Schedule Week 7
30 Sep – 4 Oct

- ***Monday – Maximum Likelihood, Multiple alignments***
- ***Wednesday – Multiple Alignments***
  - Ch 6.4 & 6.6 (Ch. 6.5 is optional)
- ***Friday - Multiple Alignments, Profile and HMM***
  - Ch 6.1-6.3 (Should have already read these)

Multiple Alignment and Trees

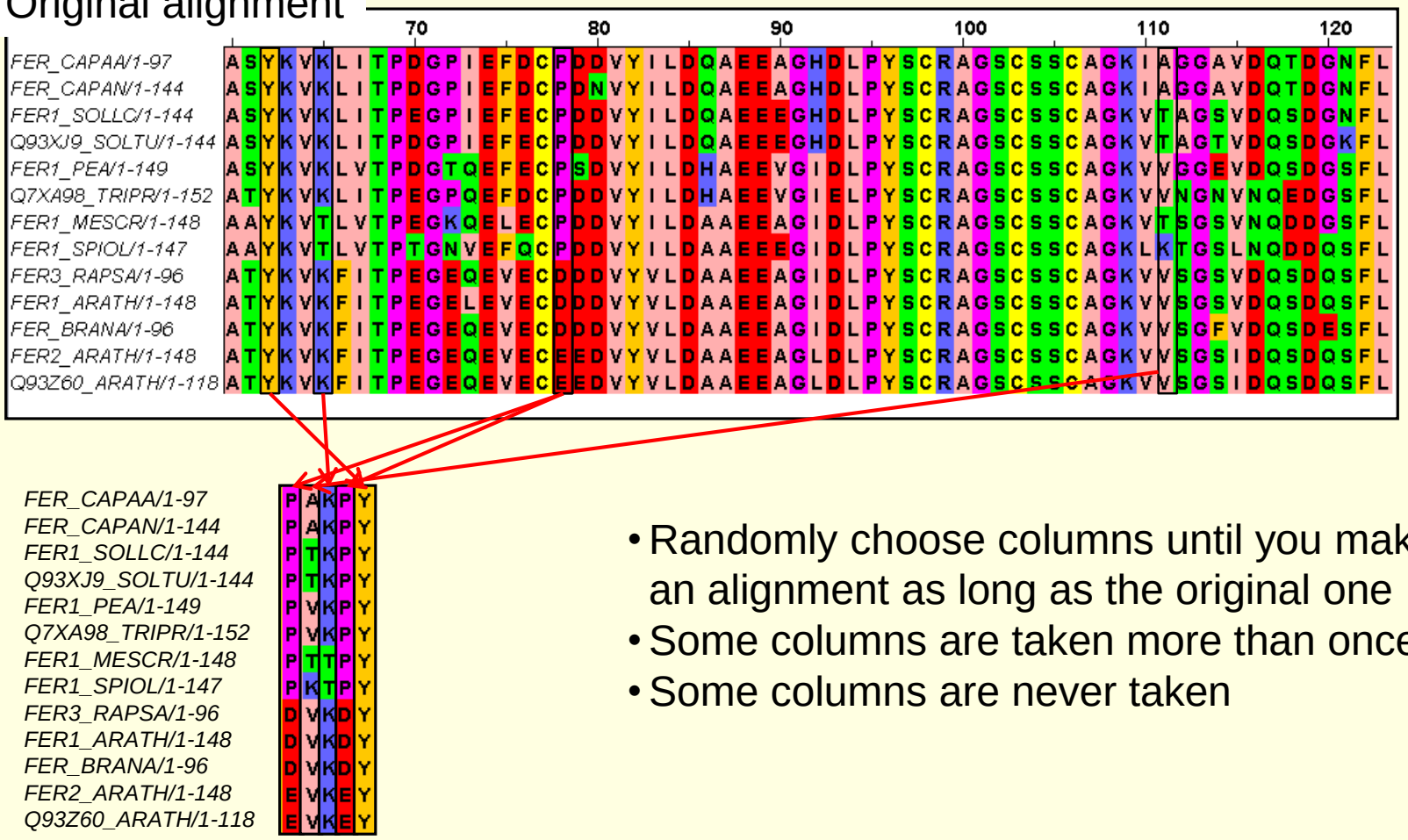
Confidence

- ***How can you tell if your tree is a "good" tree?***
- ***Bootstrap method***
- ***Each aligned column is independent from the point of view of trees.***
- ***Choose subset of positions randomly and recalculate tree***
  - Repeat many times, perhaps1000 times
  - How many times do you see the same branching pattern?
- ***Gives the confidence that the tree is consistent with the data – NOT confidence that the tree is correct***

Multiple alignments and Trees

Bootstrap resampling

Original alignment

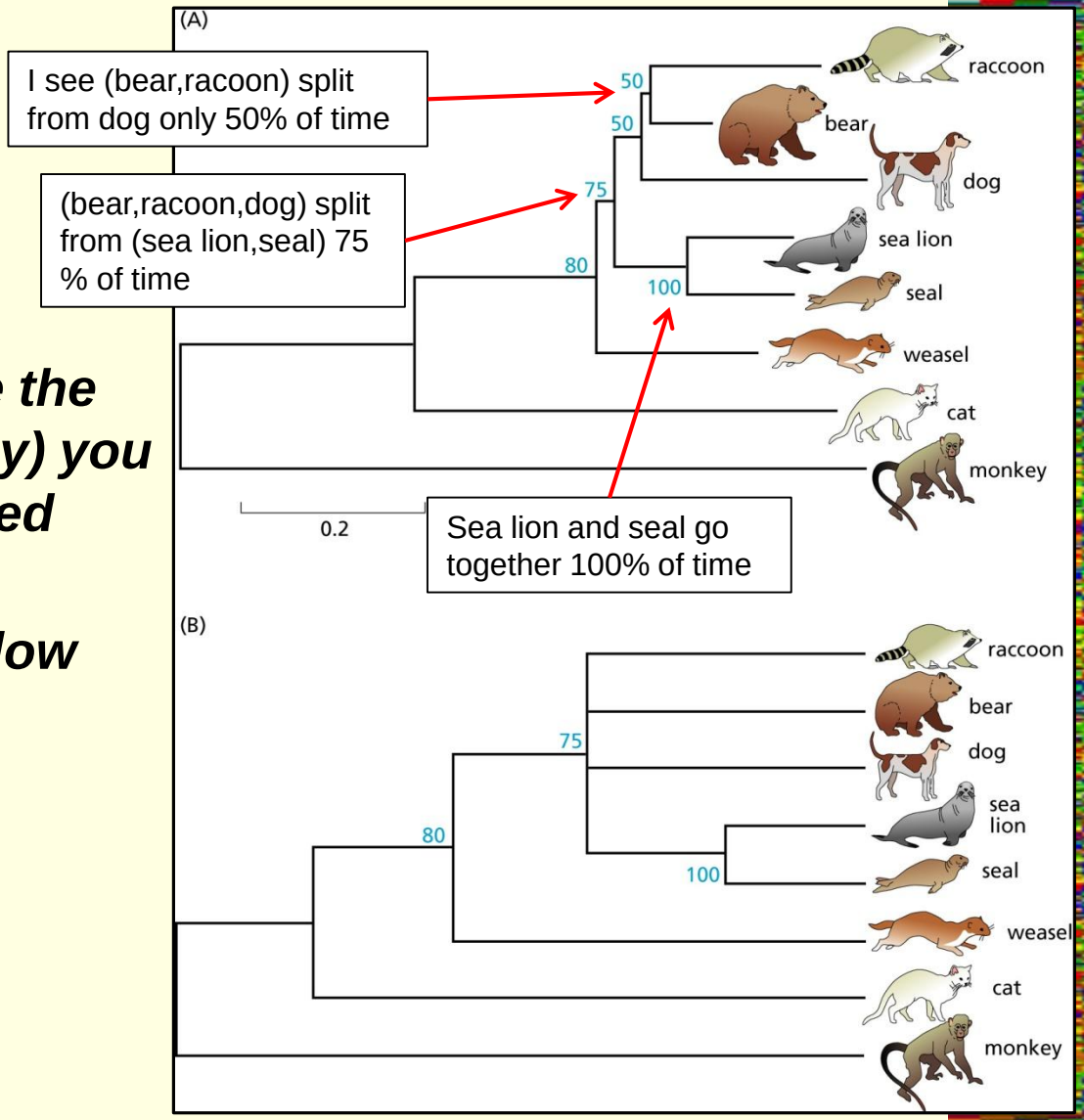


- Randomly choose columns until you make an alignment as long as the original one
- Some columns are taken more than once
- Some columns are never taken

