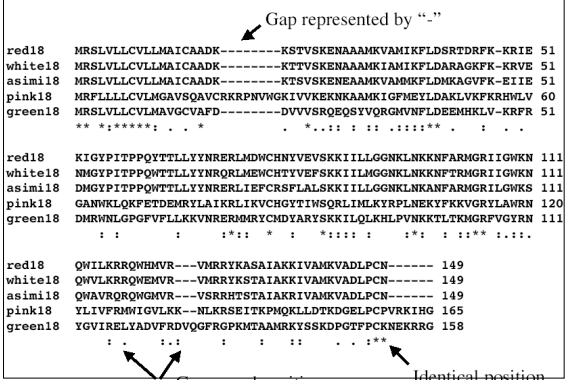
What is a multiple alignment

- Alignment of 3 or more DNA or AA sequences
- EVERY nucleotide position is assumed homologous

Example multiple alignment



What can you do with multiple alignments?

- Identify conserved and divergent regions
  - Functional inferences
- Phylogenies
  - Relationships between organisms
- Define motifs or domains
  - Identify additional family members; motif searching more sensitive than Blast.
- Secondary structure prediction
  - Protein structure

ClustalW

- Align each pair of sequences and calculate distance matrix.
- Calculate an neighbor joining phylogenetic tree.
- Progressively align sequences, starting with most similar sequences based on tree.
  - **Retain information from closely related sequences that are easy to align.**

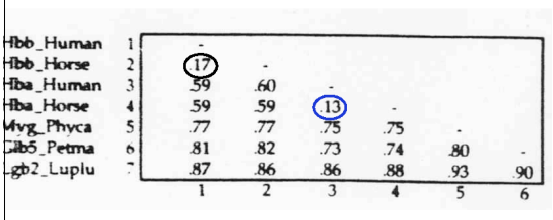
Distances

- Percent divergence of sequences.
  - Sequences aligned, number identities/number of residues, and subtract from 1

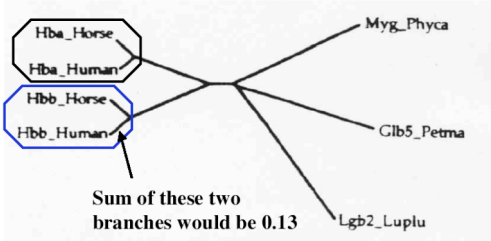
AWYSTRGHKIL
AWYTTRGHKIL

9/10 = 0.9
1 - 0.9 = 0.10

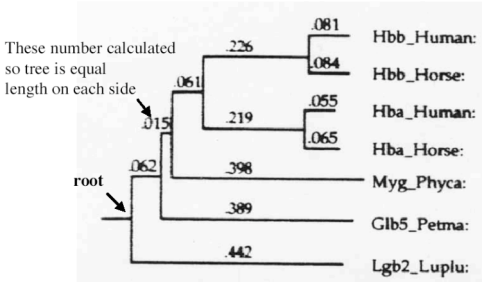
Make a matrix of all pair-wise distances. Sequences that have the lowest divergence are the most closely related and are grouped together.



Unrooted tree shows the grouping of the most closely related sequences. Length of "branch" is the distance. Make a "rooted tree" by midpoint method

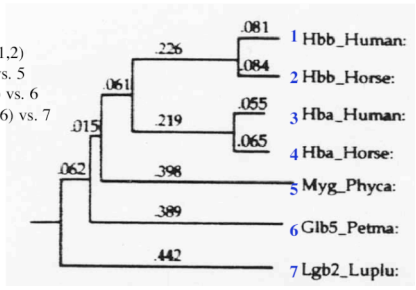


Make a "rooted" tree using midpoint method. Longest branch considered root.



Progressive alignment: retain info as you align sequences

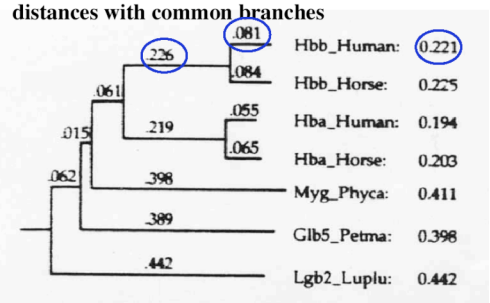
3 vs. 4
1 vs. 2
(3,4) vs. (1,2)
(3,4,1,2) vs. 5
(3,4,1,2,5) vs. 6
(3,4,1,2,5,6) vs. 7



