

# ***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, perhaps 1000 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

	70	80	90	100	110	120																																																										
FER_CAPAA/1-97	A	S	Y	K	V	K	L	I	T	P	D	G	P	I	E	F	D	C	P	D	D	V	Y	I	L	D	Q	A	E	E	A	G	H	D	L	P	Y	S	C	R	A	G	S	C	S	S	C	A	G	K	I	A	G	G	A	V	D	Q	T	D	G	N	F	L
FER_CAPAN/1-144	A	S	Y	K	V	K	L	I	T	P	D	G	P	I	E	F	D	C	P	D	N	V	Y	I	L	D	Q	A	E	E	A	G	H	D	L	P	Y	S	C	R	A	G	S	C	S	S	C	A	G	K	I	A	G	G	A	V	D	Q	T	D	G	N	F	L
FER1_SOLLC/1-144	A	S	Y	K	V	K	L	I	T	P	E	G	P	I	E	F	E	C	P	D	D	V	Y	I	L	D	Q	A	E	E	E	G	H	D	L	P	Y	S	C	R	A	G	S	C	S	S	C	A	G	K	V	T	A	G	S	V	D	Q	S	D	G	N	F	L
Q93XJ9_SOLTU/1-144	A	S	Y	K	V	K	L	I	T	P	D	G	P	I	E	F	E	C	P	D	D	V	Y	I	L	D	Q	A	E	E	E	G	H	D	L	P	Y	S	C	R	A	G	S	C	S	S	C	A	G	K	V	T	A	G	T	V	D	Q	S	D	G	K	F	L
FER1_PEA/1-149	A	S	Y	K	V	K	L	V	T	P	D	G	T	Q	E	F	E	C	P	S	D	V	Y	I	L	D	H	A	E	E	V	G	I	D	L	P	Y	S	C	R	A	G	S	C	S	S	C	A	G	K	V	V	G	G	E	V	D	Q	S	D	G	S	F	L
Q7XA98_TRIPR/1-152	A	T	Y	K	V	K	L	I	T	P	E	G	P	Q	E	F	D	C	P	D	D	V	Y	I	L	D	H	A	E	E	V	G	I	E	L	P	Y	S	C	R	A	G	S	C	S	S	C	A	G	K	V	V	N	G	N	V	N	Q	E	D	G	S	F	L
FER1_MESCR/1-148	A	A	Y	K	V	T	L	V	T	P	E	G	K	Q	E	L	E	C	P	D	D	V	Y	I	L	D	A	A	E	E	A	G	I	D	L	P	Y	S	C	R	A	G	S	C	S	S	C	A	G	K	V	T	S	G	S	V	N	Q	D	D	G	S	F	L
FER1_SPIOL/1-147	A	A	Y	K	V	T	L	V	T	P	T	G	N	V	E	F	Q	C	P	D	D	V	Y	I	L	D	A	A	E	E	E	G	I	D	L	P	Y	S	C	R	A	G	S	C	S	S	C	A	G	K	L	K	T	G	S	L	N	Q	D	D	Q	S	F	L
FER3_RAPSA/1-96	A	T	Y	K	V	K	F	I	T	P	E	G	E	Q	E	V	E	C	D	D	D	V	Y	V	L	D	A	A	E	E	A	G	I	D	L	P	Y	S	C	R	A	G	S	C	S	S	C	A	G	K	V	V	S	G	S	V	D	Q	S	D	Q	S	F	L
FER1_ARATH/1-148	A	T	Y	K	V	K	F	I	T	P	E	G	E	L	E	V	E	C	D	D	D	V	Y	V	L	D	A	A	E	E	A	G	I	D	L	P	Y	S	C	R	A	G	S	C	S	S	C	A	G	K	V	V	S	G	S	V	D	Q	S	D	Q	S	F	L
FER_BRANA/1-96	A	T	Y	K	V	K	F	I	T	P	E	G	E	Q	E	V	E	C	D	D	D	V	Y	V	L	D	A	A	E	E	A	G	I	D	L	P	Y	S	C	R	A	G	S	C	S	S	C	A	G	K	V	V	S	G	F	V	D	Q	S	D	E	S	F	L
FER2_ARATH/1-148	A	T	Y	K	V	K	F	I	T	P	E	G	E	Q	E	V	E	C	E	E	D	V	Y	V	L	D	A	A	E	E	A	G	L	D	L	P	Y	S	C	R	A	G	S	C	S	S	C	A	G	K	V	V	S	G	S	I	D	Q	S	D	Q	S	F	L
Q93Z60_ARATH/1-118	A	T	Y	K	V	K	F	I	T	P	E	G	E	Q	E	V	E	C	E	E	D	V	Y	V	L	D	A	A	E	E	A	G	L	D	L	P	Y	S	C	R	A	G	S	C	S	S	C	A	G	K	V	V	S	G	S	I	D	Q	S	D	Q	S	F	L

FER\_CAPAA/1-97  
FER\_CAPAN/1-144  
FER1\_SOLLC/1-144  
Q93XJ9\_SOLTU/1-144  
FER1\_PEA/1-149  
Q7XA98\_TRIPR/1-152  
FER1\_MESCR/1-148  
FER1\_SPIOL/1-147  
FER3\_RAPSA/1-96  
FER1\_ARATH/1-148  
FER\_BRANA/1-96  
FER2\_ARATH/1-148  
Q93Z60\_ARATH/1-118

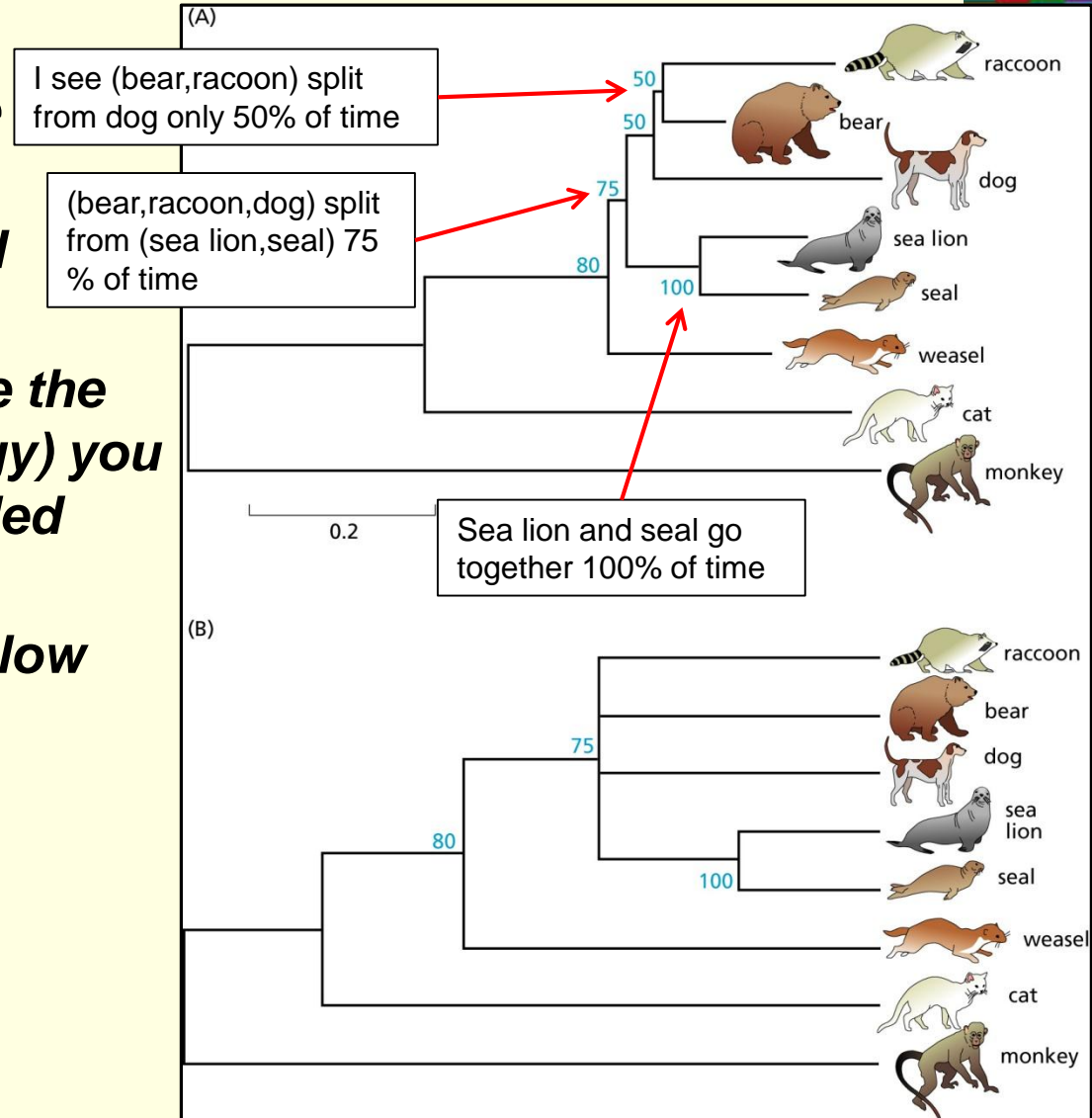
P	A	K	P	Y
P	A	K	P	Y
P	T	K	P	Y
P	T	K	P	Y
P	V	K	P	Y
P	V	K	P	Y
P	T	T	P	Y
P	K	T	P	Y
D	V	K	D	Y
D	V	K	D	Y
E	V	K	E	Y
E	V	K	E	Y

