CENTER FOR BIOLOGICAL SEQUENCE ANALYSIS

Multiple Alignment

Anders Gorm Pedersen
Molecular Evolution Group
Center for Biological Sequence Analysis
gorm@cbs.dtu.dk

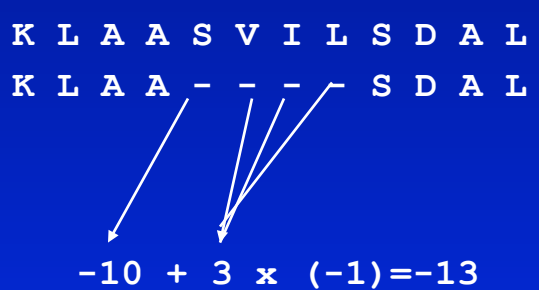
Refresher: pairwise alignments


```
43.2% identity;                    Global alignment score: 374

                10        20        30        40                50
alpha   V-LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS-----HGSA
        : :.: .:. : : ::::  .. : :.::: :... .: :. .:  : :::       :.
beta    VHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNP
               10          20        30        40        50

            60        70        80        90       100       110
alpha   QVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHL
        .::.::::  :.....::.:.. .....::.::  ::.::: ::.::.. :. .:: :.
beta    KVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHF
        60        70        80        90       100       110

           120       130       140
alpha   PAEFTPAVHASLDKFLASVSTVLTSKYR
          :::: :.:. .: .:.:...:. ::.
beta    GKEFTPPVQAAYQKVVAGVANALAHKYH
      120       130       140
```

Refresher: pairwise alignments


```
A   5
R  -2  7
N  -1 -1  7
D  -2 -2  2  8
C  -1 -4 -2 -4 13
Q  -1  1  0  0 -3  7
E  -1  0  0  2 -3  2  6
G   0 -3  0 -1 -3 -2 -3  8
.
.
.
    A  R  N  D  C  Q  E  G ...
```

- Alignment score is calculated from substitution matrix
- Identities on diagonal have high scores
- Similar amino acids have high scores
- Dissimilar amino acids have low (negative) scores



- Gaps penalized by gap-opening + gap elongation

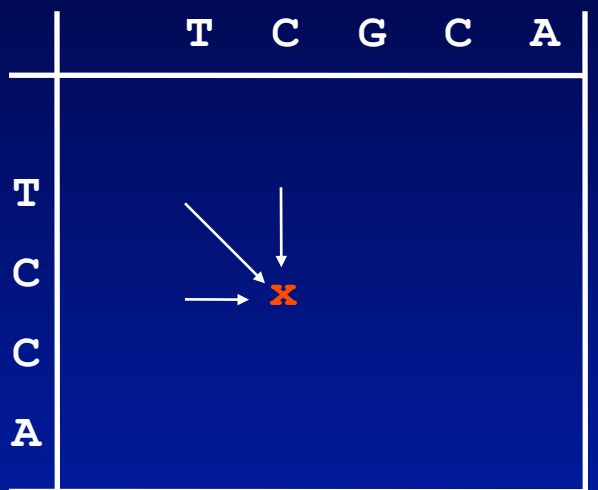
Refresher: pairwise alignments

The number of possible pairwise alignments increases explosively with the length of the sequences:

Two protein sequences of length 100 amino acids can be aligned in approximately 10^{60} different ways

