Multiple alignment, Phylogeny and Tree Building

aligning several sequences

Martin Jakt

September 9, 2024

What's this lecture really about?

- Why align lots of sequences
- Types of homologous sequences
- How to solve a *difficult* problem
- Building trees from distances

Multiple Alignment. Why?

To some extent more natural than pairwise alignment since no reason to believe that similar sequences come in pairs.

Multiple Alignment. Why?

To some extent more natural than pairwise alignment since no reason to believe that similar sequences come in pairs.

Orthologues Sequence groups that are homologous across species (i.e. same gene, but different species).

Paralogues Sequence groups that are homologous within species (i.e. several genes within a species that share an evolutionary origin).

Multiple Alignment. Why?

To some extent more natural than pairwise alignment since no reason to believe that similar sequences come in pairs.

Orthologues Sequence groups that are homologous across species (i.e. same gene, but different species).

Paralogues Sequence groups that are homologous within species (i.e. several genes within a species that share an evolutionary origin).

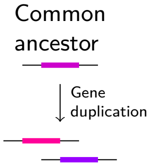
There are usually lots of these. Hence multiple alignment is more 'natural' than simple pairwise alignment.

Homology, Orthology, Paralogy

Homology, Orthology, Paralogy

Common ancestor

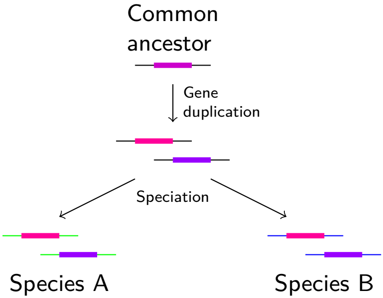
Homology, Orthology, Paralogy



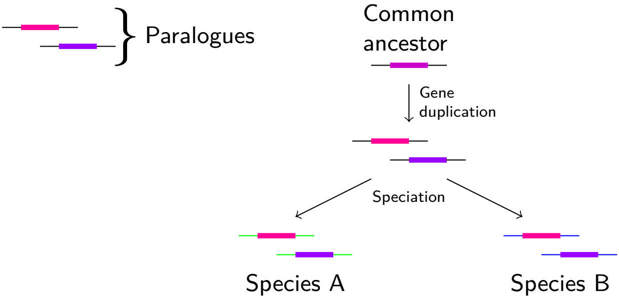
Common ancestor

Gene duplication

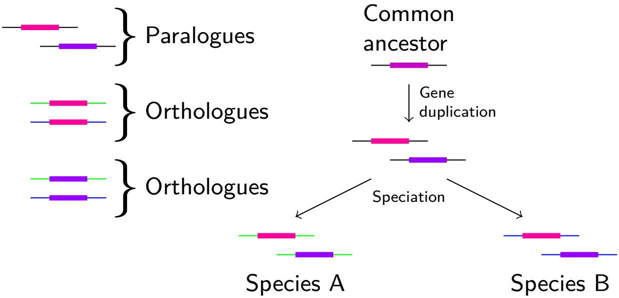
Homology, Orthology, Paralogy



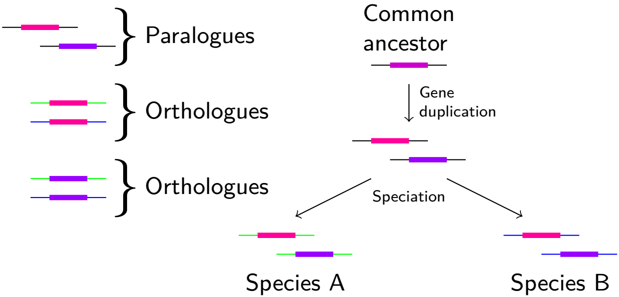
Homology, Orthology, Paralogy



Homology, Orthology, Paralogy



Homology, Orthology, Paralogy



Paralogues

Orthologues

Orthologues

Common ancestor

Gene duplication

Speciation

Species A

Species B

All homologous to each other

Orthologues and Paralogues

Orthologues Arise from speciation

Paralogues Arise from gene duplication

- Paralogues can appear and disappear as a result of gene duplication followed by loss (esp. whole genome duplication).
- Gene duplication can be followed by functional specialisation of the paralogues
  - Change of regulatory environment (i.e. when the gene is expressed)
  - Change in coding other functional sequence

Multiple Alignment. More whys

Addresses many biological questions and technical issues:

- diagnostic patterns for protein families
- detect or demonstrate homology between sequences
- help predict secondary and tertiary structures
- to suggest oligonucleotide primers for PCR
- essential prelude to molecular evolutionary analysis (allows for ancestral state inference)
- ...

Shamelessly copied from Thompson et al., NAR 1994 22(22): 4673-4680
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC308517/

How to align many sequences?

- Complexity C of optimal alignment by dynamic programming

  $$C = \prod_{i=1}^{n} l_i$$

  where $n =$ number of sequences and $l_i =$ length of the i^{th} sequence.
- Requires too much:
  - Memory
  - computation (CPU cycles)

  for more than a few sequences.
- *Heuristic* methods are used instead.
  Do not guarantee an optimal result, but provide sufficient speed.

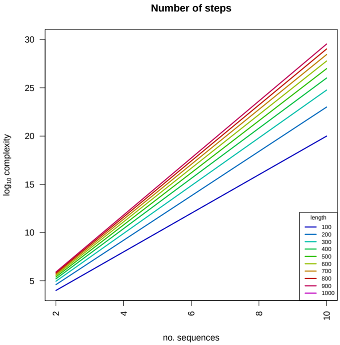
Many methods exist: we will look in detail at one of these.

Dynamic programming not possible!

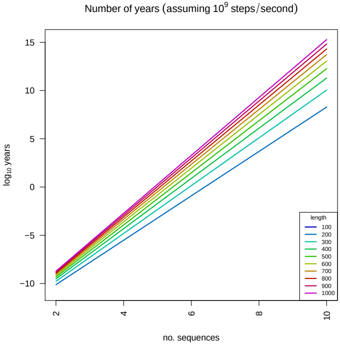
Complexity (C) scales with length (l) and number (n) of sequences:

$$C = l^n$$

For $l = 100$ and $n = 2, 3$, this is 10000 and 1000000 steps respectively.



