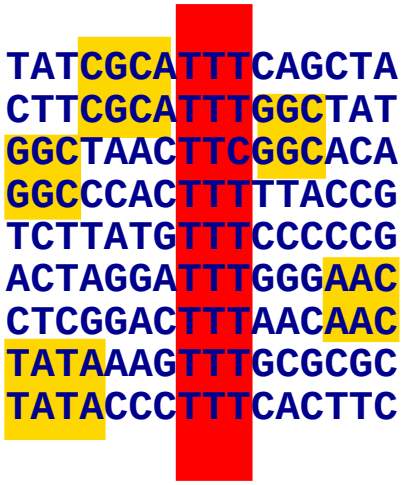


Two homologous sequences whisper ...
a full multiple alignment shouts out loud.

A. Lesk

The sum of pairs score:

Given a score system for pairwise alignment ...
... gaps are treated like any other character.

The sum of pairs score of one column is the sum over all possible pairwise comparisons in this column.

$$S(C)=\sum_{i<j} S(C(i),C(j))$$

Example: match =1, mismatch=0, space=−1

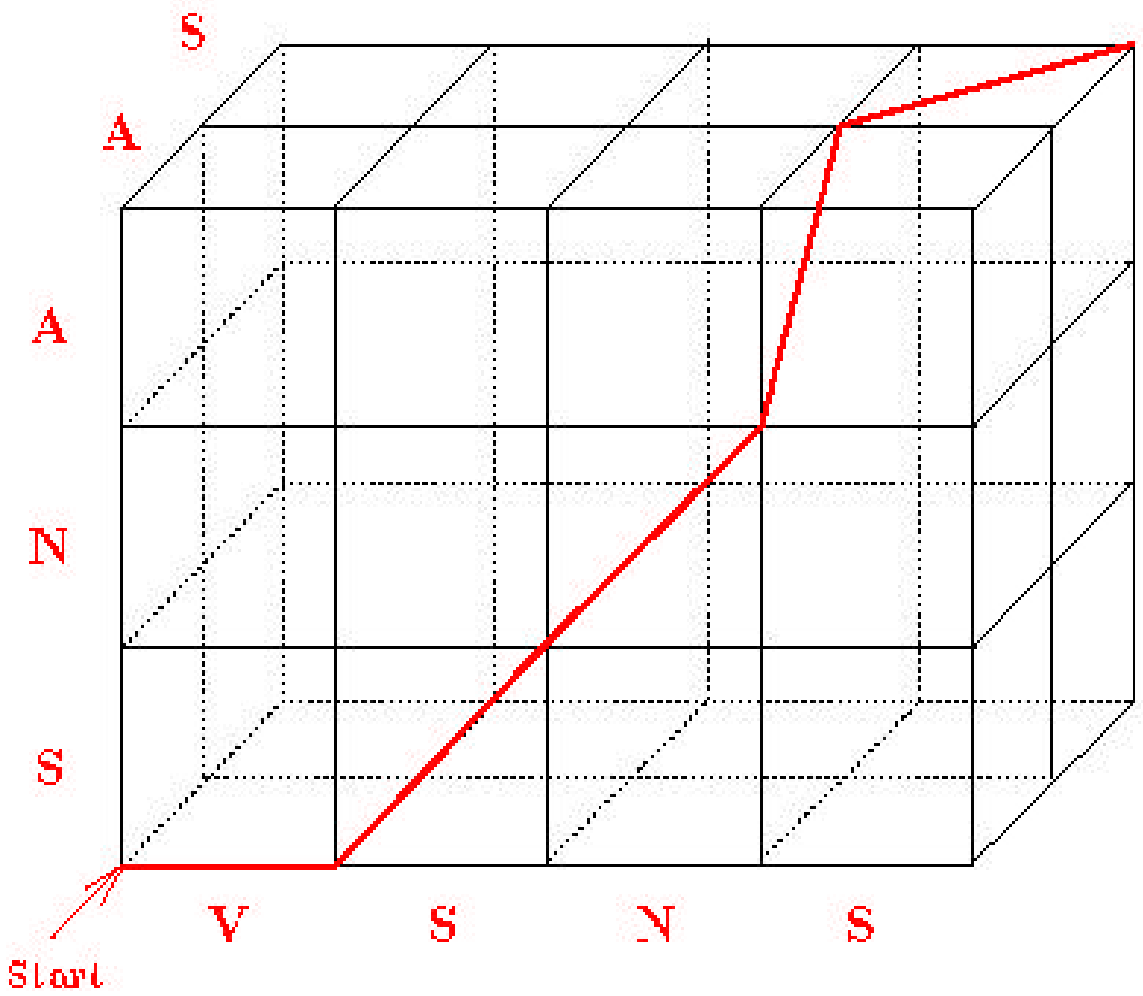
C= T
A
_
T

Score=0−1+1−1+0−1=−2.

The score of the alignment is the sum of the scores of the columns.

Three Protein Alignment
(Murata, Richardson & Sussman)

Figure 1: Alignment Path for 3 Sequences.



V S N – S
– S N A –
– – – A S

Time complexity exponential in the number of sequences:

$O(g(n1,...,nk)^k)$

NP–complete problem (Wang & Jiang 1994)

....forget it ...

... we had enough dynamic programming anyway.

Instead:
Construct multiple alignment from pairwise alignments.