10^{60} bottles of beer would fill up our entire galaxy

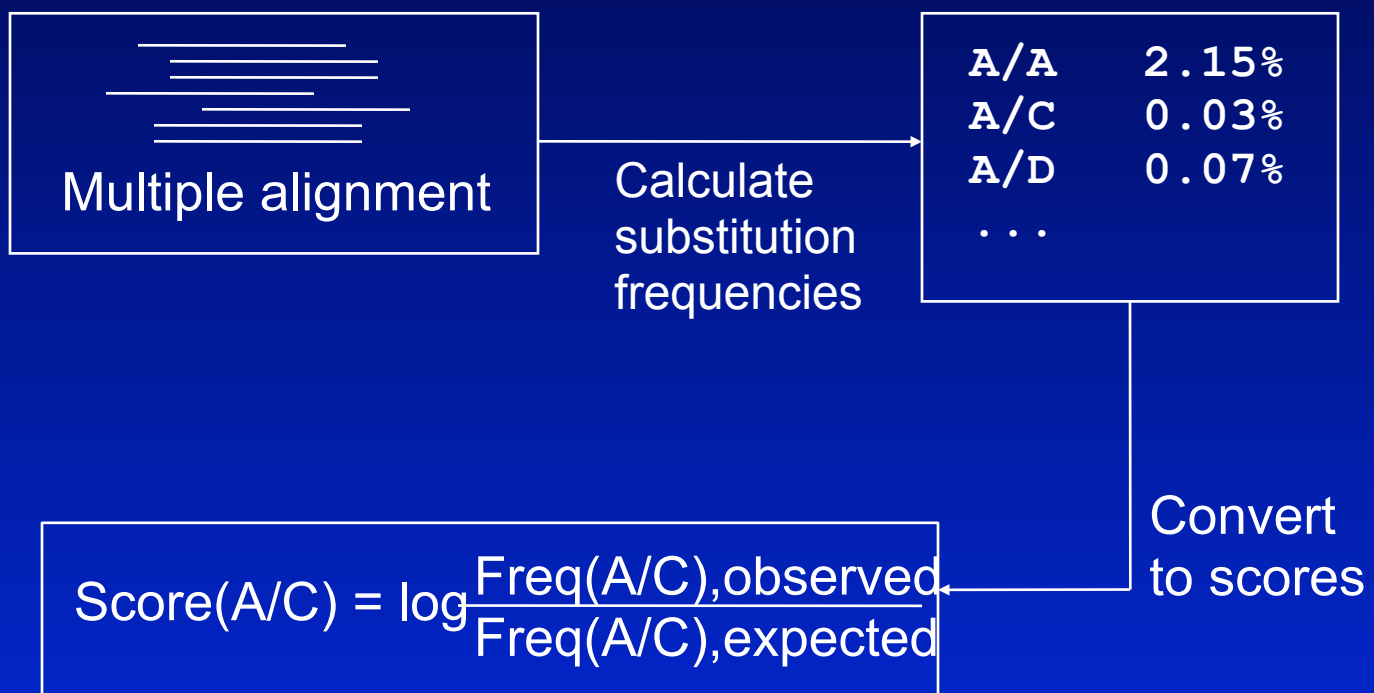
Refresher: pairwise alignments



- Solution:
  dynamic programming
- Essentially:
  the best path through any grid point in the alignment matrix must originate from one of three previous points
- Far fewer computations
- Best alignment guaranteed to be found

Refresher: pairwise alignments

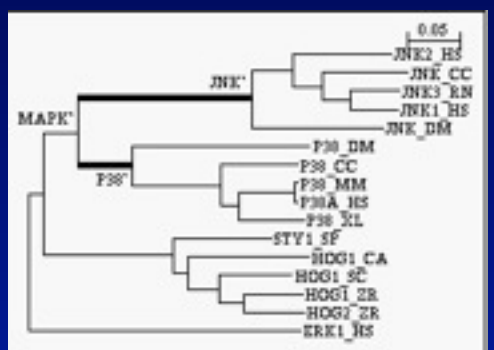
- Most used substitution matrices are themselves derived empirically from simple multiple alignments



Multiple alignment

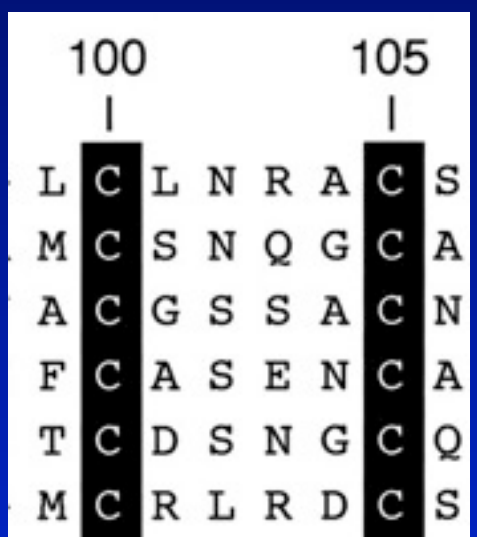
Multiple alignments: what use are they?

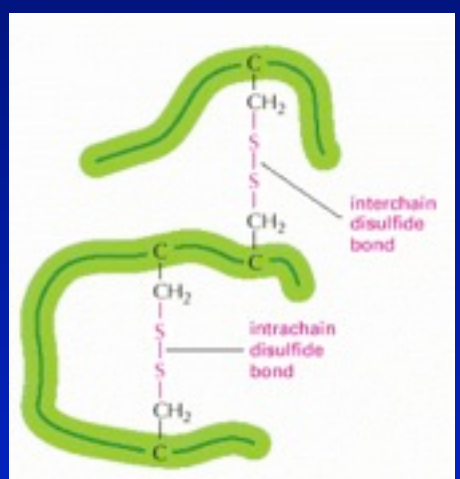
- Starting point for studies of molecular evolution



Multiple alignments: what use are they?

- Characterization of protein families:
  - Identification of conserved (functionally important) sequence regions
  - Construction of profiles for further database searching
  - Prediction of structural features (disulfide bonds, amphipathic alpha-helices, surface loops, etc.)





