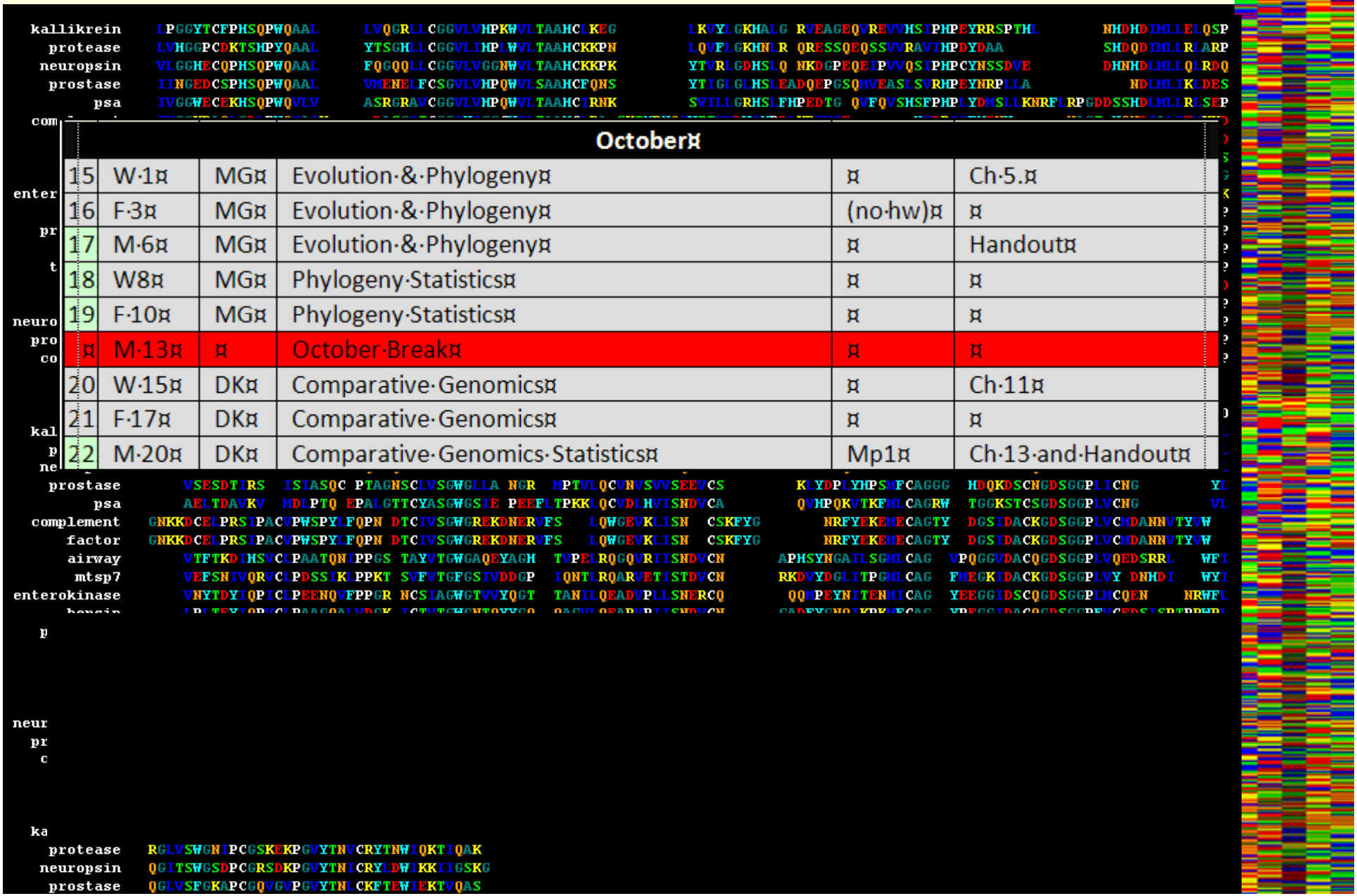
Biol 478/595 Intro to Bioinformatics

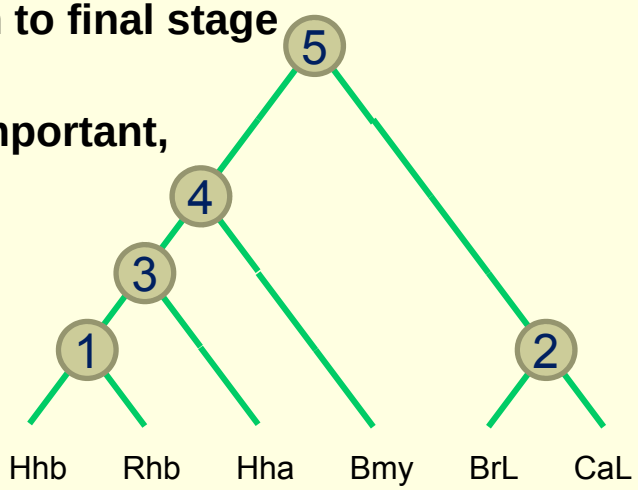


October¤					
15	W·1¤	MG¤	Evolution·&·Phylogeny¤	¤	Ch·5.¤
16	F·3¤	MG¤	Evolution·&·Phylogeny¤	(no·hw)¤	¤
17	M·6¤	MG¤	Evolution·&·Phylogeny¤	¤	Handout¤
18	W8¤	MG¤	Phylogeny·Statistics¤	¤	¤
19	F·10¤	MG¤	Phylogeny·Statistics¤	¤	¤
¤	M·13¤	¤	October·Break¤	¤	¤
20	W·15¤	DK¤	Comparative·Genomics¤	¤	Ch·11¤
21	F·17¤	DK¤	Comparative·Genomics¤	¤	¤
22	M·20¤	DK¤	Comparative·Genomics·Statistics¤	Mp1¤	Ch·13·and·Handout¤

