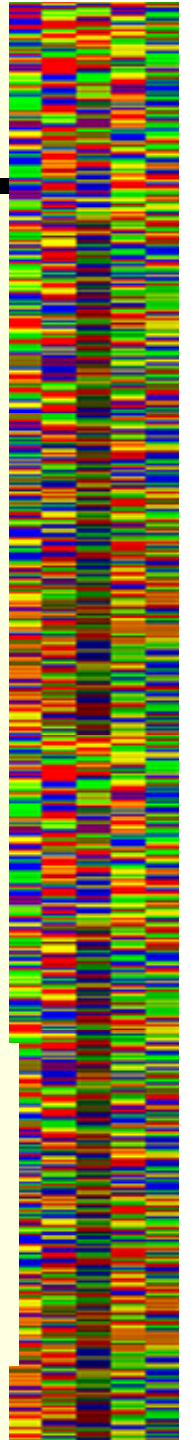


Biol 478/595 Intro to Bioinformatics

October					
1 5	W 1	MG	Evolution & Phylogeny		Ch 5.
1 6	F 3	MG	Evolution & Phylogeny	hw4 due	
1 7	M 6	MG	Evolution & Phylogeny		Handout
1 8	W 8	MG	Phylogeny Statistics		
1 9	F 10	MG	Phylogeny Statistics	Mp1	
	M 13		October Break		
2 0	W 15	DK	Comparative Genomics		Ch 11
2 1	F 17	DK	Comparative Genomics	Hw5	
2 2	M 20	DK	Comparative Genomics Statistics		Ch 13

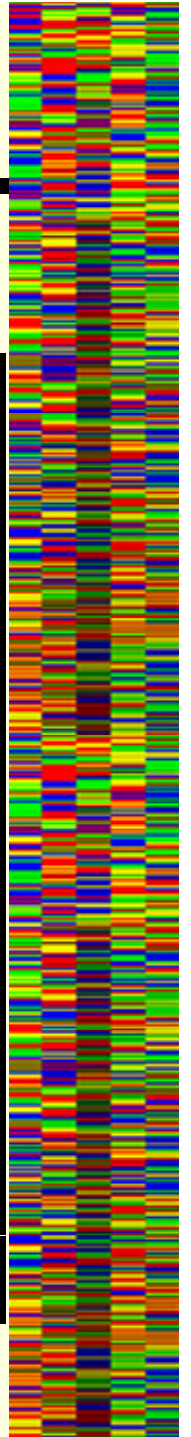
- ***Homework 4 due Friday***



Genomics - Gene Modeling

Limitation

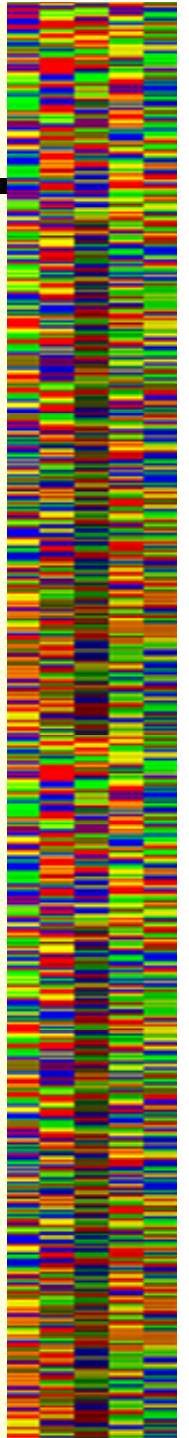
<i>Gene prediction method</i>	<i>Limitation</i>
Ab initio (Hidden Markov Model (HMM)-based) methods	Poor sensitivity and specificity, leading to whole genes or exons being missed or wrongly predicted
Similarity to existing expression sequence tags (ESTs)	Contaminating ESTs derived from unspliced mRNA, genomic DNA and nongenic transcription
Similarity to existing gene/proteins	Unable to distinguish pseudogenes (non-protein coding) and novel genes undetected
Current approaches result in	Partial genes, fragmented genes, gene fusions and spurious predictions