Projected pairwise alignments

```
1   KL-ATYMKLSC
2   KL--TYKKL--
3   KL--TYKADSA
4   KLLYEYMKLS-
```

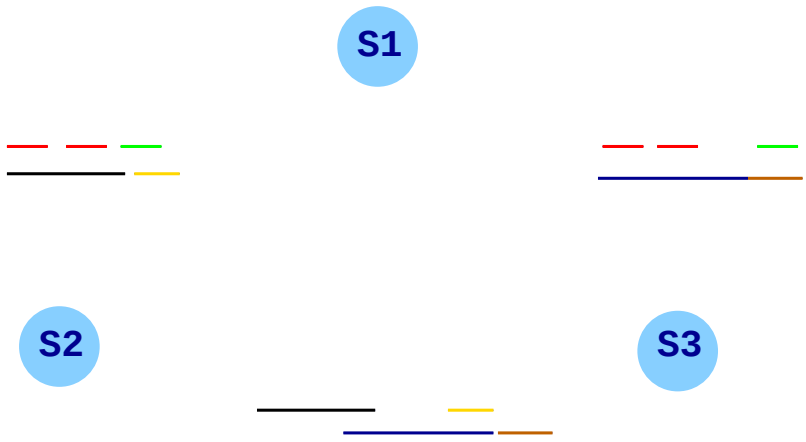
Projected alignment of 1 and 2

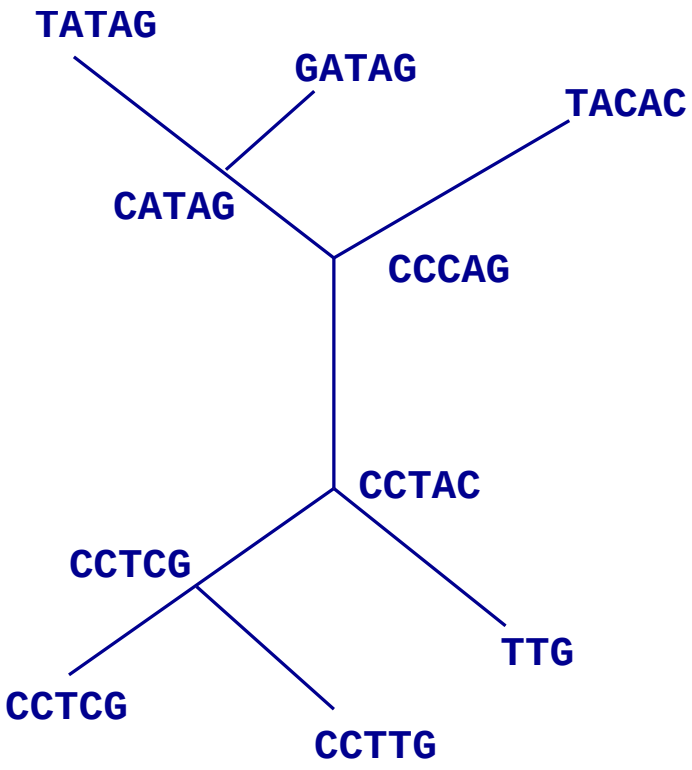
```
KLATYMKLSC
KL-TYMKLSC
```

copy the lines and delete opposite spaces

The projections of an optimal multiple alignment (sum of pairs) need not be optimal with respect to the pairwise scores ...

... and it is in general not possible to merge optimal pairwise alignments to a consistent multiple alignment.




TATAG
GATAG
TACAC
CATAG
CCCAG
CCTAC
CCTCG
TTG
CCTCG
CCTTG

Tree alignment

The multiple alignment is constructed one string at a time.

1. **Align two adjacent sequences say S1 and S2 (pairwise alignment)**

2. **Take the next sequence S3, it should be adjacent to either S1 and S2 ... assume it is connected to S1.**
   S1 might contain some gaps from the previous alignment,...leave them in and align this gapped sequence to S3. Opposite gaps have a score of zero. (pairwise alignment)

   Once a gap, always a gap

3. **Repeat until all sequences are aligned.**

If we project this multiple alignment to pairwise alignments, all pairwise alignments from sequences that are adjacent in the tree are optimal.

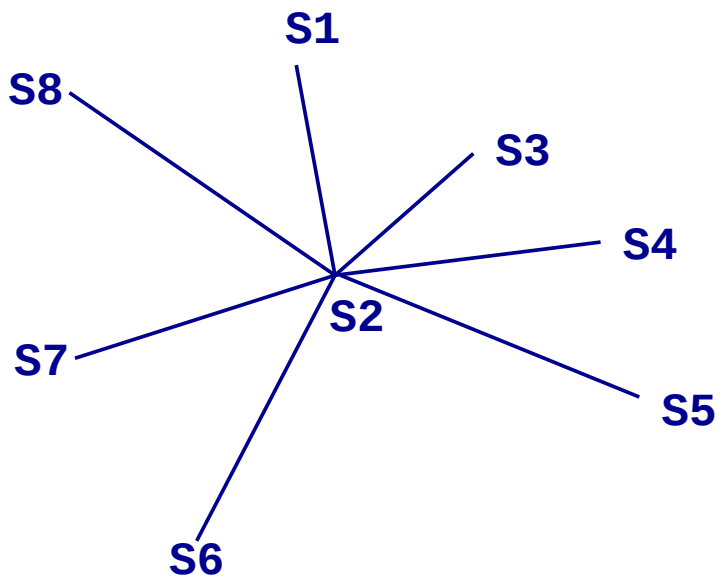
Great... but which tree do we choose ?

Score of a tree:
Sum of all pairwise alignment scores along edges of the tree.

Choose tree with optimal score among all possible trees ...

... the problem is NP-complete again.
Forget the optimal tree...