Multiple Alignment and Trees

Progressive Alignment

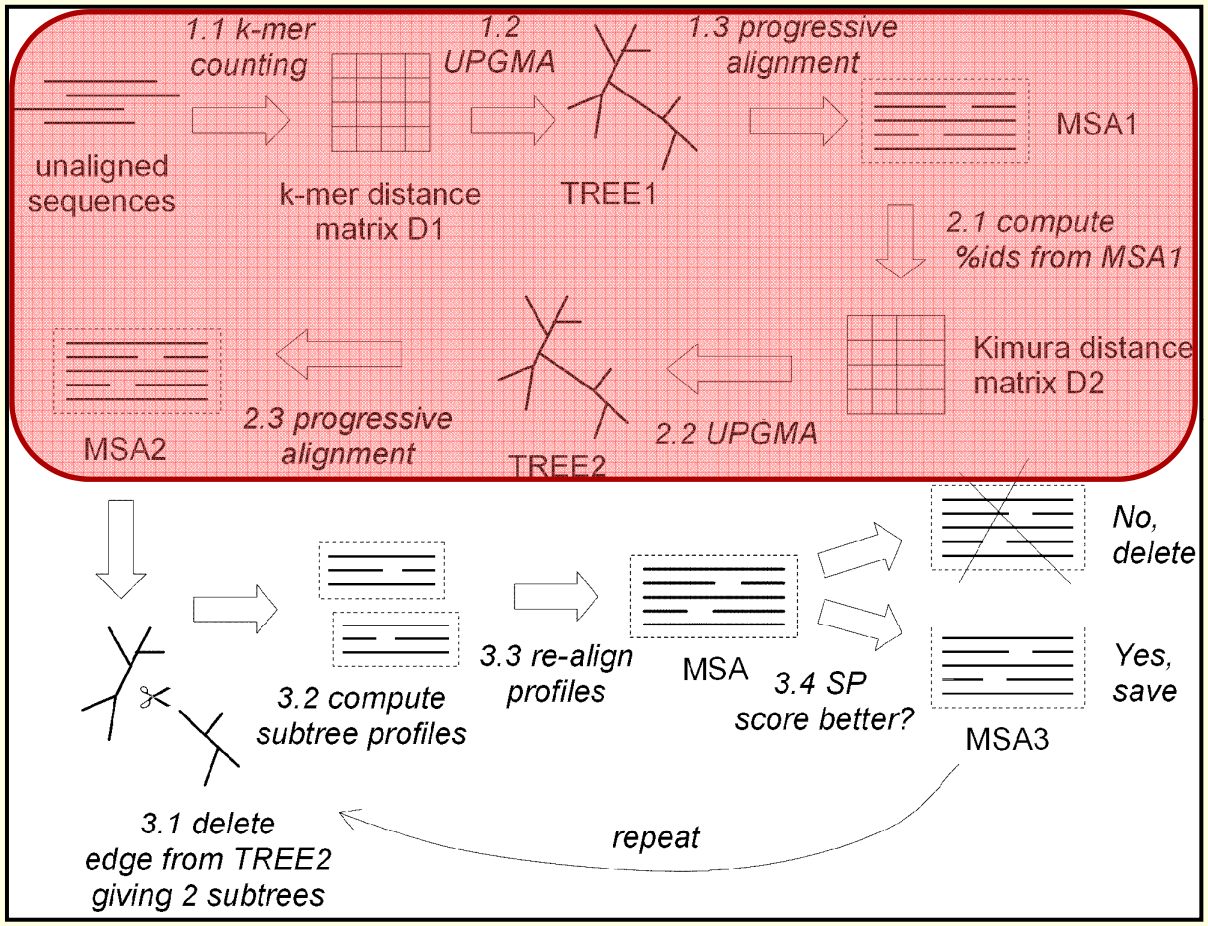
- **Practical multiple alignments are made using a progressive alignment procedure.**
- **The alignment is constructed by adding one sequences to a growing alignment in order of similarity (closest first) according to guide tree**
- **General problems with progressive alignments**
  - **Solution is not guaranteed to be optimal. The greedy strategy used in progressive alignment is highly like to be "trapped" in a local optimum.**
  - **Error in early stages propagate through to final stage - there is no error correction possible**
  - **Choices of alignment parameters are important, but appropriate settings are difficult to determine**



Multiple Alignment and Trees

Muscle (2004)

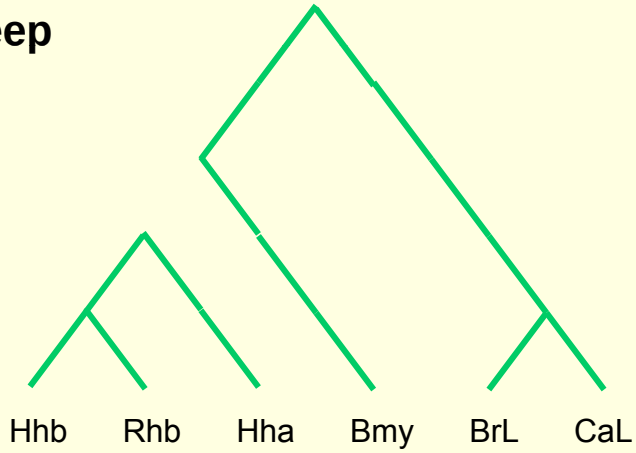
- Intially like Clustal
- Iteratively improve

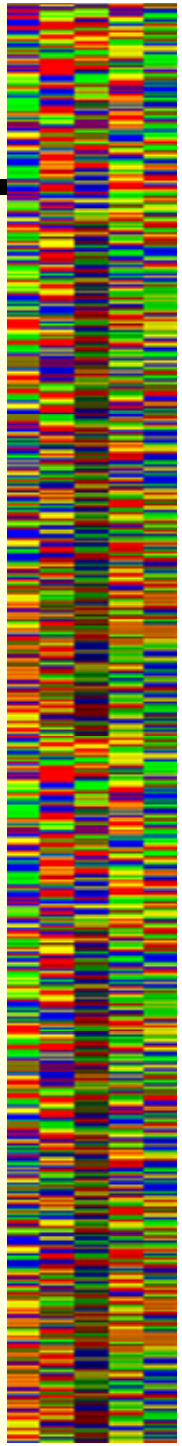


Multiple Alignment and Trees

Muscle

- **Iterative refinement**
  - **starting at root, take one edge out of tree splitting into two**
  - **align profiles for two trees**
  - **if SOP score is better, keep; otherwise keep original**
  - **Repeat until no changes, or user limit**

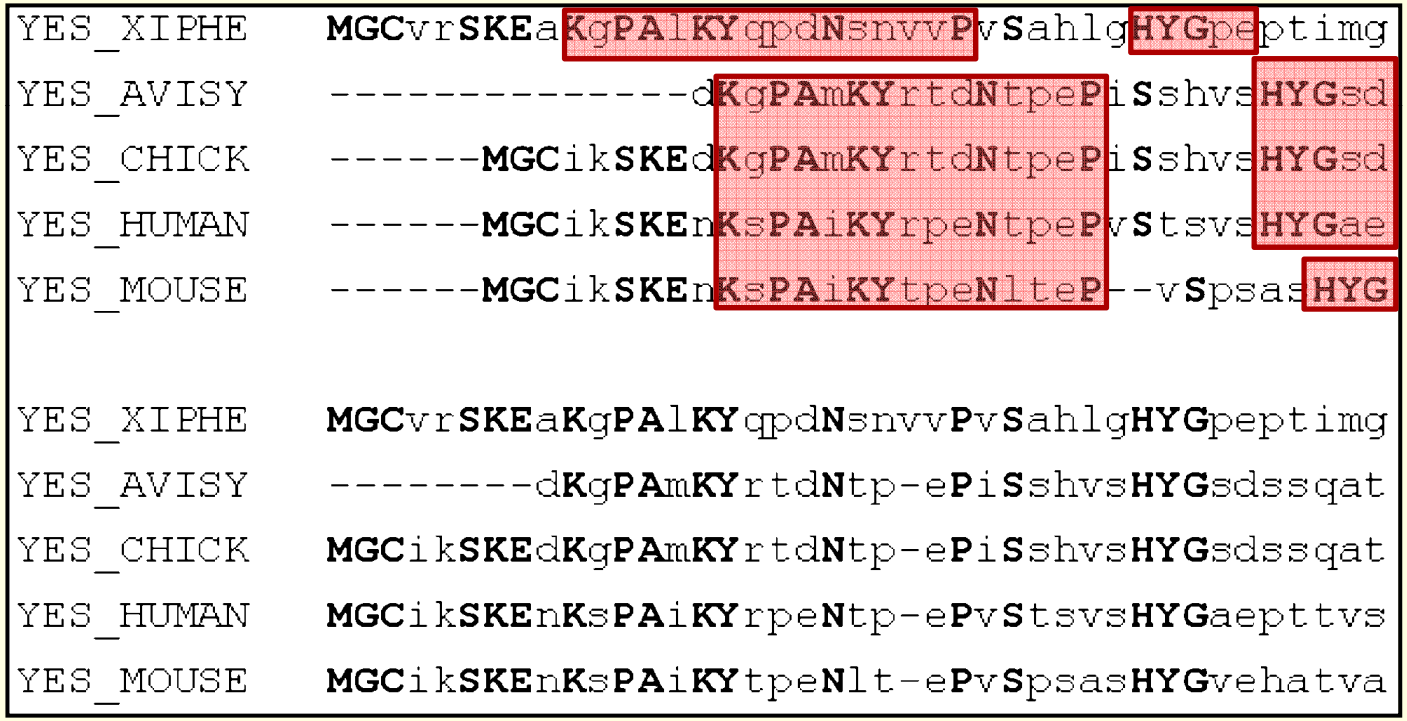




Multiple Alignment and Trees

Muscle

- **Example of improvement**



Multiple Alignment and Trees

Muscle

- **Balibase – database of alignments (mostly based on structure)**
  - **Q = quality = number of correct residue pairs divided by length of alignment**
  - **TC =Total column score, number of completely correct columns**
  - **Muscle-p skips iterative refinement**