Genomics - Gene Modeling

Things to Remember about gene modeling

- **It is, in general, organism-specific**
- **It works best on genes that are reasonably similar to something seen previously**
- **It finds protein coding regions far better than non-coding regions**
- **In the absence of external (direct) information, alternative forms will not be identified and novel genes will be missed**
- **It is imperfect! (It's biology, after all...)**



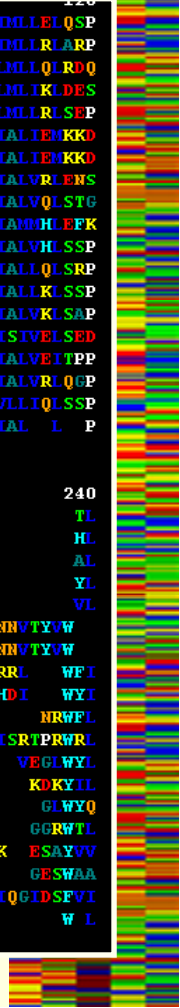
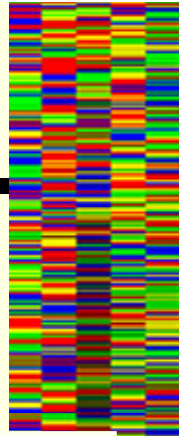
Multiple Alignments and Trees

Goals

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

	120		120		120
kallikrein	LPGGYTCFPHSQPWQAAL	LVQGRLLCGGVLVHPKWVLTAAHCKEG	IKVYLGKHALG RVEAGEQVREVVHSLPHPEYRRSPTHL	NHDHDLMLLEIQSP	
protease	LVHGGPCDKTSHPYQAAL	YTSGLL CCGVLIHPVWLTAAHCKKPN	LQVFLGKHLR QRESSQEQSSVVRVHTPDYDAA	SHQDLMLRLARP	
neuropsin	VLGGHECQPHSQPWQAAL	FQGGQQLCGGVLVGGVWLTAAHCKKPK	YTVRLGDHSLQ NKDQPEQELPVVQSLPHPCYNSSDVE	DHHDLMLLQLRQD	
protease	LINGEDCSPHSQPWQAAL	VMEHELFCGGVLVHPQWVLSAAHCFQNS	YTIGLGLHSLQADQEPGQMVVSEASLSVRHPYRNPLLA	NDLMLIKLDES	
psa	IVGGWECEKHSQPWQVLV	ASRGRAVCGGVLVHPQWVLTAAHCKRKN	SVILLGRHSLFHPEDTG QVFQVSHSFPHPLYDMSLLKNRFLRP	GDDSSHDMLLRLSEP	
complement	IVGGKRAQLGDLPWQVAIK	DASGITCGGIYIGGCWVLTAAHCLRA	SKTHRYQIWTTVVDWIHPDLKRIVIE	YVDRIIFHENY	NAGT YQNDIALIEMKKD
factor	IVGGKRAQLGDLPWQVAIK	DASGITCGGIYIGGCWVLTAAHCLRA	SKTHRYQIWTTVVDWIHPDLKRIVIE	YVDRIIFHENY	NAGT YQNDIALIEMKKD
