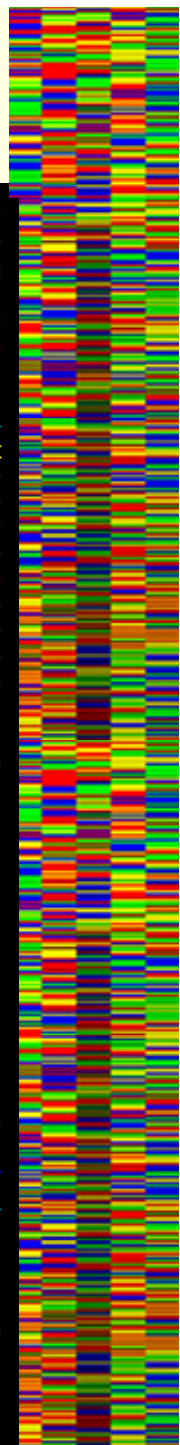


Biol 478/595 Intro to Bioinformatics

October

kallikrein	IPGGYT	CFPHSQPWQAL	LVQQRLL	CGGVLVHPKWL	TAACHLKEG	LKWLKHALG	RVEAGEQVREVVHST	IPHPEYRRSP	THL	NHDHD	IMLLEL	QSP			
protease	LVH	GGDCKTKCHPYQAAI	VTSCHI	ICGGI	TDPIWUT	TAACHCKKN	I	DFI	GKHWI	D	DFESSQFSS	WPAVTHDQVAA			
neuropsin	VLG											DHNHD	IMLLQ	RDQ	
prostate	IIIN		1	W 1	MG	Evolution & Phylogeny				Ch 5.		NDLML	IKLDES		
psa	IVG		5									RPGDD	SSHDLMLR	SEP	
complement	IVG		1	F 3	MG	Evolution & Phylogeny		hw4				NAGT	YQNDTALIE	MKKD	
factor	IVG		6					due				NAGT	YQNDTALIE	MKKD	
airway	ILG		1	M 6	MG	Evolution & Phylogeny				Handout		KSAT	HENDIALVR	LENS	
mtsp7	IVG		7									HRET	NENDIALVQL	STG	
enterokinase	IVG		1									HRRR	KONDIA	MHLFEK	
hepsin	IVG		7									HSEE	NSNDIALVH	LSSP	
prostatin	ITG		1	W8	MG	Phylogeny Statistics						QEG	SQGDIAL	QLSRP	
plasmin	VVG		8									TRKD	IALLK	LSSP	
testisin	IVG		1	F 10	MG	Phylogeny Statistics		Mp1				EN	SPYDIALVK	L	
corin	ILG		9									SRAV	VDYDTSIVEL	SED	
acrosin	IVG											NSAT	EGNDIALVE	ITPP	
neurotrypsin	IVG												SDYDIALVR	LQGP	
proteinase	IVG			M 13		October Break							ZDAENK	LNDVLLIQ	LSSP
consensus	IVG		2	W 15	DK	Comparative Genomics				Ch 11			NDIAL	L	P
			0												
	121		2	F 17	DK	Comparative Genomics		Hw5							
kallikrein	V		1									GGPLV	CNR	TL	
protease	V		2	M 20	DK	Comparative Genomics Statistics				Ch 13		GGPLV	CDG	HL	
neuropsin	V		2									GGPLV	CDG	AL	
prostate	V		2									GGPLV	CNG	YL	
psa	AELTDAVKV	MDLPTQ	EPALGTT	CYASGWGSIE	PEEFLTPKKLQ	CVDLHVISNDVCA	QVHPQKVTKFML	CAGRW	TGGKST	CSGDS	GGPLVCNG	VL			
complement	GNKKDC	ELPRSIPACV	PWSPYLQPN	DTCI	VSGWGREKDN	ERVFS	LQWGEVK	LISN	CSKFG	NRFYEKEME	CAGTY	DGSIDACKGDS	GGPLVCMDANNV	VTYVW	
factor	GNKKDC	ELPRSIPACV	PWSPYLQPN	DTCI	VSGWGREKDN	ERVFS	LQWGEVK	LISN	CSKFG	NRFYEKEME	CAGTY	DGSIDACKGDS	GGPLVCMDANNV	VTYVW	
airway	VFTFKD	IHSVCLPAATQ	NIPPGS	TAYVTG	WGAQ	QYAGH	TVPELR	QGVRIISNDV	CN	APHSYNGAIL	SGMLCAG	VPQGGVDA	CQGDSGGPLVQED	SRRLL	
mtsp7	VEFSNIV	QRVCLP	DSSIKLPPKT	SVEVTG	FSGSIVDD	GP	IQNTLR	QARVETISTD	VCN	RKDVYD	GLITP	GMLCAG	FMEGKIDACKGDS	GGPLVY	
enterokinase	VNYTDY	IQPICLPEENQ	VPPGR	NCSIAG	WGTVVYQGT	TANILQ	EADVPLLSNERCQ	QQMPEYN	ITENMICAG	YEEGGIDS	CQGDSGGPLMCQEN	NRWFL			
hepsin	LPLTEYI	IQVCLP	AAGQALVDGK	ICTVTG	WGTQYYGQ	QAGVLQ	EARVPIISNDV	CN	GADFYG	NQIKPKMFCAG	YPEGGIDAC	CQGDSGGPFV	CEDSISRT	PRWRL	
prostatin	ITFSRYI	IRPCLPAANAS	FPNGL	HCTVTG	WGWAPS	SVSLLTPKPLQ	LEVPLISRET	CNCLYNIDAK	KEEPHFVQED	MVCAG	YVEGGKDAC	CQGDSGGPLSCP	VEGLWYL		
plasmin	ITFTK	QVCLP	AVVYVDP	PLVWV	QVWV	QVWV	QVWV	QVWV	QVWV	QVWV	QVWV	QVWV	QVWV	QVWV	
testisin	ISFTGYV	RPVCLP	NEQWLEPDT	YCYITG	WGHMGNK	PFK	LQEGEVRIISLE	HCQSYFDMKT	ITTRMICAG	YESGTVDS	CMGDSGGPLVCEK	GERWTL			
corin	ISCGRF	IGCLPHL	KAGLPRGS	QSCWVAG	WGYE	EKAAPRPS	SILMEARVD	LIDL	LCN	STQWYNGRV	QPTNV	CAG	YPV	GKIDTCQGDSGGPLMCKDSK	
acrosin	EEQCAR	FSSHVLPACL	PLWRE	RPKTAS	NCYITG	WGTG	RAYSR	TLQAAIPL	LKRFCE	ERYKRF	TGRMLCAG	NLHEH	KRVDS	CQGDSGGPLM	
neurotrypsin	ANLSAS	VATVQL	QQDQ	PHGTQ	CLAMG	WRVGAH	DP	PAQVL	QELNVT	VTF	FCR	PHNICT	FVPRR	KAGI	
proteinase		I P CLP		C V GWG		LQ A V IS	C		Y I M	CAG	GG D	CQGDSGGPLV	C	W L	
consensus															
	241														
kallikrein	YGLVSWGD	FPCGQ	DRP	GVYTRV	SRYVLW	IRETIRKYET	QQQKWLK	GPQ							
protease	RGLVSWGN	IPC	GSKEK	GVYTNV	CRYTNWI	QKTIQAK									
neuropsin	QGITSWGSD	PCGRSD	KP	GVYTNV	CRYLD	WIKKLI	GSKG								
prostate	QGLV	SFGKAP	CGQ	GVYTNL	CKFT	EWIEKTV	QAS								