Table 1. BAliBASE scores and times

Method	*Q*	TC	CPU
MUSCLE	0.896	0.747	97
MUSCLE-p	0.883	0.727	52
T-Coffee	0.882	0.731	1500
NWNSI	0.881	0.722	170
CLUSTALW	0.860	0.690	170
FFTNS1	0.844	0.646	16

Average *Q* and TC scores for each method on BAliBASE are shown, together with the total CPU time in seconds. Align-m aborted on two alignments; average scores on the remainder were *Q* = 0.852 and TC = 0.670, requiring 2202 s.

Table 2. BAliBASE Q scores on subsets

Method	Ref1	Ref2	Ref3	Ref4	Ref5
MUSCLE	0.887	0.935	0.823	0.876	0.968
MUSCLE-p	0.871	0.928	0.813	0.857	0.974
T-Coffee	0.866	0.934	0.787	0.917	0.957
NWNSI	0.867	0.923	0.787	0.904	0.963
CLUSTALW	0.861	0.932	0.751	0.823	0.859
FFTNS1	0.838	0.908	0.708	0.793	0.947

The average *Q* score for each method on each BAliBASE subset is shown. Ref1 is the largest subset with 81 test sets, comprising almost 60% of the database. Other subsets are smaller. For example, Ref4 and Ref5 have 12 alignments each, and there are large variances in the individual scores from which the averages are computed. In our opinion, it is not possible to draw meaningful conclusions about the relative performance of different methods on these subsets.

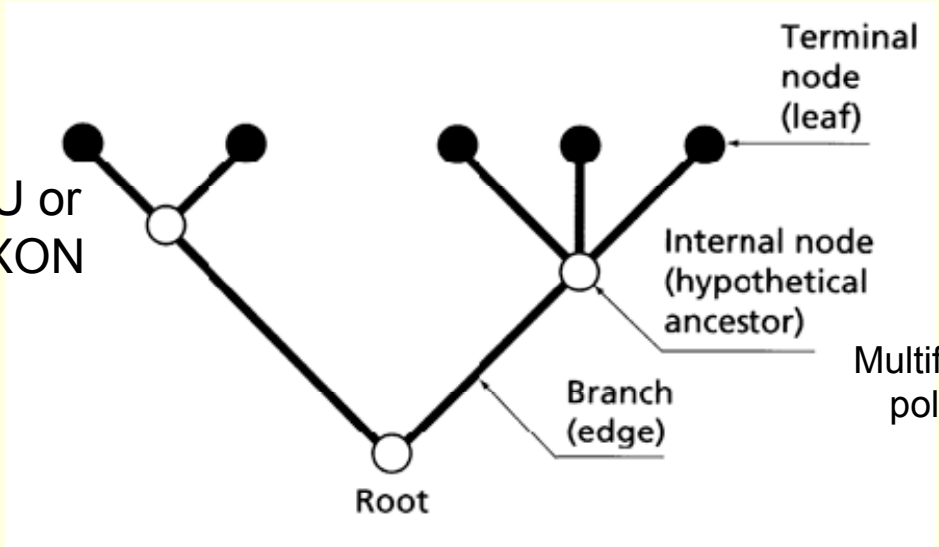
Table 3. BAliBASE TC scores on subsets

Method	Ref1	Ref2	Ref3	Ref4	Ref5
MUSCLE	0.815	0.574	0.577	0.627	0.902
MUSCLE-p	0.795	0.558	0.550	0.598	0.891
T-Coffee	0.780	0.573	0.510	0.751	0.903
NWNSI	0.788	0.514	0.514	0.742	0.859
CLUSTALW	0.782	0.579	0.470	0.542	0.638
FFTNS1	0.732	0.496	0.350	0.451	0.831

The average TC score for each method on each BAliBASE subset is shown.

Multiple Alignment and Trees

Basic tree vocabulary

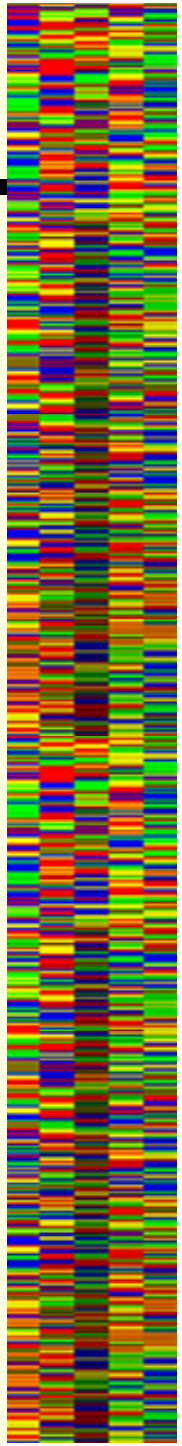


Multiple Alignment and Trees

Terminology

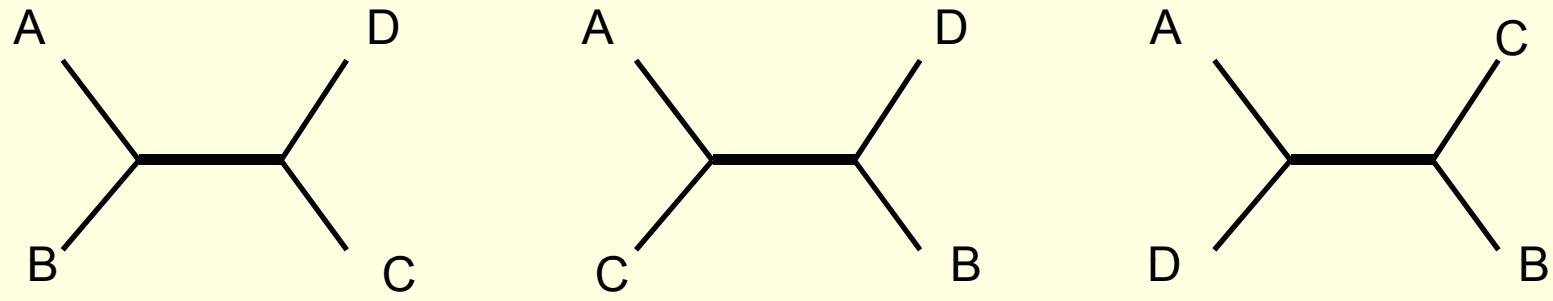
- **Topology - The branching pattern of the tree**
- **Rooted Tree - Tree in which the position of the ancestor is known**
- **Unrooted Tree - Tree with no ancestral node**

- **Taxon (Taxa) - Each leaf of the tree is a taxon (plural taxa)**
- **OTU - Operational Taxonomic Unit, a group of taxa related by a tree**
- **Clade - a group of taxa all on the same branch of a tree**