airway	ILGGTEAEQCSWPWQVSLRL	NNAHCCGSLINNWLTAAHCFR SN	SNPRDWTATSGI	STTFPKLR RVNRLIHDNY	KSAT HENDLALVRENS
mtsp7	IVQGRETAMEGEWPWQASLQLI	GSQHCCGASLISNTWLLTAAHCFW KN	KDPTQWLATFGA	TITPPAVKRNVRKILIHENY	HRET HENDLALVQLSTG
enterokinase	IVGGSNAKEGAWPWVVGLY Y	GRLLCGASLVSSDWLSAAHCVYGRN	LEPSKWTAILGLHMKS	NLTSPTQVPRLDEIVINPHY	NRRR KONDLAMHLEFK
hepsin	IVGGRTSLGRWPWQVSLR Y	DGAHL CCGSLLSGDWLTAAHCFERN	RVLSRWVFPAGAVAQA	SPHGLQLGVQAVVYHGGYLPRFR	PNSEE NSNDLALVLSPP
protease	ITGSSSAVAGQWPWQVSI TY	EGVHYCGGSLVSEQWVLSAAHCFSEH	HKEA YEVKLGAMQLD	SYSEDAKVS TLKDIIPHSYL	QEG SQGDIALIQLSRP
plasmin	VVGGCVAHPSWPWQVSLRTR	FGMHFCGGTLLIPEWVLTAAHCL EKS	PRPSYKIVLGAHQEV	NL EPHGQEVSRLELEP	TRKDIALIKLSSP
testisin	IVGGEDAELGRWPWQGSRLW	DS HVCGVSLLSHRWALTAAHCFEYSDLS	SPGSMVQFG QLT	SMPFSWSLQAYYTRYFWSNIYLSPRYLGN	SPYDIALVKSAP
corin	ILGGRTSRPGRWPWQCSLQSE	PSGHC CGVLIKAKKWLTVAHCFEGRENA	AVKVVGLGIMND	H PSVFMQTRFVKIILHPRY	SRAV VDYDISIVSESD
acrosin	IVGGKAAHQGAWPWVMSLQIFTYN	SHRYHTCGGSLINRWVLTAAHCFVGNVVD	WRLVFGAKETIYGNNKP	KAPWQERYVEKILIEKEY	NSAT EGNDLALVEIIPP
neurotrypsin	ITGGKNSLRGWWPQVSLR	KSSHGDGRLLCGVTLSSCWVLTAAHCF	KRYGNSTRSYAVRYG	DYHTLVPEVEEETGVQQIVHREYRPDR	SDYDIALVRLQGP
proteinase	IVGGHEAQPHSRPYASLQGRNPT	GSHECGGTLIHPSEVLTAAHCLRD	IPQRLVNVVLAGHNRQTQ	EPTQQHFSVAQVFLNN	YDAENKLDVLLIQLSSP
consensus	IVGG A G WPWQVSL	G H C G G L W VLTAAHCF	W V L G H	V I H Y	NDIAL L P