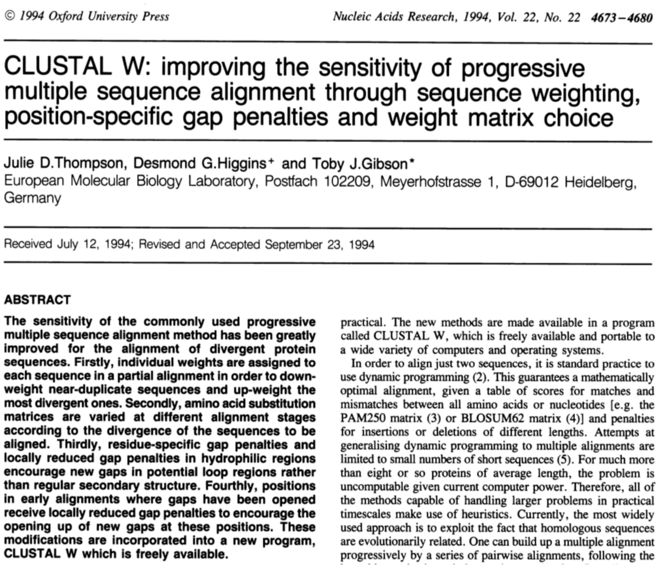
Number of steps



Number of years (assuming 10^9 steps/second)

This is a huge underestimate as the complexity of each step scales with $2^n - 1$ steps

ClustalW



© 1994 Oxford University Press

Nucleic Acids Research, 1994, Vol. 22, No. 22 4673–4680

CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice

Julie D.Thompson, Desmond G.Higgins+ and Toby J.Gibson*
European Molecular Biology Laboratory, Postfach 102209, Meyerhofstrasse 1, D-69012 Heidelberg, Germany

Received July 12, 1994; Revised and Accepted September 23, 1994

ABSTRACT

The sensitivity of the commonly used progressive multiple sequence alignment method has been greatly improved for the alignment of divergent protein sequences. Firstly, individual weights are assigned to each sequence in a partial alignment in order to down-weight near-duplicate sequences and up-weight the most divergent ones. Secondly, amino acid substitution matrices are varied at different alignment stages according to the divergence of the sequences to be aligned. Thirdly, residue-specific gap penalties and locally reduced gap penalties in hydrophilic regions encourage new gaps in potential loop regions rather than regular secondary structure. Fourthly, positions in early alignments where gaps have been opened receive locally reduced gap penalties to encourage the opening up of new gaps at these positions. These modifications are incorporated into a new program, CLUSTAL W which is freely available.

practical. The new methods are made available in a program called CLUSTAL W, which is freely available and portable to a wide variety of computers and operating systems.

In order to align just two sequences, it is standard practice to use dynamic programming (2). This guarantees a mathematically optimal alignment, given a table of scores for matches and mismatches between all amino acids or nucleotides [e.g. the PAM250 matrix (3) or BLOSUM62 matrix (4)] and penalties for insertions or deletions of different lengths. Attempts at generalising dynamic programming to multiple alignments are limited to small numbers of short sequences (5). For much more than eight or so proteins of average length, the problem is uncomputable given current computer power. Therefore, all of the methods capable of handling larger problems in practical timescales make use of heuristics. Currently, the most widely used approach is to exploit the fact that homologous sequences are evolutionarily related. One can build up a multiple alignment progressively by a series of pairwise alignments, following the

Why ClustalW

- One of the most widely used methods
- Easy to understand
- Includes phylogenetic analysis
- Paper describes the derivation and reasoning for the heuristics used nicely (eg. this can be a problem, so we tweaked this part of the method to give nicer results).
- The method extends naturally from pairwise alignment.

Paper in 'Source materials' module: clustalw_nar00046-0131.pdf
Do consider reading it.

The clustal method

For a collection of sequences:

1. Align all pairs of sequences and calculate a distance matrix (table).
2. Use the distance matrix to calculate a guide tree.
3. Align the sequences progressively according to the branch order of the guide tree.

Pairwise alignment

- Global alignment of all pairs using a modification of Needleman-Wunsch, or a faster k-tuple based heuristic method.
- Scores are calculated as: number of identities / number of residues compared (gap positions are excluded).
- Distances are are simply (1 - score)

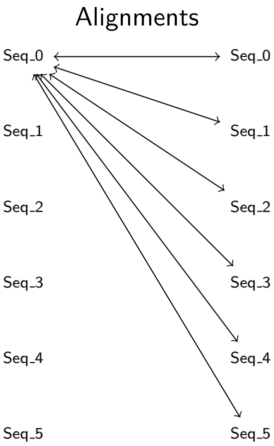
This gives an n by n distance matrix which is then used to make a guiding tree.

Pairwise alignment

Alignments

Seq_0	Seq_0
Seq_1	Seq_1
Seq_2	Seq_2
Seq_3	Seq_3
Seq_4	Seq_4
Seq_5	Seq_5

Pairwise alignment



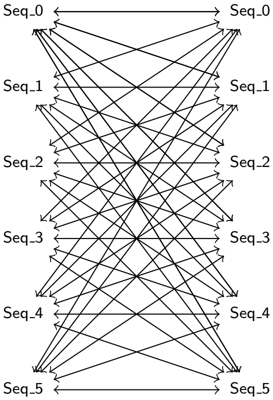
Alignments

Distance table

	S_0	S_1	S_2	S_3	S_4	S_5
S_0	$d_{0,0}$	$d_{0,1}$	$d_{0,2}$	$d_{0,3}$	$d_{0,4}$	$d_{0,5}$
S_1						
S_2						
S_3						
S_4						
S_5						

Pairwise alignment

Alignments



Distance table

	S_0	S_1	S_2	S_3	S_4	S_5
S_0	$d_{0,0}$	$d_{0,1}$	$d_{0,2}$	$d_{0,3}$	$d_{0,4}$	$d_{0,5}$
S_1	$d_{1,0}$	$d_{1,1}$	$d_{1,2}$	$d_{1,3}$	$d_{1,4}$	$d_{1,5}$
S_2	$d_{2,0}$	$d_{2,1}$	$d_{2,2}$	$d_{2,3}$	$d_{2,4}$	$d_{2,5}$
S_3	$d_{3,0}$	$d_{3,1}$	$d_{3,2}$	$d_{3,3}$	$d_{3,4}$	$d_{3,5}$
S_4	$d_{4,0}$	$d_{4,1}$	$d_{4,2}$	$d_{4,3}$	$d_{4,4}$	$d_{4,5}$
S_5	$d_{5,0}$	$d_{5,1}$	$d_{5,2}$	$d_{5,3}$	$d_{5,4}$	$d_{5,5}$

The guide tree

- Tree created from the distances to represent the similarities between the sequences and to suggest an order for the progressive alignment.
- Earlier versions used UPGMA. Newer version uses Neighbor joining algorithm.

What is a tree

- A way to represent a set of relationships (commonly distances or dis-similarities).
- Often obtained by hierarchical clustering methods from distances matrices (see below).
- Developed to represent evolutionary relationships (i.e. phylogenetic trees).
- Can be *evaluated* by maxium parsimony and likelihood methods.
- Can summarise N-dimensional data sets in general (eg. gene expression data)

A phylogenetic tree represents a *hypothesis* about how a set of species or sequences evolved.