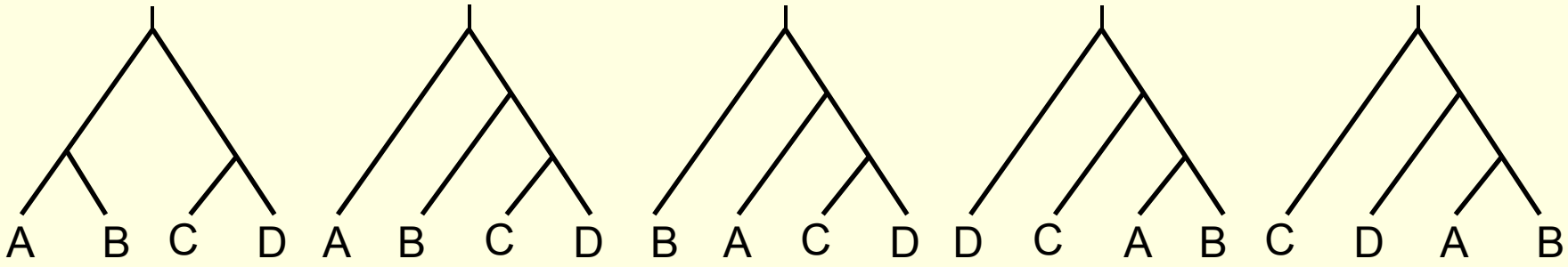


Multiple Alignment and Trees

- ***Rooted vs Unrooted Trees***



All three unrooted trees of four taxa



Five of fifteen rooted trees of four taxa, each corresponds to the unrooted tree at the left

Multiple Alignment and Trees

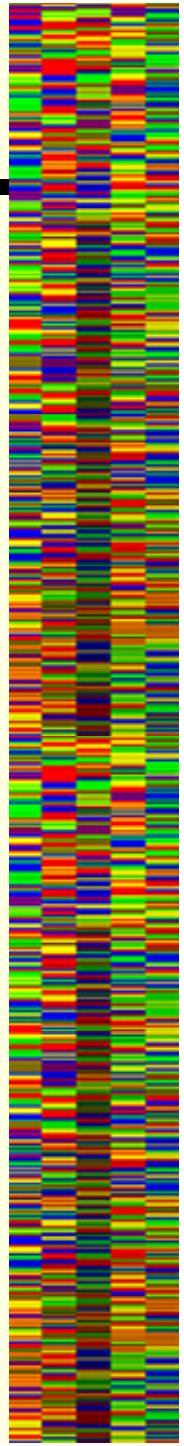
There are very many possible trees

- **One of the difficulties in constructing trees is the large number of possible trees for even relatively small numbers of taxa**

  Number_of_Unrooted_Trees = (2n-5)! / 2n-3(n-3)!

Unrooted		Rooted	
Taxa	**Trees**	**Taxa**	**Trees**
4	**3**	**7**	**945**
5	**15**	**8**	**10,395**
6	**105**	**10**	**2,027,025**

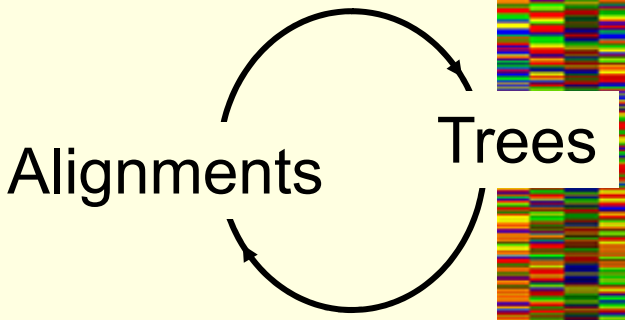
- **For large numbers of taxa it is impossible to enumerate all the trees and decide which is best**
- **There are many more rooted trees, than unrooted trees**



Multiple Alignment and Trees

Main approaches to tree construction

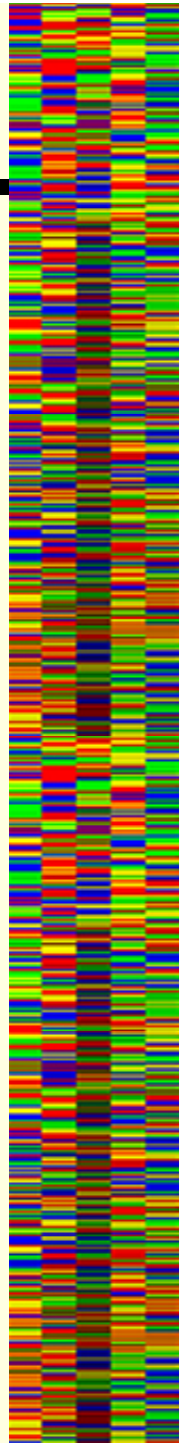
- **Approaches to constructing trees**
  - **Distance methods- Minimize difference between the realized tree and measured distances**
  - **Parsimony- Minimize the number of mutations that must be inferred**
  - **Maximum likelihood- Calculate the highest probability tree**
- **Many of these methods can also be used for other kinds of data, such as morphological characters, DNA hybridization, immunological measures, restriction sites, electrophoretic mobility, etc.**
- **Trees are built from multiple sequence alignments – multiple sequence alignments are constructed using trees**



Multiple Alignment and Trees

Distance Methods

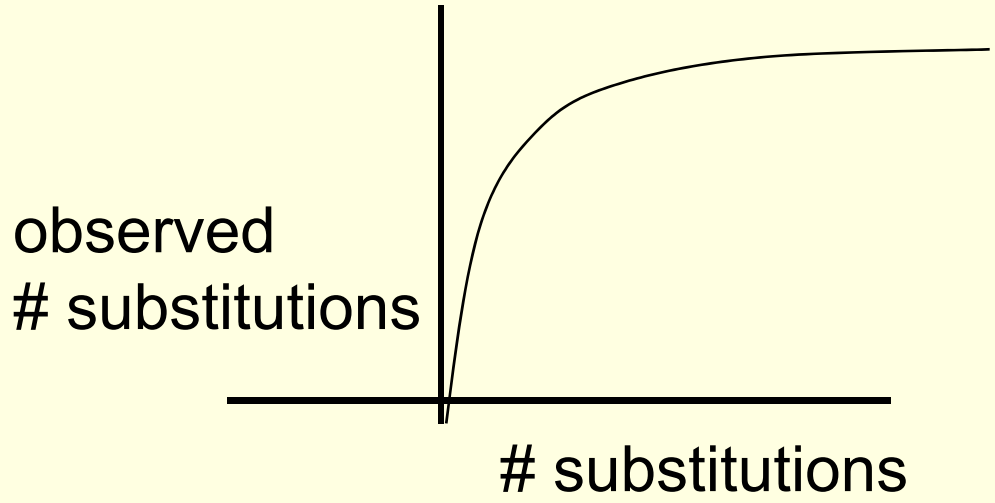
- **Require a matrix of pairwise distances. These can be distances based on alignments, or physical measurements:**
  - **Alignments**
  - **Hybridization**
  - **Complement fixation ...**
- **Try to find a tree so that the measured distances along the branches of the tree (realized tree) agree with the pairwise distance data. This is generally impossible for more than three taxa.**
- **Distance methods implicitly assume a molecular clock - that all mutations are neutral and therefore they happen at a random clocklike rate. This assumption is clearly not true.**