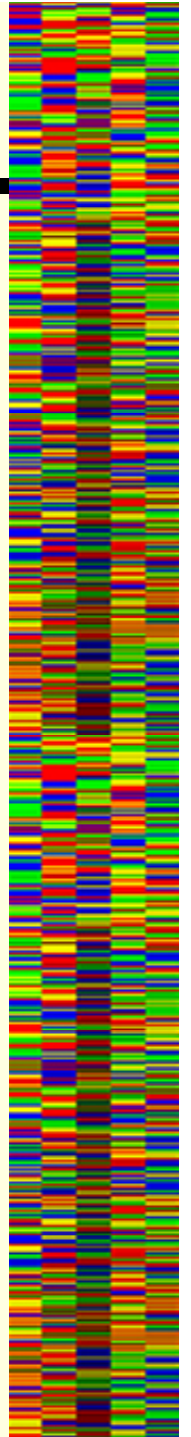
	121	Serine Proteases				1240								
kallikrein	VQLTGYIQT LPLSHNRLLTPGTT	CRVSGWGTIT	SPQVNYPKTLQCANI	QLRSDEECR	QVYPGKITDNL CAGTK	EGGKDCCEGDSGGPLVCHNR	TL							
protease	AKLSLTIQP IPLERDCSANT	TSCHILGWKTA	DG DFPD	TIQCAYIHLVSREECE	HAYPGQITQNL CAGDE	KYKDCSCGDSGGPLVCGD	HL							
neuropsin	ASLGSKVPK ISLADHCT	QPGKCVSGWGTVT	SPRENFPTL	NCAEVKIFPKKKE	DAYPGQITDGRV CAGSS	K GADTCQGDSGGPLVCDG	AL							
protease	VSESDTIRS ISIASQC	P T A G N S C L V S W G L L A	NGR	MPTVLQCVNVSVEEVC	KL Y D P L Y H P S M F C A G G G	H D Q K D S C N G D S G G P L I C N G	YL							
psa	AELTDVAVK MDLPTQ	EPALGTTCYASGWSIE	PEEFLTPKKLQ	CVDLHVISNDVCA	QVHPQKVKFM CAGRW	TGGKSTCSGDSGGPLVCHG	VL							
complement	GNKKDCEIPRSIPA	CVPWSPYLFQPN	DTICVSGWREKDERVFS	LQWGEVKLISN	CSKFGY	NRFYEKEMECAGTY	DGSDACKGDSGGPLVCMDDANNVTVW							
factor	GNKKDCEIPRSIPA	CVPWSPYLFQPN	DTICVSGWREKDERVFS	LQWGEVKLISN	CSKFGY	NRFYEKEMECAGTY	DGSDACKGDSGGPLVCMDDANNVTVW							
airway	VTFTKDIHVC	LPAAQTNI	PPGS	TAYVTEWGAQ	EYAGH	TVPELRQGVRIISNDVCN	APHSYNGAILSGRL CAG	VPQGGVDA CQGDSGGPLVQEDSRRL	WFI					
mtsp7	VEFSNIVQR	CLPSSIKLP	PKT	SVFVTEFGS	IVDDGP	IQNTLRQAR	YETISTVDCN	RKQVDYDGLITP GML CAG	FMEGKIDACKGDSGGPLVY	DHNDI	WYI			
enterokinase	VNYTDYIQPI	CLQASTFE	FNRT	DCWVTWGY	KEDEAL	PSPHTLQEVQ	VAIINNSMCNHLFL	QVPEYNI	TEMNI	CAG	YEEGGIDSCQGDSGGPLMCCEN	NRWL		
hepsin	LPLTEYI	QPVCLP	AAGQALVDGK	ICTVTEWGN	QYYGQ	QAGVLEQAR	VILSNDVCN	GADFYGNQ	IKPKVFCAG	YBEGGIDACQGDSGGPFV	CEDSISRT	PRWRL		
protease	ITFSRYIRPI	CLPAANASFP	NGL	HCVTEWGWVAP	SVSLLTPKPL	QQLLEVPLISRET	CNCLYNIDAKPEEP	PHFVQEDMV	CAG	YVEGGKDACQGDSGGPLSCP	VEGLWYL			
plasmin	AVITDKVIPA	CLPSPNYVVD	VRT	ECFITWGET	QGTFG	AGLLKEAQL	PVIEENKVCNRYEFL	NGRVQSTEL	CAG	HLAGGTDSCQGDSGGPLVCFE	KOKYIL			
testisin	VTYTKH	QPICLQAST	FEFNRT	DCWVTWGY	KEDEAL	PSPHTLQEVQ	VAIINNSMCNHLFL	KYSERK	IFGDMV	CAG	NAQGGKDACQGDSGGPLACNKN	GLWYQ		
corin	ISETGYVR	VPCLNP	EQWLEPDT	YCYITWGH	MGNKMPK	LQEGEVRIS	SLEHCQSYFDMKT	ITRMI	CAG	YESGTVDSCMGDSGGPLVCEKP	GGRWTL			
acrosin	ISCGRF	IGPCLPHLKAGL	PRGSQSCVWAG	WGYLEER	APRPS	SILHEAR	VDLIDDL	LCN	STQWYNGRV	QPTNV	CAG	YVVKIDTCQGDSGGPLMCKDSK	ESAYVV	
neurotrypsin	EEQCARESS	HLPA	CLPLWRERP	QKTASN	CYLTWGDTG	RAYSR	LQQA	APLIPKRFCE	ERYKGRFT	GRMLCAGN	LHCHKR	VDSQGDSGGPLMCEP	GESWAA	
proteinase	ANLSAS	VATVQLP	QDDQVP	PHGTQ	CLAMGWR	GAHDP	PAQVLQELN	VTVV	FFCR	PHNICTFV	PRR	KAGI	CFGDSGGPLICDGI	QGDSFVI
consensus	I P CLP	C V EWG	L Q A V I S	C	Y I M CAG	GG D	CQGDSGGPLVC	W L						



