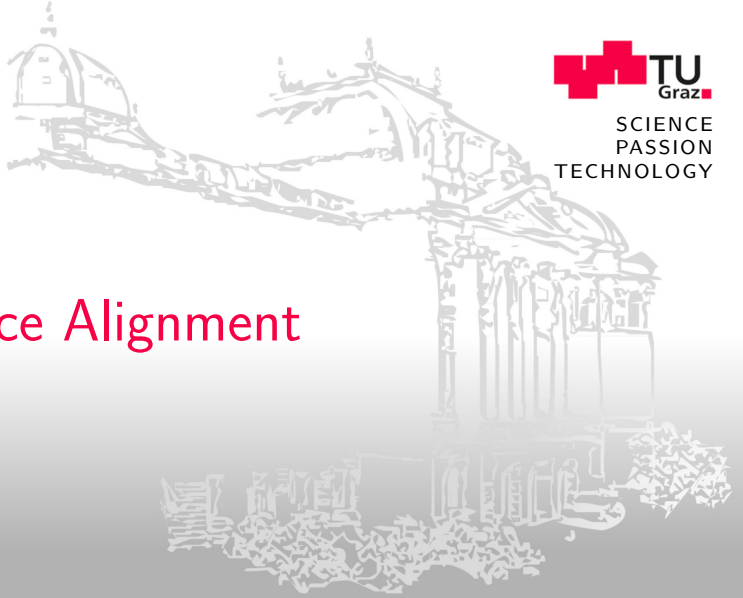


SCIENCE
PASSION
TECHNOLOGY

Multiple Sequence Alignment

Klaus Schliep, PhD

klaus.schliep@tugraz.at

Basics

- Sequence alignments
  - Global alignment
  - Local alignment (BLAST)
  - Multiple sequence alignment (based on global pairwise alignment)

Klaus Schliep, PhD, ICBT
klaus.schliep@tugraz.at

Why do we need Alignments?

- sequence search
- comparison of two nucleotide, amino acid or more general character sequences with each other
- highlight similarity between sequences and discover homologous regions
- Alignment is the procedure of writing two (or more) sequences in a way that a maximum of identical or similar characters are placed in the same column by adding gap "-" characters

Klaus Schliep, PhD, ICBT
klaus.schliep@tugraz.at

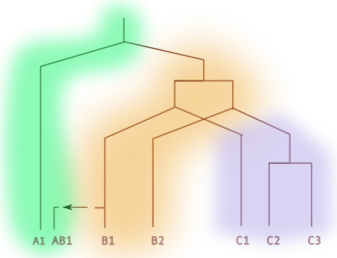
Alignment

```
Align 2 sequences: KLAUS, NIKOLAUS

Sequence 1: LAUS
            ||||
Sequence 2: LAUS
local alignment

Sequence 1: --K-LAUS
              | ||||
Sequence 2: NIKOLAUS
5 matches, 3 indels (gaps), low gap opening penalty

Sequence 1: ---KLAUS
               :||||
Sequence 2: NIKOLAUS
4 matches, 3 indels, 1 substitution (mutation), high gap opening penalty
```


Klaus Schliep, PhD, ICBT
klaus.schliep@tugraz.at

Sequence Homology

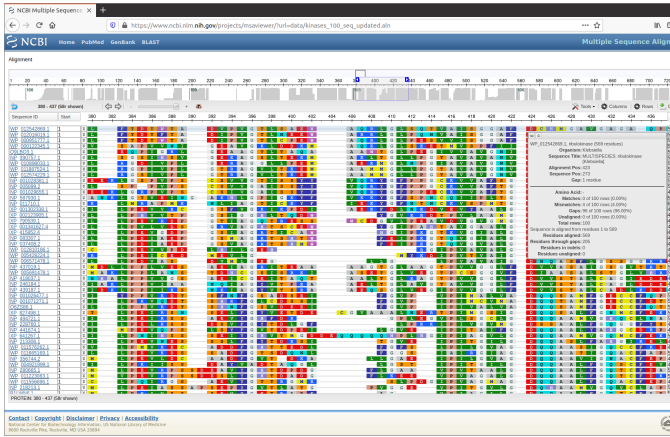
- Two sequences are homologous if they share a common ancestor. Sequences can share an ancestor because of either a speciation event (orthologs), a duplication event (paralogs) or horizontal gene transfer (xenolog) / hybridization.
- Sequences similarity does not imply homology.