Multiple Alignment and Trees

Distance methods - Calculating distances

- **Distances must be corrected for multiple changes and bias**
- **Calculating distance matrix**
- **When species/sequences are closely related, one can count mutational changes**
- **As species diverge, multiple substitutions occur in the same position and the number of changes is *underestimated* by simple counting**

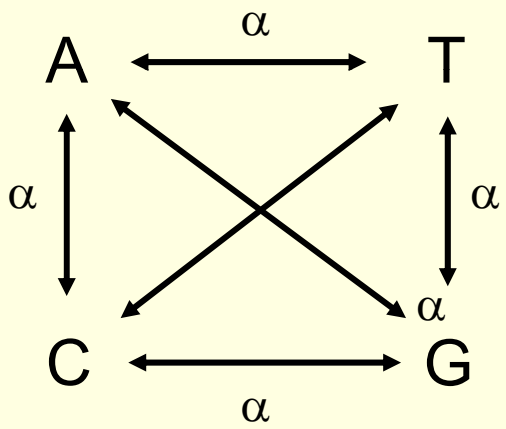


Multiple Alignment and Trees

Distance Methods - Calculating distances

- **Jukes-Cantor model**
  - **One parameter, all changes are equal**
    $P_{ij(t)} = 1/4 + (3/4)e^{-4\alpha t}$
    $P_{ij(t)} = 1/4 - (1/4)e^{-4\alpha t}$
  - $K = -3/4 \ln(1-4/3p)$

    K = substitutions per site
    p = proportion of differing bases

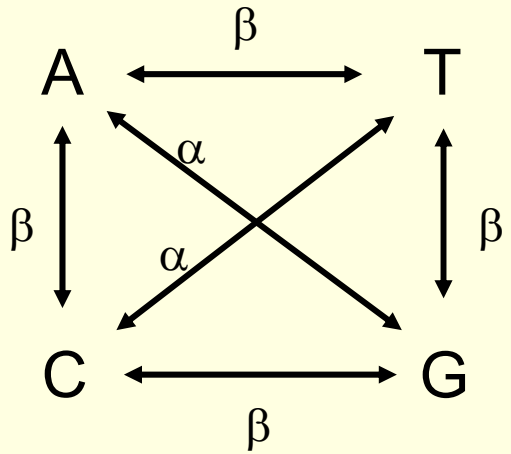


Multiple Alignment and Trees

Distance Methods - Calculating distances

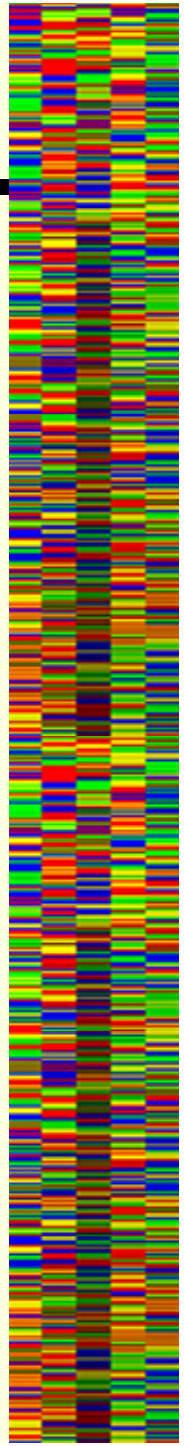
- **Kimura, 2 parameter model**
  - **transitions (A ⇔ G, C ⇔ T) are more common than transversions**
  - ***D = 2αt + 4βt***
    K = 1/2 ln a + 1/4 ln b

K = substitutions/site
a = 1/(1-2P-Q)
b = 1/(1-2Q)
P = proportion of transitions
Q = proportion of transversions



Distance Methods - Calculating distances

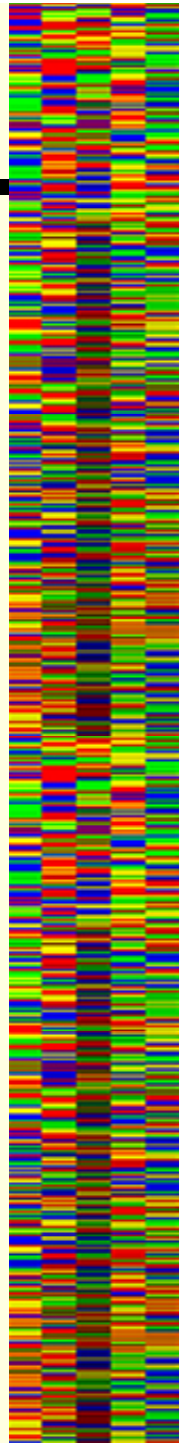
- **Proteins**
  - *Dayhoff method was developed to provide these distances*
  - *1 PAM is a distance unit*
  - *Kimura protein model Kaa = -ln(1 - P - P2/5)*
    K = substitutions per site
    P = observed proportion of differences per site



Multiple Alignment and Trees

Distance Methods – UPGMA

- **UPGMA - Unweighted Pair Group Method of Averages**
- **Assumes a clocklike distance measure**
  - **Nucleotide or amino acid substitutions, corrected for multiple changes**
  - **Other distances – DNA hybridization, immunological, etc**
- **Even though it is one of the oldest methods, it often gives good results and is still widely used today**
  - **Performs relatively well even with high errors in distance measurements**
  - **Performs poorly when evolutionary rates vary greatly between branches (long branches attract)**
- **Alternates between finding closest distance and updating distance matrix until all OTUs are joined into 1**



Multiple Alignment and Trees

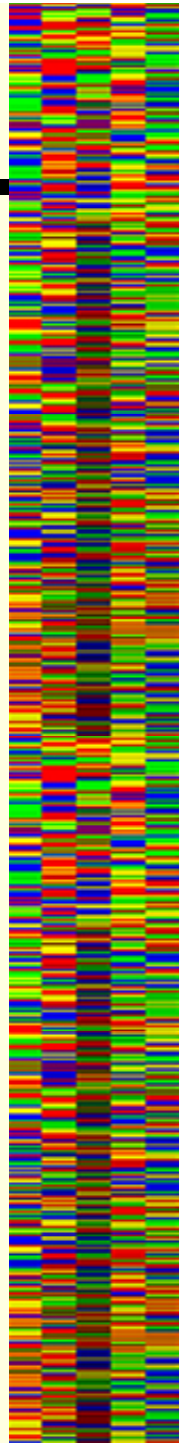
Distance Methods

- **Distance matrix is calculated from multiple alignment**

```
16 aaf79514  ~~~~~~~~~~ ~~MVIKG.MR VGKYELGRTL GEGNSAKVKF AIDTLT.GES
22 bab02040  ~~~MVRRQEE EKKAEKG.MR LGKYELGRTL GEGNFGKVKF AKDTVS.GHS
37 bab08799  ~~~~~~~~~~ ~~MGLFGTKK IGKYEIGRTI GEGNFAKVKL GYDTTN.GTY
23 bab02091  ESLPQPQNQS SPATTPAKIL LGKYELGRRL GSGSFAKVHL ARSIES.DEL
24   c71408  ES.PYPK... SPEKITGTVL LGKYELGRRL GSGSFAKVHV ARSIST.GEL
```

- **Distance between sequences is sum of distances at each aligned position.**
  - **Treatment of gaps is a problem**