Scoring a multiple alignment: the “sum of pairs” score

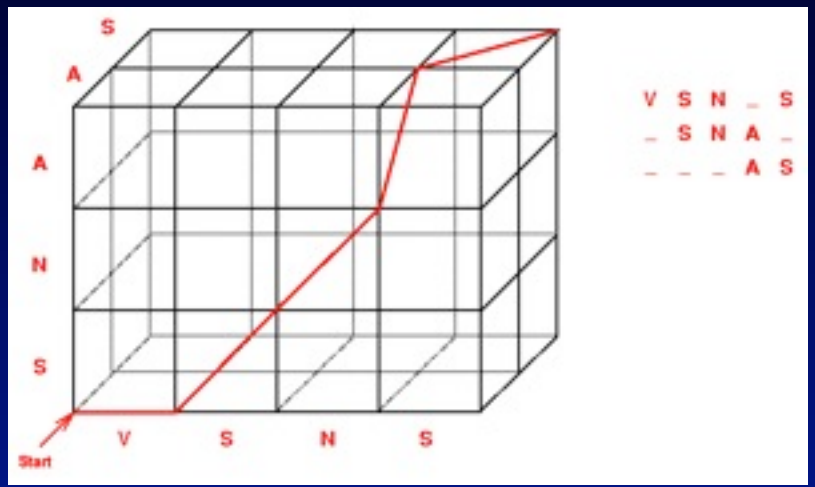
One column
from alignment

```
AA: 4, AS: 1, AT:0
AS: 1, AT: 0
ST: 1

SP- score: 4+1+0+1+0+1 = 7
```

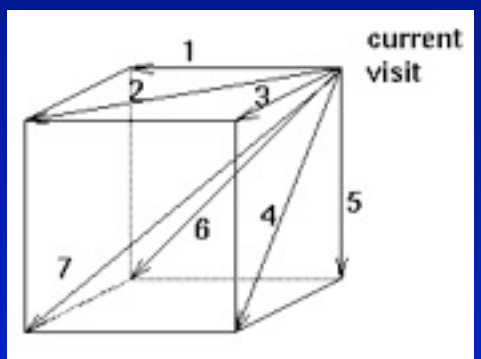
Weighted sum of pairs: each SP-score is multiplied by a weight reflecting the evolutionary distance (avoids undue influence on score by sets of very similar sequences)

=> In theory, it is possible to define an alignment score
for multiple alignments (there are several alternative scoring systems)

Multiple alignment: dynamic programming is only feasible for very small data sets



Dynamic programming matrix for 3 sequences



For 3 sequences, optimal path must come from one of 7 previous points

- In theory, optimal multiple alignment can be found by dynamic programming using a matrix with more dimensions (one dimension per sequence)
- BUT even with dynamic programming finding the optimal alignment very quickly becomes impossible due to the astronomical number of computations
- Full dynamic programming only possible for up to about 4-5 protein sequences of average length
- Even with heuristics, not feasible for more than 7-8 protein sequences
- Never used in practice

Multiple alignment: an approximate solution

- Progressive alignment (ClustalX and other programs):

1. Perform all *pairwise* alignments; keep track of sequence similarities between all pairs of sequences (construct “distance matrix”)

2. Align the most similar pair of sequences

3. Progressively add sequences to the (constantly growing) multiple alignment in order of decreasing similarity.

Progressive alignment: details

1) Perform all pairwise alignments, note pairwise distances (construct “distance matrix”)

6 pairwise alignments

	S1	S2	S3	S4
S1				
S2	3			
S3	1	3		
S4	3	2	3	

2) Construct pseudo-phylogenetic tree from pairwise distances

	S1	S2	S3	S4
S1				
S2	3			
S3	1	3		
S4	3	2	3	

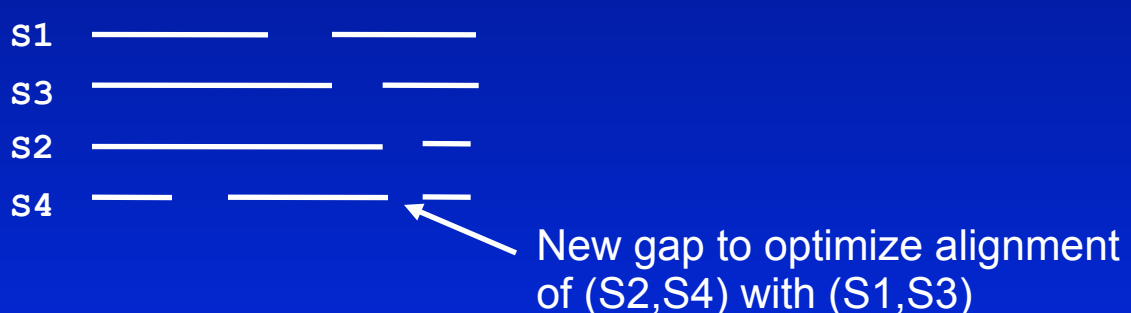
“Guide tree”

Progressive alignment: details

3) Use tree as guide for multiple alignment:
   a) Align most similar pair of sequences using dynamic programming