UPGMA: the simplest tree

Unweighted Pair Group Method with Arithmetic Mean

Distances

	A	B	C	D	E
B	2				
C	4	4			
D	6	6	6		
E	6	6	6	4	
F	8	8	8	8	8

Scheme adapted from:
`http://www.icp.ucl.ac.be/~opperd/private/upgma.html`

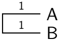
UPGMA: the simplest tree

Unweighted Pair Group Method with Arithmetic Mean

Distances

	A	B	C	D	E
B	2				
C	4	4			
D	6	6	6		
E	6	6	6	4	
F	8	8	8	8	8

join two most similar nodes



Scheme adapted from:
http://www.icp.ucl.ac.be/~opperd/private/upgma.html

UPGMA: the simplest tree

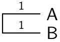
Unweighted Pair Group Method with Arithmetic Mean

Distances

	A	B	C	D	E
B	2				
C	4	4			
D	6	6	6		
E	6	6	6	4	
F	8	8	8	8	8

$\longrightarrow$

join two most similar nodes



Calculate distances between merged node (AB) and other nodes (using arithmetic mean)

$$\mathrm{dist(AB,C)} \leftarrow (\mathrm{dist(A,C)} + \mathrm{dist(B,C)})/2$$

Scheme adapted from:
`http://www.icp.ucl.ac.be/~opperd/private/upgma.html`

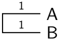
UPGMA: the simplest tree

Unweighted Pair Group Method with Arithmetic Mean

Distances

	A	B	C	D	E
B	2				
C	4	4			
D	6	6	6		
E	6	6	6	4	
F	8	8	8	8	8

join two most similar nodes



Calculate distances between merged node (AB) and other nodes (using arithmetic mean)

$\mathrm{dist}(AB,C) \leftarrow (\mathrm{dist}(A,C) + \mathrm{dist}(B,C))/2$

	AB	C	D	E
C	4			
D	6	6		
E	6	6	4	
F	8	8	8	8

Scheme adapted from:
`http://www.icp.ucl.ac.be/~opperd/private/upgma.html`

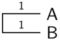
UPGMA: the simplest tree

Unweighted Pair Group Method with Arithmetic Mean

Distances

	A	B	C	D	E
B	2				
C	4	4			
D	6	6	6		
E	6	6	6	4	
F	8	8	8	8	8

$\longrightarrow$ join two most similar nodes

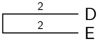


Calculate distances between merged node (AB) and other nodes (using arithmetic mean)

dist(AB,C) ← (dist(A,C) + dist(B,C))/2

	AB	C	D	E
C	4			
D	6	6		
E	6	6	4	
F	8	8	8	8

$\longrightarrow$ repeat



Scheme adapted from:
`http://www.icp.ucl.ac.be/~opperd/private/upgma.html`

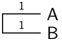
UPGMA: the simplest tree

Unweighted Pair Group Method with Arithmetic Mean

Distances

	A	B	C	D	E
B	2				
C	4	4			
D	6	6	6		
E	6	6	6	4	
F	8	8	8	8	8

join two most similar nodes

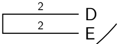


Calculate distances between merged node (AB) and other nodes (using arithmetic mean)

dist(AB,C) $\leftarrow$ (dist(A,C) + dist(B,C))/2

	AB	C	D	E
C	4			
D	6	6		
E	6	6	4	
F	8	8	8	8

repeat



	AB	C	DE
C	4		
DE	6	6	
F	8	8	8

Scheme adapted from:
`http://www.icp.ucl.ac.be/~opperd/private/upgma.html`

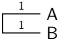
UPGMA: the simplest tree

Unweighted Pair Group Method with Arithmetic Mean

Distances

	A	B	C	D	E
B	2				
C	4	4			
D	6	6	6		
E	6	6	6	4	
F	8	8	8	8	8

join two most similar nodes

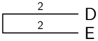


Calculate distances between merged node (AB) and other nodes (using arithmetic mean)

dist(AB,C) ← (dist(A,C) + dist(B,C))/2

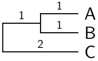
	AB	C	D	E
C	4			
D	6	6		
E	6	6	4	
F	8	8	8	8

repeat



	AB	C	DE
C	4		
DE	6	6	
F	8	8	8

and again



Scheme adapted from:
`http://www.icp.ucl.ac.be/~opperd/private/upgma.html`

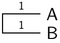
UPGMA: the simplest tree

Unweighted Pair Group Method with Arithmetic Mean

Distances

	A	B	C	D	E
B	2				
C	4	4			
D	6	6	6		
E	6	6	6	4	
F	8	8	8	8	8

→ join two most similar nodes

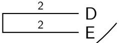


Calculate distances between merged node (AB) and other nodes (using arithmetic mean)

dist(AB,C) ← (dist(A,C) + dist(B,C))/2

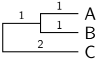
	AB	C	D	E
C	4			
D	6	6		
E	6	6	4	
F	8	8	8	8

→ repeat

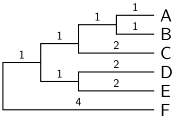


	AB	C	DE
C	4		
DE	6	6	
F	8	8	8

→ and again



eventually



Scheme adapted from:
`http://www.icp.ucl.ac.be/~opperd/private/upgma.html`

Neighbor joining algorithm

- Underlying algorithm method similar to UPGMA (i.e. progressively merge neighboring nodes until a single tree is obtained).
- Modified distance matrix used to find nearest nodes to join.
- Distances of pair members to joins are influenced by distances to external nodes.
- Does not assume equal rate of evolution
  $\Rightarrow$ neighbours have differing distances to their joining nodes.
- Better than UPGMA (?)

Saitou and Nei, Mol Biol Evol 1987, Jul;4(4);406-25

Authors of MUSCLE claim UPGMA to be better for MSA

Neighbor joining (1)

Nodes to be joined (i.e. neighbors) are chosen from a Q matrix:

$$Q_{i,j} = (n-2)d_{i,j} - \sum_{k=1}^{n} d_{i,k} - \sum_{k=1}^{n} d_{j,k}$$

$d_{i,j}$ distance between nodes i and j

n the number of nodes

Q is *only* used to rank node pairs

Outlier pairs have low Q and are joined first (i.e. pairs of nodes which are distant from the larger set).

Equations taken from Gascuel and Steel 2006; Mol. Biol. Evol. 23(11):1997-2000.

It's OK to be confused here

Gascuel and Steel 2006; Mol. Biol. Evol. 23(11):1997-2000.

It's OK to be confused here

From 'Neighbour-Joining Revealed', published almost 20 years after Saitou and Nei:

'Yet the question "what does the NJ method seek to do?" has until recently proved somewhat elusive, leading to some imprecise claims and misunderstanding'

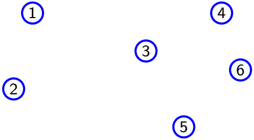
Gascuel and Steel 2006; Mol. Biol. Evol. 23(11):1997-2000.

eh??

$$Q_{i,j} = (n-2)d_{i,j} - \sum_{k=1}^{n} d_{i,k} - \sum_{k=1}^{n} d_{j,k}$$

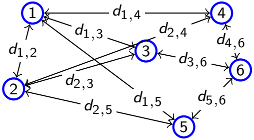
eh??

$$Q_{i,j} = (n-2)d_{i,j} - \sum_{k=1}^{n} d_{i,k} - \sum_{k=1}^{n} d_{j,k}$$



eh??

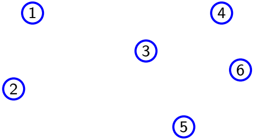
$$Q_{i,j} = (n-2)d_{i,j} - \sum_{k=1}^{n} d_{i,k} - \sum_{k=1}^{n} d_{j,k}$$



$d_{i,j}$ is the distance between nodes i and j

eh??