Multiple alignments and Trees

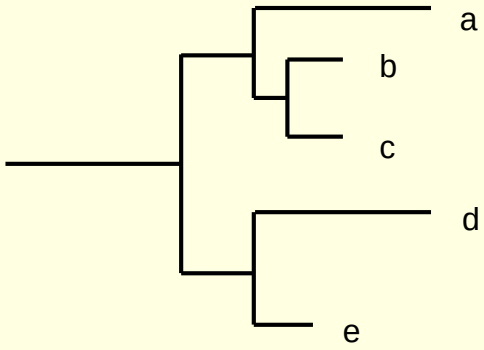
Bootstrap

- ***Make resampled sequence alignment***
- ***Make tree by same method as original***
- ***Count what percent of time the branching pattern (topology) you see is seen in the resampled trees***
- ***Merge internal nodes with low support***
  - 70% may be a good cutoff



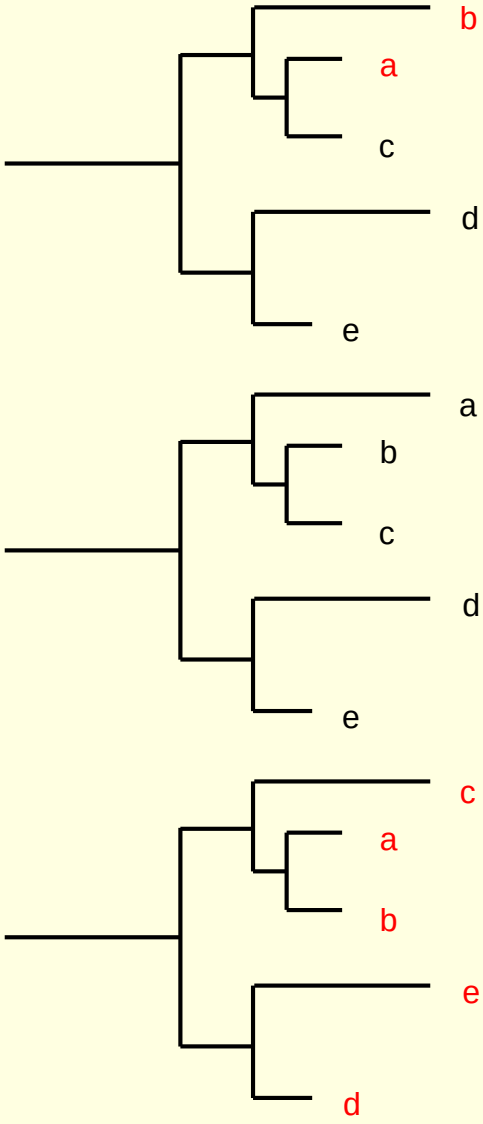
Multiple alignments and Trees

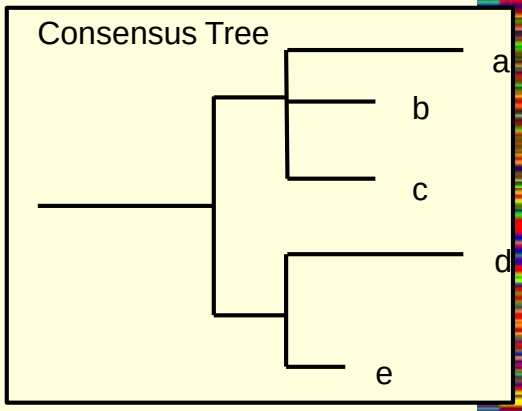
Bootstrap



Original tree

Resampled trees





Bootstrap support

b:c	33%
bc:a	33%
d:e	100%
de:abc	100%

Maximum Likelihood

- ***What is the most probability of the data given the tree***
- ***Book lists 3 ML methods (pg 251)***
  - Maximum likelihood
  - Quartet puzzling
  - Bayesian
  - The overall likelihood (probability of the data given the tree) is the product of the likelihoods at the individual sites, i

$$L = Prob(D|T) = \prod_{i=1}^{n} Prob(D_i|T)$$

Multiple Alignment and Trees

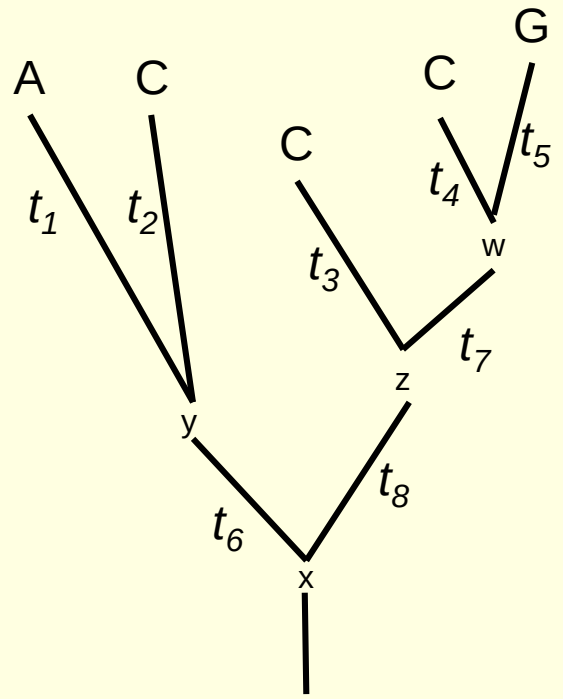
Maximum Likelihood (see eq 8.51-52 in book)

- Likelihood of the tree is the sum over all possible nucleotides that may have existed at the interior nodes, and the probabilities for each possible scenario of mutational events

$$Prob(D_i|T) = \sum_x \sum_y \sum_z \sum_w Prob(A, C, C, C, G, x, y, z, w|T)$$

- Evolution is independent on each branch, so this can be rewritten as the product of terms for each branch

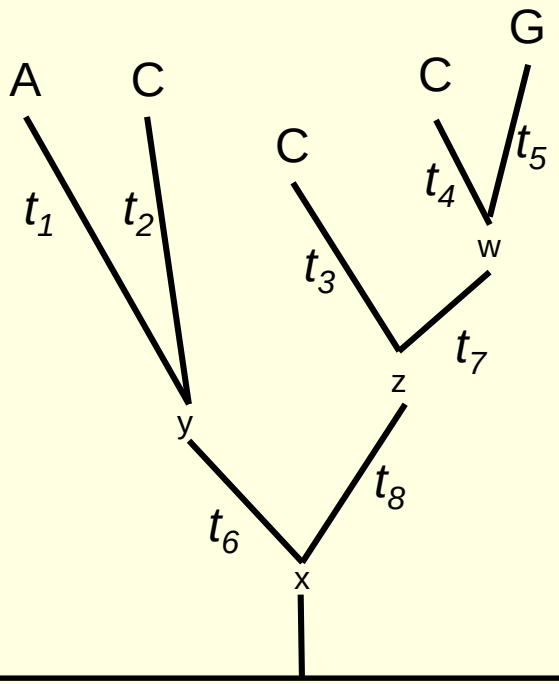
$$Prob(A, C, C, C, G, x, y, z, w|T) =$$
$$Prob(x)\ Prob(y|x, t_6)\ Prob(A|y, t_1)\ Prob(C|y, t_2)$$
$$Prob(z|x, t_8)\ Prob(C|z, t_3)$$
$$Prob(w|z, t_7)\ Prob(C|wy, t_4)\ Prob(G|w, t_5)$$