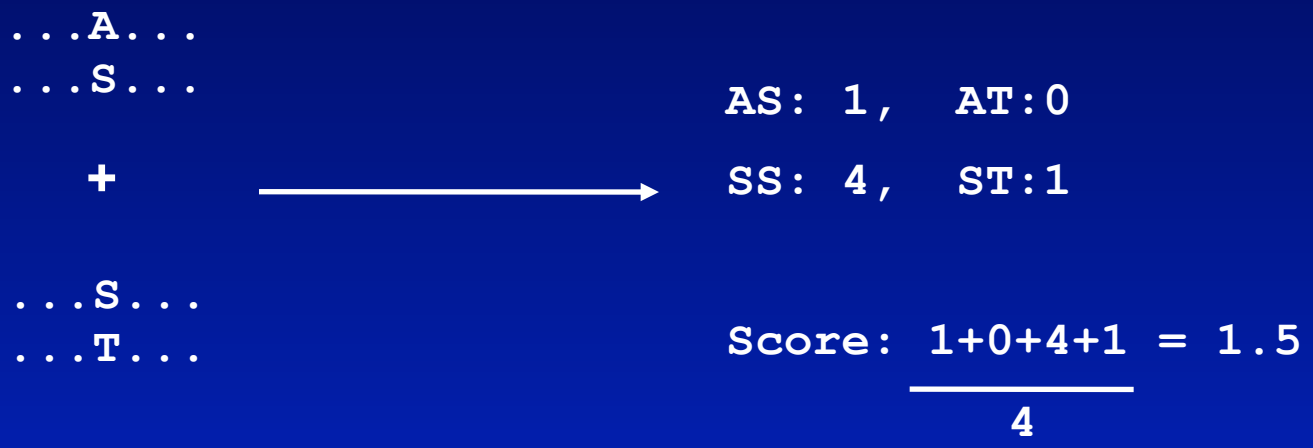
   b) Align next most similar pair

   c) Align alignments using dynamic programming - preserve gaps



Scoring profile alignments

Compare each residue in one profile to all residues in second profile. Score is average of all comparisons.