Multiple Alignments and Trees

Why make multiple alignments?

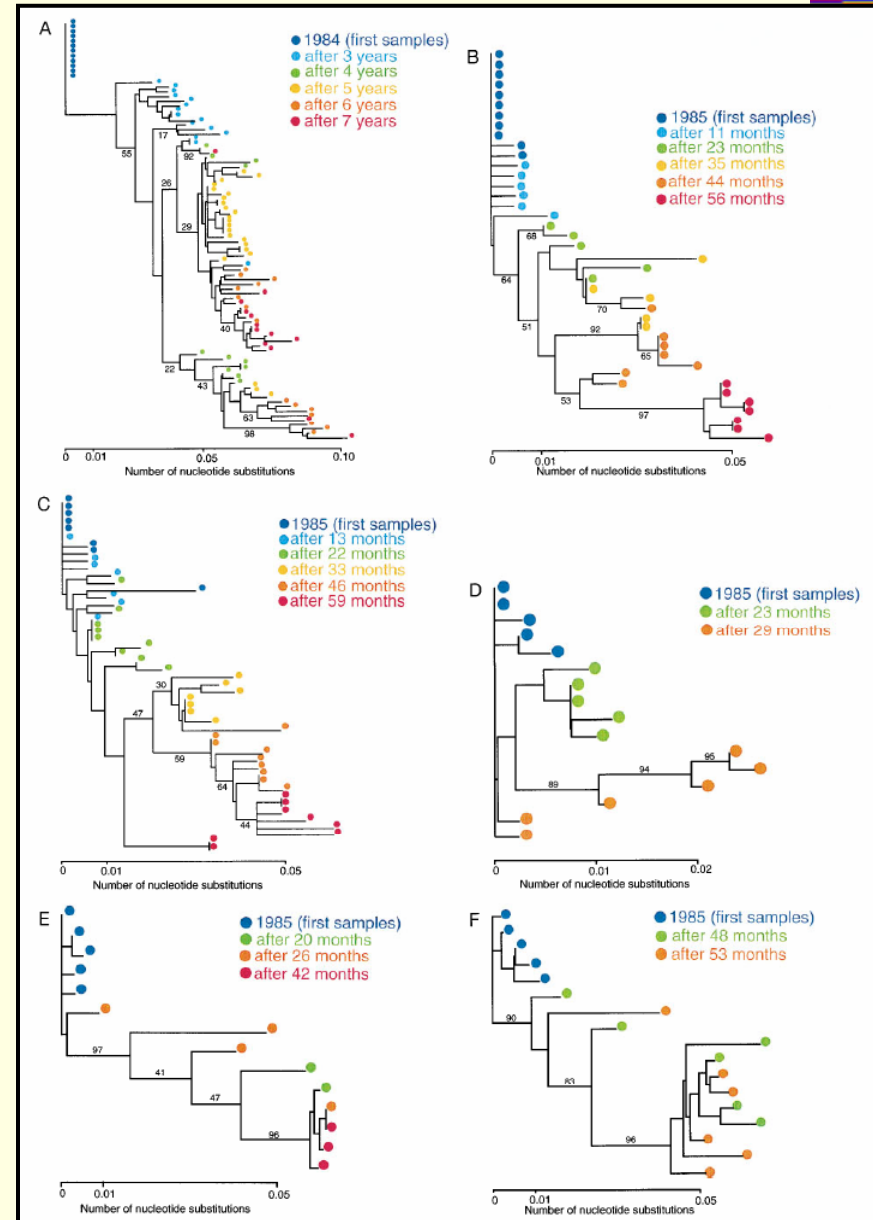
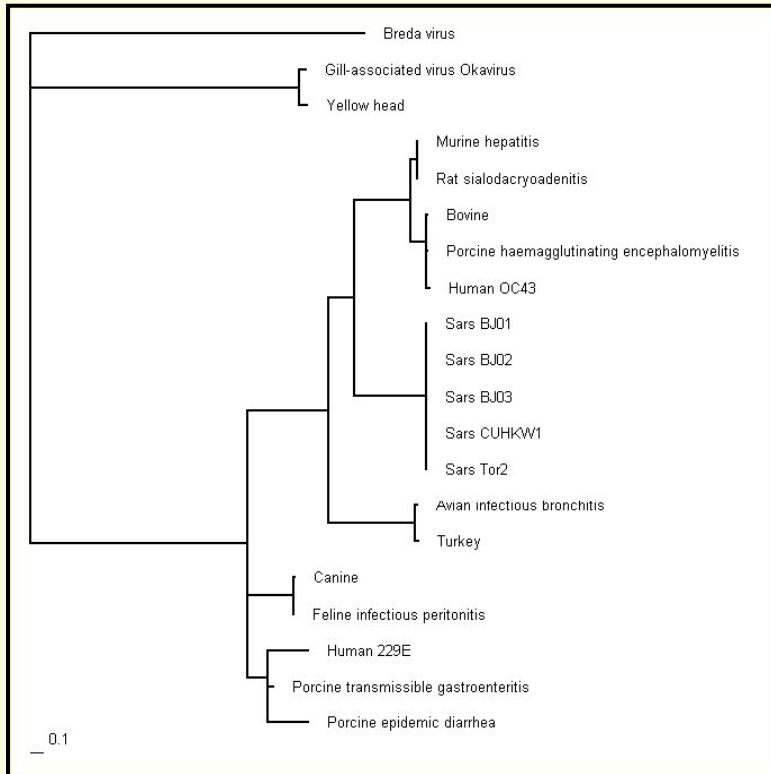
- Find evolutionarily constrained regions of proteins
 - Structural cores
 - Active sites
 - Binding surfaces
- Understand evolution of proteins
- Find origins/causes of disease
- Understand evolution of species (phylogenetics)
 - DNA used more often for species comparison