Klaus Schliep, PhD, ICBT
klaus.schliep@tugraz.at

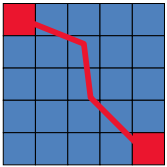
MSA

- Matrix of residues and gaps from 3+ biological sequences
- Residues in same row are from same sequence
- Residues in same column share some similarity
- Different columns contain different patterns of residue distributions

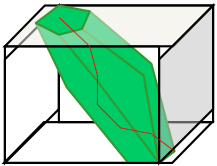


Klaus Schliep, PhD, ICBT
klaus.schliep@tugraz.at

Dynamic programming



Dynamic Programming
2 Sequences



Multi-Dimentional
Dynamic Programming
N Sequences

Klaus Schliep, PhD, ICBT
klaus.schliep@tugraz.at

Dynamic programming

- Instead of a matrix we will have a hyper-cube!
- Dynamic programming will only work for up to 10 sequences!
- Heuristic methods are needed: progressive alignment

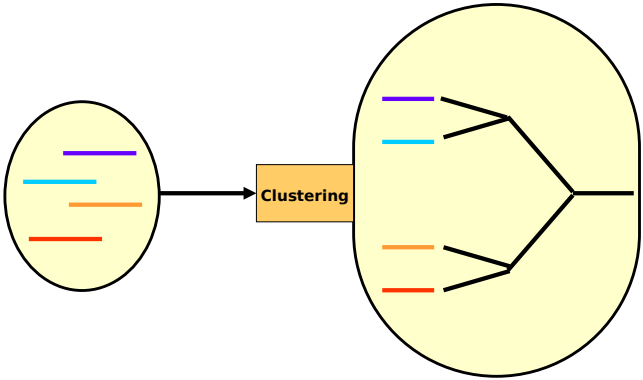
Klaus Schliep, PhD, ICBT
klaus.schliep@tugraz.at

Progressive alignment

1. Compute a pairwise distance matrix (e.g. from global alignment scores)
2. Compute a guide tree (e.g. NJ or UPGMA)
3. Progressively align along the nodes of the tree from the tips to the root
4. Maybe iterate step 2 and 3 using the alignment to improve

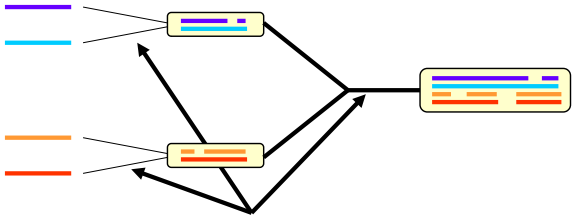
Klaus Schliep, PhD, ICBT
klaus.schliep@tugraz.at

Progressive alignment



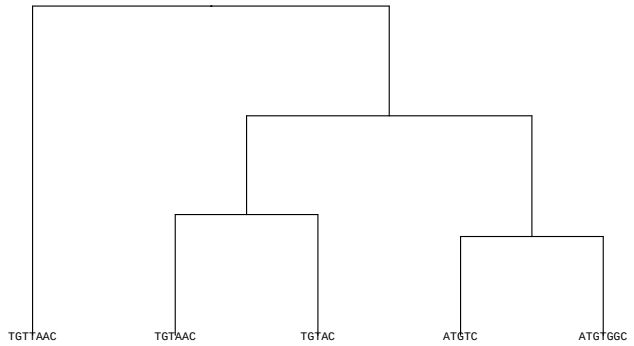
Klaus Schliep, PhD, ICBT
klaus.schliep@tugraz.at

Progressive alignment



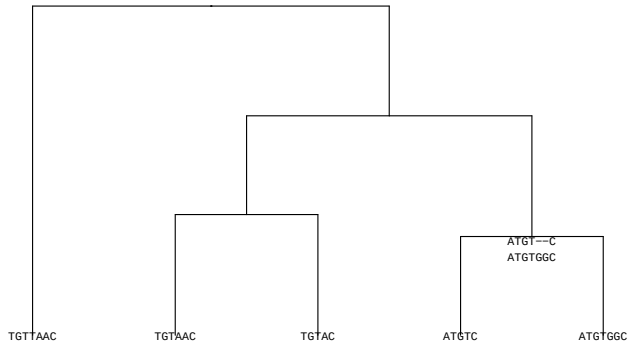
Klaus Schliep, PhD, ICBT
klaus.schliep@tugraz.at