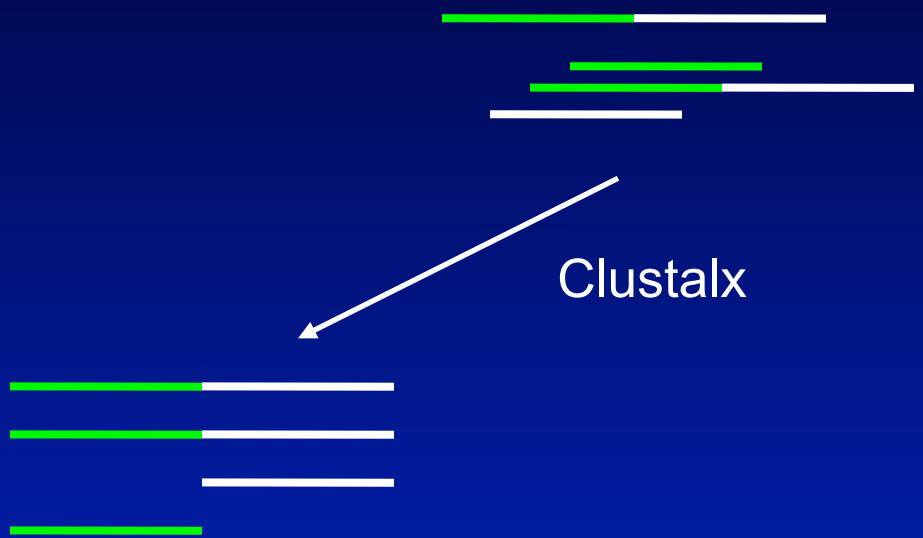
One column
from alignment

Additional ClustalX heuristics

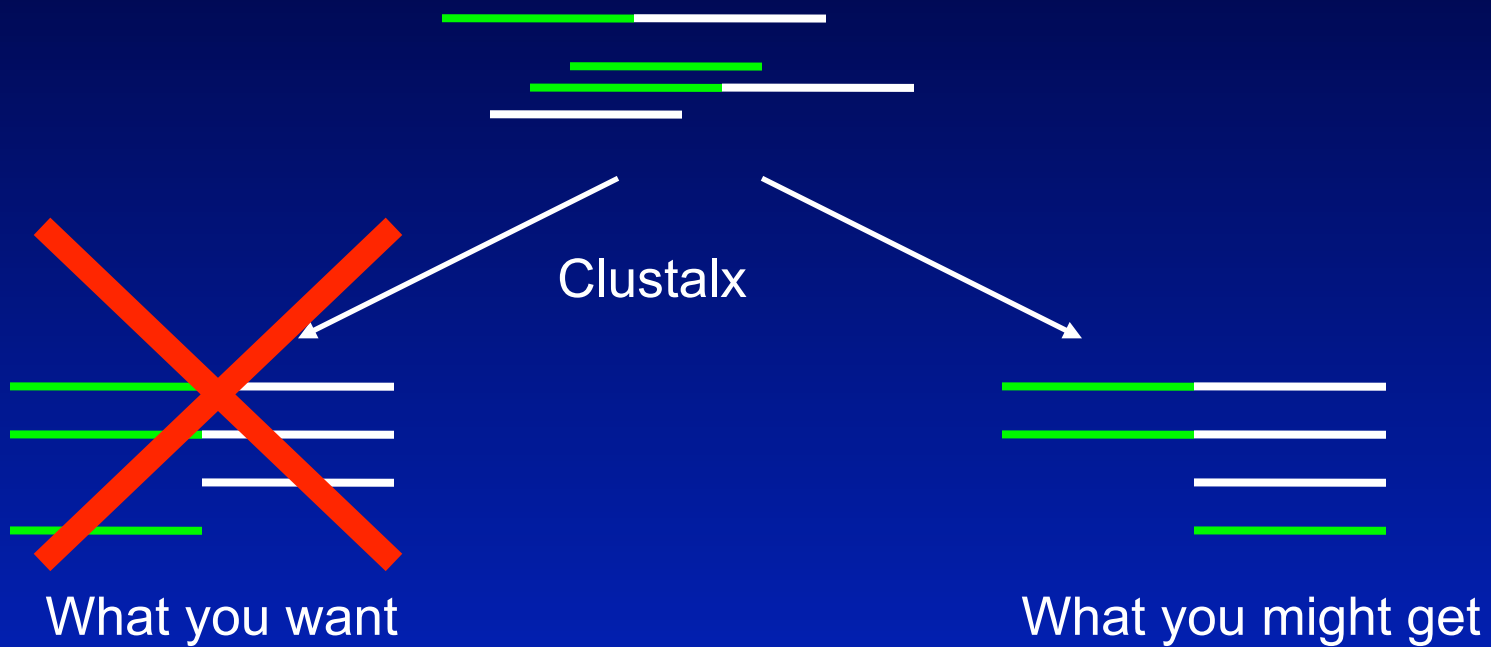
- Sequence weighting:
  - scores from similar groups of sequences are down-weighted
- Variable substitution matrices:
  - during alignment ClustalX uses different substitution matrices depending on how similar the sequences/profiles are
- Variable gap penalties:
  - gap penalties depend on substitution matrix
  - gap penalties depend on similarity of sequences
  - reduced gap penalties at existing gaps
  - increased gap penalties CLOSE to existing gaps
  - reduced gap penalties in hydrophilic stretches (presumed surface loop)
  - residue-specific gap penalties
  - and more...

Global methods (*e.g.*, ClustalX) get into trouble when data is not globally related!!!

Global methods (*e.g.*, ClustalX) get into trouble when data is not globally related!!!



Global methods (*e.g.*, ClustalX) get into trouble when data is not globally related!!!



Possible solutions:

(1) Cut out conserved regions of interest and THEN align them

(2) Use method that deals with local similarity (e.g. DIALIGN, mafft)

Other multiple alignment programs

pileup
multalign
multal
saga
hmmt
MUSCLE
ProbCons

DIALIGN
SBpima
MLpima
T-Coffee
mafft
poa
prank
...

Quantifying the Performance of Protein Sequence Multiple Alignment Programs

- Compare to alignment that is known (or strongly believed) to be correct
- Quantify by counting e.g. fraction of correctly paired residues
- Option 1: Compare performance to benchmark data sets for which 3D structures and structural alignments are available (BALiBASE, PREfab, SABmark, SMART).
  - Advantage: real, biological data with real characteristics
  - Problem: we only have good benchmark data for core regions, no good knowledge of how gappy regions really look
- Option 2: Construct synthetic alignments by letting a computer simulate evolution of a sequence along a phylogenetic tree
  - Advantage: we know the real alignment including where the gaps are
  - Problem: Simulated data may miss important aspects of real biological data