Multiple Alignments and Trees

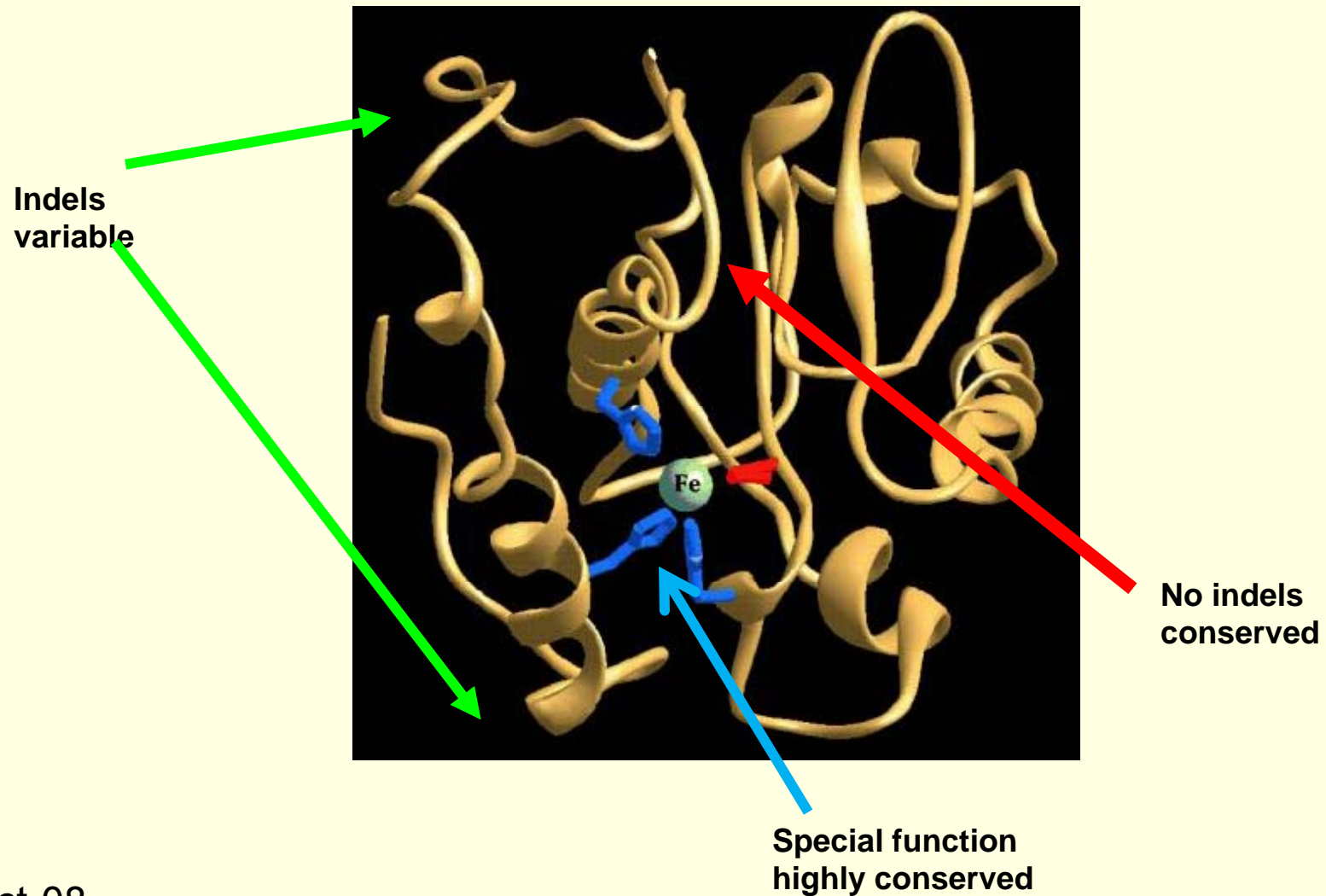
Examples of Trees

- More than just phylogeny



Multiple Alignment and Trees

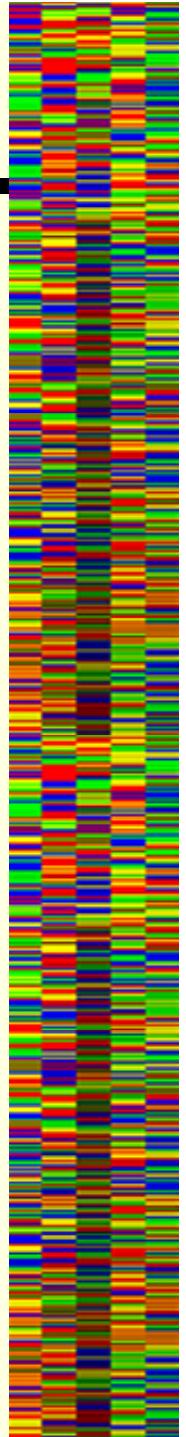
Multiple alignment and protein structure



Multiple Alignment and Trees

Kinds of Multiple alignments

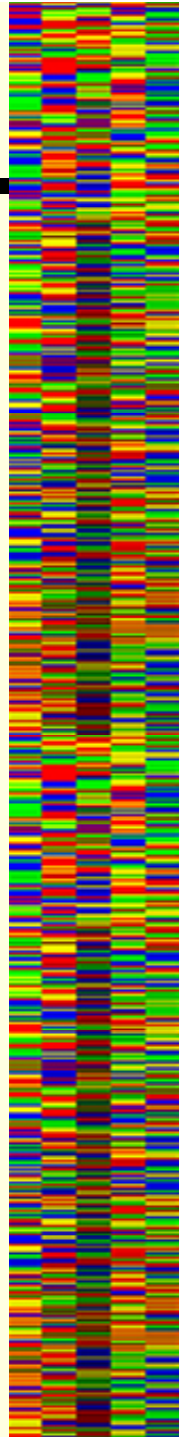
- **Leader-follower alignments**
 - All sequences aligned to one *leader* (by pairwise DP)
 - Optional format from BLAST programs
- **N-dimensional multiple alignments**
 - Simultaneous DP alignment in N dimensions
 - Rarely used due to poor scaling
- **Progressive alignment**
 - Sequential alignment 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)