Star alignment



Choose center sequence Si, such that Si maximizes

$$\sum_{i \neq j} S(Si,Sj)$$

When distances are used and the triangle inequality holds:

The sum of pairs score of the star alignment is less or equal two times the optimal score.

For a proof see Gusfield’s book.

Scoring a multiple alignment

1. The positions are not all the same and should not be treated all the same some are more conserved then others ... these should determine the alignment.

2. The sequences are not independent, but related by a phylogenetic tree.

The sum of pairs score does not care at all ...

Another Problem with the sum of pairs score:

Alignment of N sequences.

There is a certain position with only L (leucine) for some important functional reason.

s(L,L)=+5

SP−Score: 5 N(N−1)/2

By an alignment error we get 1 G (glycine) in this position and only N−1 L's.

S(G,L)=−4
SP−Score: 9 (N−1) less
Relative Loss: 9(N−1)/(5N(N−1)/2)=18/5N

The more L's we observe the more evidence we have that the G is wrong.
However, the relative loss in SP−Score is decreasing for a large number of sequences.

Pairwise comparison table

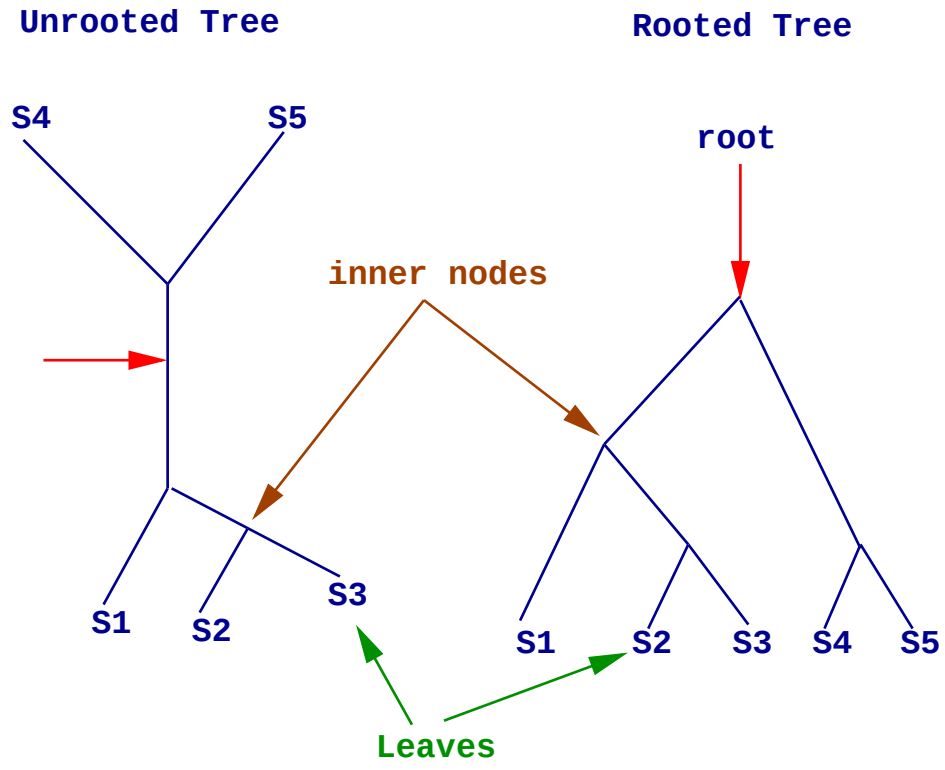
Calculate all pairwise alignment scores
and arrange them in a table

	S1	S2	S3	S4	S5
S1	●	−10	−5	4	−2
S2	−10	●	25	−8	0
S3	−5	25	●	−11	9
S4	4	−8	−11	●	−1
S5	−2	0	9	−1	●

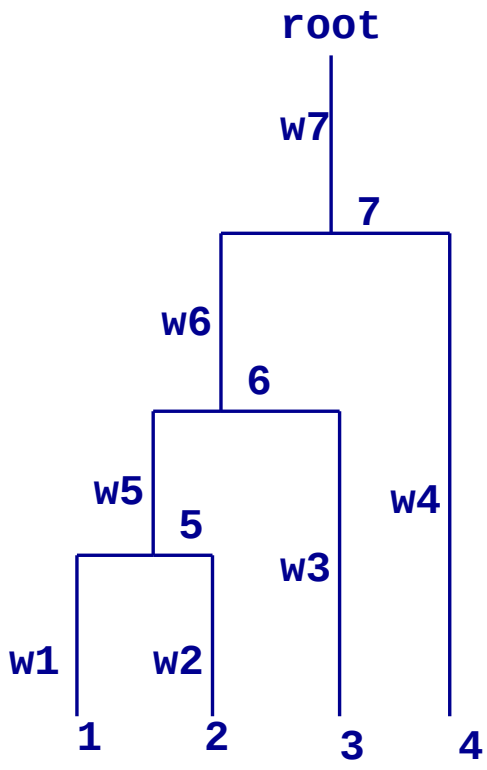
Convert all score into distances ...

1. Feng−Doolitle: D=−log(S−Srand)/(Smax−Srand)
2. Model based distances

   log det formular
   maximum likelihood


Unrooted Tree
Rooted Tree
S4
S5
root
inner nodes
S1
S2
S3
S1
S2
S3
S4
S5
Leaves

Molecular clock:
Edge Lengths represent "time of divergence"
The clock ticks at a constant rate along all paths in the tree.