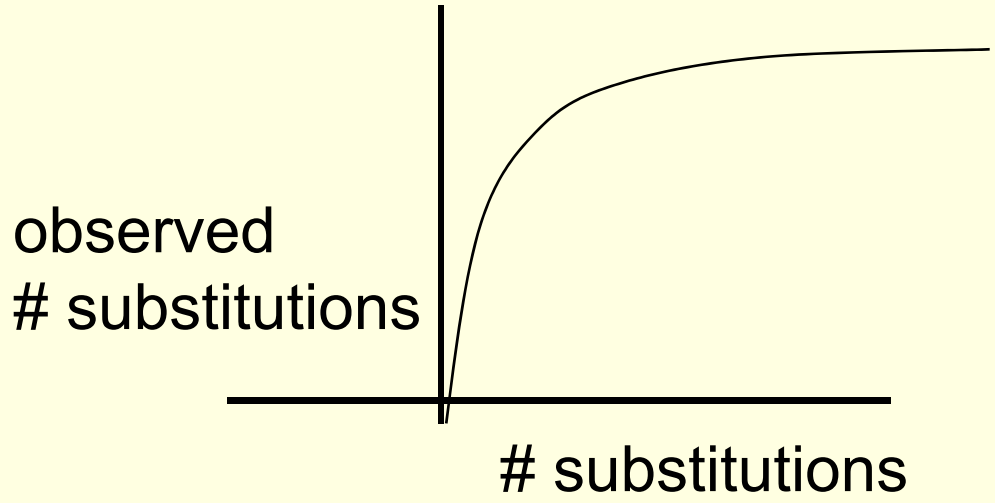
- **Adjust for multiple mutations using**
  - **Kimura formula**
  - **PAM table**
    - *Use distances from PAM scoring tables*
    - *Convert to numbers of changes*



Multiple Alignment and Trees

Distance methods - Calculating distances

- **Distances must be corrected for multiple changes and bias**
- **Calculating distance matrix**
- **When species/sequences are closely related, one can count mutational changes**
- **As species diverge, multiple substitutions occur in the same position and the number of changes is *underestimated* by simple counting**



Multiple Alignment and Trees

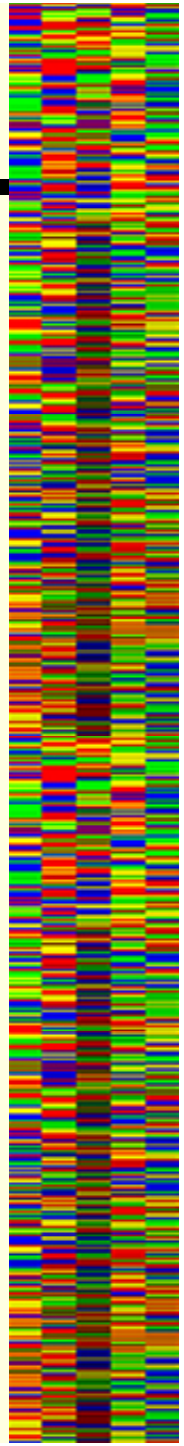
Distance Methods – UPGMA

- **Unweighted Pair-Group Method of Means**

- **Step 1 - find the two closest OTUs**

	Human	Chimp	Gorilla	Orang
Human	0	88	103	160
Chimp		0	106	170
Gorilla			0	166

- **Procedure begins with the two closest taxa, in this case Human and Chimp.**
- **These taxa are joined into an OTU with a branch length of 88/2 = 44**



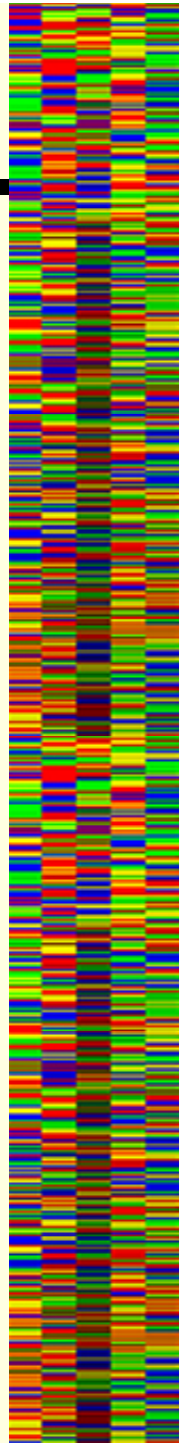
Multiple Alignment and Trees

Distance Methods – UPGMA

- **Step 2 - join closest OTUs, recalculate distance matrix**
- **After joining the taxa, the distance values are replaced in the table with their average**

	H/C	Gorilla	Orang
H/C	0	104.5	165
Gorilla		0	166

- **The next closest taxa (OTUs) are then chosen, in this case the Human/Chimp and Gorilla**

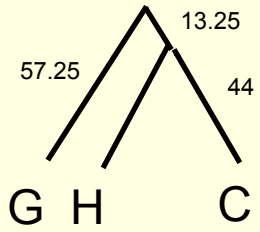


Multiple Alignment and Trees

Distance Methods – UPGMA

- **Step 1 - find next two closest OTUs**

  Again, the distance to the branch point is half of the distance between the OTUs (57.25)



- **Step 2 - Once again, the values in the distance matrix for the combined taxa are averaged**

	H/C/G	Orang
H/C/G	**0**	**165.5**

Distance Methods – UPGMA

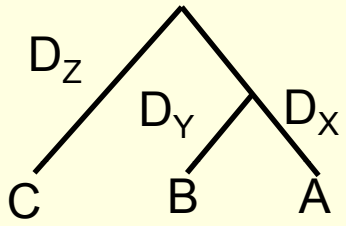
- Continue joining closest OTUs and averaging until all OTUs are joined
  - lower triangle = original distances
  - upper triangle = realized tree distances
- Note that realized distances are symmetric

	Human	*Chimp*	*Gorilla*	*Orang*
Human		*88*	*104.5*	*165.5*
Chimp	*88*		*104.5*	*165.5*
Gorilla	*103*	*106*		*165.5*
Orang	*160*	*170*	*166*	

Multiple Alignment and Trees

Distance methods - Fitch & Margoliash

- **More accurate calculation of branch lengths. Consider a simple tree**



It is easy to see that one can estimate the distance between taxon A and its immediate ancestor, D_X

$D_X = (D_{AB} + D_{AC} - D_{BC}) / 2$ **similarly,**
$D_Y = (D_{AB} + D_{BC} - D_{AC}) / 2$
$D_Z = (D_{AB} + D_{BC} - D_{AB}) / 2$

Multiple Alignment and Trees

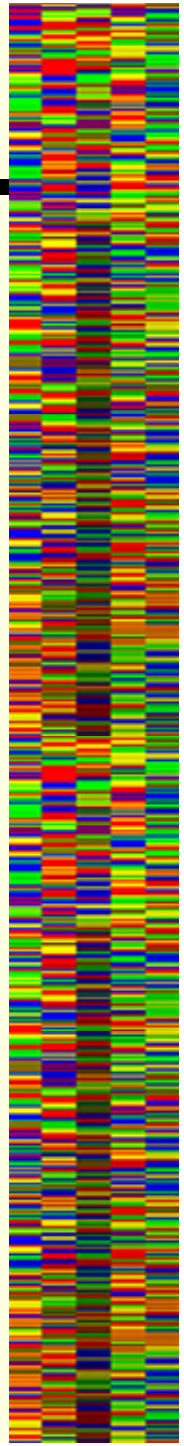
Distance Methods – UPGMA

- **How do we evaluate the fit of the realized tree to the data?**
- **One simple method is two take the sum of the squares of the differences between the measured distances and those from the tree**

  $Quality = S_{ij} (D_{ij}-d_{ij})^2$
  D_{ij} = *measured distance*
  d_{ij} = *tree distance*