Performance on BALiBASE benchmark

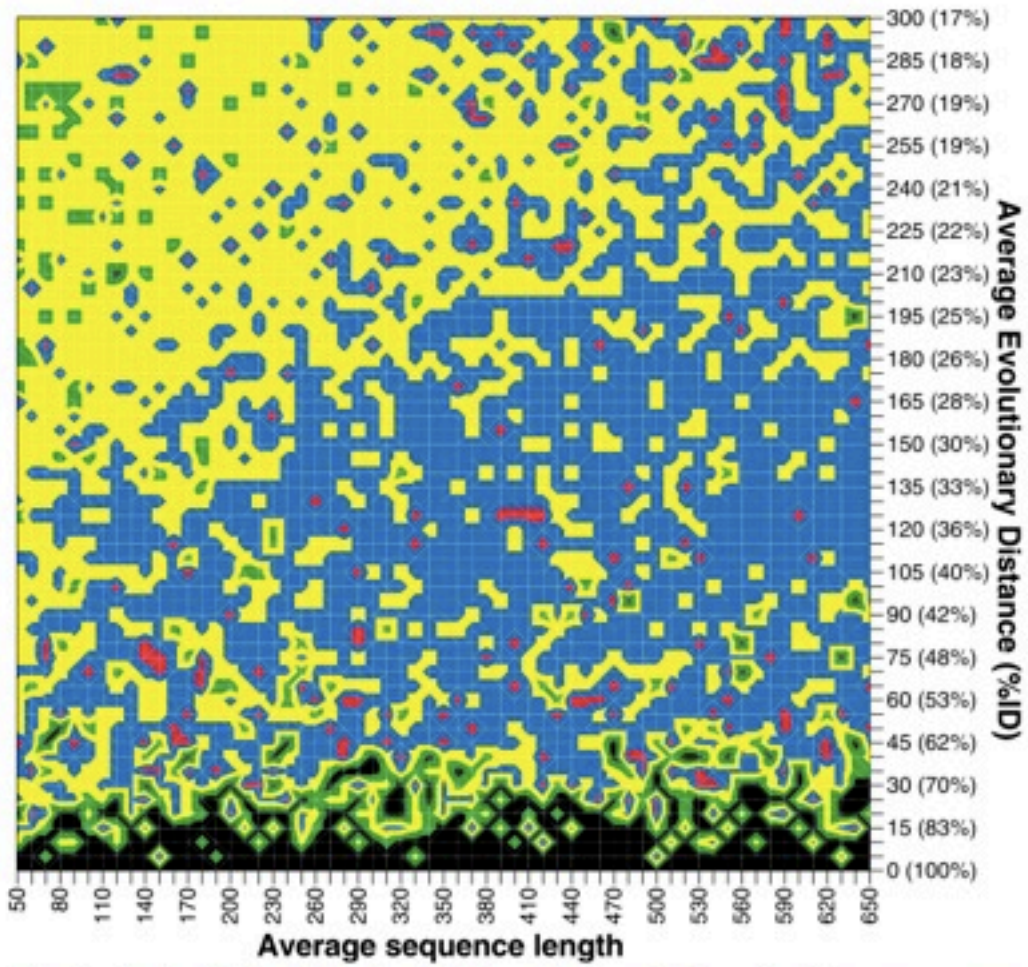


Fig. 1. Color coded matrix showing which method performed best for each pair-combination of conditions: average sequence length (x-axis) and average evolutionary distance (y-axis). The methods are Poa (green), Dialign (yellow), T-Coffee (blue) and ClustalW (red).