Guide Trees by hierachical clustering

1. Choose the pair with minimal distance
   and merge it to a new single unit.
   This unit consisting of two sequences
   is equivalent to all remaing sequences
2. Update the distance table by calculating
   the distance of this unit to all other units
   in the table.
3. repeat until there is only a single unit

This results in a rooted tree with sequences
representing leaves.

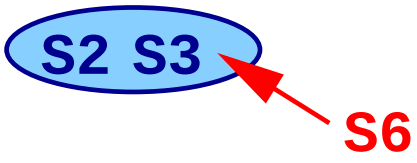
S2

S3

S4

S5

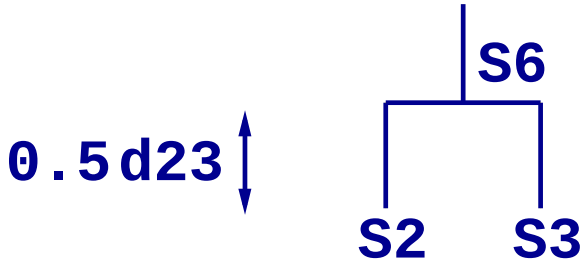
S1

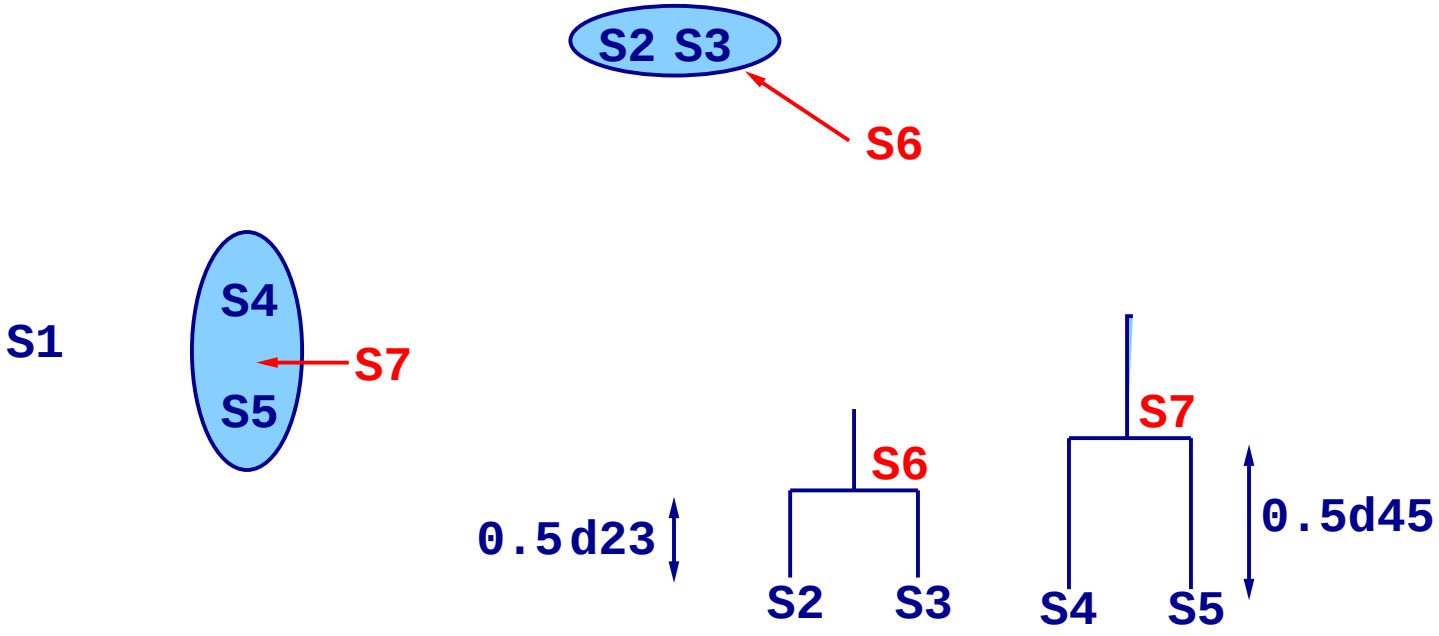