- 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

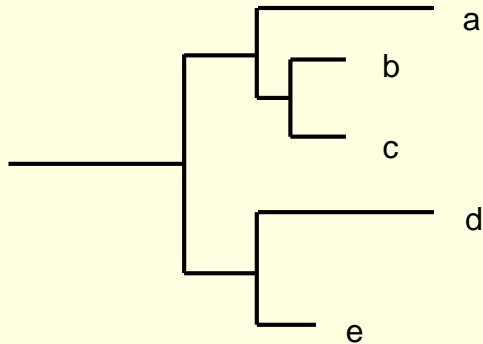
## Bootstrap

- **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

Bootstrap support

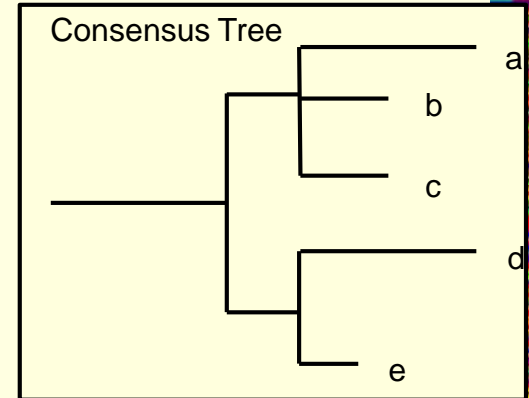
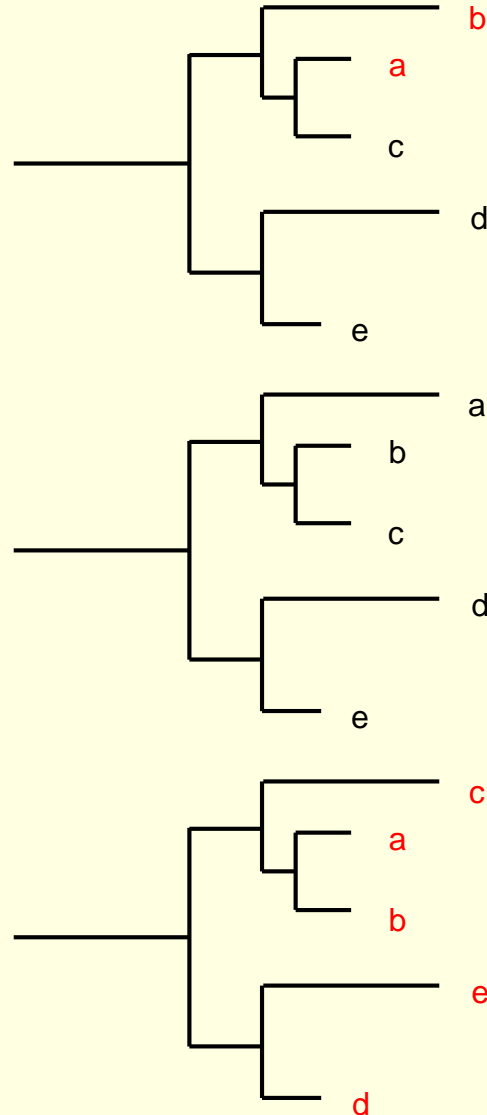
b:c 33%

bc:a 33%

d:e 100%

de:abc 100%

Resampled trees



# 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,  $i$

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

# Multiple Alignment and Trees

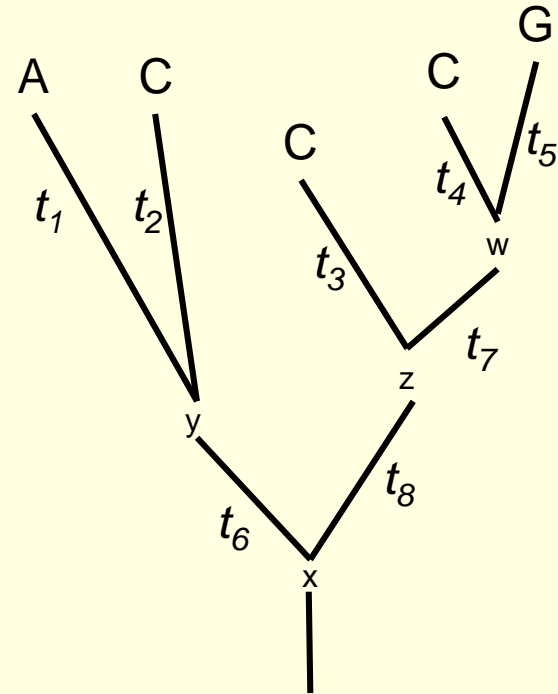
## Maximum Likelihood (see eq 8.51-52 in book)

- **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**

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

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

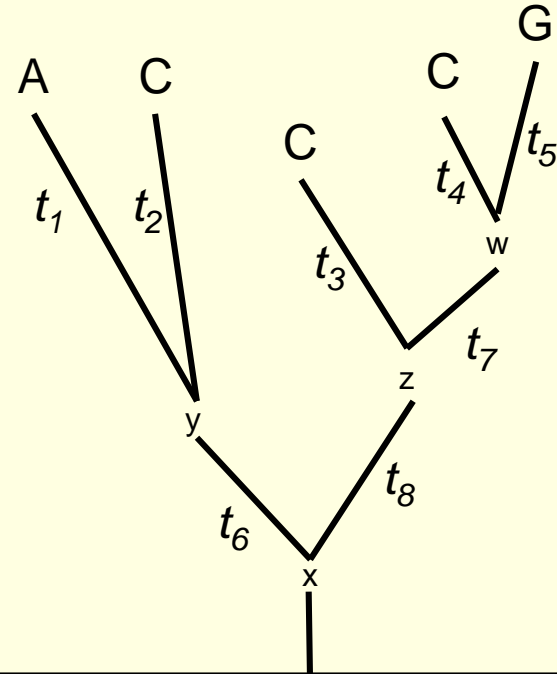
$$\begin{aligned} \text{Prob}(A, C, C, C, G, x, y, z, w|T) = & \\ & \text{Prob}(x) \text{Prob}(y|x, t_6) \text{Prob}(A|y, t_1) \text{Prob}(C|y, t_2) \\ & \text{Prob}(z|x, t_8) \text{Prob}(C|z, t_3) \\ & \text{Prob}(w|z, t_7) \text{Prob}(C|w, t_4) \text{Prob}(G|w, t_5) \end{aligned}$$



# Multiple Alignment and Trees

## Maximum Likelihood

- If evolution has been going on a long time, you can assume that  $\text{Prob}(x)$  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 ( $t_1 - t_8$ )



$$\begin{aligned} \text{Prob}(A, C, C, C, G, x, y, z, w|T) = & \\ & \text{Prob}(x) \text{Prob}(y|x, t_6) \text{Prob}(A|y, t_1) \text{Prob}(C|y, t_2) \\ & \text{Prob}(z|x, t_8) \text{Prob}(C|z, t_3) \\ & \text{Prob}(w|z, t_7) \text{Prob}(C|wy, t_4) \text{Prob}(G|w, t_5) \end{aligned}$$



# ***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,  $R \rightarrow R$  or  $Y \rightarrow Y$ , and transversion,  $R \rightarrow Y$  or  $Y \rightarrow R$ ).
  - 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***

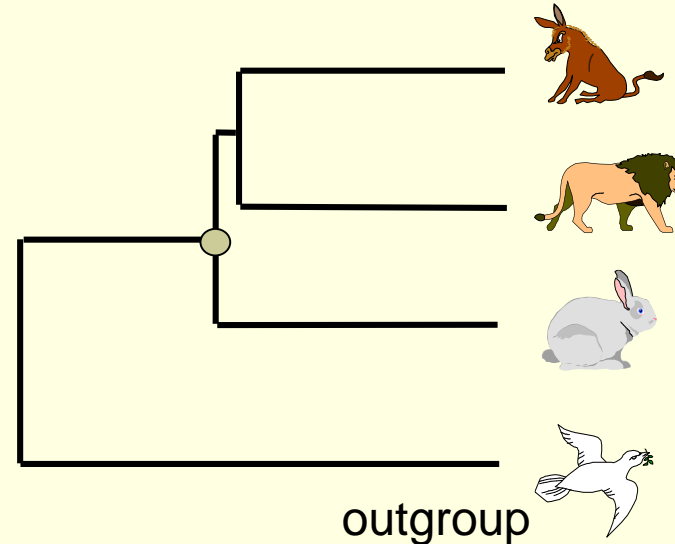
## ***Rooting trees***

- ***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:

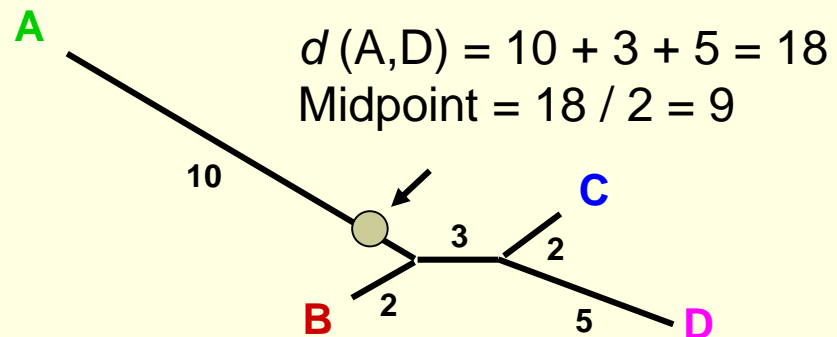
## By outgroup:

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.,  $\alpha$ -globins to root  $\beta$ -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.