Multiple Alignment and Trees

Maximum Likelihood

- ***If evolution has been going on a long time, you can assume that Prob(x) is the random probability of each base***
- ***Other probabilities come from base substitution model***
- ***Given four possible bases at each of the four internal nodes, there are 4^4 = 256 terms to calculate.***
- ***In general, for n taxa there are n-1 interior nodes and therefore 4^{n-1} terms to calculate***
  - *n*=10 262,144 terms
  - *n*=20 274,877,906,944 terms
- ***There is a recursive way to calculate this that reduces the number of calculations, but it is still a large number***
- ***Calculate for each position in the sequences***
- ***Must also consider different rates along each branch (t_1 – t_8)***



$$Prob(A,C,C,C,G,x,y,z,w|T) =$$
$$Prob(x)\ Prob(y|x,t_6)\ Prob(A|y,t_1)\ Prob(C|y,t_2)$$
$$Prob(z|x,t_8)\ Prob(C|z,t_3)$$
$$Prob(w|z,t_7)\ Prob(C|wy,t_4)\ Prob(G|w,t_5)$$

Multiple Alignment and Trees

Maximum Likelihood Methods

- ***Estimate topology and branch length viewing evolution as a random process***
- ***Requires a probability model of evolution as a function of time.***
  - For DNA one can use Jukes-Cantor model (all nucleotides have same substitution rates), or Kimura model (different rates for transitions, R→R or Y →Y, and transversion, R→Y or Y →R).
  - For proteins one can use Dayhoff, but in the probability form not the log-odds form.

Multiple Alignment and Trees

Maximum Likelihood

- ***Problems***
  - The value of *t* varies with position due to conserved and unconserved regions
  - The probability of a substitution $P_{i,j}$ is also position specific
  - Evolution may not be random
- ***ML is often used to compare several trees and choose the best one using a likelihood ratio test***

Multiple Alignment and Trees

Take home messages

- *The number of tree topologies grows factorially with the number of taxa - it is generally impossible to examine all tree topologies*
- *Trees based on molecular sequences are much more straightforward to calculate, and more reliable than those based on morphological characters*
- *Calculating real divergence times and accurate branch lengths depends on mutations acting like a molecular clock. In turn, the molecular clock assumption is only appropriate when looking at neutral mutations (Kimura's hypothesis)*
- *Often analyses will focus on apparently neutral differences such as synonymous codon changes or third position of codon changes in order to get the most clock-like data*

Multiple Alignment and Trees

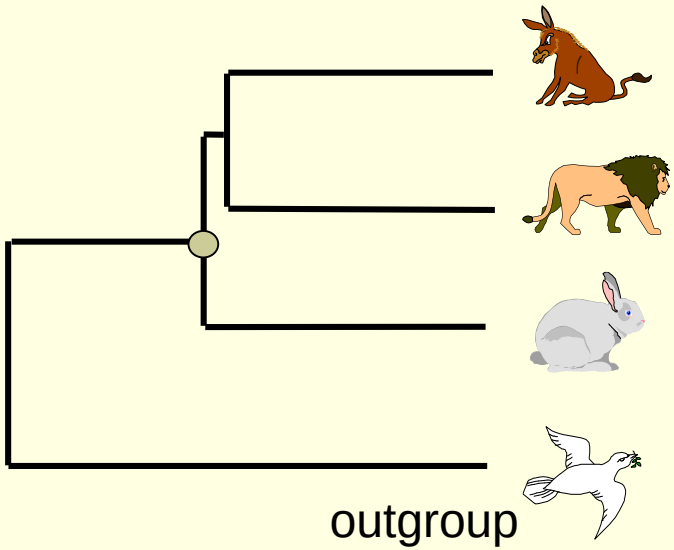
Rooting trees

- ***Rooting the tree lets you unambiguously decide what is ancestral and what is derived***
- ***Most trees are implicitly unrooted. That is, you can't tell from the data used to construct the tree where the ancestral node lies. You must have additional data to find the root of a tree.***
- ***Most common procedure is to use an outgroup, i.e. a taxon that is guaranteed to be more distant from all of the taxa of interest than any of them are from each other.***
  - for best result, the outgroup should be as close as possible to the taxa of interest while still clearly outside them
  - Orangutan can be used as outgroup for human, chimp, gorilla
  - Alligator can be used as outgroup for human, rat, dog, cow, horse
- ***Midpoint between farthest pair can be used as root (assumes a clock)***

There are two major ways to root trees:

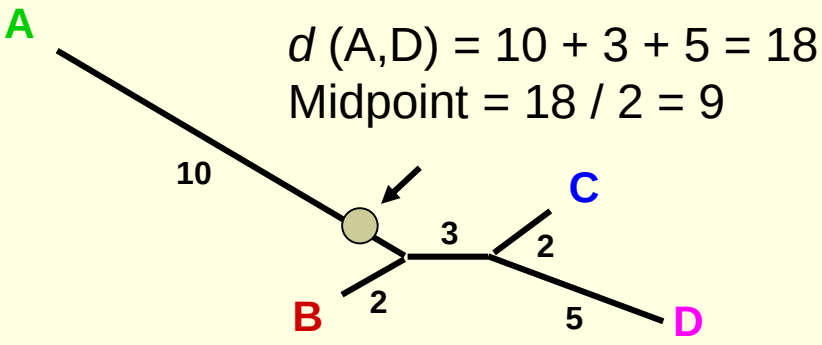
By outgroup:

Uses taxa (the "outgroup") that are known to fall outside of the group of interest (the "ingroup"). Requires some prior knowledge about the relationships among the taxa. The outgroup can either be species (*e.g.*, birds to root a mammalian tree) or previous gene duplicates (*e.g.*, α-globins to root β-globins).



By midpoint or distance:

Roots the tree at the midway point between the two most distant taxa in the tree, as determined by branch lengths. Assumes that the taxa are evolving in a clock-like manner. This assumption is built into some of the distance-based tree building methods.



d (A,D) = 10 + 3 + 5 = 18
Midpoint = 18 / 2 = 9

Multiple Alignments and Trees

Molecular Clock

- ***What are the most accurate clocks?***
- ***What is the definition of a second?***
  - Atomic Clocks are extremely accurate (0.1 nanosecond/day)
    - Based on measuring random transitions in electron energy levels
    - SI definition of a second is the duration of 9,192,631,770 cycles of transition between two energy levels of the ^{133}cesium atom.
  - Molecular clock
    - Random event is mutation
    - Assumes mutational process is constant
    - Unfortunately we cannot observe mutations, only mutations that become FIXED in a genome

Multiple Alignment and Trees

Molecular Clocks

Some trees assume or try to identify a clock – when did these species diverge?