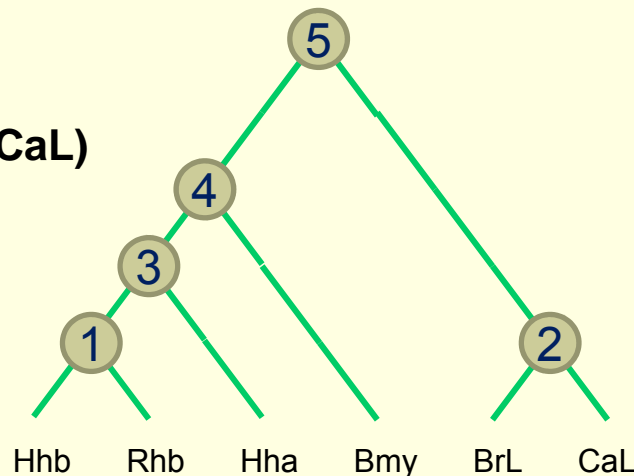
Homework 4 due today



Multiple Alignment and Trees

Progressive Alignment

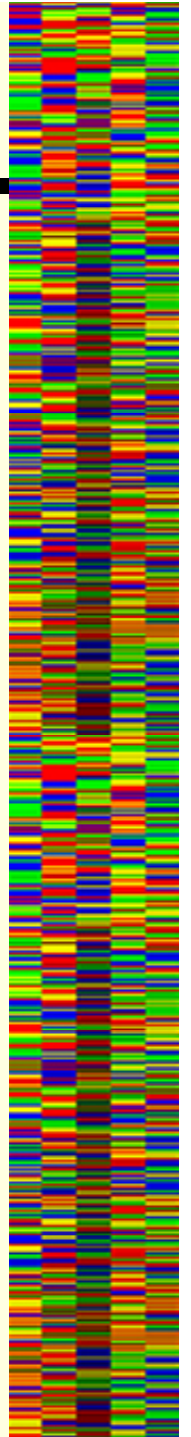
- Practical multiple alignments are made using a progressive alignment procedure.
- The alignment is constructed by adding one sequences to a growing alignment in order of similarity (closest first)
- Order of addition is determined by a guide tree, for example
 1. Align Hhb and Rhb
 2. Align BrL and CaL
 3. Align (Hhb,Rhb) and Hha
 4. Align ((Hhb,Rhb),Hha) and Bmv
 5. Align (((Hhb,Rhb),Hha),Bmy) and (BrL,CaL)



Multiple Alignment and Trees

Progressive Alignments

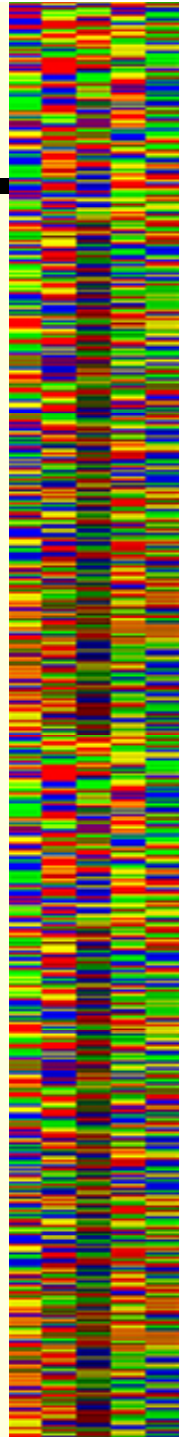
- All multiple alignment methods, including multi-dimensional dynamic programming, assume some kind of tree underlying the data!
- Tree construction methods use multiple alignments as data. The realized tree will always tend to look like the guide tree or implicit tree assumed when constructing the multiple alignment



Multiple Alignment and Trees

Progressive Alignment

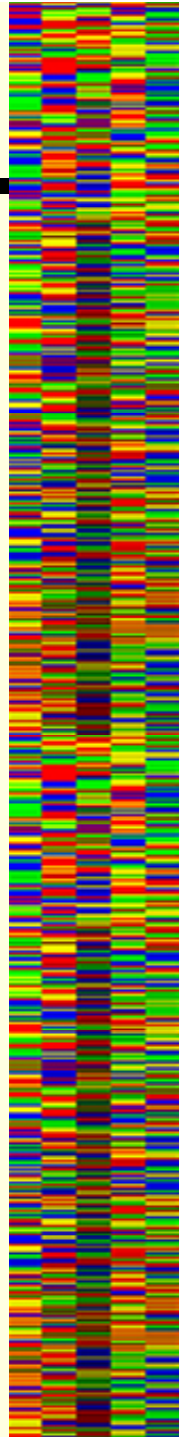
- **Progressive alignments are fast enough to allow hundreds or even thousands of sequences to be aligned**
- **Basic progressive alignment procedure**
 - **Determine distances between sequences**
 - **Use a distance-based method to construct a guide tree for the sequences (UPGMA, Neighbor Joining)**
 - **Add sequences to the growing alignment using the order given by the tree**



Multiple Alignment and Trees

Progressive Alignment

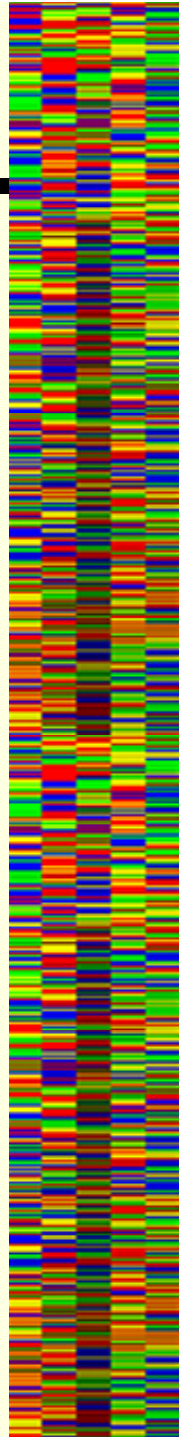
- **General problems with progressive alignments**
 - **Solution is not guaranteed to be optimal. The greedy strategy used in progressive alignment is highly likely to be “trapped” in a local optimum.**
 - **Error in early stages propagate through to final stage - there is no error correction possible**
 - **Choices of alignment parameters are important, but appropriate settings are difficult to determine**



Multiple Alignment and Trees

Progressive alignment – Guide tree

- Add sequences, one at a time, to closest other sequences according to the *guide tree*
 - Closest sequences have best information and most reliable alignment
- Alignments are frozen once they are made.
- Numbers of residues at each position are tabulated in a frequency matrix (sometimes called a profile)
- Scores are calculated between aligned positions using a scoring table, e.g., PAM or BLOSUM