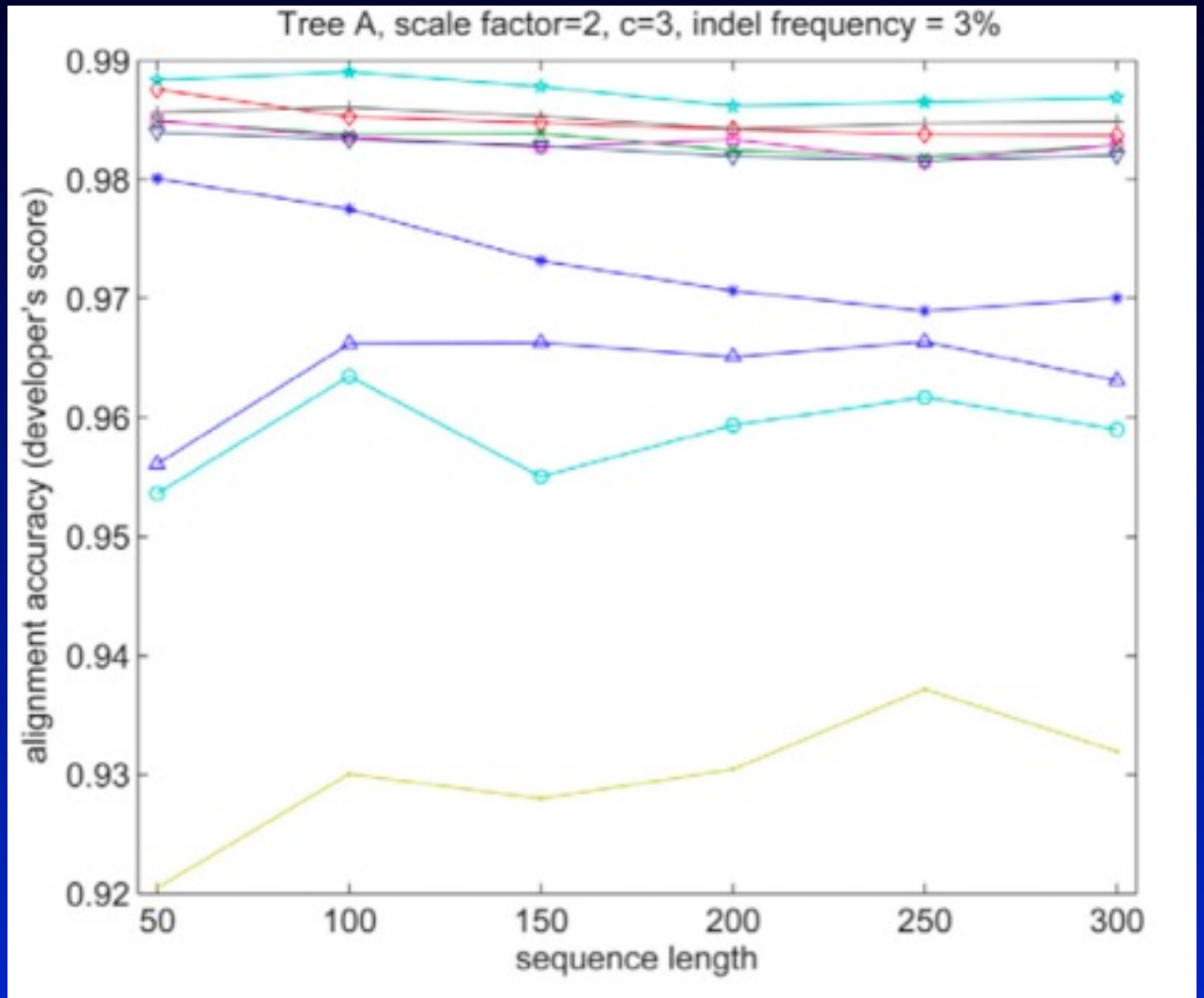
Dialign

T-Coffee

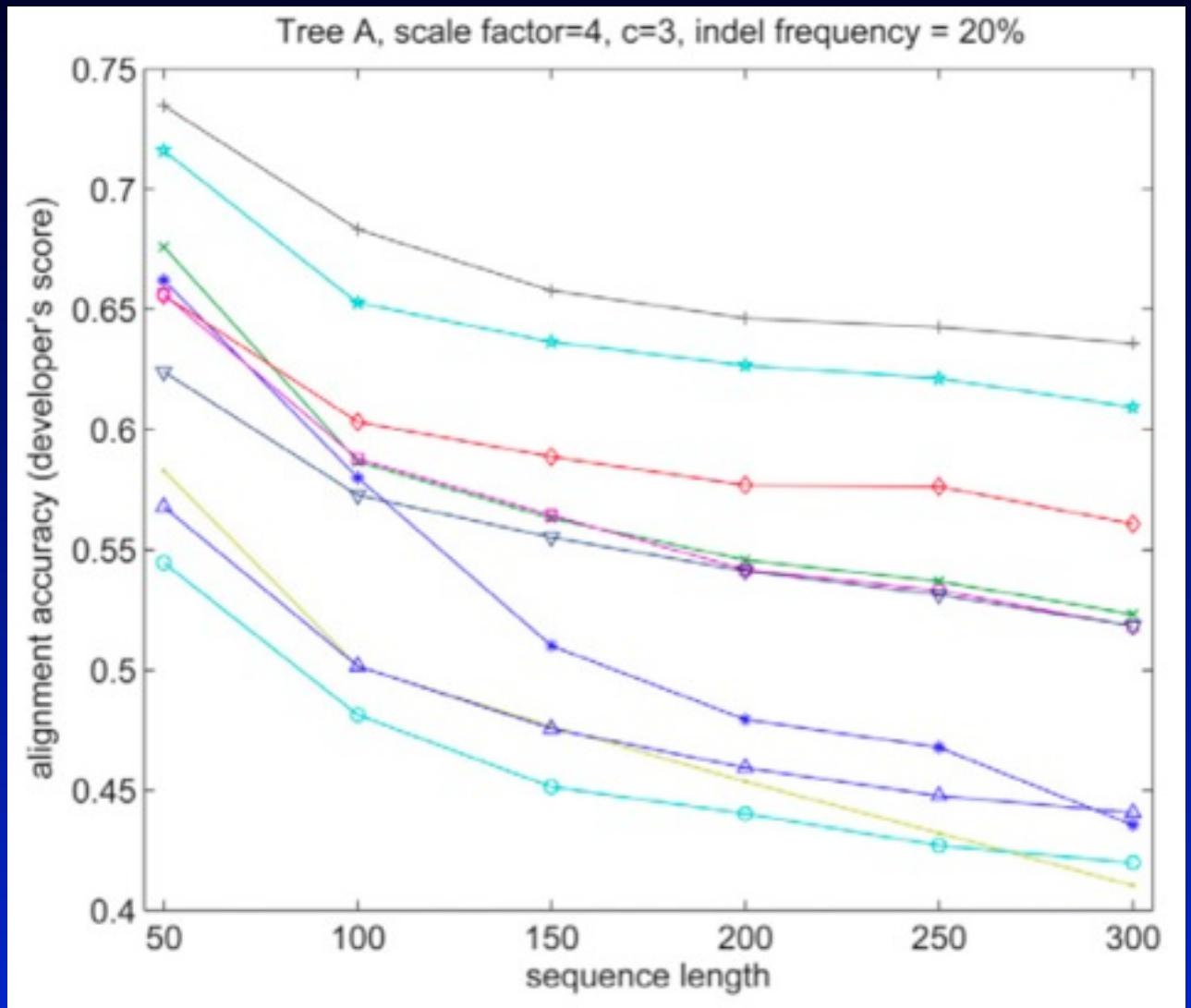
ClustalW

Poa

Performance on simulated data, few gaps



Performance on simulated data, many gaps



So which method should I choose?