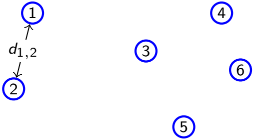
$$Q_{i,j} = (n-2)d_{i,j} - \sum_{k=1}^{n} d_{i,k} - \sum_{k=1}^{n} d_{j,k}$$



$$Q_{1,2} = (n-2)d_{1,2} - \textstyle\sum_{k=1}^{n} d_{1,k} - \sum_{k=1}^{n} d_{2,k}$$

eh??

$$Q_{i,j} = (n-2)d_{i,j} - \sum_{k=1}^{n} d_{i,k} - \sum_{k=1}^{n} d_{j,k}$$

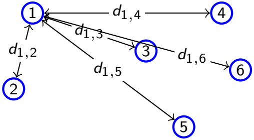


$Q_{1,2} = (n-2)d_{1,2} - \sum_{k=1}^{n} d_{1,k} - \sum_{k=1}^{n} d_{2,k}$

$(6-2) \times d_{1,2}$

eh??

$$Q_{i,j} = (n-2)d_{i,j} - \sum_{k=1}^{n} d_{i,k} - \sum_{k=1}^{n} d_{j,k}$$

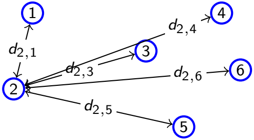


$$Q_{1,2} = (n-2)d_{1,2} - \textstyle\sum_{k=1}^{n} d_{1,k} - \sum_{k=1}^{n} d_{2,k}$$

$$(6-2) \times d_{1,2}$$
$$-(d_{1,1} + d_{1,2} + d_{1,3} + d_{1,4} + d_{1,5} + d_{1,6})$$

eh??

$$Q_{i,j} = (n-2)d_{i,j} - \sum_{k=1}^{n} d_{i,k} - \sum_{k=1}^{n} d_{j,k}$$

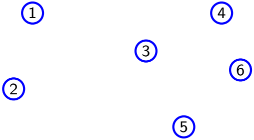


$$Q_{1,2} = (n-2)d_{1,2} - \sum_{k=1}^{n} d_{1,k} - \sum_{k=1}^{n} d_{2,k}$$

$$(6-2) \times d_{1,2}$$
$$-(d_{1,1} + d_{1,2} + d_{1,3} + d_{1,4} + d_{1,5} + d_{1,6})$$
$$-(d_{2,1} + d_{2,2}, d_{2,3} + d_{2,4} + d_{2,5} + d_{2,6})$$

eh??

$$Q_{i,j} = (n-2)d_{i,j} - \sum_{k=1}^{n} d_{i,k} - \sum_{k=1}^{n} d_{j,k}$$



$$Q_{1,2} = (n-2)d_{1,2} - \textstyle\sum_{k=1}^{n} d_{1,k} - \sum_{k=1}^{n} d_{2,k}$$

$$\begin{aligned}
&(6-2) \times d_{1,2} \\
&-(d_{1,1} + d_{1,2} + d_{1,3} + d_{1,4} + d_{1,5} + d_{1,6}) \\
&-(d_{2,1} + d_{2,2}, d_{2,3} + d_{2,4} + d_{2,5} + d_{2,6})
\end{aligned}$$

This favours nodes close to each other, but which are far from the others; i.e. pairs of outliers.

Neighbor joining (2)

The pair of nodes (f, g) with the lowest Q value are joined through a new node u.
The distances between the new node u and f and g are:

$$\begin{aligned} \delta_{f,u} &= \frac{1}{2} d_{f,g} + \frac{1}{2(n-2)} \left[\sum_{k=1}^{n} d_{f,k} - \sum_{k=1}^{n} d_{g,k} \right] \\ \delta_{g,u} &= d_{f,g} - \delta_{f,u} \end{aligned} \qquad (1)$$

$\delta_{f,u}$ and $\delta_{g,u}$ are adjusted such that they are proportional to their respective distances to the remaining nodes.

ehh again?

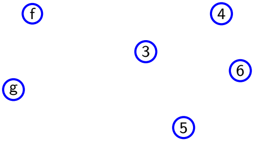
$$\begin{aligned}\delta_{f,u} &= \frac{1}{2}d_{f,g} + \frac{1}{2(n-2)}\left[\sum_{k=1}^{n} d_{f,k} - \sum_{k=1}^{n} d_{g,k}\right] \\ \delta_{g,u} &= d_{f,g} - \delta_{f,u}\end{aligned} \tag{2}$$

Here f and g are equivalent to 1 and 2

Note that, $d_{f,1}$ and $d_{g,2}$ are 0 and that $d_{f,2} = d_{g,1}$ and so cancel each other.

ehh again?

$$
\begin{aligned}
\delta_{f,u} &= \frac{1}{2} d_{f,g} + \frac{1}{2(n-2)} \left[\sum_{k=1}^{n} d_{f,k} - \sum_{k=1}^{n} d_{g,k} \right] \\
\delta_{g,u} &= d_{f,g} - \delta_{f,u}
\end{aligned}
\tag{2}
$$



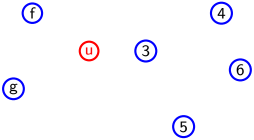
Here f and g are equivalent to 1 and 2

Note that, $d_{f,1}$ and $d_{g,2}$ are 0 and that $d_{f,2} = d_{g,1}$ and so cancel each other.

ehh again?

$$\delta_{f,u} = \frac{1}{2} d_{f,g} + \frac{1}{2(n-2)} \left[\sum_{k=1}^{n} d_{f,k} - \sum_{k=1}^{n} d_{g,k} \right] \tag{2}$$

$$\delta_{g,u} = d_{f,g} - \delta_{f,u}$$



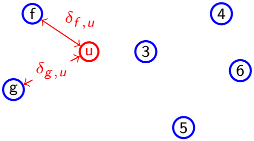
Here f and g are equivalent to 1 and 2

Note that, $d_{f,1}$ and $d_{g,2}$ are 0 and that $d_{f,2} = d_{g,1}$ and so cancel each other.

ehh again?

$$\delta_{f,u} = \frac{1}{2} d_{f,g} + \frac{1}{2(n-2)} \left[\sum_{k=1}^{n} d_{f,k} - \sum_{k=1}^{n} d_{g,k} \right] \tag{2}$$

$$\delta_{g,u} = d_{f,g} - \delta_{f,u}$$

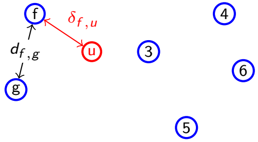


Here f and g are equivalent to 1 and 2

Note that, $d_{f,1}$ and $d_{g,2}$ are 0 and that $d_{f,2} = d_{g,1}$ and so cancel each other.

ehh again?