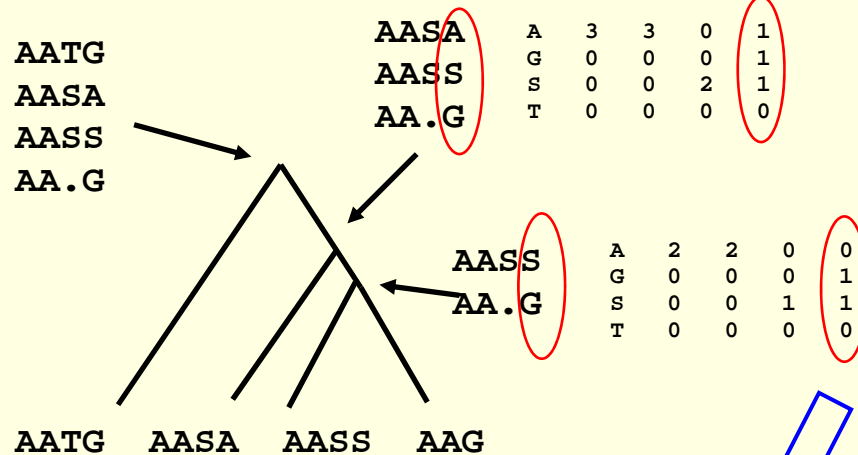
Multiple Alignment and Trees

Progressive alignment

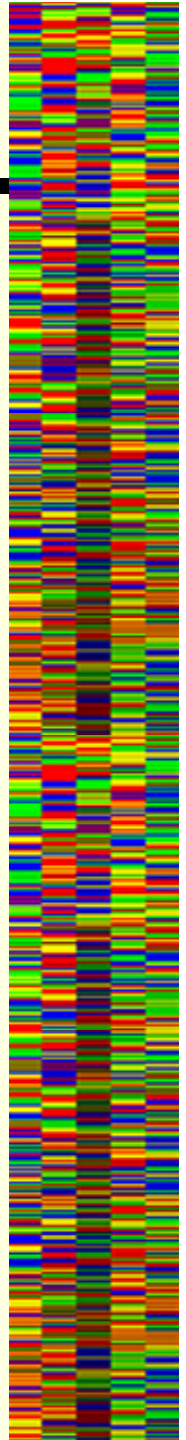
H	-4.0	1.0	-3.0	10.0
G	3.0	-2.0	6.0	
C	-2.0	14.0		
A	8.0			
	A	C	G	H

Scoring table

A	4	4	0	1
G	0	0	0	2
S	0	0	2	1
T	0	0	1	0



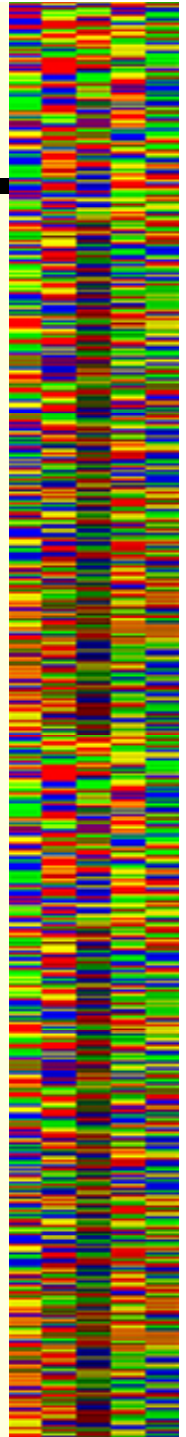
$$S_{ij} = 1 \times G:A + 1 \times S:A$$



Multiple Alignment and Trees

Progressive alignments

- **What goes wrong**
 - alignment errors in early stages are frozen into alignment
 - just as with pairwise alignments in leader-follower alignment, alignments can be erroneous where the similarity is weakest
 - poorest alignments tend to be near
 - *insertions and deletions*
 - *duplications*
 - *repetitive sequence*



Multiple Alignment and Trees

Progressive alignments

```
Human HBB 1 MVHLTPEEKSAVTALWGKVV--DEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK 60
Rana HBB 1 MVHWTAEKAVINSVWQKVDV--EQDGHEALTRLFIVYPWTQRYFSTFGDLSSPAAIAGNPK 60
Human HBA 3 LSPADKTNVKAAWGKVGAAHAGEYGAELERMFLSFPPTTKTYFPHFDLSH-----GSAQ 55
Bovine MY 3 LSDGEWQLVLNAWGKVEADVAGHGQEV LIRLFTGHPETLEKFDKFKHLKTEAEMKASED 61
BrNapa HB 6 FTEKQEALVKESWEILKQDIPKYSLHFFSQILEIAPAAKDMFSFLRD--TDEVHNP NPK 62
CanLi LB 1 MGA FSEKQESLVKSSWEAFKQNPVPHSAVFYTLILEKAPAAQNMFSFL----SNGVDPNNPK 54
```

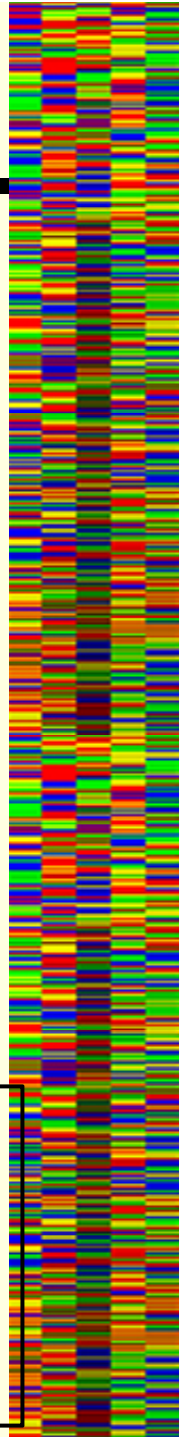
```
Human HBB 61 VKAHGKKVLGAFSDGLAHL DNLKG-----TFATLSELHCDKLHVDPENFRLLGNVLVCV 114
Rana HBB 61 VHAHGKKILGAIDNAIHNLD DVGK-----TLHDLSEEHANELHVDPENFRRLGEVLIVV 114
Human HBA 56 VKGHGKKVADALTNVAHVDDMPN-----ALSALSDLHAHKL RVDPNFKLLSHCLLV T 109
Bovine MY 62 LKKHGNVLTALGGI LKKKGHHEA-----EVKHLAESHANKHKIPVKYLEFISDAI IHV 115
BrNapa HB 63 LKAHAVKVFKMTCE TAIQLREK GKVVADTTLQYLG SVHF KSGVLD P-HFEVVKEALVRT 121
CanLi LB 55 LKAHA EKVFKMTVDSAVQLRAKGEVVLADPTL---GSVHVQKGVLD P-HFLVVKEALLKT 114
```

```
Human HBB 115 LAH HFGKEFTPPVQAAYQKV VAGVANALAHKYH 147 Structural alignment
Rana HBB 115 LGAKLGKAFSPQVQHVWEKFIAVLVDALSHSYH 147 shaded regions show alpha
Human HBA 110 LAAHLPAEFTPAVHASLDKFLASVSTVLT SKYR 141 helix
Bovine MY 116 LHAKHPSDFGADAQAAMSKALELFRNDMAAQYKVLGFHG 147
BrNapa HB 122 LKEGLGEKYNEEVEGAWSKAYDHLALAIKAEMKQEDSQKP 149
CanLi LB 115 FKEAVGDKW NDELGNAWEVAYDELA AAIKKAMGSA 147
```

```
Human HBB 1 MVHLTPEEKSAVTALWG--KVVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN 58
CANLI LB 1 MGA FSEKQESLVKSSWEAFKQNPVPHSAVFYTLILEKAPAAQNMFSFLSNGVDPN----N 56

Human HBB 59 PKVKAHGKKVLGAFSDGLAHL DNLKGTFA----TLSELHCDKLHVDPENFRLLGNVLVCV 114
CANLI LB 57 PKLKAHA EKVFKMTVDSAVQL-RAKGEVVLADPTL GSVHVQKGVLD P-HFLVVKEALLKT 114

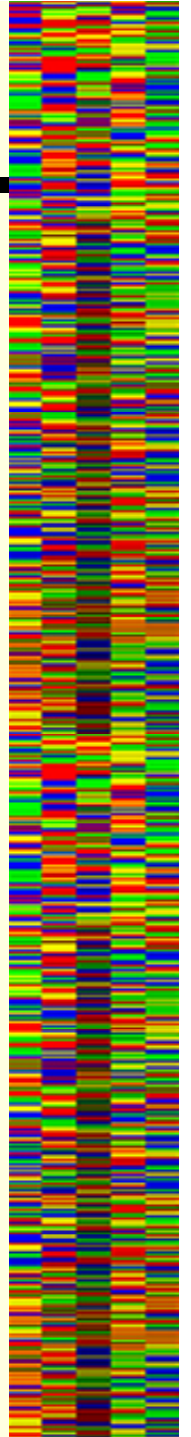
Human HBB 115 LAH HFGKEFTPPV----QAAYQKV VAGVANALA 143 DP pairwise alignment
CANLI LB 115 FKEAVGDKW NDELGNAWEVAYDELA AAIKKAMG 147
```