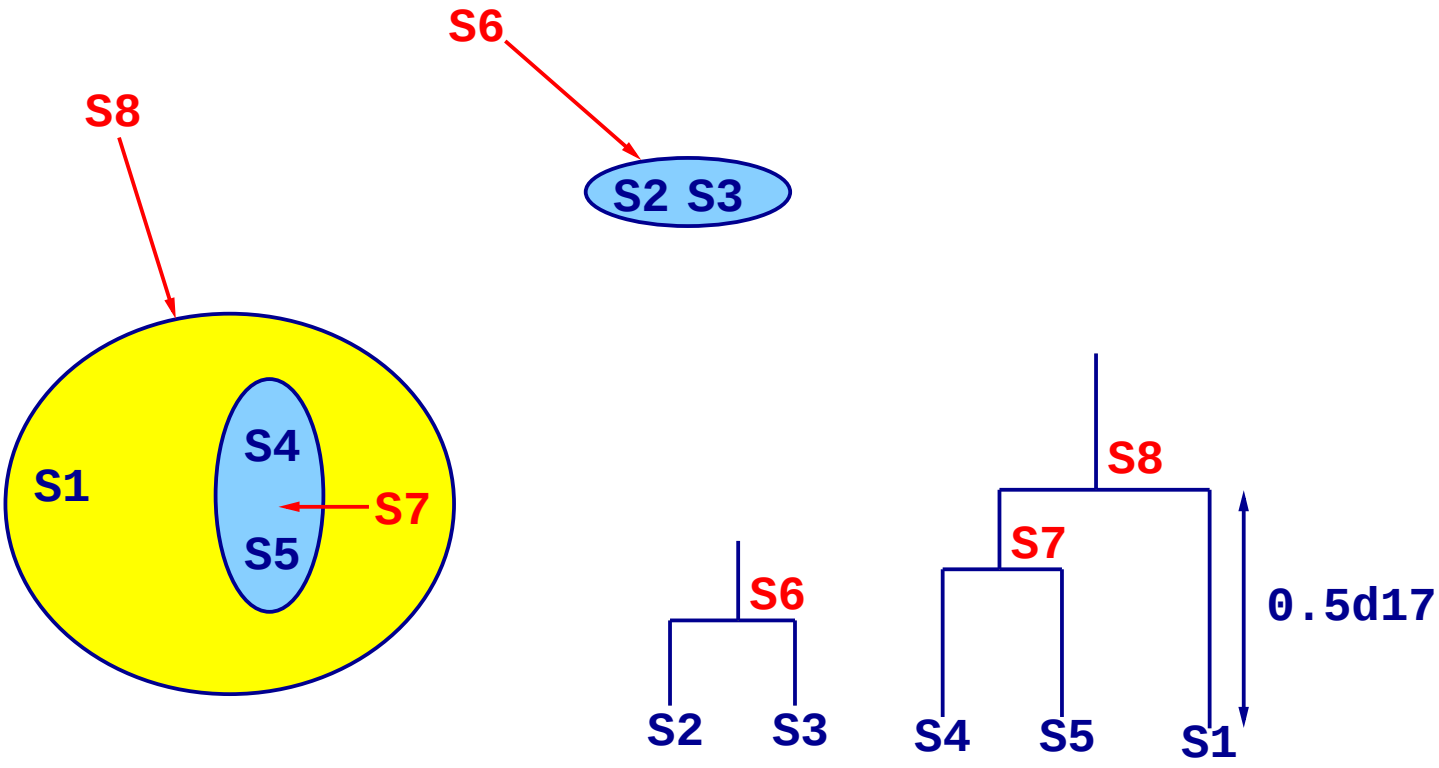
S4

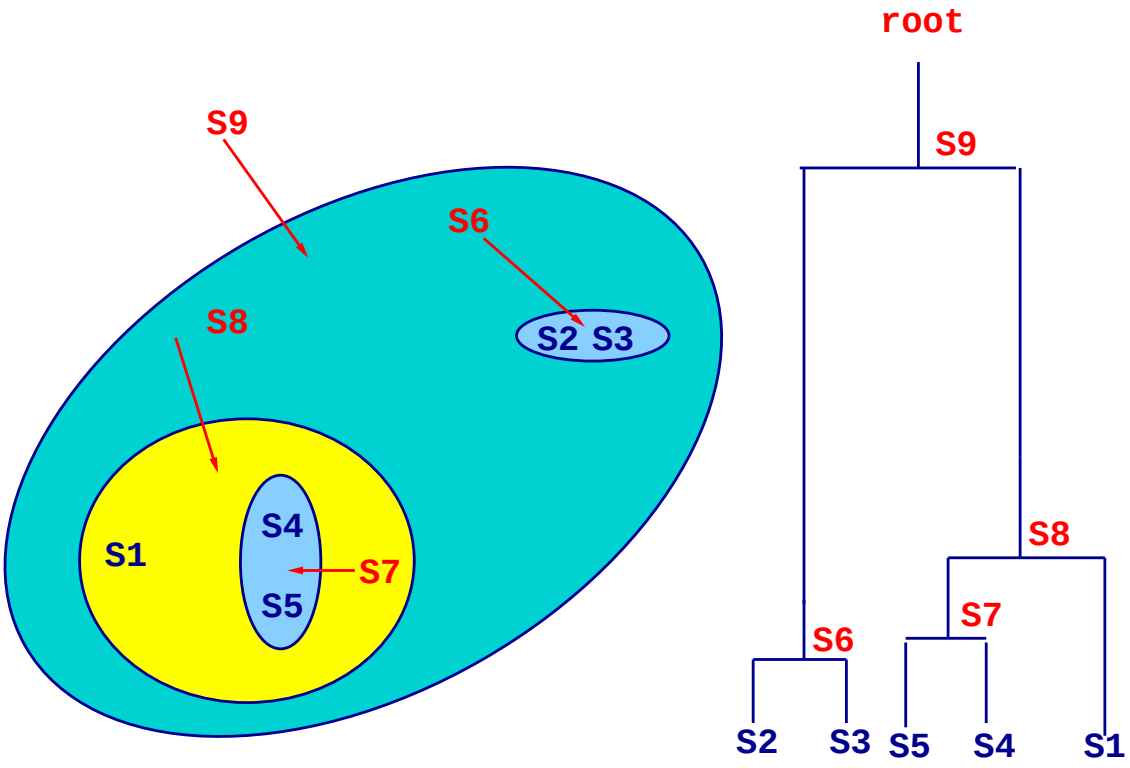
S1

S5

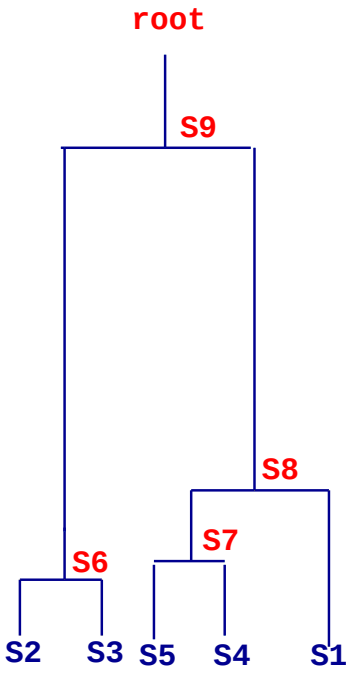



S2 S3
S6
S1
S4
S7
S5
0.5d23
S6
S2
S3
S7
0.5d45
S4
S5


S6
S8
S2 S3
S1
S4
S7
S5
S8
S7
S6
0.5d17
S2
S3
S4
S5
S1


S9
S6
S8
S2 S3
S1
S4
S7
S5
root
S9
S8
S7
S6
S2
S3
S5
S4
S1

The Guide Tree

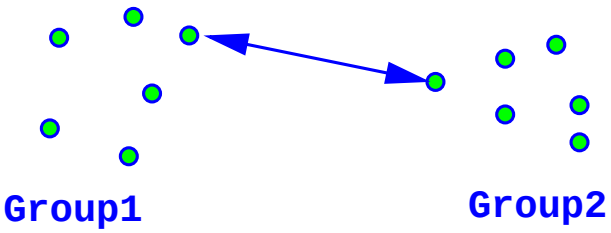


How do we calculate the score of groups of sequences?

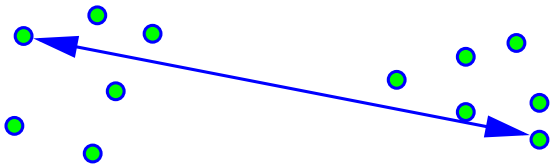
Group1: S1,...,Sn
Group2: T1,...,Tm

Single linkage: $S(Group1,Group2)=\max_{i,j} S(Si,Tj)$

the most similar pair of sequences