$$\begin{aligned} \delta_{f,u} &= \frac{1}{2} d_{f,g} + \frac{1}{2(n-2)} \left[\sum_{k=1}^{n} d_{f,k} - \sum_{k=1}^{n} d_{g,k} \right] \\ \delta_{g,u} &= d_{f,g} - \delta_{f,u} \end{aligned} \tag{2}$$



$$\delta_{f,u} = \tfrac{1}{2} d_{f,g} +$$

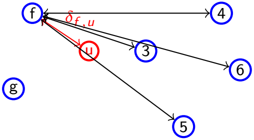
Here f and g are equivalent to 1 and 2

Note that, $d_{f,1}$ and $d_{g,2}$ are 0 and that $d_{f,2} = d_{g,1}$ and so cancel each other.

ehh again?

$$\delta_{f,u} = \frac{1}{2}d_{f,g} + \frac{1}{2(n-2)}\left[\sum_{k=1}^{n} d_{f,k} - \sum_{k=1}^{n} d_{g,k}\right] \tag{2}$$

$$\delta_{g,u} = d_{f,g} - \delta_{f,u}$$



$$\delta_{f,u} = \tfrac{1}{2}d_{f,g} +$$

$$\tfrac{1}{2\times(6-2)}\times$$
$$((d_{f,1} + d_{f,2} + d_{f,3} + d_{f,4} + d_f f, 5 + d_{f,6})$$

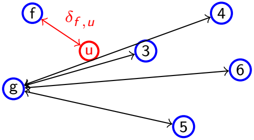
Here f and g are equivalent to 1 and 2

Note that, $d_{f,1}$ and $d_{g,2}$ are 0 and that $d_{f,2} = d_{g,1}$ and so cancel each other.

ehh again?

$$\delta_{f,u} = \frac{1}{2} d_{f,g} + \frac{1}{2(n-2)} \left[\sum_{k=1}^{n} d_{f,k} - \sum_{k=1}^{n} d_{g,k} \right] \tag{2}$$

$$\delta_{g,u} = d_{f,g} - \delta_{f,u}$$



$$\delta_{f,u} = \tfrac{1}{2} d_{f,g} +$$

$$\tfrac{1}{2\times(6-2)} \times$$
$$((d_{f,1} + d_{f,2} + d_{f,3} + d_{f,4} + d_f f, 5 + d_{f,6})$$
$$-(d_{g,1} + d_{g,2}, d_{g,3} + d_{g,4} + d_f g, 5 + d_{g,6}))$$

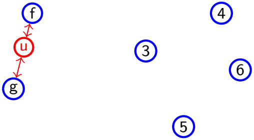
Here f and g are equivalent to 1 and 2

Note that, $d_{f,1}$ and $d_{g,2}$ are 0 and that $d_{f,2} = d_{g,1}$ and so cancel each other.

ehh again?

$$\delta_{f,u} = \frac{1}{2} d_{f,g} + \frac{1}{2(n-2)} \left[\sum_{k=1}^{n} d_{f,k} - \sum_{k=1}^{n} d_{g,k} \right] \tag{2}$$

$$\delta_{g,u} = d_{f,g} - \delta_{f,u}$$



$$\delta_{f,u} = \tfrac{1}{2} d_{f,g} +$$

$$\tfrac{1}{2\times(6-2)}\times$$
$$((d_{f,1} + d_{f,2} + d_{f,3} + d_{f,4} + d_f f, 5 + d_{f,6})$$
$$-(d_{g,1} + d_{g,2}, d_{g,3} + d_{g,4} + d_f g, 5 + d_{g,6}))$$

Here f and g are equivalent to 1 and 2

Note that, $d_{f,1}$ and $d_{g,2}$ are 0 and that $d_{f,2} = d_{g,1}$ and so cancel each other.

Neighbor joining (3)

The distances of the remaining nodes to the joining node u are set as:

$$\delta_{u,k} = \frac{1}{2}[d_{f,k} + d_{g,k} - d_{f,g}]$$

u	joining node
k	a remaining node
f, g	the joined nodes

This assures that the total distance within the tree is consistent.

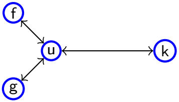
should be: $k \neq f, g$

These equations taken from the Wikipedia page:
`https://en.wikipedia.org/wiki/Neighbor_joining`
They differ slightly in their form, but not substance from the original publication.

Neighbor joining (3)

The total distance of the tree is consistent:

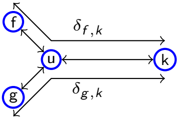
$$\delta_{u,k} = \frac{1}{2}[d_{f,k} + d_{g,k} - d_{f,g}]$$



Neighbor joining (3)

The total distance of the tree is consistent:

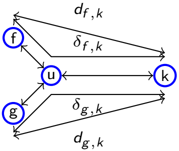
$$\delta_{u,k} = \frac{1}{2}[d_{f,k} + d_{g,k} - d_{f,g}]$$



Neighbor joining (3)

The total distance of the tree is consistent:

$$\delta_{u,k} = \frac{1}{2}[d_{f,k} + d_{g,k} - d_{f,g}]$$

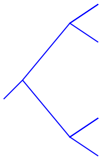


$$d_{f,k} + d_{g,k} = \delta_{f,k} + \delta_{g,k}$$

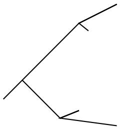
Neighbor joining: putting it together

1. Determine the Q matrix based on the current distance matrix.
2. Find the pair of nodes with the smallest Q value.
3. Create a new node that connects this pair.
4. Determine the distances of all the nodes to this new joining node.
5. Replace the neighbour pair with the new node and update the distance matrix.
6. Repeat from (1) until the tree is fully connected.

UPGMA vs Neighbor joining



UPGMA

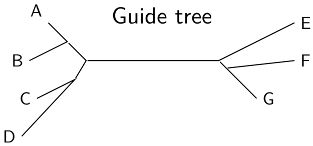


Neighbor Joining

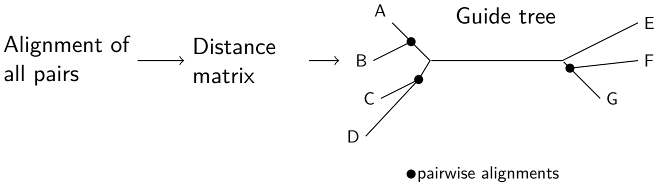
Neighbor joining does not assume equal rate of evolution when joining nodes.