- ***Most of our clocks are deterministic, they make "ticks" at precise intervals***
- ***A stochastic clock is probabilistic, it makes ticks at a certain probability in each unit of time. Ticks may not be evenly spaced.***
- ***Stochastic clocks are not necessarily inaccurate – atomic clocks are stochastic clocks, however for them to be accurate you must have a single process underlying the "ticking", and a large number of "ticks"***

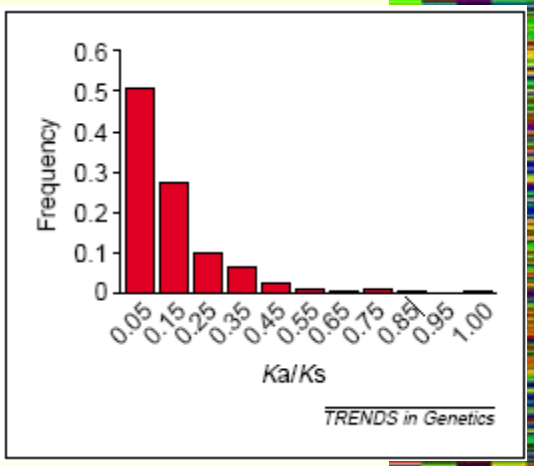
Multiple Alignment and Trees

Neutral Mutations

- ***Neutral mutations - Most mutations are neither highly advantageous or deleterious - they are effectively neutral (Kimura theory).***
- ***Neutral mutations should be the most clocklike because the represent the random accumulation of changes over time.***
  - Very common attempt to find neutral mutations is to use the third codon position (usually synonymous) or only DNA sequences where the encoded residue does not change
- ***Correcting distances - distances are often corrected for multiple mutational events so that they have a linear relationship to time.***
- ***One of the reasons behind the original formulation of the Dayhoff mutational distance matrix was to provide a mapping from amino acid residue changes to a clock – 1 PAM is a time unit in this sense.***

Multiple Alignments and Trees

Identifying selection - K_a/K_s ratio

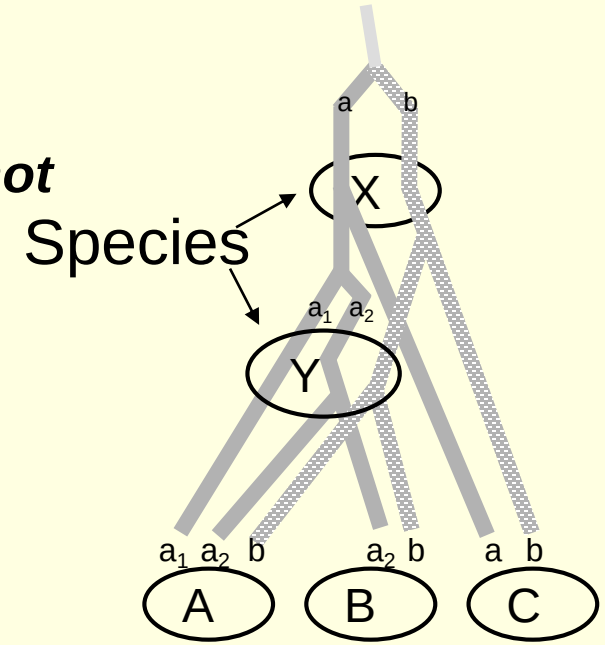


- ***Ratio of non-synonymous (K_a, changed amino acid) to synonymous (K_s, unchanged amino acid) changes in coding gene DNA***
- ***K_s measures clocklike behavior***
- ***If K_a is much larger, it is changing in a not clock-like way – it is being selected***
  - Has to be adjusted for multiple hits if time period is long
  - Has to be adjusted for 2, 4, and 6 codon families
- ***Usually $K_a << K_s$ (most non-synonomous changes are bad)***
- ***$K_a/K_s > 1$ is usually taken to indicate positive selection***

Multiple Alignment and Trees

Gene Trees vs Species Trees

- *Genes usually diverge before species, i.e., gene trees overestimate distance to common ancestor*
- *In speciation, two species would be expected to get different populations of alleles*
- *Alleles are copies of genes that are already diverging in a species*
- *species X has a and b*
- *species Y has a_1, a_2, and b*
- *if some alleles are lost, A or B may not have all three*



Multiple Alignment and Trees

Orthology and Paralogy

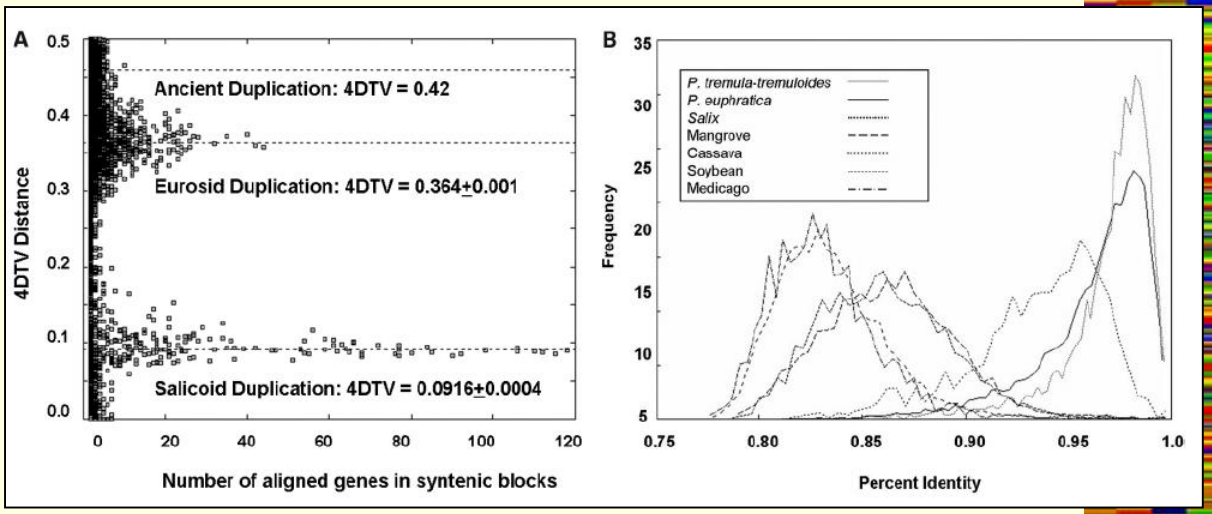
- ***Tree construction seeks to understand the evolutionary relationship between taxa (groups of organisms).***
- ***In sequence based methods, one must distinguish between gene trees and species trees due to the presence of orthologous and paralogous genes***
  - Orthologous - homologous genes that are truly the same, e.g. myoglobin in sharks and humans (arise from speciation).
  - Paralogous - homologous genes that resulted from a gene duplication, e.g. hemoglobin and myoglobin.
  - Xenologous - horizontally transferred genes
- ***These concepts are often used, perhaps overused, in making inferences about gene function – i.e., that orthologous genes have the same function and paralogous genes have different functions. This is an oversimplification at best due to continuous creation of paralogous genes by duplication and stochastic loss of genes by deletion.***

Multiple Alignment and Trees

Timing gene duplications

4DTV distance - fourfold synonymous third-codon transversion

- Identify possibly duplicated genes (BLAST)
- Calculate 4DTV based on alignments
- third codon position is not usually selected (should be clocklike)
- Find pairs of genes with same distances



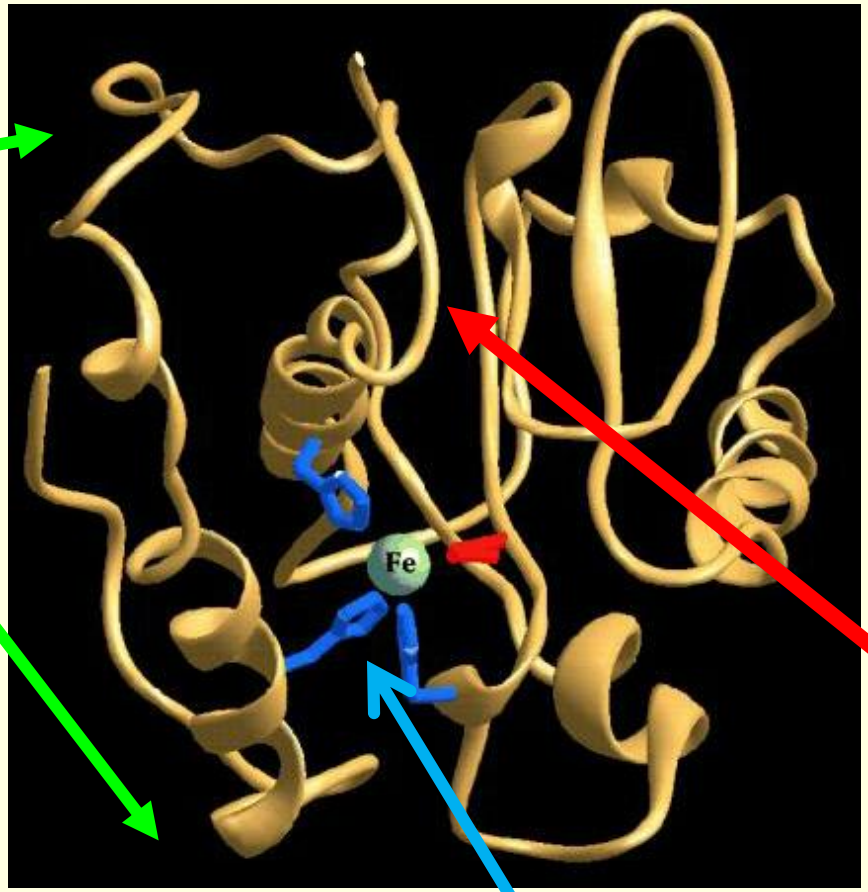
Multiple Alignments and Trees

Why make multiple alignments?