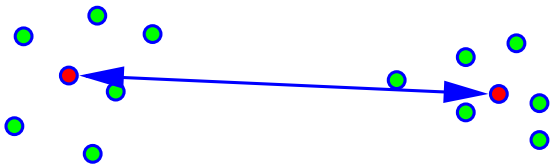
Total linkage: S(Group1,Group2)=min S(Si,Tj)
i,j



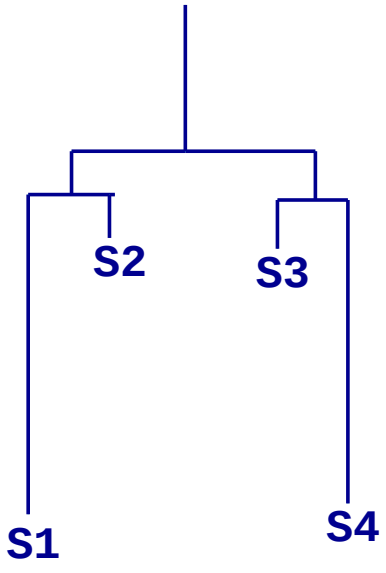
Group1 **Group2**

the least similar neighbor

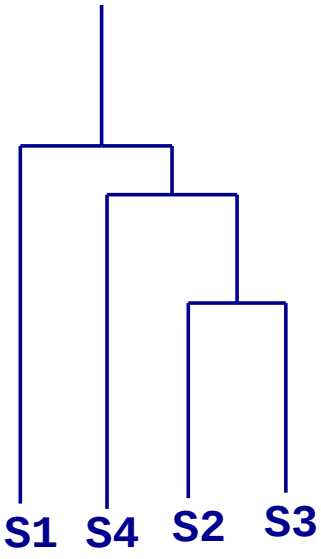
Average linkage: S(Group1,Group2)=mean S(Si,Tj)



Group1 **Group2**



Phylogenetic Tree



Guide Tree

Progressive Alignment

Build partial multiple alignments using only a subset of the sequences.

Start with the most similar sequences

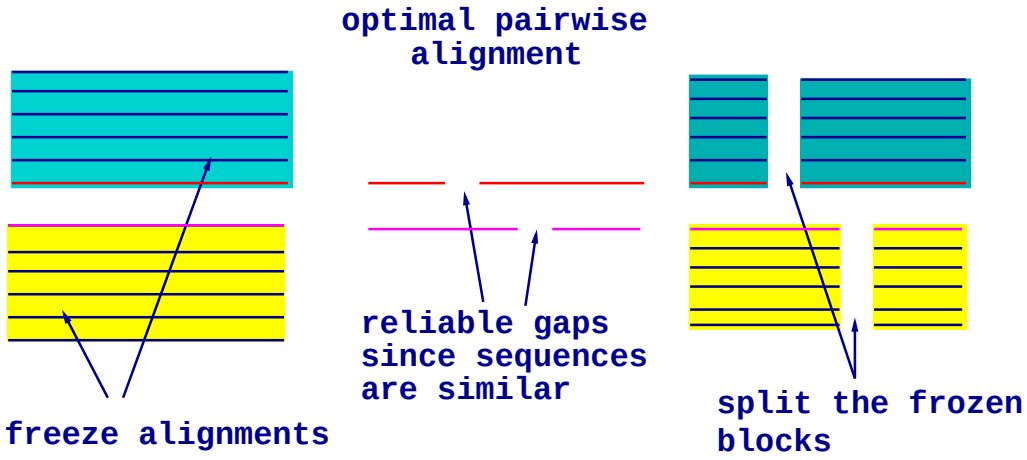
Join partial alignments using the "once a gap always a gap strategy"