# 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  $^{133}\text{cesium}$  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

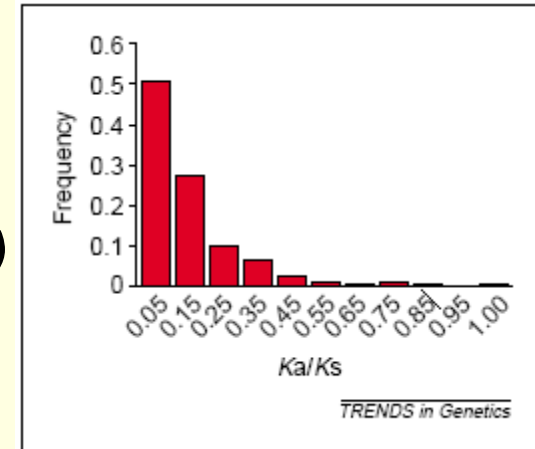
- **Neutral mutations - Most mutations are neither highly advantageous or deleterious - they are effectively neutral (Kimura theory).**
- **Neutral mutations should be the most clocklike because they represent the random accumulation of changes over time.**
  - Very common attempt to find neutral mutations 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

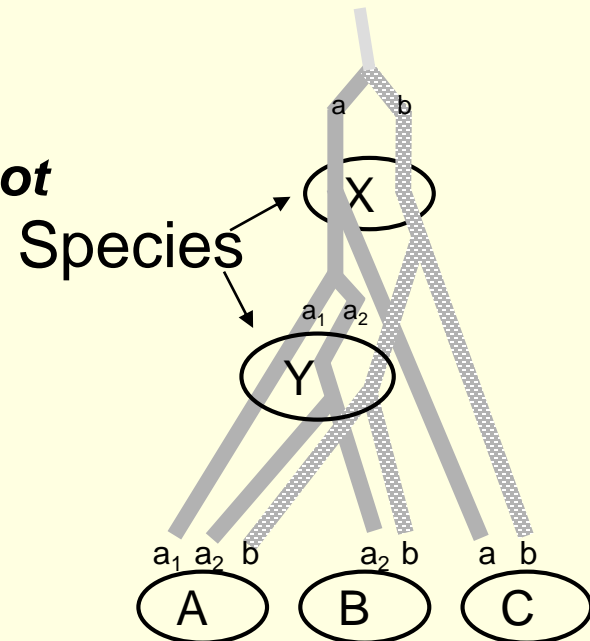
- **Ratio of non-synonymous ( $K_a$ , changed amino acid) to synonymous ( $K_s$ , unchanged amino acid) changes in coding gene DNA**
- **$K_s$  measures clocklike behavior**
- **If  $K_a$  is much larger, it is changing in a not clock-like way – it is being selected**
  - Has to be adjusted for multiple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

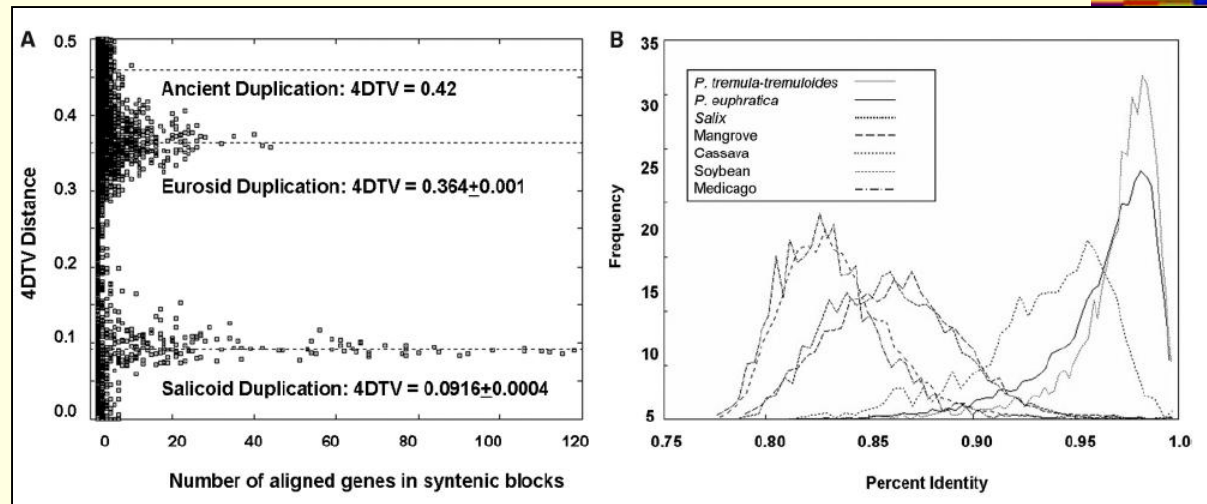
- **Tree construction seeks to understand the evolutionary relationship between taxa (groups of organisms).**
- **In sequence based methods, one must distinguish between gene trees and species trees due to the presence of orthologous and paralogous genes**
  - Orthologous - homologous genes that are 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

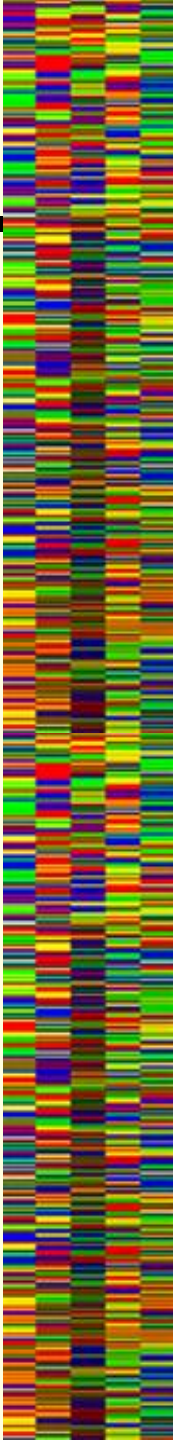


# ***Multiple Alignments and Trees***

---

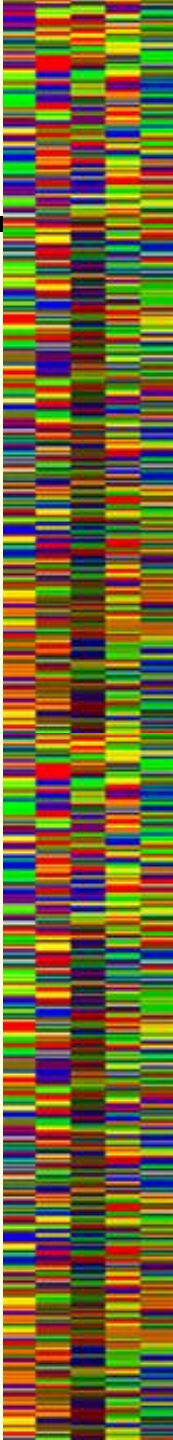
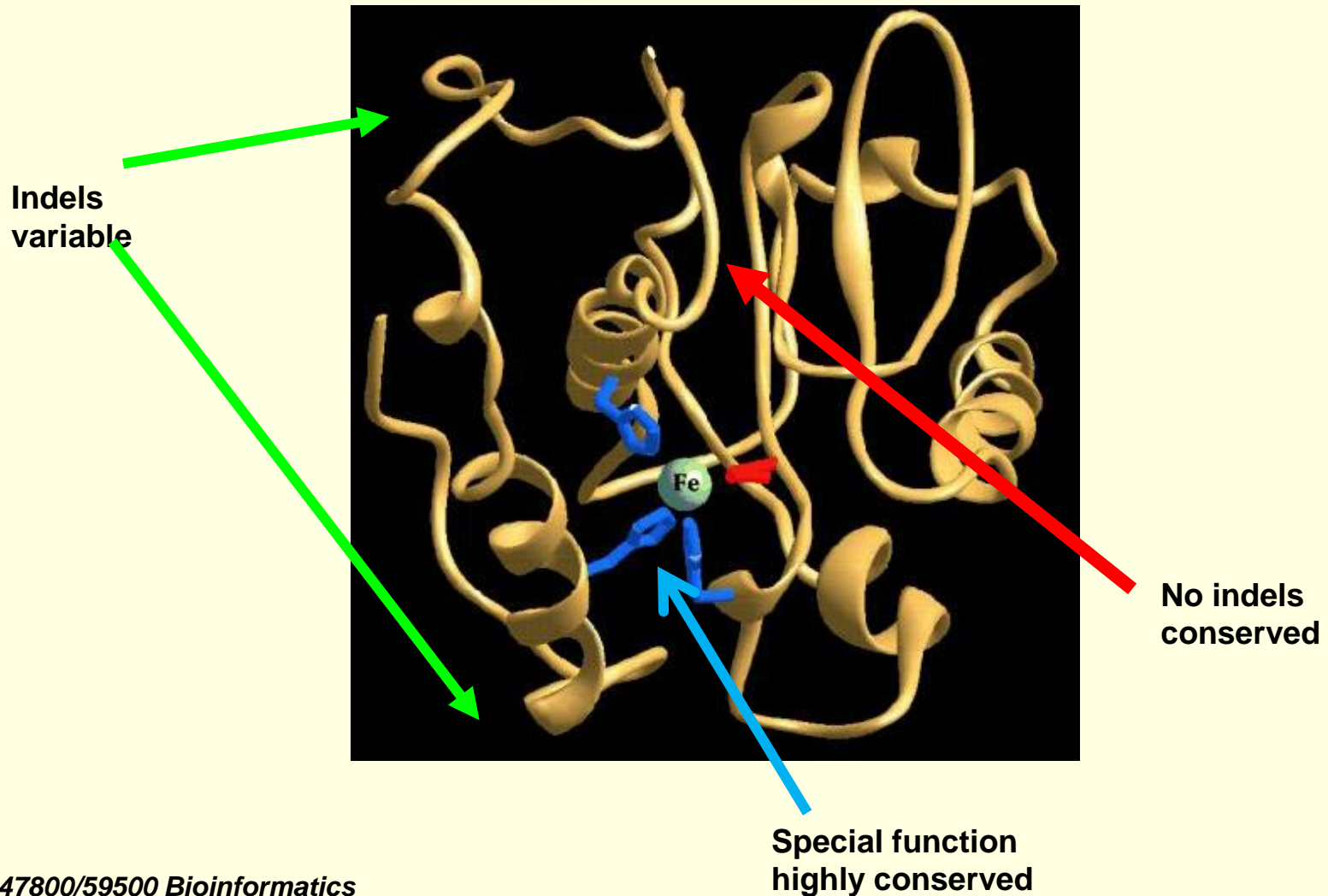
## ***Why make multiple alignments?***