- ***Find evolutionarily constrained regions of proteins***
  - Structural cores
  - Active sites
  - Binding surfaces
- ***Understand evolution of proteins***
- ***Find origins/causes of disease***
- ***Understand evolution of species (phylogenetics)***
  - DNA used more often for species comparison

Multiple Alignment and Trees

Multiple alignment and protein structure



Indels variable

No indels conserved

Special function highly conserved

Multiple Alignments and Trees

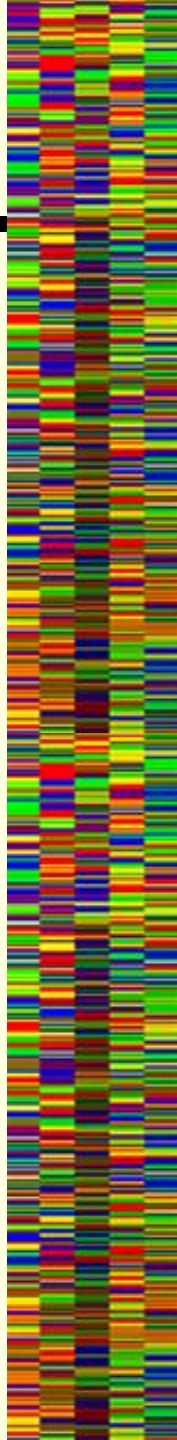
Goals

- *what is a good alignment?*
- *how do you find it?*

Multiple Alignment and Trees

Kinds of Multiple alignments

- ***Leader-follower alignments***
  - All sequences aligned to one *leader* (by pairwise DP)
  - Optional format from BLAST programs
- ***N-dimensional multiple alignments***
  - Simultaneous DP alignment in N dimensions
  - Rarely used due to poor scaling
- ***Progressive alignment***
  - Sequential alignment of single sequences to growing alignment
  - Most commonly used – probably best quality
- ***Profile alignment***
  - Alignment of each sequence to profile or HMM



Multiple Alignment and Trees

Pairwise alignments to leader (Human HBB)

```
Human HBB   1  MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK  60
Rana HBB    1  MVHWTAEEKAVINSVWQKVDVEQDGHEALTRLFIVYPWTQRYFSTFGDLSSPAAIAGNPK  60

Human HBB 61   VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG  120
Rana HBB  61   VHAHGKKILGAIDNAIHNLDDVKGTLHDLSEEHANELHVDPENFRRLGEVLIVVLGAKLG  120

Human HBB 121  KEFTPPVQAAYQKVVAGVANALAHKYH  147
Rana HBB  121  KAFSPQVQHVWEKFIAVLVDALSHSYH  147
```

```
Human HBB 4    LTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV  61
Human HBA 3    LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF------DLSHGSAQV  56

Human HBB 62   KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK  121
Human HBA 57   KGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPA  116

Human HBB 122  EFTPPVQAAYQKVVAGVANALAHKY  146
Human HBA 117  EFTPAVHASLDKFLASVSTVLTSKY  141
```

```
Human HBB 4    LTPEEKSAVTALWGKVNVDEVG--GEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV  61
Bovine MY 3    LSDGEWQLVLNAWGKVEADVAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASEDL  62

Human HBB 62   KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK  121
Bovine MY 63   KKHGNTVLTALGGILKKKGHHEAEVKHLAESHANKHKIPVKYLEFISDAIIHVLHAKHPS  122

Human HBB 122  EFTPPVQAAYQKVVAGVANALAHKY  146
Bovine MY 123  DFGADAQAAMSKALELFRNDMAAQY  147
```

Multiple Alignment and Trees

Pairwise alignments to leader (Human HBB)

```
Human HBB 4     LTPEEKSAVTALWG--KVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV  61
BrNapa HB 6     FTEKQEALVKESWEILKQDIPKYSLHFFSQILEIAPAAKDMFSFLRD--TDEVPHNNPKL  63

Human HBB 62    KAHGKKVLGAFSDGLAHLDNLKG-------TFATLSELHCDKLHVDPENFRLLGNVLVCV  114
BrNapa HB 64    KAHAVKVFKMTCETAIQLRE-KGKVVVADTTLQYLGSVHFKSGVLDP-HFEVVKEALVRT  121

Human HBB 115   LAHHFGKEFTPPVQAAYQKVVAGVANAL  142
BrNapa HB 122   LKEGLGEKYNEEVEGAWSKAYDHLALAI  149
```

```
Human HBB 1     MVHLTPEEKSAVTALWG--KVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN  58
CANLI LB  1     MGAFSEKQESLVKSSWEAFKQNVPHHSAVFYTLILEKAPAAQNMFSFLSNGVDPN----N  56

Human HBB 59    PKVKAHGKKVLGAFSDGLAHLDNLKGTFA----TLSELHCDKLHVDPENFRLLGNVLVCV  114
CANLI LB  57    PKLKAHAEKVFKMTVDSAVQL-RAKGEVVLADPTLGSVHVQKGVLDP-HFLVVKEALLKT  114

Human HBB 115   LAHHFGKEFTPPV----QAAYQKVVAGVANALA  143
CANLI LB  115   FKEAVGDKWNDELGNAWEVAYDELAAAIKKAMG  147
```