Multiple Alignments and Trees

Progressive alignments

- **No computational alignment is error free**
 - Inferences based on alignments are only as good as the alignment
 - Using automated alignments to build trees guarantees disaster – the guide tree tends to be reproduced because the alignment has been optimized to fit that tree
- **Manual improvements are necessary**
 - Shift residues from one side of gap to the other
 - “Balance out” indels of equal length
 - Use all sequences to find improved positions for specific residues

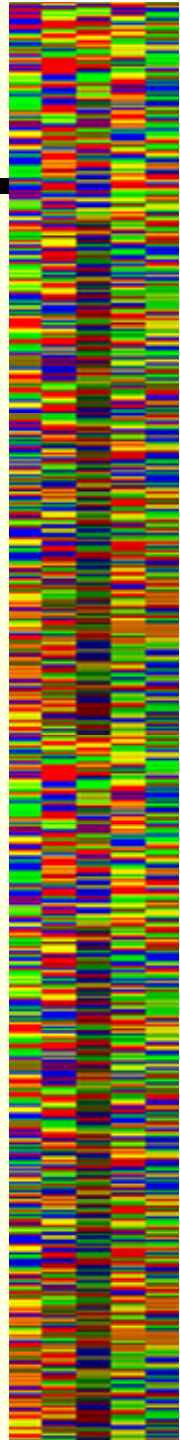


Multiple Alignments and Trees

Manual Correction

kallikrein	LPGGYTCFPHSQPWQAAL	LVQGRLLCGGVLVHPKWLTAACHLKY	LKVVYLGKHALG RVEAGEQVREVVHSIPHPYRRSPHLL	NRNDTMLLELQSP	
protease	LVHGGPCDKTSHPYQAAL	YTSCHLLCGGVLHPLWLTAAHCKFPN	LQVFLGKHNLR QRESSQEQSSVRAVHPDYDAA	SHDQIMLLRLARP	
neuropsin	VLGGHECOPHSQPWQAAL	FQGGQLLGGVLVGGNWLTAACHKPK	YTVRLGDHSLQ NKDGPEQELPVVQSIPHPCYNSSDVE	DHMHMLLQLRDQ	
prostase	IINGEDCSPHSQPWQAAL	VMENELFCGVLVHPQWVLSAAHCQNS	YTIGLGLHSLEADQEPGSMVEASLSVRHPEYMRPLLA	NDMLIKLDES	
psa	IVGGWECCKHSQPWQVLV	ASRRRAVCGGVLVHPQWVLTAAHCRNK	SVILLGRHSLFHPEDTG QVFQVSHSFPHPLYDMSELLKNRF	RPDDSSHDMLLRSEF	
complement	IVGGKRAQLGDLPWQVAIK	DASGITCGGIYIGGCWILTAAHCR	RA SKTHRYQIWTTVVDWIHPDLKRIVIE	YVDRIFHENY NAGT YQND ALTEMKKD	
factor	IVGGKRAQLGDLPWQVAIK	DASGITCGGIYIGGCWILTAAHCR	RA SKTHRYQIWTTVVDWIHPDLKRIVIE	YVDRIFHENY NAGT YQND ALLEBKKD	
airway	ILGGTEAEEGSPWQVSLRL	NNAHHCGGSLINNMWILTAAHCR	SN SNPRDWIATSGI	STTFPKLRVRNLIHNNY KSAT HEND ALVRENS	
mtsp7	IVQGRETAMEGEPWPQASLQLI	GSQHCGGASLISNTWLLTAACHW	KN KDPQTWIATFGA	TIITPPAVKRNVRKILHENY HRET NEND ALVQLSTG	
enterokinase	IVGGSNAKEGAWPVVWGLY Y	GGRLCGASLVSSDWLVSAACHYGRN	LEP SKWTAIILGLHNS	NLTSPTVPRIDEIVINPHY NRRR KOND AMMHCEFK	
hepsin	IVGGRDTSLRWPWQVSLR Y	DGAHLGGSLISGDWLTAAHCPERN	RVLSRWRFAGAVAQA	SPHGLQLGVQAVVYHGGYLPFRDPNSEE NSND ALVHLSPP	
prostasin	ITGSSAVAGQWPWQVSI TY	EGVHVCGGSLVSEQWVLSAAHC	PSEH HKEA YEVKLGALD	SYSEDAKVS TLKQIHPHSYL QEG SQGD ALLQLSRP	
plasmin	VVGGCVAPHSPWPQVSLRTR	FGMHFCGGTLISPEWLTAAHC	ETYS SYPSSYKVLGAHQEV	NL EPHGQIEVSRIFLEP TRKD ALLKLSPP	
testisin	IVGGDAELGRWPWQVSLRLW	DS HVCVGLLSHRWALTAACH	ETYS DLSDPGGMVQFG QLT	SMPSEWSLQAYYTRYFVSNLYLSPRYLGN	SPYD ALVKLSAP
corin	ILGGRTSRPGRWPWQVSLQSE	PSCHICGCVLIARWLTVAHCEGR	ENA AVWRVVLGINLH	H P SVFQTRFVKIILHPRY	SRAV VDYD SILVELSD
acrosin	IVGGKAAQHGAWPWPVSLQIFTYNSHRVYHT	CGGSLISRWLTAAHC	VGKNNVHD	WRVLPFAKEITYGNNKPKAPV	QERYVEKIIHKEY NSAT EGND ALVEITPP
neurotrypsin	IIGGKNSLRGGWPWQVSLRLKSSHGDRLL	CGVTLSSCWLTAAHC	KRYGNSTRSYAVRVG	DYHTLVPVEFEELGVQQIVIHREYRPR	SDYD ALVRLQGP
proteinase	IVGGHEAQPHSRPYMASLQMRGNP	ESHFCGGTLIHPFVLTAAHCRD	IPQRVNWVLGAHVRTQ	EPTQQHPSVAQVFLNN	YDAENK NLLIQLSSP
consensus	IVGG A G WPWQVSL	G H CGG L WLTAAHCF	W V I G H	V I H Y	ADIAL L P