- ***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



# Multiple Alignments and Trees

## Goals

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

kallikrein	LPGGYTCFPHSQPWQAAL	LVQGRLLCGGVLVHPKWLTAACHLCKGG	LKVVYLGKHALG RVEAGEQVREVVHSIPHPFYRRSPTHL	NHDHDMLELQSP
protease	LVHGGPCKTSHPYQAAL	YTSGHLLCGGVLIHPWLTAACHCKKPN	LQVPLGKHNR QRESSQEQSSVRAVTHPYDAA	SHDQDMLLRARP
neuropsin	VLGGHCQPHSQPWQAAL	FQGGQLICGGVLVGGNWLTAACHCKPK	YTVRLGDHSLQ NKDQPEQELPVVQSIPHPCYNSSDVE	DHNHDLMLLQLRQ
protease	LIINGEDCSPHSQPWQAAL	VMENELFCGGVLVHPQWVLSAAHCFQNS	YTIGLGLHSLHADQEPGQMVVEASLSVRHPFYNRPLLA	NDLMLTKDES
psa	IVGGWCEKHSQPWQVLV	ASRFRAVCGGVLVHPQWVLSAAHCLRK	SVILLGRHSLFHPEDTG QVFQVSHSFPHPYDMSLLKNRFRP	GDDSSHDMLRLSEP
complement	IVGGKRALGLDLPWQVAIK	DASGITCGGIYIGGCWILTAACHLRA	SKTHRYQIWTVVVDWIHPDLKRIVIE	YVDRIIFHENY
factor	IVGGKRALGLDLPWQVAIK	DASGITCGGIYIGGCWILTAACHLRA	SKTHRYQIWTVVVDWIHPDLKRIVIE	YVDRIIFHENY
airway	ILGGTEAEEGSPWQVSLRL	NNAHHCGGSLINNMWILTAACHCFR SN	SNPRDWIATSGI	STTFPKLRTRRNILIHNNY
mtsp7	IVQGRETAMEGEWPWQASLQLI	CSGHCCGASLISNTWLLTAACHCFW KN	KDPTQWIATPGA	TIITPPAVKRNVRKIILHENY
enterokinase	IVGGSNAKEGAWPWVVGGLY Y	GGRLICGASLVSSDVLVSAAHCVYGRN	LEPSKWTAILGLHMS	NLTSPTVPRIDEIVINPHY
hepsin	IVGGERDTSLGRWPWQVSLR Y	DGAHLCCGSLSGDWLTAACHCFPERN	RVLSRWRFAGAVAQA	SPHGLQLGQAVVYHGGYLPFRDPNSEE
proctasin	ITGGSSAVAGQWPWQVSI TY	EGVHVCGGSLVSEQWLSAAHCFPSEH	HKEA YEVKLGAKHLD	SYSEDAKVS TLKDIIPHPSYL
plasmin	VVGGCVAHPSWPWQVSLRTR	FGMHFCGGTLISPEWVLTAAHCL EKS	PRPSSYKVLGAHQEV	NL EPHGQETEVSRFLFEP
testisin	IVGGEDAELGRWPWQVSLRLW	DS HVCCGVSLLSRWALTAACHCFETYSDLSDPGGMVQFG QLT		SHPSFWSLQAYYRYFVSNITXLSPRYLGN
corin	ILGERTSRPGRWPWQCSIQSE	PSGHHCQCVLIADKKWVLTVAHCFEGRENA	AVWVVLGGINND	H PVSVMQTRFVKLILHPRY
acrosin	IVGGKAAQHGAWPWAIVSLQIFTYNSHRYHT	CGGSLNLSRWLTAACHCFGKGNVHD	WRVFGAKELTYGNKKPVKAPVQERYEKILIHKEY	NSAT EGDNDIALVETTP
neurotrypsin	IIGGKNSLRGGWPWQVSLRLKSSHGDERLL	CGVITLLSCWLTAAHCF KRYGNSTRSYAVRVG	DYHTLVVPEEETEGVQIVIHREYRPDR	SDYDIALVRLQSP
proteinase	IVGGHEAQPHSRPYMASLQWRGNP	GSHFCGGTLIHPSEVLTAAHCLRD	IPQRLVNVVLAGHNVRTQ	EPTQQHPSVAQVFLNN
consensus	IVGG A G WPWQVSL	G H CCG L WLTAAHCF	W V LG H	V I H Y

**Serine Proteases**

	121				240
kallikrein	VQLTGXIQT	LPLSHNRILPGTTCRVSGWGTIT	SPQVNYPKTIQCANIQLRSDPCR	QWYPRKITDMLCAGTK	EGGKDCSGDSSGGLVLCNR
protease	AKLSELIQP	LPLERDCSANT TSCHLGWKTA DG	DFPDITQCAIHLVSRREECE	HAYPGQITQMLCAGDE	KYKDCSCGDSGGLVLCGD
neuropsin	ASLGSKVKP	ISLADHCT QPGQKCTVSGWGTVT	SPRENFPITNCAEVKIFPKKCE	DAYPGQITDGMVCASS	K GADTCQDSSGGLVLCGD
protease	VSESDTIRS	ISIASQC PTAGRNSCLVSGWGLLA	NGR MPVILQCWNVSVEEVC	KLYDPLVHPSMFCAGGG	HQKDCSCNGDSSGGLVLCNG
psa	AELTDAVKV	MDLPTQ EPALGTTICVAGWGSIE	PEEFLTPKKIQCVDLHVISNDVCA	QVHPQVTKFMLCAGRW	TGGKSTCSGDSGGLVLCNG
complement	GNKKDCLEPRSIACVPSYPYLFQPN	DTCLVSGWGREKDNRFPS	LQWGEVKLISN CSKFGY	NRFYEKEMECAGTY	DGSDACKGDSGGLVLCMDANNVTYVW
factor	GNKKDCLEPRSIACVPSYPYLFQPN	DTCLVSGWGREKDNRFPS	LQWGEVKLISN CSKFGY	NRFYEKEMECAGTY	DGSDACKGDSGGLVLCMDANNVTYVW
airway	VTFTKDHSVVC	PAATQNIIPPGS	TAYVTGWAQYAGH	TVPELRQGGVRIISNDVGN	APHSYNGAILSGMLCAG
mtsp7	VVEFSNIVQRVCLPDSISKLPKPT	SUVFVIGFGSIVDDGP	IQNTLRQARVETISTDVGN	RKQVYDGLITPGLCAG	FMEGKIDACKGDSGGLVLCMDANNVTYVW
enterokinase	VNYTDYIQPIC	PEENQVFPFR	NCSIAGWGTVYVQGT	TANILQEAQVPLSNERCQ	QQMPEYNITENMLCAG
hepsin	LPLTEYIQPIC	PAAGQALVDGK	ICFVTGWCNTQYQGQ	QAGVQLQEARVPIISNDVGN	GADFYCNQIKPKFCAG
proctasin	ITFSRYIRPIC	PAANASFPNGL	HCTVIGWGHVAPSYSLLTPKPI	QQLEVPILISRET	CNCLYMDACKPEEPHFVQEDMVCAG
plasmin	AVITDRVTPAC	PSPNYVADRT	ECFITGWGET	QGTFG AGLIKEAQVPIENK	CHRYEFL
testisin	VTYTKHIQPIC	QASTFEFENRT	DCWVTGWGYI	KEDEALPSPHITQEVQVALINNS	CNHLFL
corin	ISCTGVYRPFVCL	PNPEQWLEPDT	YCYITGWGHMGNKIFPK	IQEGEVRITISLEHCQSYEDMKT	ITTRMTCAG
acrosin	ISCFRFIPGPCI	PHIKAGLPRGCSQ	SCWVAGWGYIEEKAPRPS	SILMEARVDLIDLCLN	STQWYNGRVPPTNVCAG
neurotrypsin	EEQCAREFSHWLPA	CLWRERPKTASNCYITGWGDTG	RAYSRILQQAAILPLKPRFCE	ERYKRFTRMLCAGHL	HEHKRVDS
proteinase	AHLSAS	VATVQ	PQQDQVPVHGTO	CLAMGWRVGAHPD	PAQVILQENLVITVTFPCR
consensus	I P C L P	C V G W G	I Q A V I S C	Y I M C A G	G G D C Q G D S G G L V L C W L