Progressive alignment

Alignment of all pairs ⟶ Distance matrix ⟶



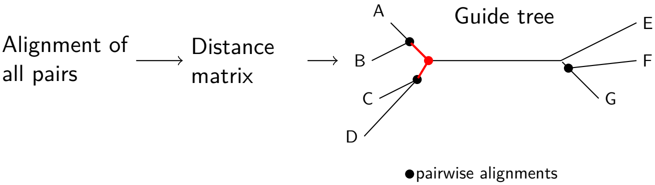
Progressive alignment



Alignment of all pairs → Distance matrix → Guide tree

● pairwise alignments

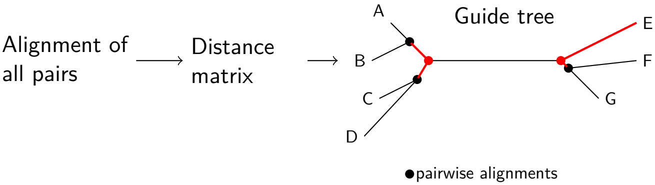
Progressive alignment



Alignment of all pairs → Distance matrix →

1. Align AB to CD

Progressive alignment

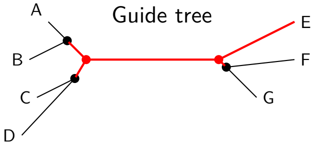


Alignment of all pairs → Distance matrix →

1. Align AB to CD
2. Align E to FG

Progressive alignment

Alignment of all pairs → Distance matrix →

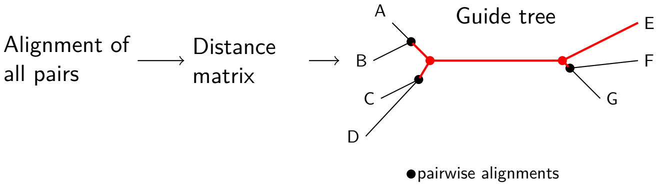


Guide tree

●pairwise alignments

1. Align AB to CD
2. Align E to FG
2. Align ABCD to EFG

Progressive alignment



Alignment of all pairs → Distance matrix →

1. Align AB to CD
2. Align E to FG
2. Align ABCD to EFG

How to align two alignments?

Aligning alignments

Modify the scoring function to use several sequences.

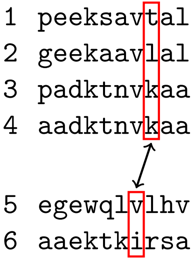
Match score is set to the mean of all independent pairs:

```
1  peeksavtal
2  geekaavlal
3  padktnvkaa
4  aadktnvkaa

5  egewqlvlhv
6  aaektkirsa
```

Aligning alignments

Modify the scoring function to use several sequences.

Match score is set to the mean of all independent pairs:

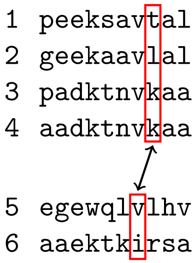


```
1  peeksavtal
2  geekaavlal
3  padktnvkaa
4  aadktnvkaa

5  egewqlvlhv
6  aaektkirsa
```

Aligning alignments

Modify the scoring function to use several sequences.

Match score is set to the mean of all independent pairs:



```
1  peeksavtal
2  geekaavlal
3  padktnvkaa
4  aadktnvkaa

5  egewqlvlhv
6  aaektkirsa
```

$$\text{Score} = \frac{\begin{array}{rc} & M(t,v) \\ + & M(t,i) \\ + & M(l,v) \\ + & M(l,i) \\ + & M(k,v) \\ + & M(k,i) \\ + & M(k,v) \\ + & M(k,i) \end{array}}{8}$$

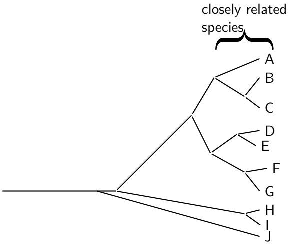
Where M(i,j) is are values taken from the given substitution matrix

Refinements

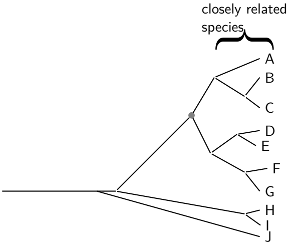
- Weighting of sequences to correct for unequal sampling across evolutionary distances in the data set (greater weight to outlier sequences)
- Dynamic variation of gap penalties (to mimic known tendencies in proteins)
  - Increase gap opening penalty within 8 amino acid of a gap opening
  - Decrease gap opening penalty in hydrophilic stretches (associated with loops)
  - Decreased gap opening penalties at positions of gaps in early alignments.
- Dynamic use of substitution matrices: starting with substitution matrices suitable for closely related sequences and moving to divergent matrices.

Thompson et al., NAR 1994 22(22): 4673-4680
`http://www.ncbi.nlm.nih.gov/pmc/articles/PMC308517/`

Why weight sequences

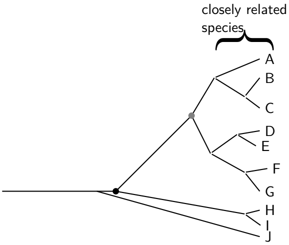


Why weight sequences



A. GKSWKALTPP
B. GKSWKSLSPS
C. GKSWKSLSTS
D. GKSWKSLSPS
E. GKSWKSLSPS
F. GKSWRALSPS
G. GRSWKSLSPS

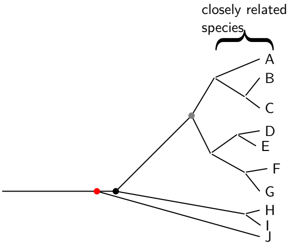
Why weight sequences



```
A. GKSWKALTPP
B. GKSWKSLSPS
C. GKSWKSLSTS
D. GKSWKSLSPS
E. GKSWKSLSPS
F. GKSWRALSPS
G. GRSWKSLSPS

H. PKSWRASSPS
I. PRSWKSSSPS
```

$$S = \frac{7 \times M_{P,G} + 7 \times M_{P,G}}{14}$$

Why weight sequences

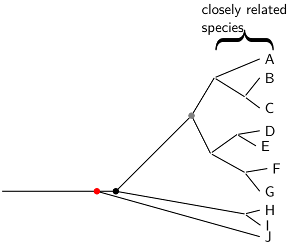


```
A. GKSWKALTPP
B. GKSWKSLSPS
C. GKSWKSLSTS
D. GKSWKSLSPS
E. GKSWKSLSPS
F. GKSWRALSPS
G. GRSWKSLSPS


H. PKSWRASSPS
I. PRSWKSSSPS


J. PKSWRALSPS
```

$$S = \frac{7 \times M_{P,G} + 2 \times M_{P,P}}{9}$$

$[A - G]$ and $[H - I]$ represent single lineages. But:
S Dominated by $M_{P,G}$

Why weight sequences



```
A. GKSWKALTPP
B. GKSWKSLSPS
C. GKSWKSLSTS
D. GKSWKSLSPS
E. GKSWKSLSPS
F. GKSWRALSPS
G. GRSWKSLSPS


H. PKSWRASSPS
I. PRSWKSSSPS


J. PKSWRALSPS
```

$$S = \frac{7\times M_{P,G}+2\times M_{P,P}}{9}$$

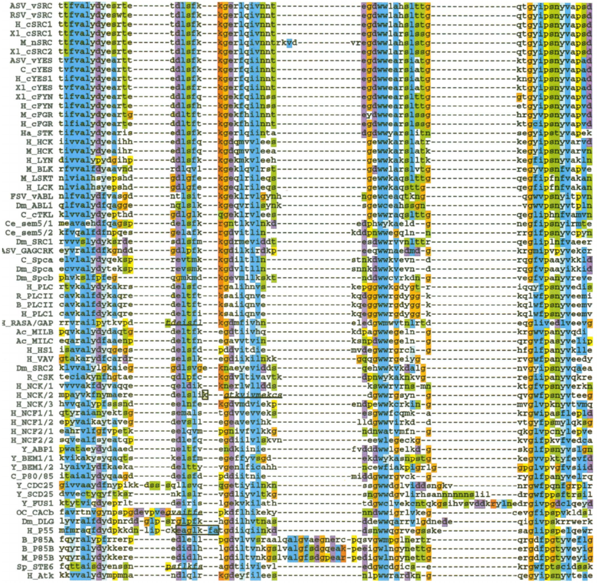
$[A - G]$ and $[H - I]$ represent single lineages. But:
S Dominated by $M_{P,G}$

May also wish to weight by branch length
(leads to maximum likelihood)

Problems?