Multiple Alignment and Trees

Alignments to Leader

```
Human HBB    1  MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK  60
Rana HBB     1  MVHWTAEEKAVINSVWQKVDVEQDGHEALTRLFIVYPWTQRYFSTFGDLSSPAAIAGNPK  60
Human HBB    4     LTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV  61
Human HBA    3     LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF------DLSHGSAQV  56
Human HBB    4     LTPEEKSAVTALWGKVNVDEVG--GEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV  61
Bovine MY    3     LSDGEWQLVLNAWGKVEADVAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASEDL  62
Human HBB    4     LTPEEKSAVTALWG--KVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV  61
BrNapa HB    6     FTEKQEALVKESWEILKQDIPKYSLHFFSQILEIAPAAKDMFSFLRD--TDEVPHNNPKL  63
Human HBB    1  MVHLTPEEKSAVTALWG--KVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN  58
CanLi LB     1  MGAFSEKQESLVKSSWEAFKQNVPHHSAVFYTLILEKAPAAQNMFSFLSNGVDPN----N  56

Human HBB   61    VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG  120
Rana HBB    61    VHAHGKKILGAIDNAIHNLDDVKGTLHDLSEEHANELHVDPENFRRLGEVLIVVLGAKLG  120
Human HBB   62     KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK  121
Human HBA   57     KGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPA  116
Human HBB   62     KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK  121
Bovine MY   63     KKHGNTVLTALGGILKKKGHHEAEVKHLAESHANKHKIPVKYLEFISDAIIHVLHAKHPS  122
Human HBB   62     KAHGKKVLGAFSDGLAHLDNLKG-------TFATLSELHCDKLHVDPENFRLLGNVLVCV  114
BrNapa HB   64     KAHAVKVFKMTCETAIQLRE-KGKVVVADTTLQYLGSVHFKSGVLDP-HFEVVKEALVRT  121
Human HBB   59  PKVKAHGKKVLGAFSDGLAHLDNLKGTFA----TLSELHCDKLHVDPENFRLLGNVLVCV  114
CanLi LB    57  PKLKAHAEKVFKMTVDSAVQL-RAKGEVVLADPTLGSVHVQKGVLDP-HFLVVKEALLKT  114

Human HBB  121         KEFTPPVQAAYQKVVAGVANALAHKYH  147
Rana HBB   121         KAFSPQVQHVWEKFIAVLVDALSHSYH  147
Human HBB  122          EFTPPVQAAYQKVVAGVANALAHKY  146
Human HBA  117          EFTPAVHASLDKFLASVSTVLTSKY  141
Human HBB  122          EFTPPVQAAYQKVVAGVANALAHKY  146
Bovine MY  123          DFGADAQAAMSKALELFRNDMAAQY  147
Human HBB  115  LAHHFGKEFTPPVQAAYQKVVAGVANAL  142
BrNapa HB  122  LKEGLGEKYNEEVEGAWSKAYDHLALAI  149
Human HBB  115  LAHHFGKEFTPPV----QAAYQKVVAGVANALA  143
CanLi LB   115  FKEAVGDKWNDELGNAWEVAYDELAAAIKKAMG  147
```

Multiple Alignment and Trees

Propagate Gaps vs Leader

```
Human HBB    1  MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK      60
Rana HBB     1  MVHWTAEEKAVINSVWQKVDVEQDGHEALTRLFIVYPWTQRYFSTFGDLSSPAAIAGNPK      60
Human HBB    4     LTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV   61
Human HBA    3     LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF------DLSHGSAQV   56
Human HBB    4     LTPEEKSAVTALWGKVNVDEVG--GEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV   61
Bovine MY    3     LSDGEWQLVLNAWGKVEADVAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASEDL   62
Human HBB    4     LTPEEKSAVTALWG--KVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV   61
BrNapa HB    6     FTEKQEALVKESWEILKQDIPKYSLHFFSQILEIAPAAKDMFSFLRD--TDEVPHNNPKL   63
Human HBB    1  MVHLTPEEKSAVTALWG--KVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN      58
CanLi LB     1  MGAFSEKQESLVKSSWEAFKQNVPHHSAVFYTLILEKAPAAQNMFSFLSNGVDPN----N      56
```

```
Human HBB    1  MVHLTPEEKSAVTALWG++KV++NVDEVG++GEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK    60
Rana HBB     1  MVHWTAEEKAVINSVWQ++KV++DVEQDG++HEALTRLFIVYPWTQRYFSTFGDLSSPAAIAGNPK    60
Human HBB    4     LTPEEKSAVTALWG++KV--NVDEVG++GEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV   61
Human HBA    3     LSPADKTNVKAAWG++KVGAHAGEYG++AEALERMFLSFPTTKTYFPHF------DLSHGSAQV   56
Human HBB    4     LTPEEKSAVTALWG++KV++NVDEVG--GEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV   61
Bovine MY    3     LSDGEWQLVLNAWG++KV++EADVAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASEDL   62
Human HBB    4     LTPEEKSAVTALWG--KV++NVDEVG++GEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV   61
BrNapa HB    6     FTEKQEALVKESWEILKQ++DIPKYS++LHFFSQILEIAPAAKDMFSFLRD--TDEVPHNNPKL   63
Human HBB    1  MVHLTPEEKSAVTALWG--KV++NVDEVG++GEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN      58
CanLi LB     1  MGAFSEKQESLVKSSWEAFKQ++NVPHHS++AVFYTLILEKAPAAQNMFSFLSNGVDPN----N      56
```

Multiple Alignment and Trees

Remove redundant leader sequences

```
Human HBB    1  MVHLTPEEKSAVTALWG++KV++NVDEVG++GEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK  60
Rana HBB     1  MVHWTAEEKAVINSVWQ++KV++DVEQDG++HEALTRLFIVYPWTQRYFSTFGDLSSPAAIAGNPK  60
Human HBB    4     LTPEEKSAVTALWG++KV--NVDEVG++GEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV  61
Human HBA    3     LSPADKTNVKAAWG++KVGAHAGEYG++AEALERMFLSFPTTKTYFPHF------DLSHGSAQV  56
Human HBB    4     LTPEEKSAVTALWG++KV++NVDEVG--GEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV  61
Bovine MY    3     LSDGEWQLVLNAWG++KV++EADVAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASEDL  62
Human HBB    4     LTPEEKSAVTALWG--KV++NVDEVG++GEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV  61
BrNapa HB    6     FTEKQEALVKESWEILKQ++DIPKYS++LHFFSQILEIAPAAKDMFSFLRD--TDEVPHNNPKL  63
Human HBB    1  MVHLTPEEKSAVTALWG--KV++NVDEVG++GEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN     58
CanLi LB     1  MGAFSEKQESLVKSSWEAFKQ++NVPHHS++AVFYTLILEKAPAAQNMFSFLSNGVDPN----N     56
```

```
Human HBB    1  MVHLTPEEKSAVTALWG--KV--NVDEVG--GEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK  60
Rana HBB     1  MVHWTAEEKAVINSVWQ--KV--DVEQDG--HEALTRLFIVYPWTQRYFSTFGDLSSPAAIAGNPK  60
Human HBA    3     LSPADKTNVKAAWG--KVGAHAGEYG--AEALERMFLSFPTTKTYFPHF------DLSHGSAQV  56
Bovine MY    3     LSDGEWQLVLNAWG--KV--EADVAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASEDL  62
BrNapa HB    6     FTEKQEALVKESWEILKQ--DIPKYS--LHFFSQILEIAPAAKDMFSFLRD--TDEVPHNNPKL  63
CanLi LB     1  MGAFSEKQESLVKSSWEAFKQ--NVPHHS--AVFYTLILEKAPAAQNMFSFLSNGVDPN----N     56
```

Multiple alignment and Trees

Do we need all the gaps?

Multiple Alignment and Trees

Which is best?

```
Human HBB    1   --KV--NVDEVG--
Rana HBB     1   --KV--DVEQDG--
Human HBA    3   --KVGAHAGEYG--
Bovine MY    3   --KV--EADVAGHG
BrNapa HB    6   ILKQ--DIPKYS—
CanLi LB     1   AFKQ--NVPHHS--

Human HBB    1   KV--NVDEVG
Rana HBB     1   KV--DVEQDG
Human HBA    3   KVGAHAGEYG
Bovine MY    3   KVEADVAGHG
BrNapa HB    6   ILKQDIPKYS
CanLi LB     1   AFKQNVPHHS
```

Qualitative

- I'm an expert, the bottom one is better, trust me

- ***Quantitative***
  - Sum of pairs score

$$\sum_{columns} \sum_{i=1}^{n} \sum_{j=i+1}^{n} s_{ij}$$

for the cmoparison scores (s_{ij}) for letters in all pairs of sequences *i* and *j*

  - Number of mutations (as in PAM calculation)
    - is there an overcount due to relationships between the sequences?

Multiple Alignment and Trees

Which is best?