Progressive alignment example



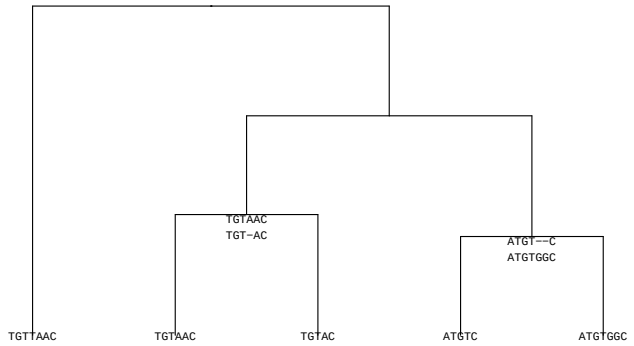
Klaus Schliep, PhD, ICBT
klaus.schliep@tugraz.at

Progressive alignment example



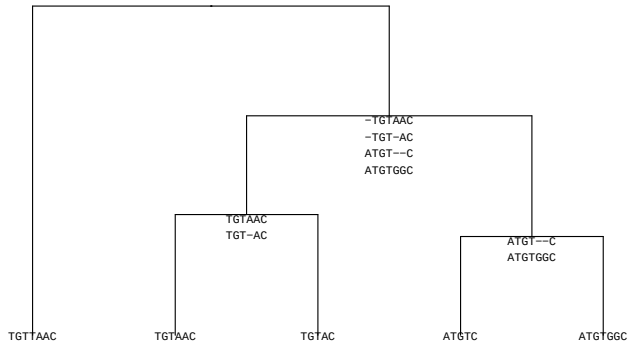
Klaus Schliep, PhD, ICBT
klaus.schliep@tugraz.at

Progressive alignment example



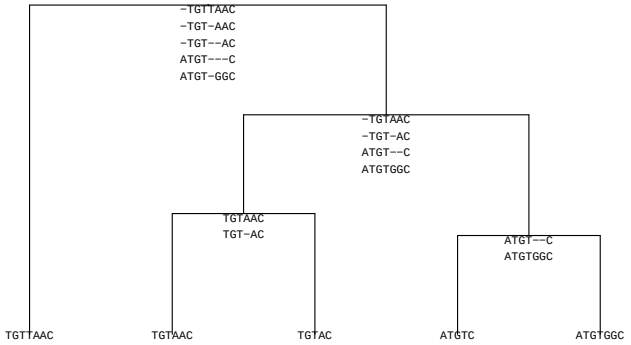
Klaus Schliep, PhD, ICBT
klaus.schliep@tugraz.at

Progressive alignment example



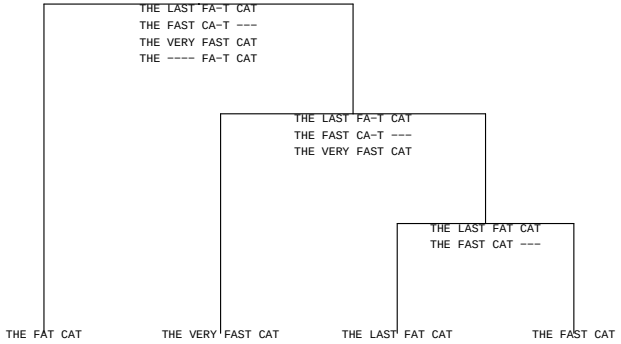
Klaus Schliep, PhD, ICBT
klaus.schliep@tugraz.at

Progressive alignment example



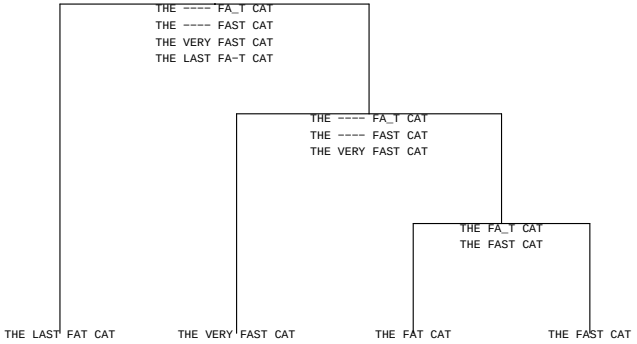
Klaus Schliep, PhD, ICBT
klaus.schliep@tugraz.at

Order is important



Klaus Schliep, PhD, ICBT
klaus.schliep@tugraz.at

Order is important



Klaus Schliep, PhD, ICBT
klaus.schliep@tugraz.at

Chicken and egg problem

guide $\longrightarrow$ multiple
tree $\longleftarrow$ alignment

We compute the tree from the alignment, and construct the alignment using a tree!