- All alignments are global alignments and it may be necessary to trim sequences to give reasonable alignments.
- The guide tree is based on a matrix of distances of separately aligned sequences and may not be reliable. This may lead to mistakes early in the merging process that cannot be corrected later.

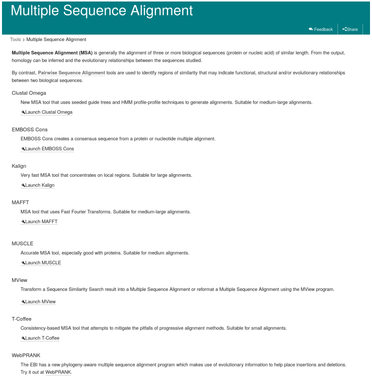
works ok!



```
  ASV_vSRC ttfvalydyesrte-------tdlsfk---kgerlqivnnt----------------egdwwlahslttg---------------qtgyipsnyvapsd
  RSV_vSRC ttfvalydyeswte-------tdlsfk---kgerlqivnnt----------------egdwwlahslttg---------------qtgyipsnyvapsd
   H_cSRC1 ttfvalydyesrte-------tdlsfk---kgerlqivnnt----------------egdwwlahslstg---------------qtgyipsnyvapsd
  Xl_cSRC1 ttfvalydyesrte-------tdlsfk---kgerlqivnnt----------------egdwwlarslssg---------------qtgyipsnyvapsd
    M_nSRC ttfvalydyesrte-------tdlsfk---kgerlqivnntrkvd----------vregdwwlahslstg---------------qtgyipsnyvapsd
  Xl_cSRC2 ttfvalydyesrte-------tdlsfr---kgerlqivnnt----------------egdwwlarslssg---------------qtgyipsnyvapsd
  ASV_vYES tvfvalydyeartt-------ddlsfk---kgerfqiinnt----------------egdwwearsiatg---------------ktgyipsnyvapad
    C_cYES tvfvalydyeartt-------ddlsfk---kgerfqiinnt----------------egdwwearsiatg---------------ktgyipsnyvapad
   H_cYES1 tifvalydyeartt-------edlsfk---kgerfqiinnt----------------egdwwearsiatg---------------kngyipsnyvapad
   Xl_cYES tvfvalydyeartt-------edlsfr---kgerfqiinnt----------------egdwwearsiatg---------------ktgyipsnyvapad
   Xl_cFYN tlfvalydyearte-------ddlsfq---kgekfqilnss----------------egdwwearslttg---------------gtgyipsnyvapvd
    H_cFYN tlfvalydyearte-------ddlsfh---kgekfqilnss----------------egdwwearslttg---------------etgyipsnyvapvd
    M_cFGR tifvalydyeartg-------ddltft---kgekfhilnnt----------------eydwwearslssg---------------hrgyvpsnyvapvd
    H_cFGR tlfialydyearte-------ddltft---kgekfhilnnt----------------egdwwearslssg---------------ktgcipsnyvapvd
    Ha_STK tifvalydyearis-------edlsfk---kgerlqiinta----------------dgdwwyarslitn---------------segyipstyvapek
     H_HCK iivvalydyeaihh-------edlsfq---kgdqmvvlees-----------------gewwkarslatr---------------kegyipsnyvarvd
     M_HCK tivvalydyeaihr-------edlsfq---kgdqmvvleea-----------------gewwkarslatk---------------kegyipsnyvarvn
     H_LYN divvalypydgihp-------ddlsfk---kgekmkvleeh-----------------gewwkakslltk---------------kegfipsnyvakln
     M_BLK rfvvalfdyaavnd-------rdlqvl---kgeklqvlrst-----------------gdwwlarslvtg---------------regyvpsnfvapve
    M_LSKT nlvialhsyepshd-------gdlgfe---kgeqlrileqs-----------------gewwkaqslttg---------------qegfipfnfvakan
     H_LCK nlvialhsyepshd-------gdlgfe---kgeqlrileqs-----------------gewwkaqsttg----------------qegfipfnfvakan
  FSV_vABL nlfvalydfvasgd-------ntlsit---kgeklrvlgynh---------------ngewceaqtkng-----------------qgwvpsnyitpvn
   Dm_ABL1 qlfvalydfqagge-------nqlslk---kgeqvrilsynk---------------sgewceahssgn-----------------vgwvpsnyvtpln
    C_cTKL klvvalydyepthd-------gdlglk---qgeklrvlees-----------------gewwraqslttg---------------qegliphnfvamvn
 Ce_sem5/1 meavaehdfqagsp-------delsfk---rgntlkvlnkd---------------edphwykaeld--g---------------negfipsnyirmte
 Ce_sem5/2 kfvqalfdfnpqes-------gelafk---rgdvitlin----------------kddpnwwegqln--n---------------rrgifpsnyvcpyn
   Dm_SRC1 rvvvslydyksrde-------sdlsfm---kgdrmeviddt----------------esdwwrvvnlttr---------------qeglipInfvaeer
ASV_GAGCRK eyvralfdfkgndd-------gdlpfk---kgdilkirdkp----------------eeqwwnaedmd-g---------------krgmipvpyvekcr
    C_Spca elvlalydyqeksp-------revtmk---kgdiltlln----------------stnkdwwkvevn--d---------------rqgfvpaayvkkld
   Dm_Spca ecvvalydyteksp-------revsmk---kgdvltlln----------------snnkdwwkvevn--d---------------rqgfvpaayikkid
   Dm_Spcb phvkslfpfeg----------qgmkmd---kgevmllkskt----------------nddwwcvrkdn-g---------------vegfvpanyvreve
     H_PLC rtvkalydykakrs-------delsfc---rgalihnvs----------------kepggwwkgdygt-r---------------iqqyfpsnyvedis
   R_PLCII cavkalfdykaqre-------deltft---ksaiiqnve----------------kqdggwwrgdygg-k---------------kqlwfpsnyveemi
   B_PLCII cavkalfdykaqre-------deltft---ksaiiqnve----------------kqeggwwrgdygg-k---------------kqlwfpsnyveemv
    H_PLC1 cavkalfdykaqre-------deltfi---ksaiiqnve----------------kqeggwwrgdygg-k---------------kqlwfpsnyveemv
H_RASA/GAP rrvrailpytkvpd------tdeisfl---kgdmfivhn----------------eledgwmwvtnlrtd---------------eqglivedlveevg
   Ac_MILB pqvkalydydaqtg-------deltfk---egdtiivhq----------------kdpagwwegeln--g---------------krgwvpanyvqdi
   Ac_MILC eqaralydfaaenp-------deltfn---egavvtvin----------------ksnpdwwegeln--g---------------qrgvfpasyvelip
     H_HS1 isavalydyqgegs-------delsfd---pddvitdie----------------mvdegwwrgrch--g---------------hfglfpanyvklle
     H_VAV gtakarydfcardr-------selslk---egdiikilnkk---------------gqqgwwrgeiyg-----------------rvgwfpanyveedy
   Dm_SRC2 klvvalylgkaieg-------gdlsvge--knaeyevidds----------------qehwwkvkdalg----------------nvgyipsnyvqaea
     R_CSK teciakynfhgtae-------qdlpfc---kgdvltivavtk---------------dpnwykaknkvg----------------regiipanyvqkre
   H_NCK/1 vvvvakfdyvaqqe-------qeldik---knerlwlldds-----------------kswwrvrns-mn---------------ktgfvpsnyverkn
   H_NCK/2 mpayvkfnymaere-------delslik---gtkvivmekcs----------------dgwwrgsyn--g---------------qvgwfpsnyvteeg
   H_NCK/3 hvvqalypfsssnd-------eelnfe---kgdvmdviekp----------------endpewwkcrkin-g---------------mvglvpknyvtvmq
  H_NCF1/1 qtyraianyektsg-------semals---tgdvvevveks----------------esgwwfcqmk--a---------------krgwipasflepld
  H_NCF1/2 epyvaikaytaveg-------devsll---egeavevihk----------------lldgwwvirkd--d---------------vtgyfpsmylqksg
  H_NCF2/1 eahrvlfgfvpetk-------eelqvm---pgnivfvlkkg----------------ndnwatvmfn--g---------------qkglvpcnylepve
  H_NCF2/2 sqvealfsyeatqp-------edlefq---egdiilvlskvn----------------eewlegeckg-----------------kvgifpkvfvedca
    Y_ABP1 pwataeydydaaed-------neltfv---endkiinie----------------fvdddwwlgelkd-g---------------skglfpsnyvslgn
  Y_BEM1/1 kvikakysyqaqts-------kelsfm---egeffyvsgd-----------------ekdwykasnpstg---------------kegvvpktyfevfd
  Y_BEM1/2 lyaivlydfkaeka-------deltty---vgenlficahh----------------ncewfiakpigrlg--------------gpglvpvgfvsiid
  C_P80/85 itaialydyqaagd-------deisfd---pddiitnie----------------middgwwrgvck--g---------------ryglfpanyvelrq
   Y_CDC25 givvaaydfnypikk-dss-sqllsvq---ggetiyilnkn----------------ssgwwdglviddsngkv-----------nrgwfpqnfgrplr
   Y_SCD25 dvvectyqyftksr-------nklslr---vgdliyvltkg----------------sngwwdgvlirhsannnnnnslil----drgwfppsftrsil
    Y_FUS1 ktytviqdyeprlt-------deiris---lgekvkilath----------------tdgwclvekcntqkgsihvsvddkrylnedrgivpgdclqeyd
   OC_CACb favrtnvgynpspgdevpvegvaitfe---pkdflhikeky----------------nndwwigrlvkeg--------------cevgfipspvkldsl
    Dm_DLG lyvralfdydpnrdd-glp-srglpfk---hgdilhvtnas----------------ddewwqarrvlgdnede-----------qigivpskrrwerk
     H_P55 mfmraqfdydpkkdn-lip-ckeaglk-fatgdiiqiinkd----------------dsnwwqgrvegsske-------------saglipspelqewr
    B_P85A fqyralypfrrerp-------edlell---pgdvlvvsraalqalgvaegnerc-pqsvgwmpglnertr---------------qrgdfpgtyveflg
    B_P85B yqyralydykkere-------edidlh---lgdiltvnkgslvalgfsdgqeak-peeigwlngynettg---------------ergdfpgtyveyig
    M_P85B yqyralydykkere-------edidlh---lgdiltvnkgslvalgfsdgpear-pedigwlngynettg---------------ergdfpgtyveyig
   Sp_STE6 fqttaisdyenssn------psflkfs---agdtiivievl----------------edgwcdgics--e---------------krgwfptscidssk
     H_Atk kkvvalydympmna-------ndlqlr---kgdeyfilees----------------nlpwwrardkn-g---------------qegyipsnyvteae
```