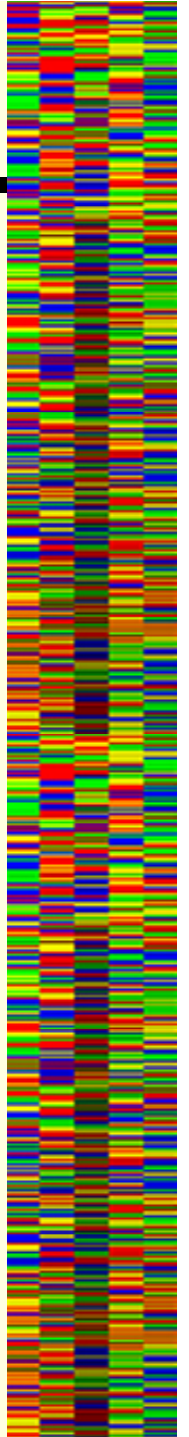
kallikrein	VQLTGYIQT	LPLSHNNRLTPGTCRVSQWGTIT	SPQVNPYKTLQCANIQRSDEE	CR	QVYPEKITDN	LCAGTK	EGEKDSCGDSGGPLVCNR	TL	
protease	AKLSELIQP	LPLERDCSANT	TSCHILGWGKTA DG	DPPDTIQCAIHLVSRDEE	CE	HAYPGQITQNL	LCAGDE	KYEKDSCQDSSGGPLVCGD	HL
neuropsin	ASLGSKVKP	ISLADHCT	QPGGKCTVSGWGTV	SPRENFDTLNCAEVKIFPKKCE		DAYPGQITDGL	VCAGSS	KHADTCQDSSGGPLVCGD	AL
prostase	VSESDTIRS	ISIASQC	PTAGNSCLVSGWGLLA	NGR MPTVQLCVNVSVVSEEVCS		KLYDPIYHPS	FCAGGG	HDKDCSNGDSGGPLICNG	YL
psa	AELTDVAVK	MDLPTQ	EPALGTTICVYASGWSIE	PEEFLTPKKLQCVDLHVVISNDVCA		QVHPQKVTKEF	LCAGRW	TGKSTCSGDSGGPLVCNG	VL
complement	GNKKDCELPRSPACVPSYLFQPN	DTCIVSGWREKONERVS	LQWGEVKLISN	CSKPYG	NRFYEKE	ECAGTY	DGSDACKGDSGGPLVCDANNVTYVW		
factor	GNKKDCELPRSPACVPSYLFQPN	DTCIVSGWREKONERVS	LQWGEVKLISN	CSKPYG	NRFYEKE	ECAGTY	DGSDACKGDSGGPLVCDANNVTYVW		
airway	VITFKDIHSVCLPAATQNIHPPGS	TAYVTGWAQAYACH	TVPBLRQGVRIISNDV	CN	APHSYNGAILSG	LCAG	VPQGVADACQDSSGGPLVQEDSRRRL	WFI	
mtsp7	VEFSNIVQRVCLPSSIKPKPT	SVFTGFGSIVDDGP	IQNTLRQAVETISTDY	CN	RKQVYDGLITPG	LCAG	FMEGKIDACKGDSGGPLVYDNHDI	WYI	
enterokinase	VNYTDYIQPICLPEENQVFPPER	NCSTAGWTVVYQGT	TANILQEADVPLLSNERCQ		QQMPEYNI	TERLCAG	YEEGSDSCQDSSGGPLMCCQEN	NRWF	
hepsin	LPLTEYIQPICLPAAGQALVDGK	ICTVTGWTNTQYQGQ	QAGVLQEARVPIISNDV	CN	GADFYGNQ	IKPKFCAG	YPEGFDACQDSSGGPFVCDSTSRIPRWR		
prostasin	ITFSRYIRPICLPAANASFPNGL	HCTVTGWHVAPSLSLLTPKLLQOLEVPLISRET	CNCLYNDAKPEEPHFVQEDV	CAG	YVEGKDACQDSSGGPLSCP		VEGLWYL		
plasmin	AVITDKVIPACLPSFNIVADRT	ECFITGWGET	QGTFGAGLLKEAQLPVIENK	CNRYEFL	NGRQST	LCAG	HLAGFDSCQDSSGGPLVCFE	KOKYIL	
testisin	VITYTKHIQPICLQASTFEFENRT	DCWVTGWTNTQYQGEAL	PSPHTLQEVQVAIINNS	CNHLFL	KYSFRKDFG	VCAG	NAQEKDACQDSSGGPLACMKN	GLWYQ	
corin	ISCTGYRVPVCLPNEQWLEPDT	YCYITGWHHGNK	PFK LQEGEVRIISLEHCQSYFDMKT		ITTR	LCAG	YESGVDSCMDSGGPLVCEKP	GGRWTL	
acrosin	ISCGRIQPGCLPHKAGLPRGSQ	SCWVAGWGYIEEKAPRS	SILMEARVDLIDLCN		STQWYNGRV	QPFCAG	YPVGIKDTQDSSGGPLMCKDSK	ESAYVY	
neurotrypsin	EEQCARESSHVLPACLPLWRERPQKTASH	CYITGWDTC	RAYSRTLQQAAPLIPKRF	CE	ERYKRF	TGR	LCAGNHEHRRDSCQDSSGGPLMCCRP	GESWAA	
proteinase	ANLSAS	VATVQLPQQDQVPHGTQ	CLAMGWRVGAHP	PAQVQLQELNVTVVT	FFCR	PHICTFV	PRRKAGI	CFGDSGGPLICDGIIGIDSFVI	
consensus	I P C L P	C V G W G	L Q A V I S C		Y I R C A G		G D C Q D S G G P L V C	W L	