# 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	MVHLTPPEEKSAVTALWGKVVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
Rana HBB	1	MVHWTAEEKAVINSVWQKVDVEQDGHEALTRLFIVYPWTQRYFSTFGDLSSPAAIAGNPK	60
Human HBB	61	VKAHGKKVLGAFSDGLAHL DNLKGT FATLSELHCDKLHVDPENFRLLGNVLCVLAHFFG	120
Rana HBB	61	VHAHGKKILGAIDNAIHNLDDVKGTLHDLSEEHANELHVDPENFRRLGEVLIVVLGAKLG	120
Human HBB	121	KEFTPPVQAAYQKVVAGVANALAHKYH	147
Rana HBB	121	KAFSPQVQHVWEKFI AVLVDALSHSYH	147
<hr/>			
Human HBB	4	LTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV	61
Human HBA	3	LSPADKTNVKAAWGKVG AHAGEYGA EALERMFLSFPTTKTYFPHF-----DL SHGSAQV	56
Human HBB	62	KAHGKKVLGAFSDGLAHL DNLKGT FATLSELHCDKLHVDPENFRLLGNVLCVLAHFFGK	121
Human HBA	57	KGHGKKVADAL TNAVAHVDDMPNAL SALSDLHAHKL RVD PVNFKLLSHCLLVTLAHLPA	116
Human HBB	122	EFTPPVQAAYQKVVAGVANALAHKY	146
Human HBA	117	EFTPAVHASLDKFLASVSTVLT SKY	141
<hr/>			
Human HBB	4	LTPEEKSAVTALWGKVVNDEVG--GEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV	61
Bovine MY	3	LSDGEWQLVLNAWGKVEADVAGHGQEV LIRLFTGHPETLEKFDKFKHLKTEAEMKASEDL	62
Human HBB	62	KAHGKKVLGAFSDGLAHL DNLKGT FATLSELHCDKLHVDPENFRLLGNVLCVLAHFFGK	121
Bovine MY	63	KKHGNTVLTALGGILKKKGHHEAEVKHLAESHANKHKI PVKYLEFISDAI IHVLHAKHPS	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--KVN	VDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV	61
BrNapa HB	6	FTEKQEQEALVKESWEILKQDIPK	YSLHFFSQILEIAPAAKDMFSFLRD--TDEVPHNNPKL	63
Human HBB	62	KAHGKKVLGAFSDGLAHL	DNLKG-----TFATLSELHCDKLHVDPENFRLLGNVLVCV	114
BrNapa HB	64	KAHAVKVKFMTCETAIQLRE-KG	KVVVADTTLQYLGSVHFKSGVLDP-HFEVVKEALVRT	121
Human HBB	115	LAHHFGKEFTPPVQAAYQKVVAG	VANAL	142
BrNapa HB	122	LKEGLGEKYNEEVEGAWSKAYDHL	LALAI	149
<hr/>				
Human HBB	1	MVHLTPEEKSAVTALWG--KVN	VDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN	58
CANLI LB	1	MGAFSEKQESLVKSSWEAFKQNV	PHHSAVFYTLILEKAPAAQNMFSFLSNGVDPN----N	56
Human HBB	59	PKVKAHGKKVLGAFSDGLAHL	DNLKGTF-----TLSELHCDKLHVDPENFRLLGNVLVCV	114
CANLI LB	57	PKLKAHAEKVFKMTVD SAVQL-	RAKGEVVLADPTLGSVHVQKGVLDP-HFLVVKEALLKT	114
Human HBB	115	LAHHFGKEFTPPV----QAAYQKVVAG	VANALA	143
CANLI LB	115	FKEAVGDKWDELGN AWEVAYDELA	AAAIKKAMG	147

# Multiple Alignment and Trees

## Alignments to Leader

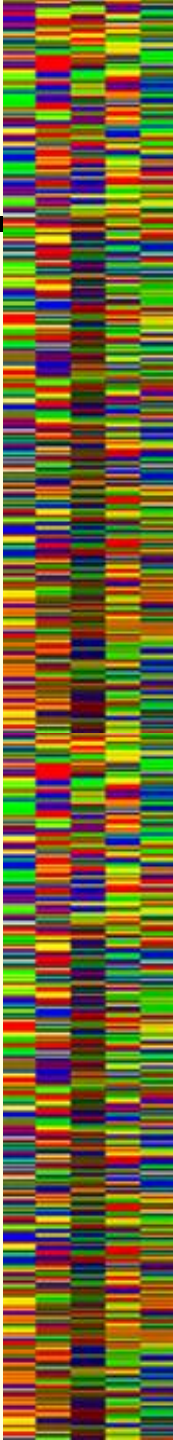
Human HBB	1	MVH <b>LT</b> PEEKSAVTALWGK <b>VNV</b> DEVG <b>GE</b> ALGRLLVVYPWTQR <b>FF</b> ESFGDLSTPDAVMGNPK	60
Rana HBB	1	MVHWTAEKAVINSVWQKVDVEQDGHEALTRLFIVYPWTQRYFSTFGDLSSPAAIAGNPK	60
Human HBB	4	<b>LT</b> PEEKSAVTALWGKV <b>--NV</b> DEVG <b>GE</b> ALGRLLVVYPWTQR <b>FF</b> ESFGDLSTPDAVMGNPKV	61
Human HBA	3	LSPADKTNVKAAGKVGVAHAGEYGAEALERMFLSFPTTKTYFPHF-----DLSHGSAQV	56
Human HBB	4	<b>LT</b> PEEKSAVTALWGK <b>VNV</b> DEVG <b>--GE</b> ALGRLLVVYPWTQR <b>FF</b> ESFGDLSTPDAVMGNPKV	61
Bovine MY	3	LSDGEWQLVLNAWGKVEADVAGHGQEVLIIRLFTGHPETLEKFDKFKHLKTEAEMKASEDL	62
Human HBB	4	<b>LT</b> PEEKSAVTALWG <b>--KVNV</b> DEVG <b>GE</b> ALGRLLVVYPWTQR <b>FF</b> ESFGDLSTPDAVMGNPKV	61
BrNapa HB	6	FTEKQEALVKESWEILKQDIPKYSLHFFSQILEIAPAAKDMFSFLRD--TDEVPHNNPKL	63
Human HBB	1	MVH <b>LT</b> PEEKSAVTALWG <b>--KVNV</b> DEVG <b>GE</b> ALGRLLVVYPWTQR <b>FF</b> ESFGDLSTPDAVMGN	58
CanLi LB	1	MGAFSEKQESLVKSSWEAFKQNVPHHSVVFYTLILEKAPAAQNMFSFSLNGVDPN----N	56
Human HBB	61	<b>VKAHG</b> KKVLGAFSDGLAHL <b>DNLKGT</b> F <b>ATL</b> SELHCDKLH <b>VD</b> ENFRLLGNVLV <b>CVLA</b> HHFG	120
Rana HBB	61	VHAHGKKILGAIDNAIHNLDDVKGTLHDLSEEHANELHVDPENFRRLGEVLIVVLGAKLG	120
Human HBB	62	<b>KAHG</b> KKVLGAFSDGLAHL <b>DNLKGT</b> F <b>ATL</b> SELHCDKLH <b>VD</b> ENFRLLGNVLV <b>CVLA</b> HHFGK	121
Human HBA	57	KGHGKKVADALTNVAHAVDDMPNALSALSDLHAHKLRLVDPVNFKLLSHCLLVTLAAHLP	116
Human HBB	62	<b>KAHG</b> KKVLGAFSDGLAHL <b>DNLKGT</b> F <b>ATL</b> SELHCDKLH <b>VD</b> ENFRLLGNVLV <b>CVLA</b> HHFGK	121
Bovine MY	63	KKHGNTVLTALGGILKKKGHHEAEVKHHLAESHANKHKIPVKYLEFISDAI IHVLHAKHPS	122
Human HBB	62	<b>KAHG</b> KKVLGAFSDGLAHL <b>DNLK</b> ----- <b>TFATL</b> SELHCDKLH <b>VD</b> ENFRLLGNVL <b>VCV</b>	114
BrNapa HB	64	KAHAVKVFKMTCEITAIQLRE-KGKVVVADTTLQYLGSVHFKSGVLDP-HFEVVKEALVRT	121
Human HBB	59	<b>PKVKAHG</b> KKVLGAFSDGLAHL <b>DNLKGT</b> F <b>---</b> TLSELHCDKLH <b>VD</b> ENFRLLGNVL <b>VCV</b>	114
CanLi LB	57	PKLKAHAEKVFKMTVDSAVQL-RAKGEVVLADPTLGSVHVQKGVLDP-HFLVVKEALKKT	114
Human HBB	121	<b>KEFT</b> PPVQ <b>AA</b> Y <b>QK</b> VVAGVANALAHKYH	147
Rana HBB	121	KAFSPQVQHVWEKFI AVLVDALSHSYH	147
Human HBB	122	<b>EFT</b> PPVQ <b>AA</b> Y <b>QK</b> VVAGVANALAHKY	146
Human HBA	117	EFTP AVHASLDKFLASVSTVLTSKY	141
Human HBB	122	<b>EFT</b> PPVQ <b>AA</b> Y <b>QK</b> VVAGVANALAHKY	146
Bovine MY	123	DFGADAQAAMSKALELFRNDMAAQY	147
Human HBB	115	<b>LAH</b> HFG <b>KEFT</b> PPVQ <b>AA</b> Y <b>QK</b> VVAGVANAL	142
BrNapa HB	122	LKEGLGEEKYNEEVEGAWSKAYDHLALAI	149
Human HBB	115	<b>LAH</b> HFG <b>KEFT</b> PPV <b>---</b> Q <b>AA</b> Y <b>QK</b> VVAGVANALA	143
CanLi LB	115	FKEAVGDKWNDELGNAVEVAYDELAIAIKKAMG	147