- Performance depends on way of measuring and on nature of data set
- No single method performs best under all conditions (although mafft and ProbCons look quite good)
- To be on the safe side, you ought to check that results are robust to alignment uncertainty (try a number of methods, check conclusions on each alignment)
- Future perspectives: Bayesian techniques, alignment inferred along with rest of analysis, conclusions based on probability distribution over possible alignments.

Special purpose alignment programs

Nucleic Acids Research, 2003, Vol. 31, No. 13 **3537–3539**
DOI: 10.1093/nar/gkg609

RevTrans: multiple alignment of coding DNA from aligned amino acid sequences

Rasmus Wernersson and Anders Gorm Pedersen*

Center for Biological Sequence Analysis, BioCentrum-DTU, Technical University of Denmark, Building 208, DK-2800, Lyngby, Denmark

- RevTrans: alignment of coding DNA based on information at protein level
- Codon-codon boundaries maintained

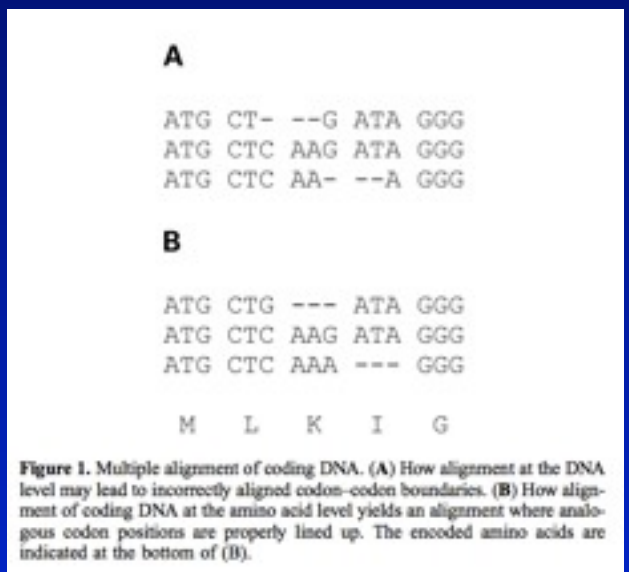


Figure 1. Multiple alignment of coding DNA. (**A**) How alignment at the DNA level may lead to incorrectly aligned codon–codon boundaries. (**B**) How alignment of coding DNA at the amino acid level yields an alignment where analogous codon positions are properly lined up. The encoded amino acids are indicated at the bottom of (B).

Special purpose alignment programs

BMC Bioinformatics

BioMed Central

Methodology article

Open Access

MaxAlign: maximizing usable data in an alignment

Rodrigo Gouveia-Oliveira*, Peter W Sackett and Anders G Pedersen

Address: Center for Biological Sequence Analysis, Technical University of Denmark, Building 208, DK-2800 Lyngby, Denmark

Email: Rodrigo Gouveia-Oliveira* - rodrigo@cbs.dtu.dk; Peter W Sackett - pws@cbs.dtu.dk; Anders G Pedersen - gorm@cbs.dtu.dk

* Corresponding author

- MaxAlign: remove subset of sequences to get fewer gapped columns
- Detect non-homologous or mis-aligned sequences

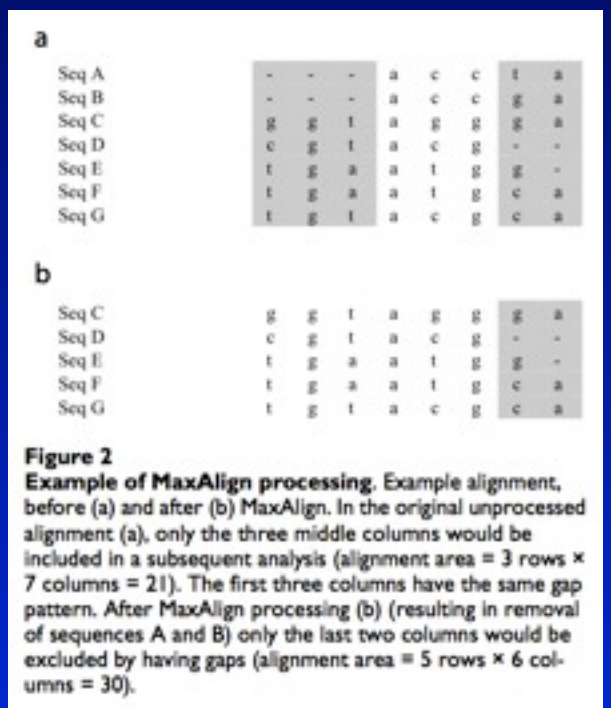


Figure 2
Example of MaxAlign processing. Example alignment, before (a) and after (b) MaxAlign. In the original unprocessed alignment (a), only the three middle columns would be included in a subsequent analysis (alignment area = 3 rows × 7 columns = 21). The first three columns have the same gap pattern. After MaxAlign processing (b) (resulting in removal of sequences A and B) only the last two columns would be excluded by having gaps (alignment area = 5 rows × 6 columns = 30).