Thompson *et al*, NAR 1994, 22; 4673-4680

Other methods



Multiple Sequence Alignment

Tools > Multiple Sequence Alignment

Multiple Sequence Alignment (MSA) is generally the alignment of three or more biological sequences (protein or nucleic acid) of similar length. From the output, homology can be inferred and the evolutionary relationships between the sequences studied.

By contrast, Pairwise Sequence Alignment tools are used to identify regions of similarity that may indicate functional, structural and/or evolutionary relationships between two biological sequences.

Clustal Omega

New MSA tool that uses seeded guide trees and HMM profile-profile techniques to generate alignments. Suitable for medium-large alignments.

Launch Clustal Omega

EMBOSS Cons

EMBOSS Cons creates a consensus sequence from a protein or nucleotide multiple alignment.

Launch EMBOSS Cons

Kalign

Very fast MSA tool that concentrates on local regions. Suitable for large alignments.

Launch Kalign

MAFFT

MSA tool that uses Fast Fourier Transforms. Suitable for medium-large alignments.

Launch MAFFT

MUSCLE

Accurate MSA tool, especially good with proteins. Suitable for medium alignments.

Launch MUSCLE

MView

Transform a Sequence Similarity Search result into a Multiple Sequence Alignment or reformat a Multiple Sequence Alignment using the MView program.

Launch MView

T-Coffee

Consistency-based MSA tool that attempts to mitigate the pitfalls of progressive alignment methods. Suitable for small alignments.

Launch T-Coffee

WebPRANK

The EBI has a new phylogeny-aware multiple sequence alignment program which makes use of evolutionary information to help place insertions and deletions. Try it out at WebPRANK.

https://www.ebi.ac.uk/Tools/msa/

- Clustal Omega
- EMBOSS Cons
- Kalign
- MAFFT
- MUSCLE
- MView
- T-Coffee
- WebPRANK

Other methods

- DCA. Semi-exhaustive, divide and conquer algorithm. Breaks sequences into segments based on local similarity. Segments are aligned by dynamic programming and then joined. bibiserv.techfak.uni-bielefeld.de/dca?id=dca_view_webservice
- Poa (Partial order alignments). Uses a graph representation of the multiple alignment that can be aligned by dynamic programming. bioinformatics.oxfordjournals.org/content/18/3/452.short
- Dialign (and Dialign2, Dialign-TX). Compares segments of sequences rather than individual residues without using gap penalties. Good for comparing sequences with only local similarity. mobyle.pasteur.fr/cgi-bin/portal.py?#forms::dialign
- Can also be achieved by probabilistic methods (eg. Hidden Markov Models), though these are normally used to model (describe) sequences rather than to perform the actual alignment. But see hmmbuild: hmmer.janelia.org/

Many more available. Method usually determined by the sheep algorithm (i.e. use what others in the field are using).