# Multiple Alignment and Trees

## Propagate Gaps vs Leader

Human HBB	1	MVHLTPEEKSAVTALWG <b>KV</b> NVDEVG <b>GE</b> ALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
Rana HBB	1	MVHWTAEKAVINSVWQKVDVEQDGHEALTRLFIVYPWTQRYFSTFGDLSSPAAIAGNPK	60
Human HBB	4	LTPEEKSAVTALWG <b>KV</b> --NVDEVG <b>GE</b> ALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV	61
Human HBA	3	LSPADKTNVKAAWGKVGAGAHAGEYGAEALERMFLSFPTTKTYFPHF-----DLSHGSAQV	56
Human HBB	4	LTPEEKSAVTALWG <b>KV</b> NVDEVG-- <b>GE</b> ALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV	61
Bovine MY	3	LSDGEWQLVLNAWGKVEADVAGHGQEVLRIRLFTGHPETLEKFDKFKHLKTEAEMKASEDL	62
Human HBB	4	LTPEEKSAVTALWG-- <b>KV</b> NVDEVG <b>GE</b> ALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV	61
BrNapa HB	6	FTEKQEALVKESWEILKQDIPKYSLHFFSQILEIAPAADMFSFLRD--TDEVPHNNPKL	63
Human HBB	1	MVHLTPEEKSAVTALWG-- <b>KV</b> NVDEVG <b>GE</b> ALGRLLVVYPWTQRFFESFGDLSTPDAVMGN	58
CanLi LB	1	MGAFSEKQESLVKSSWEAFKQNPVPHHSVAVFYTLILEKAPAAQNMFSFLSNGVDPN----N	56

Human HBB	1	MVHLTPEEKSAVTALWG <b>++KV++</b> NVDEVG <b>++GE</b> ALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
Rana HBB	1	MVHWTAEKAVINSVWQ <b>++KV++</b> DVEQDG <b>++HE</b> ALTRLFIVYPWTQRYFSTFGDLSSPAAIAGNPK	60
Human HBB	4	LTPEEKSAVTALWG <b>++KV--</b> NVDEVG <b>++GE</b> ALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV	61
Human HBA	3	LSPADKTNVKAAWG <b>++KVG</b> AGAHAGEYG <b>++AE</b> ALERMFLSFPTTKTYFPHF-----DLSHGSAQV	56
Human HBB	4	LTPEEKSAVTALWG <b>++KV++</b> NVDEVG-- <b>GE</b> ALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV	61
Bovine MY	3	LSDGEWQLVLNAWG <b>++KV++</b> EADVAGHGQEVLRIRLFTGHPETLEKFDKFKHLKTEAEMKASEDL	62
Human HBB	4	LTPEEKSAVTALWG-- <b>KV++</b> NVDEVG <b>++GE</b> ALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV	61
BrNapa HB	6	FTEKQEALVKESWEILKQ <b>++DIP</b> KYS <b>++LH</b> FFSQILEIAPAADMFSFLRD--TDEVPHNNPKL	63
Human HBB	1	MVHLTPEEKSAVTALWG-- <b>KV++</b> NVDEVG <b>++GE</b> ALGRLLVVYPWTQRFFESFGDLSTPDAVMGN	58
CanLi LB	1	MGAFSEKQESLVKSSWEAFKQ <b>++NP</b> VPHHS <b>++AV</b> FYTLILEKAPAAQNMFSFLSNGVDPN----N	56



# Multiple Alignment and Trees

## Remove redundant leader sequences

Human HBB	1	MVHLTPEEKSAVTALWG++KV++NVDEVG++GEALGRLLVVYPWTQRRFFESFGDLSTPDAVMGNPK	60
Rana HBB	1	MVHWTAEKAVINSVWQ++KV++DVEQDG++HEALTRLFIVYPWTQRYFSTFGDLSSPAAIAGNPK	60
Human HBB	4	LTPEEKSAVTALWG++KV--NVDEVG++GEALGRLLVVYPWTQRRFFESFGDLSTPDAVMGNPKV	61
Human HBA	3	LSPADKTNVKAAG++KVG AHAGEYG++AEALERMFLSFPTTKTYFPHF-----DLSHGSAQV	56
Human HBB	4	LTPEEKSAVTALWG++KV++NVDEVG--GEALGRLLVVYPWTQRRFFESFGDLSTPDAVMGNPKV	61
Bovine MY	3	LSDGEWQLVLNAWG++KV++EADVAGHGQEVLI RLFTHGHPETLEKFDKFKHLKTEAEMKASEDL	62
Human HBB	4	LTPEEKSAVTALWG--KV++NVDEVG++GEALGRLLVVYPWTQRRFFESFGDLSTPDAVMGNPKV	61
BrNapa HB	6	FTEKQEALVKESWEILKQ++DIPKYS++LHFFSQILEIAPA AKDMFSFLRD--TDEVPHNNPKL	63
Human HBB	1	MVHLTPEEKSAVTALWG--KV++NVDEVG++GEALGRLLVVYPWTQRRFFESFGDLSTPDAVMGN	58
CanLi LB	1	MGAFSEKQESLVKSSWEAFKQ++NVP HHS++AVFYTLILEKAPAAQNMFSFLSNGVDPN----N	56
Human HBB	1	MVHLTPEEKSAVTALWG--KV--NVDEVG--GEALGRLLVVYPWTQRRFFESFGDLSTPDAVMGNPK	60
Rana HBB	1	MVHWTAEKAVINSVWQ--KV--DVEQDG--HEALTRLFIVYPWTQRYFSTFGDLSSPAAIAGNPK	60
Human HBA	3	LSPADKTNVKAAG--KVG AHAGEYG--AEALERMFLSFPTTKTYFPHF-----DLSHGSAQV	56
Bovine MY	3	LSDGEWQLVLNAWG--KV--EADVAGHGQEVLI RLFTHGHPETLEKFDKFKHLKTEAEMKASEDL	62
BrNapa HB	6	FTEKQEALVKESWEILKQ--DIPKYS--LHFFSQILEIAPA AKDMFSFLRD--TDEVPHNNPKL	63
CanLi LB	1	MGAFSEKQESLVKSSWEAFKQ--NVP HHS--AVFYTLILEKAPAAQNMFSFLSNGVDPN----N	56