The order in which sequences are included is determined by a Guide Tree

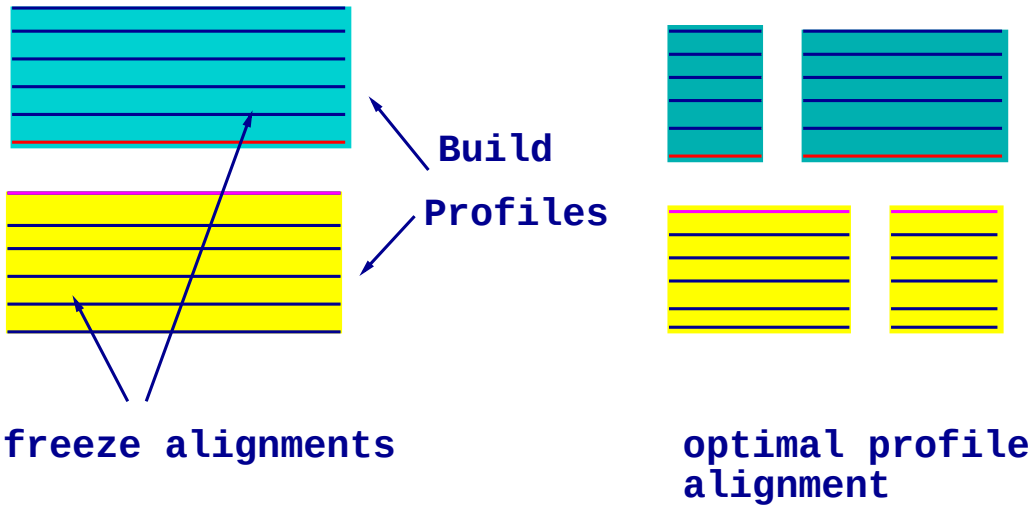
Feng Doolitle 1987

Join two partial multiple alignments according to the optimal pairwise alignment of the two most similar sequences in the partial alignments.

once a gap always a gap



Progressive Profile Alignment



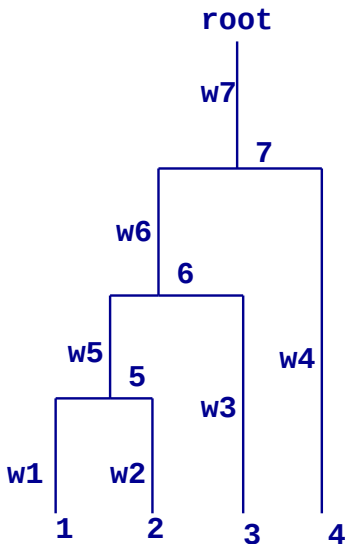
This set of very similar sequences will dominate the alignment.

Normally we do not have a random sample of family members. There are subfamilies that caught more attention in science than others ... hence, more data from this families is available.

Idea: Introduce sequence weights and limit the influence these clusters of similar sequences have in the alignment.

Rank them down ... and protect the rights of minorities.

Progressive Profile Alignment gives us a rooted tree with edge weights.



How do edge weights translate into weights for the sequences?

Gernstein, Sonnhammer and Chothia:

Work up the tree from the leaves to the root, incrementing the weights.

Let wi denote edge weights and li denote sequence weights.

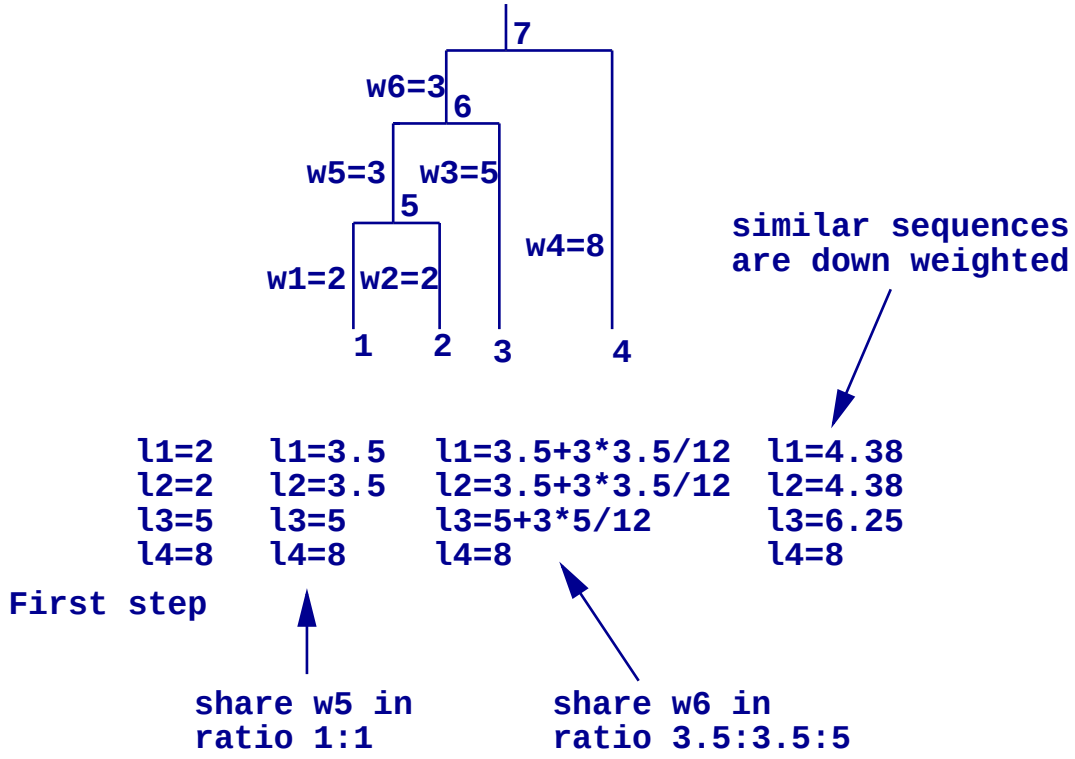
(1) Assign to each sequence the weight of the edge right above it.