- ***Quantitative***
  - Sum of pairs score
  - Use BLOSUM or PAM
    (for example purposes, I used identities)

$$\sum_{columns} \sum_{i=1}^{n} \sum_{j=i+1}^{n} s_{ij}$$

```
Human HBB    1   --KV--NVDEVG--
Rana HBB     1   --KV--DVEQDG--
Human HBA    3   --KVGAHAGEYG--
Bovine MY    3   --KV--EADVAGHG
BrNapa HB    6   ILKQ--DIPKYS--
CanLi LB     1   AFKQ--NVPHHS--

Human HBB    1   KV--NVDEVG
Rana HBB     1   KV--DVEQDG
Human HBA    3   KVGAHAGEYG
Bovine MY    3   KVEADVAGHG
BrNapa HB    6   ILKQDIPKYS
CanLi LB     1   AFKQNVPHHS
```

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Total	Gaps
Original	0	0	15	7	0	0	2	4	2	2	2	7	0	0	41	6 (52)
De-gapped	X	X	6	6	X	X	1	2	4	6	1	2	X	X	28	2 (2)

Multiple Alignments and Trees

SOP score

- ***More gaps gives higher SOP score***
  - Lower alignment is not clearly better

```
Human HBB    1  --KV--NVDEVG--
Rana HBB     1  --KV--DVEQDG--
Human HBA    3  --KVGAHAGEYG--
Bovine MY    3  --KV--EADVAGHG
BrNapa HB    6  ILKQ--DIPKYS--
CanLi LB     1  AFKQ--NVPHHS--
```

SOP=42
gaps: 6 (52)

```
Human HBB    1  --KV--NVD---EV-G
Rana HBB     1  --KV--DV----EQDG
Human HBA    3  --KVGA---HAGE-YG
Bovine MY    3  --KVEADV--AGH--G
BrNapa HB    6  ILKQ--DIPK----YS
CanLi LB     1  AFKQ--NVPH--H--S
```

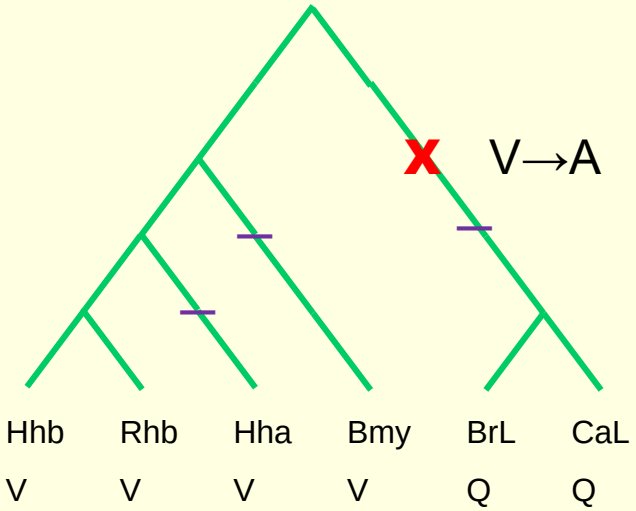
SOP=49
gaps: 13 (57)

Multiple Alignments and Trees

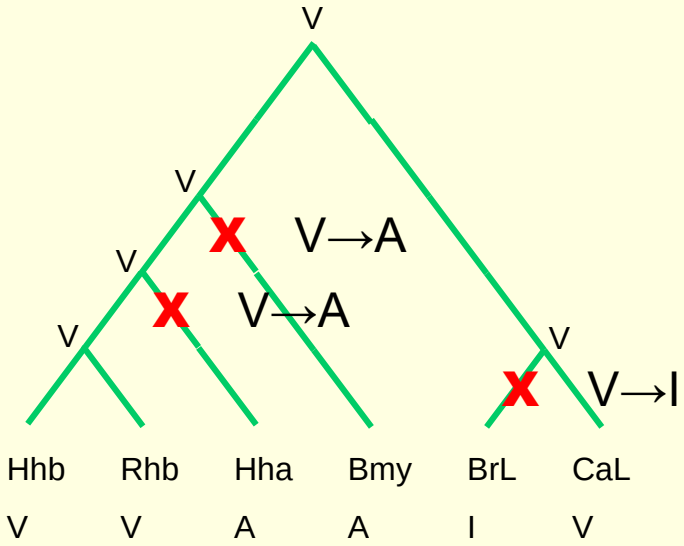
Mutational Transitions

- ***3 indels, each in one branch (–)***

```
Human HBB     1   --KV--NVDEVG--
Rana HBB      1   --KV--DVEQDG--
Human HBA     3   --KVGAHAGEYG--
Bovine MY     3   --KV--EADVAGHG
BrNapa HB     6   ILKQ--DIPKYS--
CanLi LB      1   AFKQ--NVPHHS--
```



SOP=7
mutations=1



SOP=4
mutations=3

Multiple Alignment and Trees

Final leader-follower alignment

```
Human HBB    1  MVHLTPEEKSAVTALWG--KV--NVDEVG--GEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK  60
Rana HBB     1  MVHWTAEEKAVINSVWQ--KV--DVEQDG--HEALTRLFIVYPWTQRYFSTFGDLSSPAAIAGNPK  60
Human HBA    3     LSPADKTNVKAAWG--KVGAHAGEYG--AEALERMFLSFPTTKTYFPHF------DLSHGSAQ  55
Bovine MY    3     LSDGEWQLVLNAWG--KV--EADVAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASED  61
BrNapa HB    6     FTEKQEALVKESWEILKQ--DIPKYS--LHFFSQILEIAPAAKDMFSFLRD--TDEVPHNNPK  62
CanLi LB     1  MGAFSEKQESLVKSSWEAFKQ--NVPHHS--AVFYTLILEKAPAAQNMFSFLSNGVDPN----NPK  54

Human HBB   61  VKAHGKKVLGAFSDGLAHLDNLKG-------TFA----TLSELHCDKLHVDPENFRLLGNVLVCV  114
Rana HBB    61  VHAHGKKILGAIDNAIHNLDDVKG-------TLH----DLSEEHANELHVDPENFRRLGEVLIVV  114
Human HBA   56  VKGHGKKVADALTNAVAHVDDMPN-------ALS----ALSDLHAHKLRVDPVNFKLLSHCLLVT  109
Bovine MY   62  LKKHGNTVLTALGGILKKKGHHEA-------EVK----HLAESHANKHKIPVKYLEFISDAIIHV  115
BrNapa HB   63  LKAHAVKVFKMTCETAIQLRE-KGKVVVADTTLQ----YLGSVHFKSGVLDP-HFEVVKEALVRT  121
CanLi LB    55  LKAHAEKVFKMTVDSAVQL-RAKG-------EVVLADPTLGSVHVQKGVLDP-HFLVVKEALLKT  114

Human HBB  115  LAHHFGKEFTPPV----QAAYQKVVAGVANALAHKYH  147
Rana HBB   115  LGAKLGKAFSPQV----QHVWEKFIAVLVDALSHSYH  147
Human HBA  110  LAAHLPAEFTPAV----HASLDKFLASVSTVLTSKY  141
Bovine MY  116  LHAKHPSDFGADA----QAAMSKALELFRNDMAAQY  147
BrNapa HB  122  LKEGLGEKYNEEV----EGAWSKAYDHLALAI  149
CanLi  LB  115  FKEAVGDKWNDELGNAWEVAYDELAAAIKKAMG  147
```