Multiple Alignment and Trees

Progressive Alignment - ClustalW

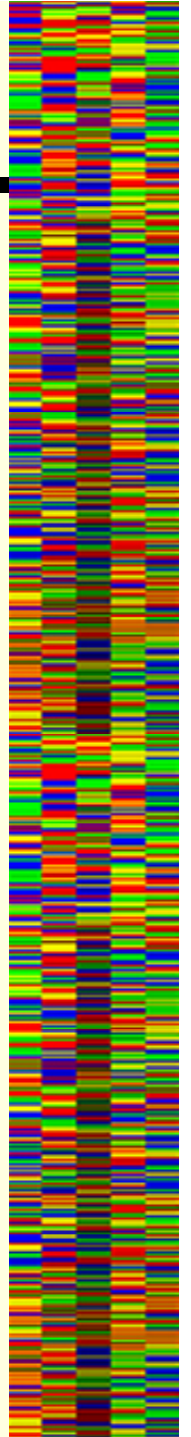
- Most widely used program for multiple alignment
- Incorporates many heuristics in an attempt to improve alignments
- Deals well with length differences
- Progressive alignment using neighbor joining method for *guide tree*
- W stands for weighting due to position specific weighting of parameters used in alignments
- Calculate initial set of pairwise dynamic programming alignments to get distances
- Use distance matrix to calculate neighbor joining *guide tree*
 - Correct distances using Kimura correction
$$K_{aa} = -\ln(1 - D - D^2/5)$$
 $D = \text{observed differences/site}$
and interpolation from Dayhoff Model
- Use scoring matrices for different distances as tree grows
- Use position specific gap penalties



Multiple Alignment and Trees

Progressive Alignment – ClustalW

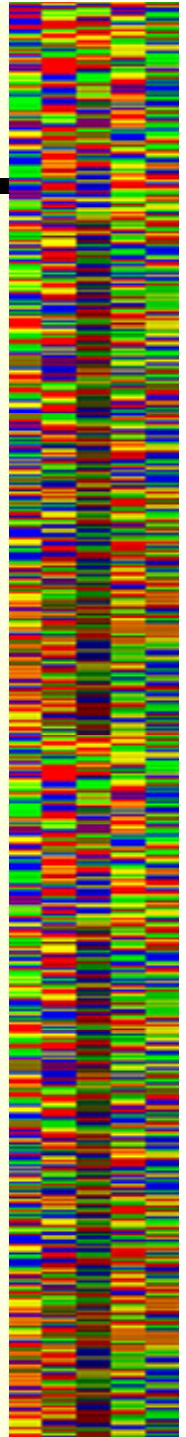
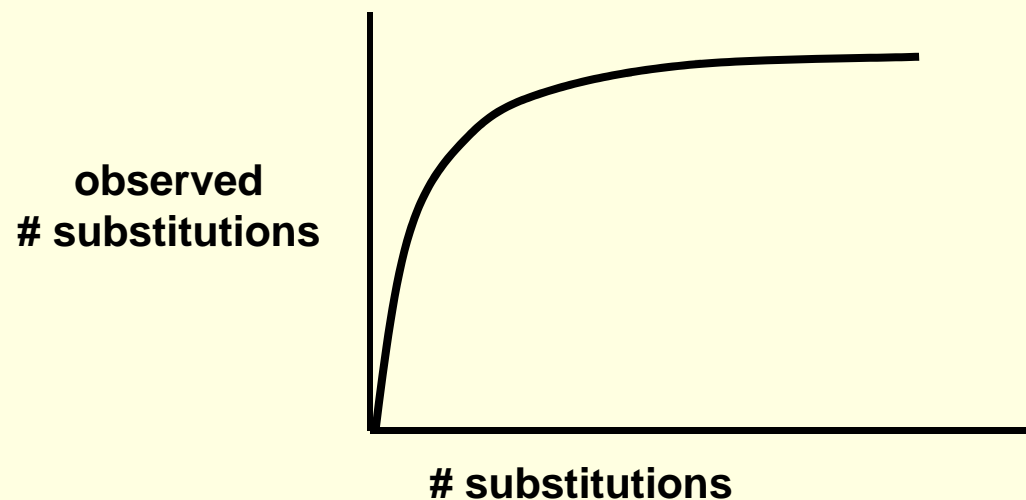
- **Features of ClustalW**
 - **Fast (word matching) and slow (DP) methods for initial distances**
 - **Multiple scoring tables to match evolutionary distance between sequences being aligned**
 - **Position dependent gap penalties**
 - **Sequence weighting to correct counts for closely related sequences**
 - **Tree based on final alignment can be produced with gapped regions omitted**
 - **Bootstrapping to evaluate final tree**
- **Drawbacks**
 - **Tree/alignment is not global optimum**
 - **Errors introduced early in alignment procedure can't be fixed**
 - **Many (heuristic) parameters**



Multiple Alignment and Trees

Scoring Matrix Distance Correction

- 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

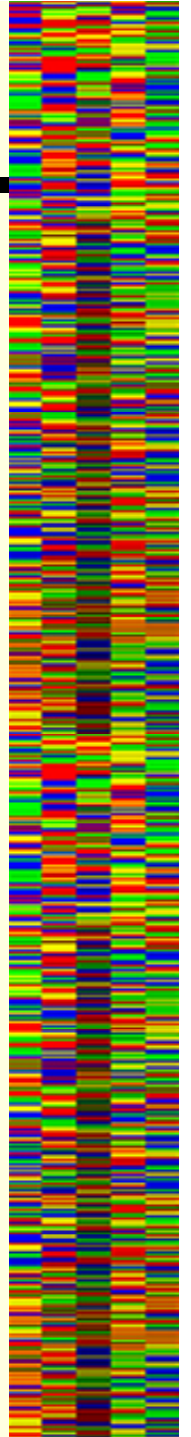
Progressive Alignment – ClustalW

- Dayhoff table should be adjusted for distance between sequences so that target frequencies agree with sequences
- ClustalW interpolates from number of differences to correct PAM distance.
- This scoring file is used for comparing the sequences and changes for each node in the tree.
- Gap penalties need to be adjusted for scoring table and for differences in length

$$GOP = A * B * \{GOP_{default} + \log[\min(N,M)]\}$$

$$GEP = GEP_{default} * [1.0 + |\log(N/M)|]$$

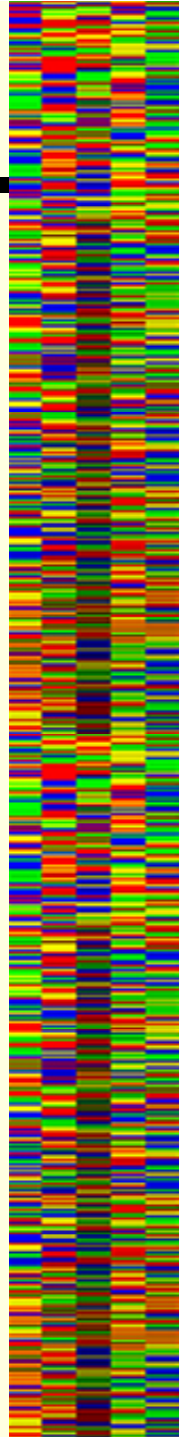
- where N and M are the lengths of the sequences, A is the average value of a mismatch in the scoring matrix, and B is the percent similarity between the two sequences.