li=wi

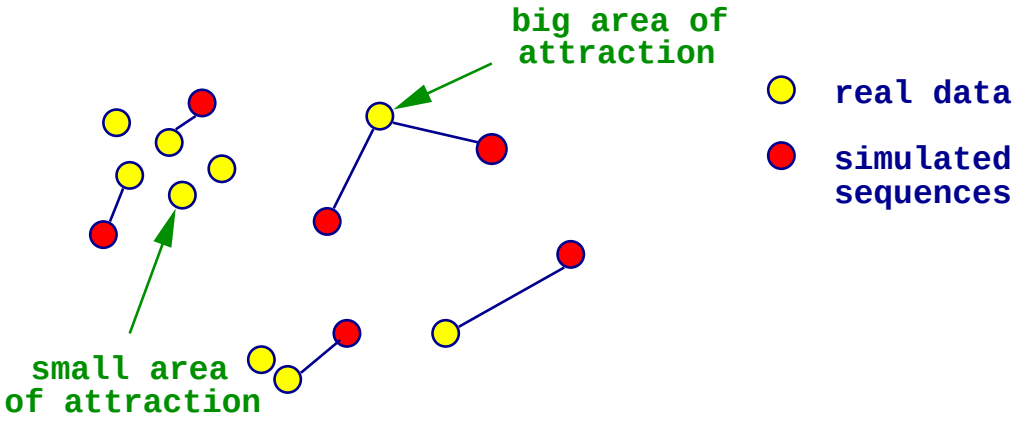
(2) Now suppose node n has been reached

Share the weight wn among all the sequences that are below n and increase there weights accordingly

$$\Delta li = wi * li \Big/ \sum_{k \text{ below } n} lk$$


7
w6=3
6
w5=3
w3=5
5
w4=8
w1=2
w2=2
1
2
3
4
similar sequences
are down weighted
l1=2
l2=2
l3=5
l4=8
l1=3.5
l2=3.5
l3=5
l4=8
l1=3.5+3*3.5/12
l2=3.5+3*3.5/12
l3=5+3*5/12
l4=8
l1=4.38
l2=4.38
l3=6.25
l4=8
First step
share w5 in
ratio 1:1
share w6 in
ratio 3.5:3.5:5

... or weights without a tree:



Simulate thousands of sequences from a profile type model and identify the sequence in the family that is most similar. Icrease the count of this sequence by one. Use counts as weights.

We need sequence weights to calculate the multiple alignment

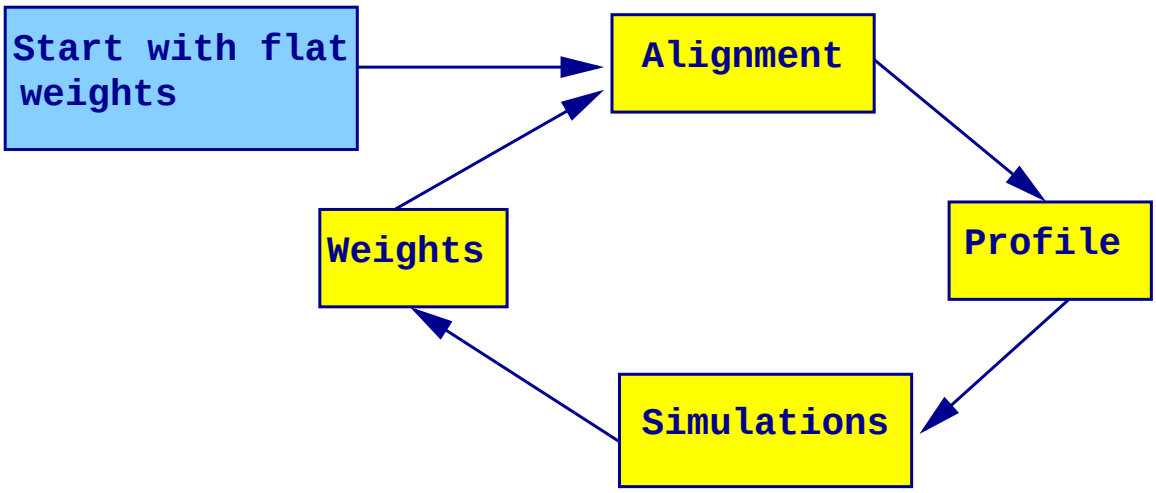
We need simulations to calculate weights

We need a model for the simulations

We need a profile for the model

We need a multiple alignment for the profile

Iterative updating of the components.



Check convergence!

Other weighting methods

- **Gaussian branching processes (used in ClustalW)**
- **Maximum Discrimination (used by hidden Markov models)**
- **Maximum Entropy**

see Durbin et al.

...along the same lines:

Early decissions might be wrong since there was little profile information available at this time.

Remove the first sequence from the alignment and realign it to the almost full alignment using sequence to profile alignment.

Continue with the second, third ,... sequence.

Iterate for some time.

(not implemented in ClustalW)