Multiple Alignment and Trees

Final leader-follower alignment

```
Human HBB     1  MVHLTPEEKSAVTALWG--KV--NVDEVG--GEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK  60
Rana HBB      1  MVHWTAEEKAVINSVWQ--KV--DVEQDG--HEALTRLFIVYPWTQRYFSTFGDLSSPAAIAGNPK  60
Human HBA     3     LSPADKTNVKAAWG--KVGAHAGEYG--AEALERMFLSFPTTKTYFPHF------DLSHGSAQ  55
Bovine MY     3     LSDGEWQLVLNAWG--KV--EADVAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASED  61
BrNapa HB     6     FTEKQEALVKESWEILKQ--DIPKYS--LHFFSQILEIAPAAKDMFSFLRD--TDEVHNPNPK  62
CanLi LB      1  MGAFSEKQESLVKSSWEAFKQ--NVPHHS--AVFYTLILEKAPAAQNMFSFLSNGVDPN----NPK  54
                                *                       *

Human HBB    61  VKAHGKKVLGAFSDGLAHLDNLKG-------TFA----TLSELHCDKLHVDPENFRLLGNVLVCV  114
Rana HBB     61  VHAHGKKILGAIDNAIHNLDDVKG-------TLH----DLSEEHANELHVDPENFRRLGEVLIVV  114
Human HBA    56  VKGHGKKVADALTNAVAHVDDMPN-------ALS----ALSDLHAHKLRVDPVNFKLLSHCLLVT  109
Bovine MY    62  LKKHGNTVLTALGGILKKKGHHEA-------EVK----HLAESHANKHKIPVKYLEFISDAIIHV  115
BrNapa HB    63  LKAHAVKVFKMTCETAIQLRE-KGKVVVADTTLQ----YLGSVHFKSGVLDP-HFEVVKEALVRT  121
CanLi LB     55  LKAHAEKVFKMTVDSAVQL-RAKG-------EVVLADPTLGSVHVQKGVLDP-HFLVVKEALLKT  114
                   *                                       *

Human HBB   115  LAHHFGKEFTPPV----QAAYQKVVAGVANALAHKYH  147
Rana HBB    115  LGAKLGKAFSPQV----QHVWEKFIAVLVDALSHSYH  147
Human HBA   110  LAAHLPAEFTPAV----HASLDKFLASVSTVLTSKY   141
Bovine MY   116  LHAKHPSDFGADA----QAAMSKALELFRNDMAAQY   147
BrNapa HB   122  LKEGLGEKYNEEV----EGAWSKAYDHLALAI----   149
CanLi  LB   115  FKEAVGDKWNDELGNAWEVAYDELAAAIKKAMG      147
```

agrees with structural
unclear in structure
disagrees with structural

Multiple Alignment and Trees

Leader makes a big difference

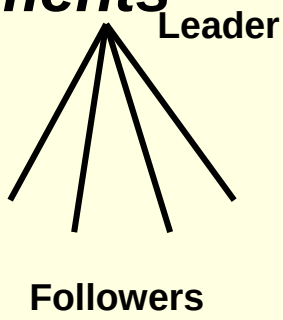
- **Human HBB vs CanLi as Leader**

```
Human HBB    1  MVHLTPEEKSAVTALWG--KV--NVDEVG--GEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK  60
Rana HBB     1  MVHWTAEEKAVINSVWQ--KV--DVEQDG--HEALTRLFIVYPWTQRYFSTFGDLSSPAAIAGNPK  60
Human HBA    3     LSPADKTNVKAAWG--KVGAHAGEYG--AEALERMFLSFPTTKTYFPHF------DLSHGSAQ  55
Bovine MY    3     LSDGEWQLVLNAWG--KV--EADVAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASED  61
BrNapa HB    6     FTEKQEALVKESWEILKQ--DIPKYS--LHFFSQILEIAPAAKDMFSFLRD--TDEVPHNNPK  62
CanLi LB     1  MGAFSEKQESLVKSSWEAFKQ--NVPHHS--AVFYTLILEKAPAAQNMFSFLSNGVDPN----NPK  54

CanLi  LB 7          KQESLVKSSWEAFKQNVP-HHSAVFYTLILEKAPAAQ--NMFS-F-L--SNGVDP----N----NPK
Human HBB 7           EEKSAVTALWG--KVNVD-EVGGEALGRLLVVYPWTQ--RFFE-S-F--GDLSTP----DAVMGNPK
Rana HBB  4         WTAEEKAVINSVWQKVDVEQD-GHEAL--TRLFIVYPWTQ--RYFS-T-F--GDLSSP----AAIAGNPK
Human HBA 4          SPADKTNVKAAWGKVGAHAG-EYGAEALERMFLSFPTTK--TYFPHFDL--SHG----------SAQ
Bovine MY 1    MG---LSDGEWQLVLNAWGKVEADVAGHGQEVLIRLFTGHPETLEKFDKFK-H-L--KTEAEM----K---ASED
BrNapa HB 1    MGEIVFTEKQEALVKESWEILKQDIP-KYSLHFFSQILEIAPAAK--DMFS-F-LRDTDEVPH----N----NPK
```

Multiple Alignment and Trees

Problems revealed by leader-follower alignments

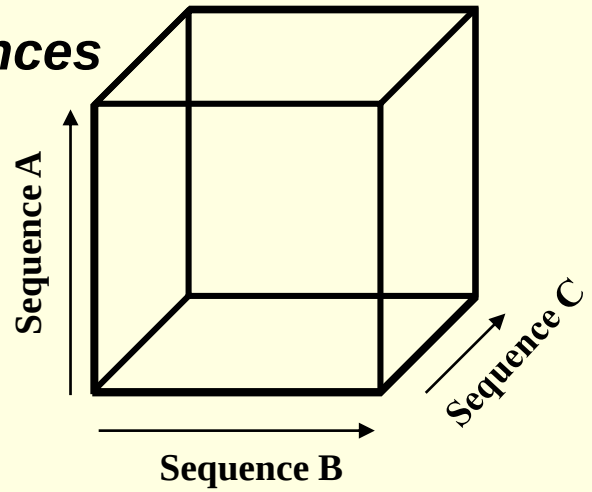


- ***Alignment depends on leader***
  - Model is not biologically realistic (star tree)
- ***Inaccuracy in pairwise alignments compromises multiple alignment***
- ***Some alignments are better defined than others (the ones where the sequences are close are less ambiguous)***

Multiple Alignment and Trees

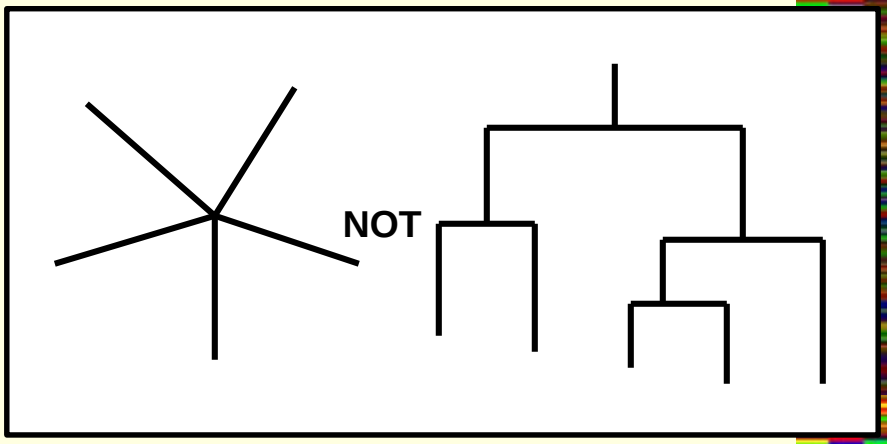
Dynamic Programming in Multiple Dimensions

- ***Three sequences can be aligned using the same dynamic programming procedure used with two sequences.***
  - The score matrix that must be filled is cube rather than a square.
  - Time required is thus L^3 where L is the length of the sequences.
- ***For more than three sequences the problem (time and memory) scale exponentially with the number of sequences,***
  - N sequences require L^N time.
- ***Impractical for large numbers of sequences***



Multiple Alignment and Trees

Dynamic Programming in Multiple Dimensions



- ***Carrillo-Lipman algorithm***
  - Tries to optimize SOP score
    - Sum-of-pairs implies a "star topology"
  - Works by using pairwise alignments to restrict N-dimensional alignment space
    - Score for a pair of sequences in a multiple alignment can only be less than or equal to the pairwise alignment, and should be within some distance ε
  - Will handle a small number ~10 of average length proteins sequences
  - One implementation: Lipman, Altschul, Kececioglu, PNAS 86:4412-4415 (1989)
  - Software
    - server: http://searchlauncher.bcm.tmc.edu/multi-align/multi-align.html