# Multiple alignment and Trees

## Do we need all the gaps?

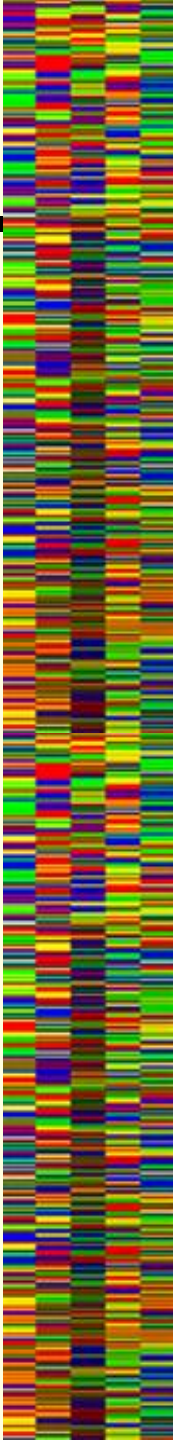
Human HBB	1	MVHLTPEEKSAVTALWG	--KV--NVDEVG--	GEALGRLLVVY	PWTQRFFESFGDLSTPDAVMGNPK	60
Rana HBB	1	MVHWTAEKAVINSVWQ	--KV--DVEQDG--	HEALTRLFIVY	PWTQRYFSTFGDLSSPAAIAGNPK	60
Human HBA	3	LSPADKTNVKAAWG	--KVGAHAGEYG--	AEALERMFLSF	PTTKTYFPHF-----DLSHGSAQV	56
Bovine MY	3	LSDGEWQLVLNAWG	--KVEADVAGHG	QEVLI RLF	TGHPETLEKFDKFKHLKTEAEMKASEDL	62
BrNapa HB	6	FTEKQEALVKESWE	ILKQ--DIPKYS--	LHFFSQILEIA	PAAKDMFSFLRD--TDEVPHNNPKL	63
CanLi LB	1	MGAFSEKQESLVKSSWE	AFKQ--NVPHHS--	AVFYTLILEKA	PAAQNMFSFLSNGVDPN----N	56



Human HBB	1	MVHLTPEEKSAVTALWG	--KV--NVDEVG	GEALGRLLVVY	PWTQRFFESFGDLSTPDAVMGNPK	60
Rana HBB	1	MVHWTAEKAVINSVWQ	--KV--DVEQDG	HEALTRLFIVY	PWTQRYFSTFGDLSSPAAIAGNPK	60
Human HBA	3	LSPADKTNVKAAWG	--KVGAHAGEYG	AEALERMFLSF	PTTKTYFPHF-----DLSHGSAQV	56
Bovine MY	3	LSDGEWQLVLNAWG	--KVEADVAGHG	QEVLI RLF	TGHPETLEKFDKFKHLKTEAEMKASEDL	62
BrNapa HB	6	FTEKQEALVKESWE	ILKQ--DIPKYS	LHFFSQILEIA	PAAKDMFSFLRD--TDEVPHNNPKL	63
CanLi LB	1	MGAFSEKQESLVKSSWE	AFKQ--NVPHHS	AVFYTLILEKA	PAAQNMFSFLSNGVDPN----N	56



Human HBB	1	MVHLTPEEKSAVTALWG	KV--NVDEVG	GEALGRLLVVY	PWTQRFFESFGDLSTPDAVMGNPK	60
Rana HBB	1	MVHWTAEKAVINSVWQ	KV--DVEQDG	HEALTRLFIVY	PWTQRYFSTFGDLSSPAAIAGNPK	60
Human HBA	3	LSPADKTNVKAAWG	KVGAHAGEYG	AEALERMFLSF	PTTKTYFPHF-----DLSHGSAQV	56
Bovine MY	3	LSDGEWQLVLNAWG	KVEADVAGHG	QEVLI RLF	TGHPETLEKFDKFKHLKTEAEMKASEDL	62
BrNapa HB	6	FTEKQEALVKESWE	ILKQDIPKYS	LHFFSQILEIA	PAAKDMFSFLRD--TDEVPHNNPKL	63
CanLi LB	1	MGAFSEKQESLVKSSWE	AFKQNVPHHS	AVFYTLILEKA	PAAQNMFSFLSNGVDPN----N	56



# Multiple Alignment and Trees

## Which is best?

### Qualitative

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

### • Quantitative

- Sum of pairs score

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

for the comparison 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?

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

# Multiple Alignment and Trees

## Which is best?

- **Quantitative**

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

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

```

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
```

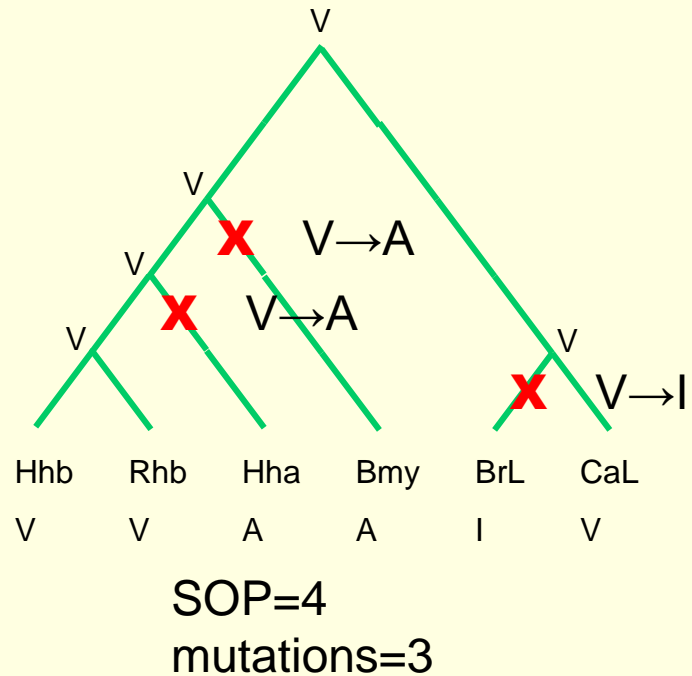
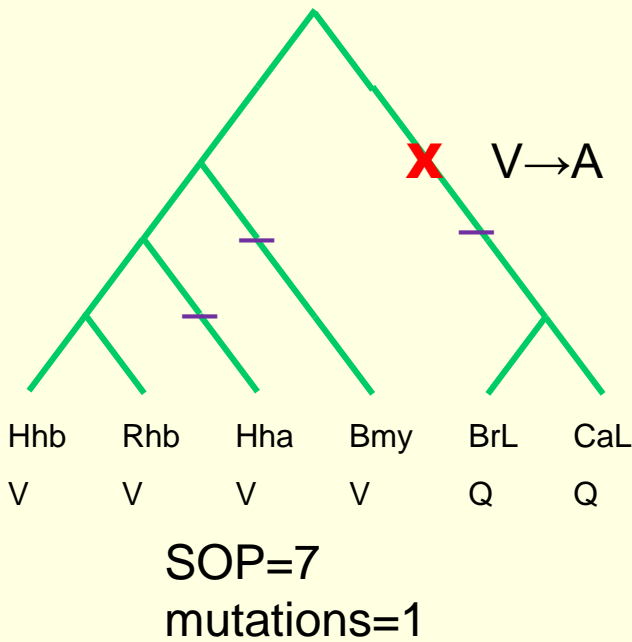
SOP=49  
gaps: 13 (57)

# 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	--KVG AHAGEYG--
Bovine MY	3	--KV--EADVAGHG
BrNapa HB	6	ILKQ--DIPKYS--
CanLi LB	1	AFKQ--NVPHHS--