Sequence weighting

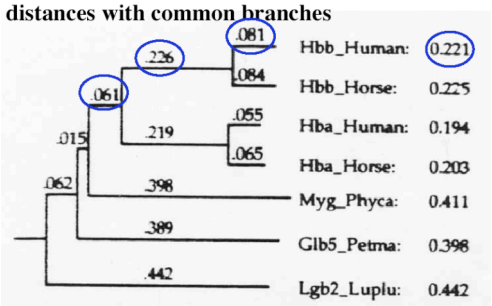
- Way to down-weight duplicate information
- Gives more weight to rare, divergent sequences

Sequence weighting: distance to root, but share distances with common branches



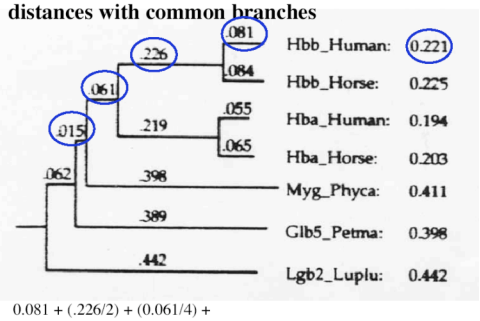
0.081 +

Sequence weighting: distance to root, but share distances with common branches



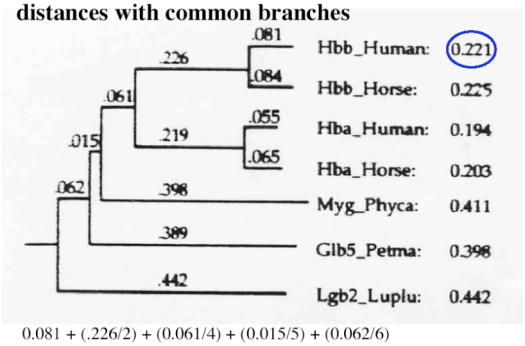
0.081 + (.226/2) +

Sequence weighting: distance to root, but share distances with common branches



0.081 + (.226/2) + (0.061/4) +

Sequence weighting: distance to root, but share distances with common branches



0.081 + (.226/2) + (0.061/4) + (0.015/5) + (0.062/6)

Gap penalties

- Not just gap open and gap extend penalty.
- Use information in alignment to guide choice of gap penalties.
- Position specific gap penalties.

Initial Gap penalty

- Just like Blast
  - Gap open
  - Gap extend

Sequence similarity

- Closely related sequences less likely to have gaps, reduce the penalty.
- Gap open penalty decreases on linear scale for more divergent sequences

Matrix used

- Different matrices have different optimal gap open penalties.
  - Similar to sequence divergence
- Take average score for mismatched residues as a scaling factor.

Lowered gap penalties for existing alignments

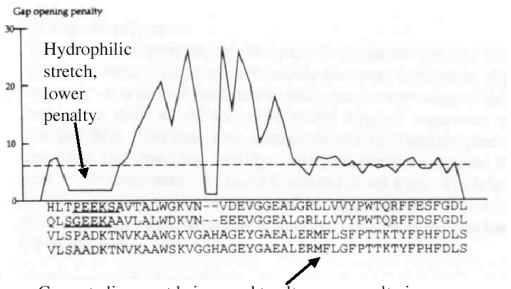
- If a gap already occur in the progressive alignment, it is likely that it will occur in the same position in subsequent alignments
- Reduced based upon proportion of sequences with a gap in that position

Increased penalty near existing gaps.

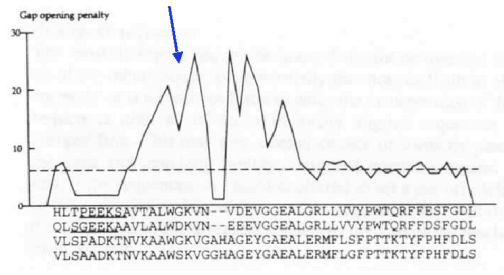
- Force gaps to occur in same position in the alignment
- Increase gap penalties in region 8 residues form existing gaps

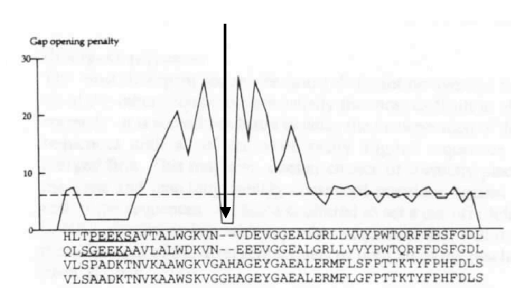
Hydrophilic stretches

- Reduce gap in a run of 5 hydrophilic residues
- Hydrophilic residues tend to be in loops, where most gaps occur.



Current alignment being used to alter gap penalty in a position specific manner.