Multiple Alignment and Trees

Progressive Alignment - ClustalW

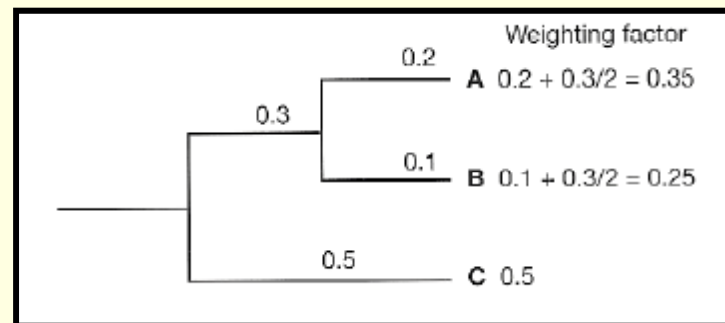
- Variable gap penalties – accounts for preferences of gaps for surface regions of proteins
- Try to guess surface loops
 - Use lower gap penalties where gaps occur in one or more sequences
 - $GOP = GOP_{default} * 0.3 * (W/N)$
 - where W is the number of sequences without gaps and N is the number of sequences (range $0.3 \rightarrow 0.0 * GOP$)
 - Increase penalty adjacent to gaps
 - $GOP = GOP_{default} * (2 + (8-D^2)/8)$ where D is the distance from the gap (range $1.0 \rightarrow 3.0 * GOP$)
 - Reduce gap penalties in stretches of 5 or more hydrophilic residues (DEGKNQPRS)
 - $GOP = GOP/3$
 - Increase or decrease gap penalties depending on residue specific gap propensity
 - $GOP = GOP * P_i$ where P_i is the gap propensity of residue i



Multiple Alignment and Trees

Progressive Alignment – ClustalW

- Sequence weighting – corrects for uneven sampling of sequences
- Guide trees are used to weight sequences
- Weights are related to distance from root to each leaf node (sequence)
 - Each sequence gets weight equal to the branch leading to it, plus $1/n$ for branches shared with others



Multiple Alignment and Trees

ClustalW

Adjust counts based on sequence weights

H	-4.0	1.0	-3.0	10.0
G	3.0	-2.0	6.0	
C	-2.0	14.0		
A	8.0			
	A	C	G	H

Scoring table

A	4	4	0	1
G	0	0	0	2
S	0	0	2	1
T	0	0	1	0

AATG
AASA
AASS
AA.G

AASA
AASS
AA.G

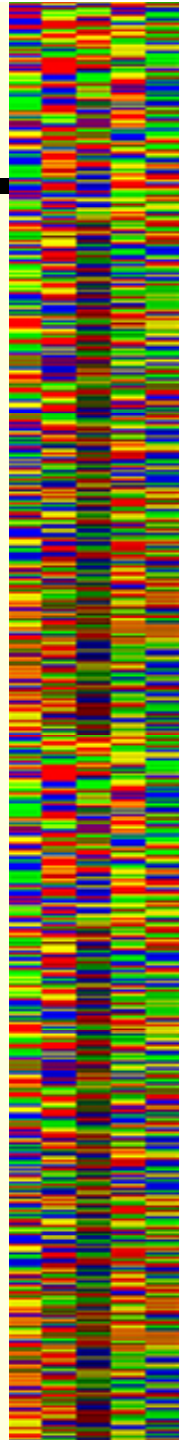
A	3	3	0	1
G	0	0	0	1
S	0	0	2	1
T	0	0	0	0

AASS
AA.G

A	2	2	0	0
G	0	0	0	1
S	0	0	1	1
T	0	0	0	0

AATG AASA AASS AAG

$$S_{ij} = w_g w_a \times G:A + w_s w_a \times S:A$$



Multiple Alignment and Trees

ClustalW

CLUSTAL W (1.8) multiple sequence alignment

```

HBB3_RANCA    MVH--WTAEKAVINSVWQKVDVE--QDGHEALTRLFIVYPWTQRYFSTFGDLSSPAAI A
HBB_HUMAN     MVH--LTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFESFGDLSTPDAVM
HBA_HUMAN     MV---LSPADKTNVKAAGWGVGAHAGEYGAELERMFLSFPPTTKTYPPHF--DLS-----H
HBL2_BRANA    MG EIVFTEKQEQEALVKESWEILKQDIPKYSLHFFSQILEIAPA AKDMFSFLRDTD--EVPH
LGB_CANLI     MG--AFSEKQESLVKSSWEAFKQNVPHHS AVFYTLILEKAPAAQNMFSFLSN----GVDP
MYG_BOVIN     MG---LSDGEWQLVLNAWGKVEADVAGHGQEV LIRLFTGHPETLEKFDKFKHLKTEAEMK
*      :      :      :      *      .      .      .      :      :      *      :      *      :
  
```

```

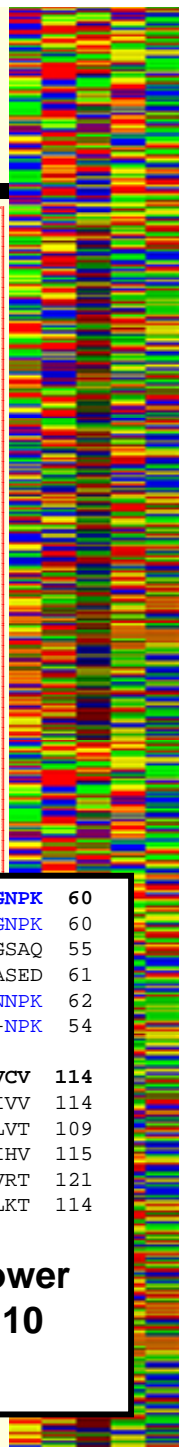
HBB3_RANCA    GNPKVHANGKKILGAI DNAIHNLDDVKGTLHD-----LSEEHANELHVDPENFRRLG EV
HBB_HUMAN     GNPKVKAHGKKVLGAFSDGLAHL DNLKGTTFAT-----LSELHCDKLHVDPENFRLLGNV
HBA_HUMAN     GSAQVKGHGKQVADALTN AVAHVDDMPNALS A-----LSDLHAHKL RVDPVNFKLLSHC
HBL2_BRANA    NNPKLKAHAVKVFKMT CETAIQLREKGVVADTTLQYLG SVHFKSGVLDP-HFEVVKEA
LGB_CANLI     NNPKLKAHA EKVFKMTVDSAVQLRAKGEVVLADPT---LGSVHVQKGVLDP-HFLVVKEA
MYG_BOVIN     ASEDLKKHGNTVLTALGGIL KKKGHHEAEVKH-----LAESHANKHKIPVKYLEPISDA
. . . : * . . :      :      . . . * . . :      :      :
  
```

```

HBB3_RANCA    LIVVLGAKLGKAFSPQVQHV
HBB_HUMAN     LVCVLAHHFGKEFTPPVQAAY
HBA_HUMAN     LLVTLAAHLPAEFTPAVHASI
HBL2_BRANA    LVRTLKEGLG EKYN EEEV EGA
LGB_CANLI     LLKTFKEAVGDKWDELGNAM
MYG_BOVIN     IIHVLHAKHPSDFGADAQAAM
:: :      :
  