- **For the UPGMA tree shown, counting only unique distances we get**

  $Q = (160-165.5)^2 + (170-165.5)^2 + (166-160.5)^2 + (103-104.5)^2 + (103-104.5)^2 + (88-88)^2$
  $Q = 85.25$

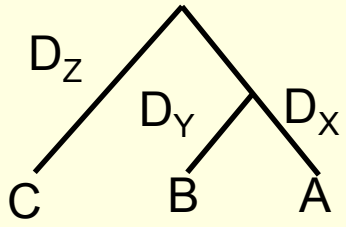


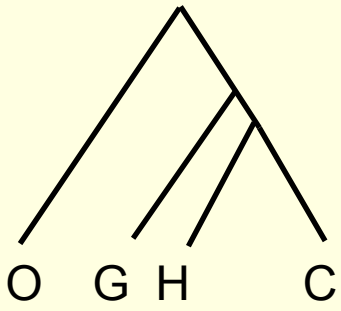
Multiple Alignment and Trees

Distance methods - Fitch & Margoliash

- **When there are more than three taxa, the third OTU is simply the average of the distances that share a common branch. Here Orangutan (O) and Gorilla (G) both contribute to D_x, and can be average together in place of C.**

$$D_X = (D_{AB} + D_{AC} - D_{BC}) / 2$$
$$D_X = [D_{HC} + (D_{OC} + D_{GC})/2 - (D_{OH} + D_{GH})/2] / 2$$

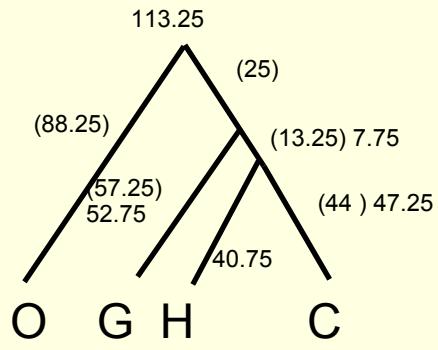




Multiple Alignment and Trees

Distance methods - Fitch & Margoliash

- **For the Human/Chimp/Gorilla/Orangutan tree shown before, we get the following distances**



	Human	*Chimp*	*Gorilla*	*Orang*
Human		***(88)*** ***88***	***(104.5)*** ***101.25***	***(165.5)*** ***161.75***
Chimp	***88***		***(104.5)*** ***107.75***	***(165.5)*** ***168.25***
Gorilla	***103***	***106***		***(165.5)*** ***166***
Orang	***160***	***170***	***166***	

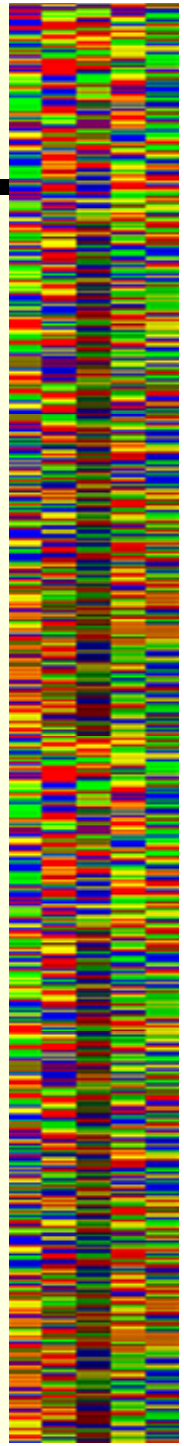
Multiple Alignment and Trees

Distance methods – Fitch-Margoliash

- **Least Squares**

  $Q = (160-161.75)^2 + (170-168.25)^2 + (166-166)^2 + (103-101.25)^2 + (106-107.75)^2 + (88-88)^2$

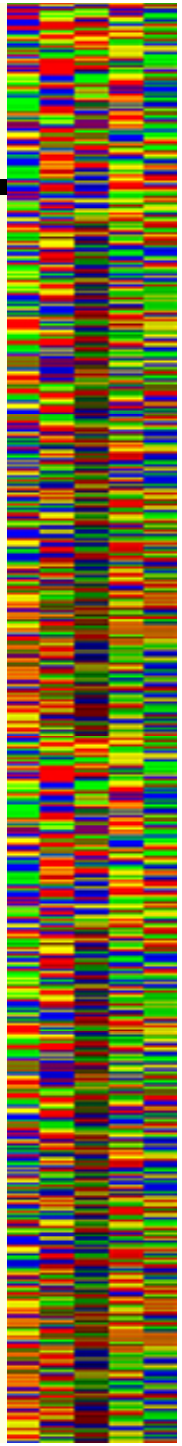
  $Q = 12.25$



Multiple Alignment and Trees

Distance Methods - Neighbor Joining

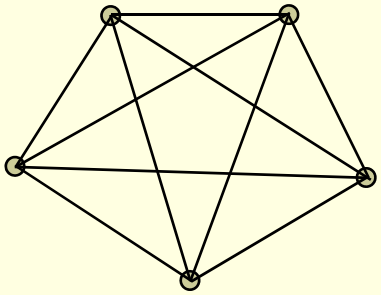
- **Works on pairs of taxa (OTUs), trying to find the pair that are closer to each other than to all other taxa**
- **Method is similar to Fitch-Margoliash method for trees/branch lengths**
- **Produces single unique tree**

