC
AF289029_4_I     TGAAGGAGTCATCACACAAGATGACATGCACAACCCAAAAGGGTTGAAAGAAAGA-----GTTGAGA
AY618993_4_II    TGAAGGAGTCATCACACAAGACGACATGCAGAACCCAAAAGGGTTGAAAGAAAGA-----GTCGAGA
NC_002640_4_II   TGAAGGAGTCATCACACAAGATGACATGCAGAACCCAAAAGGGTTGAAAGAAAGA-----GTTGAGA
AY618988_4_III   TGAAGGAGTTATCACACAAGATGACATGCAAAACCCAAAAGGGCTGAAAGAAAGA-----GTTGAGA
EF457906_4_Syl   TGAGGGAGTCATCACACAAGATGATATGCAGAACCCAAAAGGGCTGAAGGAAAAA-----GTTGAGA
```

- **The final decision on what to include or exclude is yours.**

Sensitivity analysis
(check the impact of your decisions)

References

- The Phylogenetic Handbook: A Practical Approach to DNA and Protein Phylogeny. Marco Salemi, Anne-Mieke Vandamme (Eds). Cambridge University Press. (2009).
- Katoh, Standley 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* 30:772-780.
- Clustal W and Clustal X version 2.0. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. Bioinformatics. 2007 Nov1;23(21):2947-8.
- Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004 Mar 19;32(5):1792-7.
- BLAST® Help. Bethesda (MD): National Center for Biotechnology Information (US); 2008 (http://www.ncbi.nlm.nih.gov/books/NBK1762/).
- Talavera, G., and Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Systematic Biology 56, 564-577. http://molevol.cmima.csic.es/castresana/Gblocks.html