Clustal using different matrices during alignment (remember difference between Blosum62 and Blosum 45 in BLAST)

- Matrix varied throughout alignment dependent upon %ID
  - BLOSUM or PAM (BLOSUM default)
- Sequences 80% - 100% identical
  - BLOSUM80
- Sequences 60% - 80% identical
  - BLOSUM62 (Blast default)
- Sequences 30% - 60% identical
  - BLOSUM45
- Sequences 0% - 30% identical
  - BLOSUM30

Multiple Fasta formatted unaligned sequences

```
>red18
MRSLVLLCVLLMAICAADKKSTVSKENAAAMKVAMIKFLDSRTDRFKKRIE
KIGYPITPPQYTTLLYYNRERLMDWCHNYVEVSKKIILLGGNKLNKKNFARMGRIIGWKN
QWILKRRQWHMVRVMRRYKASAIAKKIVAMKVADLPCN
>white18
MRSLVLLCVLLMAICAADKKTTVSKENAAAMKIAMIKFLDARAGKFKKRVE
NMGYPITPPQWTTLLYYNRQRLMEWCHTYVEFSKKIILMGGNKLNKKNFTRMGRIIGWKN
QWVLKRRQWEMVRVMRRYKSTAIAKKIVAMKVADLPCN
>asimi18
MRSLVLLCVLLMAICAADKKTSVSKENEAAMKVAMMKFLDMKAGVFKEIIE
DMGYPITPPQWTTLLYYNRERLIEFCRSFLALSKKIILLGGNKLNKANFARMGRILGWKS
QWAVRQRQWGMVRVSRRHTSTAIAKRIVAMKVADLPCN
>pink18
MRFLLLLCVLMGAVSQAVCRKRPNVWGKIVVKEKNKAAMKIGFMEYLDAKLVKFKRHWLV
GANWKLQKFETDEMRYLAIKRLIKVCHGYTIWSQRLIMLKYRPLNEKYFKKVGRYLAWRN
YLIVFRMWIGVLKKNLKRSEITKPMQKLLDTKDGELPCPVRKIHG
>green18
MRSLVLLCVLMAVGCVAFDDVVVSRQEQSYVQRGMVNFLDEEMHKLVKRFR
DMRWNLGPGFVFLLKKVNRERMMRYCMDYARYSKKILQLKHLPVNKKTLTKMGRFVGYRN
YGVIRELYADVFRDVQGFRGPKMTAAMRKYSSKDPGTFPCKNEKRRG
```

Clustal Demo

ClustalW Parameters: for most uses the default works fine

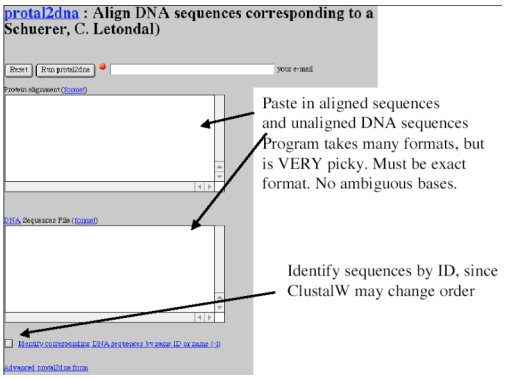
- Output format
  - Phylip, PIR, etc.
- Gap open/extend
- Matrix
  - BLOSUM is best
- Ktup
  - Word size

ClustalW will give you a guide tree. There are better methods for constructing phylogenies, and we will go over them later

Aligning Protein coding regions

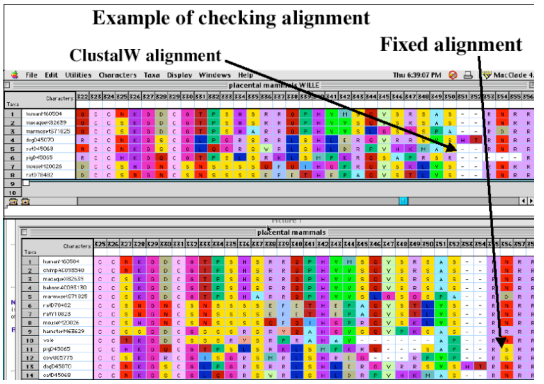
- Align protein sequence
  - Gaps are easier to deal with; wobble
  - AA information regarding substitutions
- Use aligned protein sequence to align DNA
  - Use program Protal2DNA

Protal2DNA Demo



You must check alignments by eye

- Each position must be homologous.
- Focus on regions around gaps.
- Look for conservative AA changes, as indicated by sequence editor color.
- If doing protein coding region, check both protein and DNA alignment (silent changes).
- Do not over edit. If sequence is so far out of alignment you should try changing gap parameters in Clustal.



Use Se-Al to check alignment

- Look at Protein
- Look at DNA (minimize silent changes)
- Export to different file formats
- Color coded amino acids
- Has some problems, occasionally crashes the computer (but it is free). Save work as you go.

Se-al example