```

Human HBB	1	MVHLTPEEKSAVTALWG--KV--NVDEVG--GEALGRLLVVYEW TQRFESFGDLSTPDAVMGNPK	60
Rana HBB	1	MVHWTAEKAVINSVWQ--KV--DVEQDG--HEALTRLFIVYEW TQRYFSTFGDLSSPAAIAGNPK	60
Human HBA	3	LSPADKTNVKAAGW--KVG AHAGEYGAELERMFLSFPPTTKTYPPHF-----DLSHGSAQ	55
Bovine MY	3	LSDGEWQLVLNAWG--KV--EADVAGHGQEV LIRLFTGHPETLEKFDKFKHLKTEAEMKASED	61
BrNapa HB	6	FTEKQEQEALVKESWEILKQ--DIPKYS--LHFFSQILEIAPA AKDMFSFLRDTD--TDEVPHNPK	62
CanLi LB	1	MGAFSEKQESLVKSSWEAFKQ--NVP HHS--AVFYTLILEKAPAAQNMFSFLSNGVDPN----NPK	54
Human HBB	61	VKAHGKKVLGAFSDGLAHL DNLKGTTFAT-----TFA----TLSELHCDKLHVDPENFRLLGNVLCV	114
Rana HBB	61	VHAHGKKILGAI DNAIHNLDDVKG-----TLH----DLSEEHANELHVDPENFRRLG EVLIVV	114
Human HBA	56	VKGHGKKVADALTN AVAHVDDMPN-----ALS----ALSDLHAHKL RVDPVNFKLLSHCLLVT	109
Bovine MY	62	LKKHGNTVLTALGGIL KKKGHHEA-----EVK----HLAESHANKHKIPVKYLEPISDAIHV	115
BrNapa HB	63	LKAHAVKVFKMT CETAIQLRE-KGKVVADTTLQ----YLG SVHFKSGVLDP-HFEVVKEALVRT	121
CanLi LB	55	LKAHA EKVFKMTVDSAVQL-RAKGEVVLADPTLGSVHVQKGVLDP-HFLVVKEALLKT	114
Human HBB	115	LAHHFGKEFTPPV----QAAYQKV VAGVANALAHKYH	147
Rana HBB	115	LGAKLGKAFSPQV----QHVWEKFI AVLVDALSHSYH	147
Human HBA	110	LAHLPAEFTPAV----HASLDKFLASVSTVLT SKY	141
Bovine MY	116	LHAKHPSDFGADA----QAAMSKALELFRNDMAAQY	147
BrNapa HB	122	LKEGLGEKYNEEV----EGAWSKAYDHLALAI	149
CanLi LB	115	FKEAVGDKWDELGN AVEVAYDELA AAIK KAMG	147

Leader-Follower
from slide 2.10



Multiple Alignment and Trees

ClustalW

- Final tree
 - Often use 2. Exclude positions with gaps to omit regions with poorest alignment (but watch out for fragment sequences).
 - Omits all columns where any sequence has a gap
 - Usually bootstrap if you want to know confidence in tree

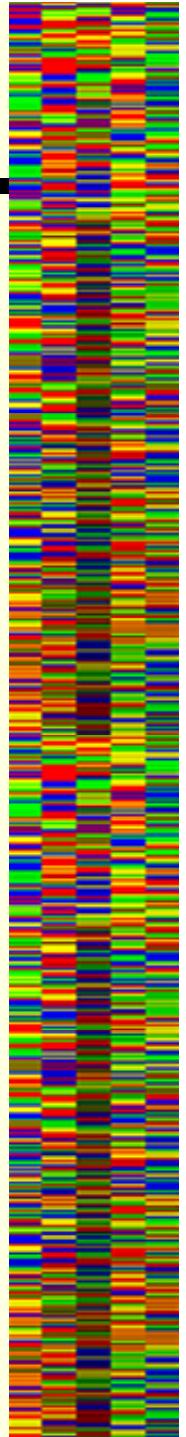
```
Your choice: 4

***** PHYLOGENETIC TREE MENU *****

1.  Input an alignment
2.  Exclude positions with gaps?      = OFF
3.  Correct for multiple substitutions? = OFF
4.  Draw tree now
5.  Bootstrap tree
6.  Output format options

S.  Execute a system command
H.  HELP
or press [RETURN] to go back to main menu

Your choice: █
```



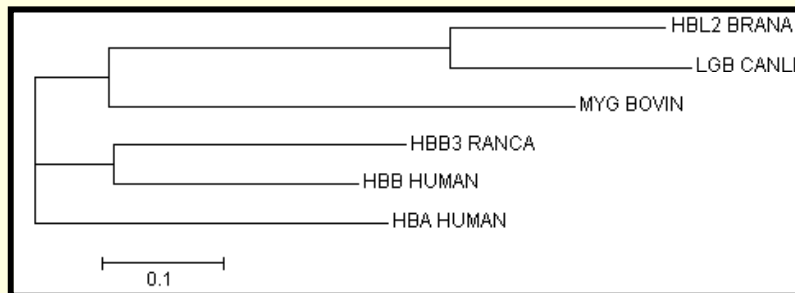
Multiple Alignment and Trees

ClustalW

- Trees - Newick format

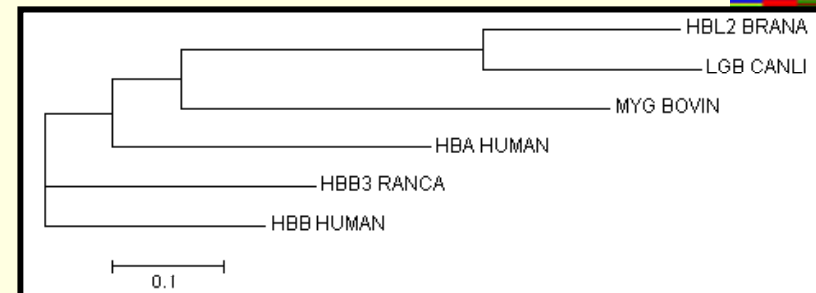
```
(
(
(
HBL2_BRANA:0.17743,
LGB_CANLI:0.19841)
:0.28051,
MYG_BOVIN:0.38318)
:0.05975,
(
HBB3_RANCA:0.24072,
HBB_HUMAN:0.20145)
:0.06376,
HBA_HUMAN:0.28909);
```

Guide Tree



```
(
(
(
HBL2_BRANA:0.17872,
LGB_CANLI:0.19712)
:0.27231,
MYG_BOVIN:0.38788)
:0.06196,
HBA_HUMAN:0.28747)
:0.05930,
HBB3_RANCA:0.24405,
HBB_HUMAN:0.19813);
```

Final Tree



Multiple Alignment and Trees

ClustalW

- **Bootstrap**
 - sample columns from final alignment to make new alignment
 - make tree
 - count how many times the branching pattern seen in first tree reoccurs
 - Measures how well the data support the tree (not whether the tree is correct)
 - Same tree from previous page in radial format with bootstrap values (1000 trials)