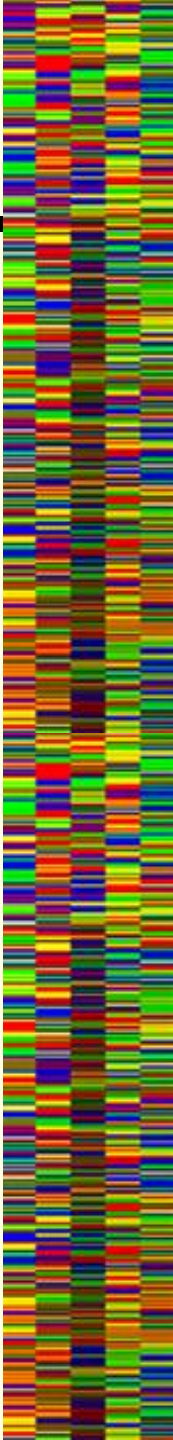
# Multiple Alignment and Trees

## Final leader-follower alignment

```
Human HBB 1 MVHLTPEEKSAVTALWG--KV--NVDEVG--GEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK 60
Rana HBB 1 MVHWTAEKAVINSVWQ--KV--DVEQDG--HEALTRLFIVYPWTQRYFSTFGDLSSPAAIAGNPK 60
Human HBA 3 LSPADKTNVKAAWG--KVG AHAGEYG--AEALERMFLSFPTTKTYFPHF-----DL SHGSAQ 55
Bovine MY 3 LSDGEWQLVLNAWG--KV--EADVAGHGQEV LIRLFTGH PETLEKFDKFKHLKTEAEMKASED 61
BrNapa HB 6 FTEKQEALVKESWEILKQ--DIPKYS--LHFFSQILEIAPA AKDMFSFLRD--TDEVPHN NPK 62
CanLi LB 1 MGAFSEKQESLVKSSWEAFKQ--NVP HHS--AVFYTLILEKAPAAQNMFSFLSNGVDPN----NPK 54

Human HBB 61 VKAHGKKVLGAFSDGLAHLDNLKG-----TFA----TLSELHCDKLHVDPENFRLLGNVLCV 114
Rana HBB 61 VHAHGKKILGAIDNAIHNLDDVKG-----TLH----DLSEEHANELHVDPENFRRLGEVLIVV 114
Human HBA 56 VKGHGKKVADALTNVAHVDDMPN-----ALS----ALSDLHAHKL RVDPVNFKLLSHCLLVT 109
Bovine MY 62 LKKGHNTVLTALGGILKKGHHEA-----EVK----HLAESHANKHKIPVKYLEFISDAIIHV 115
BrNapa HB 63 LKAHAVKVFKMTCE TAIQLRE-KGKV VVADTTLQ----YLGSVHFKSGVLDP-HFEVVKEALVRT 121
CanLi LB 55 LKAHAEKVFKMTVDSAVQL-RAKG-----EVVLADPTLGSVHVQKGVLDP-HFLVVKEALLKT 114

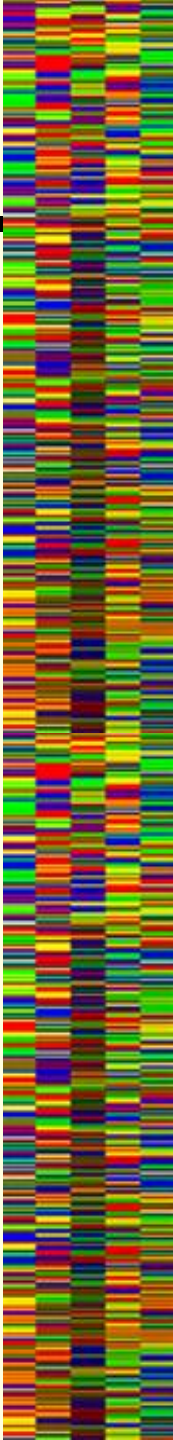
Human HBB 115 LAHFFGKEFTPPV----QAAYQKV VAGVANALAHKYH 147
Rana HBB 115 LGAKLGKAFSPQV----QHVWEKFI AVLVDALSHSYH 147
Human HBA 110 LAAHLPAEFTPAV----HASLDKFLASVSTVLT SKY 141
Bovine MY 116 LHAKHPSDFGADA----QAAMSKALELFRNDMAAQY 147
BrNapa HB 122 LKEGLGEKYNEEV----EGAWSKAYDHLALAI 149
CanLi LB 115 FKEAVGDKWNDELGN AWEVAYDELA AAIKKAMG 147
```



# Multiple Alignment and Trees

## Final leader-follower alignment

Human HBB	1	MVHLTPEEKSAVTALWG--KV--NVDEVG--GEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60	
Rana HBB	1	MVHWTAEKAVINSVWQ--KV--DVEQDG--HEALTRLFIVYPWTQRYFSTFGDLSSPAAIAGNPK	60	
Human HBA	3	LSPADKTNVKAAWG--KVG AHAGEYG--AEALERMFLSFPPTTKTYFPHF-----DLSHGSAQ	55	
Bovine MY	3	LSDGEWQLVLNAWG--KV--EADVAGHGQEV LIRLFTGHPETLEKFDKFKHLKTEAEMKASED	61	
BrNapa HB	6	FTEKQEALVKESWEILKQ--DIPKYS--LHFFSQILEIAPA AKDMFSFLRD--TDEVHNP NPK	62	
CanLi LB	1	MGAFSEKQESLVKSSWEAFKQ--NVP HHS--AVFYTLILEKAPAAQNMFSFLSNGVDPN----NPK	54	
		* * *		
Human HBB	61	VKAHGKKVLGAFSDGLAHL DNLKG-----TFA----TLSELHCDKLHVDPENFRLLGNVLCV	114	
Rana HBB	61	VHAHGKKILGAIDNAIHNLD DVKG-----TLH----DLSEEHANELHVDPENFRRLGEVLIVV	114	
Human HBA	56	VKGHGKKVADALTNVAHVDDMPN-----ALS----ALSDLHAHKLRVDPVNFKLLSHCLLVV	109	
Bovine MY	62	LKKHGNTVLTALGGILKKKG HHEA-----EVK----HLAESHANKHKIPVKYLEFISDAIIVV	115	
BrNapa HB	63	LKAHAVKVFKMTCE TAIQLRE-KGKV VVADTTLQ----YLGSVHFKSGVLDP-HFEVVKEALVRT	121	
CanLi LB	55	LKAHA EKVFKMTVDSAVQL-RAKG-----EVVLADPTLGSVHVQKGVLDP-HFLVVKEALLKT	114	
		* * *		
Human HBB	115	LAH HFGKEFTPPV----QAAYQKV VAGVANALAHKYH	147	agrees with structural
Rana HBB	115	LGAKLGKAFSPQV----QH VWEKFI AVLVDALSHSYH	147	unclear in structure
Human HBA	110	LA AHLPAEFTPAV----HASLDKFLASVSTVLT SKY	141	disagrees with structural
Bovine MY	116	LHAKHPSDFGADA----QAAMSKALELFRNDMAAQY	147	
BrNapa HB	122	LKEGLGEKYNEEV----EGAWSKAYDHLALAI----	149	
CanLi LB	115	FKEAVGDKWNDELGN AWEVAYDELA AAIKKAMG	147	



# Multiple Alignment and Trees

## Leader makes a big difference

- *Human HBB vs CanLi as Leader*

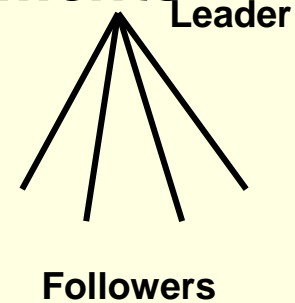
Human HBB	1	MVHLTPEEKSAVTALWG-- <b>KV</b> --NVDEVG--GEALGRLLVVY <b>PWT</b> QRFFESFGDLSTPDAVMGNPK	60
Rana HBB	1	MVHWTAEEKAVINSVWQ-- <b>KV</b> --DVEQDG--HEALTRLFIVY <b>PWT</b> QRYFSTFGDLSSPAAIAGNPK	60
Human HBA	3	LSPADKTNVKAAWG-- <b>KV</b> GAHAGEYG--AEALERMFLSF <b>PTT</b> KTYFPHF-----DLSHGSAQ	55
Bovine MY	3	LSDGEWQLVLNAWG-- <b>KV</b> --EADVAGHGQEV LIRLFTGH <b>PET</b> LEKFDKFKHLKTEAEMKASED	61
BrNapa HB	6	FTEKQEALVKESWEIL <b>KQ</b> --DIPKYS--LHFFSQILEIA <b>PAAK</b> DMFSFLRD--TDEVPHNNPK	62
CanLi LB	1	MGAFSEKQESLVKSSWEAF <b>KQ</b> --NVPHHS--AVFYTLILEKA <b>PAA</b> QNMFSFLSNGVDPN----NPK	54

<b>CanLi LB 7</b>		<b>KQESLVKSSWEAF<b>KQ</b>NVP-HHSAVFYTLILEKA<b>PAA</b>Q--NMFS-F-L--SNGVDP----N----NPK</b>	
Human HBB 7		EEKSAVTALWG-- <b>KV</b> NVD-EVGGEALGRLLVVY <b>PWT</b> Q--RFFE-S-F--GDLSTP----DAVMGNPK	
Rana HBB 4		WTAEKAVINSVWQ <b>KV</b> DVEQD-GHEAL--TRLFIVY <b>PWT</b> Q--RYFS-T-F--GDLSSP----AAIAGNPK	
Human HBA 4		SPADKTNVKAAWG <b>KV</b> GAHAG-EYGAEALERMFLSF <b>PTT</b> K--TYFPHF DL--SHG-----SAQ	
Bovine MY 1		MG---LSDGEWQLVLNAWG <b>KV</b> EADVAGHGQEV LIRLFTGH <b>PET</b> LEKFDKFK-H-L--KTEAEM----K---ASED	
BrNapa HB 1		MGEIVFTEKQEALVKESWEIL <b>KQ</b> DIP-KYSLHFFSQILEIA <b>PAAK</b> --DMFS-F-LRDTDEVPH----N----NPK	

# Multiple Alignment and Trees

## Problems revealed by leader-follower alignments

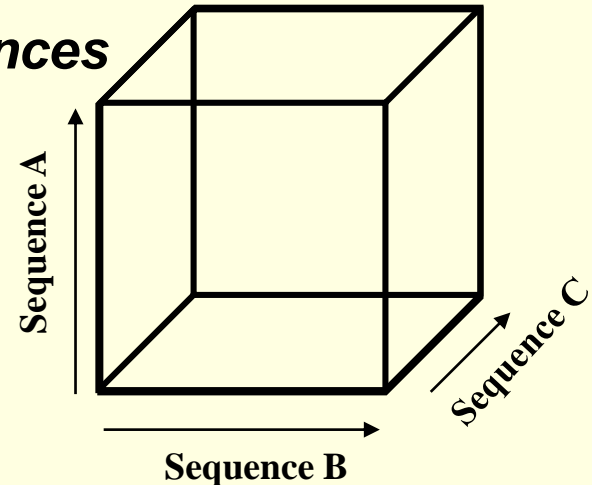
- **Alignment depends on leader**
  - Model is not biologically realistic (star tree)
- **Inaccuracy in pairwise alignments compromises multiple alignment**
- **Some alignments are better defined than others (the ones where the sequences are close are less ambiguous)**



# 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  $\epsilon$
- Will handle a small number  $\sim 10$  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>