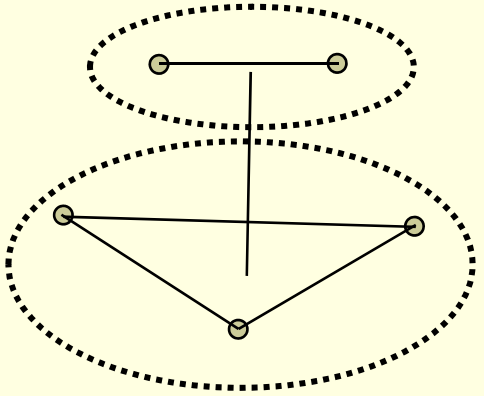


Multiple Alignment and Trees

Distance Methods – Neighbor Joining

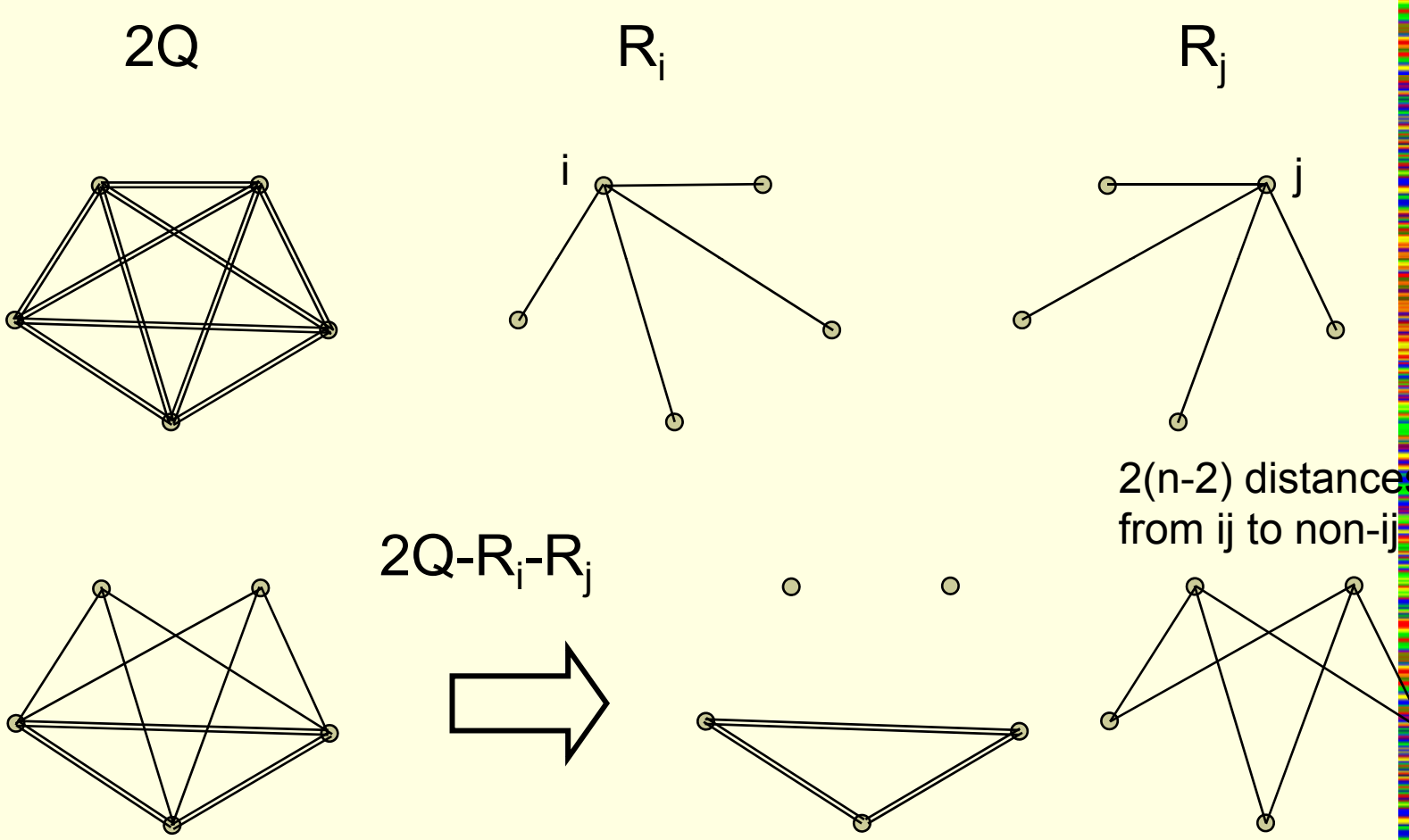
- **Consider a group of taxa and all of the distances between them.**
- **We want to find the two taxa that are closer to each other than to anyone else, I.e., two split off two taxa so that the the two groups are both as compact as possible**
- **Use all distances as in Fitch-Margoliash**





Multiple Alignment and Trees

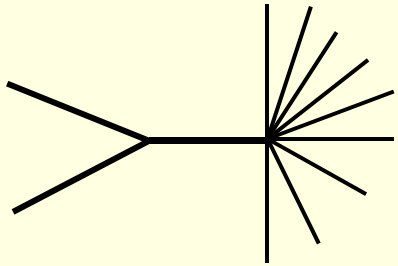
Neighbor Joining



Multiple Alignment and Trees

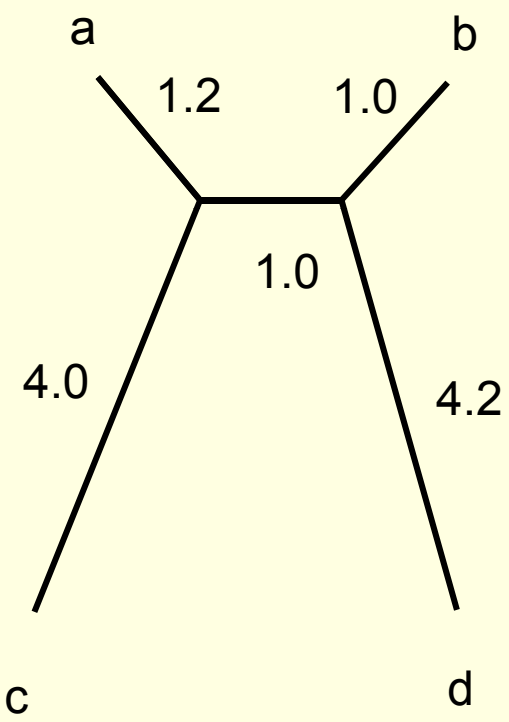
Distance methods - Neighbor Joining

- **Find pair of sequences i,j that minimize S**
  $S_{ij} = D_{ij}/2 + [2Q - R_i - R_j] / 2(n-2)$
  where
  $Q = \Sigma_{ij} D_{ij}$ $R_j = \Sigma_i D_{ij}$ $R_i = \Sigma_j D_{ij}$
- **Replace distances in matrix by average values**
- **Iterate as in UPGMA, finding best pair to link at each stage until all are linked.**
- **Determine branch lengths by Fitch-Margoliash procedure**



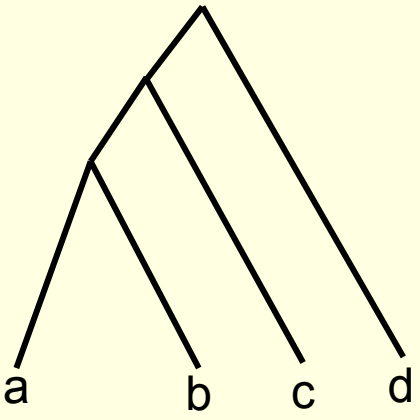
Multiple Alignment and Trees

Long Branches Attract



Distance Matrix

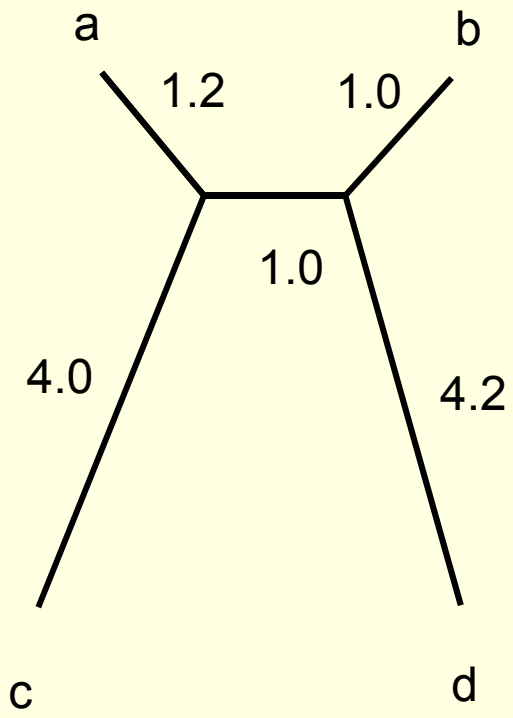
	a	b	c	d
a	-	3.2	5.2	6.4
b		-	6.0	5.2
c			-	9.2
d				-



UPGMA Tree

Multiple Alignment and Trees

Neighbor Joining and long branches



Distance Matrix

	a	b	c	d
a	-	3.2	5.2	6.4
b		-	6.0	5.2
c			-	9.2
d				-

Neighbor joining:

Q = 35.2 R_a = 14.8 R_b = 14.4
R_c = 20.4 R_d =20.8

S_{ab} = 3.2/2 + (70.4-14.8-14.4)/4 = 14.5
S_{ac} = 5.2/2 + (70.4 -14.8-20.4)/4 = 11.5
S_{ad} = 6.4/2 + (70.4 -14.8-20.8)/4 = 11.7
S_{bc} = 6/2 + (70.4 -14.4-20.4)/4 = 11.9
S_{bd} = 5.2/2 + (70.4 -14.4-20.8)/4 = 11.4
S_{cd} = 9.2/2 + (70.4 -20.4-20.8)/4 = 11.9