Klaus Schliep, PhD, ICBT
klaus.schliep@tugraz.at

ClustalW / ClustalΩ

- The classic progressive aligner
- Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research*, 22(22), 4673-4680.
- One of the most cited papers!
- ClustalW is superseded by ClustalΩ

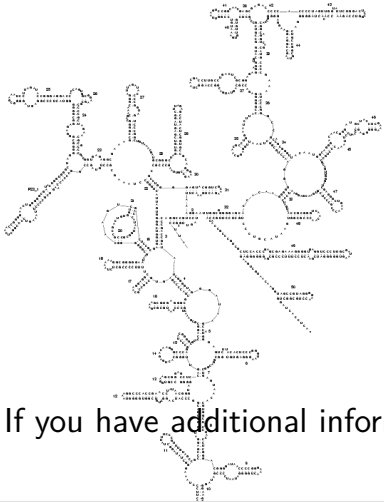
Klaus Schliep, PhD, ICBT
klaus.schliep@tugraz.at

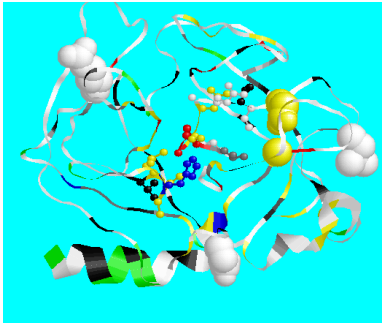
T-Coffee

- Notredame, C., Higgins, D. G., & Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology*, 302(1), 205-217.
- Combines local and global alignment
- Can incorporate PDB structure files during the alignment

Klaus Schliep, PhD, ICBT
klaus.schliep@tugraz.at

T-Coffee





If you have additional information, make use of it!

Klaus Schliep, PhD, ICBT
klaus.schliep@tugraz.at

T-Coffee variants

- M-coffee: Combination of Multiple Sequence Alignment Packages
- R-coffee: Aligns RNA sequences using predicted secondary structures

Klaus Schliep, PhD, ICBT
klaus.schliep@tugraz.at

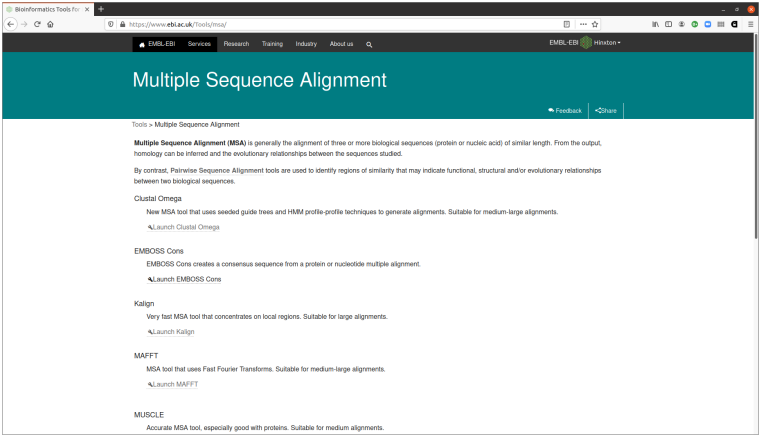
Mafft and MUSCLE

- Fast, state of the art progressive aligner
- Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4), 772-780.
- Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5), 1792-1797.
- Edgar, Robert C. (2022). Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nature Communications*, 13(1), 6968.
- Mafft is used to align the 100000+ Covid sequences

Klaus Schliep, PhD, ICBT
klaus.schliep@tugraz.at

MSA online https://www.ebi.ac.uk/jdispatcher/msa

All the common multiple sequence alignment tools



Klaus Schliep, PhD, ICBT
klaus.schliep@tugraz.at

What makes a good alignment?

- The fewer indels, the better
- Nice ungapped blocks separated with indels
- Different classes of residues within a block
  - Completely conserved
  - Conserved for hydropathy or size
- It's an art (subjective): matter of personal judgment, experience and knowledge

Klaus Schliep, PhD, ICBT
klaus.schliep@tugraz.at

Summary

- Compare different methods, parameters
- Progressive alignment propagates gaps
- Do not trust gappy regions
- Take with a large dose of skepticism

Sequence alignment is just a good guess,
but some guesses are better than others

Klaus Schliep, PhD, ICBT
klaus.schliep@tugraz.at