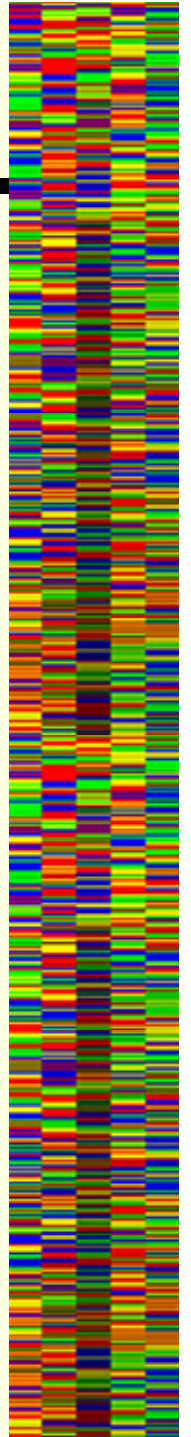
Human HBB	1	MVHLTPEEKSAVTALWGKVVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
Rana HBB	1	MVHWTAEKAVINSVWQKVDVEQDGHEALTRLFIVYPWTQRYFSTFGDLSSPAAIAGNPK	60
Human HBB	61	VKAHGKKVLGAFSDGLAHLNLIKGTFFATLSELHCDKLHVDPENFRLLGNVLCVLAHFFG	120
Rana HBB	61	VHAHGKKILGAIDNAIHNLDDVKGTLHDLSEEHANELHVDPENFRRLGEVLIVVLGAKLG	120
Human HBB	121	KEFTPPVQAAYQKVVAGVANALAHKYH	147
Rana HBB	121	KAFSPQVQHVWEKFI AVLVDALSHSYH	147
<hr/>			
Human HBB	4	LTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV	61
Human HBA	3	LSPADKTNVKAAWGKVGGAHAGEYGAEALERMFSLFPTTKTYFPHF-----DLSHGSAQV	56
Human HBB	62	KAHGKKVLGAFSDGLAHLNLIKGTFFATLSELHCDKLHVDPENFRLLGNVLCVLAHFFGK	121
Human HBA	57	KGHGKKVADALTNAVAHVDDMPNALSA LSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPA	116
Human HBB	122	EFTPPVQAAYQKVVAGVANALAHKY	146
Human HBA	117	EFTPAVHASLDKFLASVSTVLTISKY	141
<hr/>			
Human HBB	4	LTPEEKSAVTALWGKVVNDEVG--GEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV	61
Bovine MY	3	LSDGEWQLVLNAWGKVEADVAGHGQEVLI RLFTHGPETLEKFDKFKHLKTEAEMKASEDL	62
Human HBB	62	KAHGKKVLGAFSDGLAHLNLIKGTFFATLSELHCDKLHVDPENFRLLGNVLCVLAHFFGK	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--KVVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV	61
BrNapa HB	6	FTEKQEALVKESWEILKQDIPKYSLHFFSQILEIAPAADMFSFLRD--TDEVPHNNPKL	63
Human HBB	62	KAHGKKVLGAFSDGLAHL DNLKG-----TFATLSELHCDKLHVDPENFRLLGNVLVCV	114
BrNapa HB	64	KAHAVKVFKMT CETAIQLRE-KGKVVVADTTLQYLGSVHFKSGVLDP-HFEVVKEALVRT	121
Human HBB	115	LAHHFGKEFTPPVQAAYQKV VAGVANAL	142
BrNapa HB	122	LKEGLGEKYNEEVEGAWSKAYDHLALAI	149
<hr/>			
Human HBB	1	MVHLTPEEKSAVTALWG--KVVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN	58
CANLI LB	1	MGAFSEKQESLVKSSWEAFKQNVPHHS AVFYTLILEKAPAAQNMFSFLSNGVDPN----N	56
Human HBB	59	PKVKAHGKKVLGAFSDGLAHL DNLKGTFA----TLSELHCDKLHVDPENFRLLGNVLVCV	114
CANLI LB	57	PKLKAHAEKVFKMTVDSAVQL-RAKGEVVLADPTLGSVHVQKGVLDP-HFLVVKEALLKT	114
Human HBB	115	LAHHFGKEFTPPV----QAAYQKV VAGVANALA	143
CANLI LB	115	FKEAVGDKWDELGN AWEVAYDELAAA IKKAMG	147



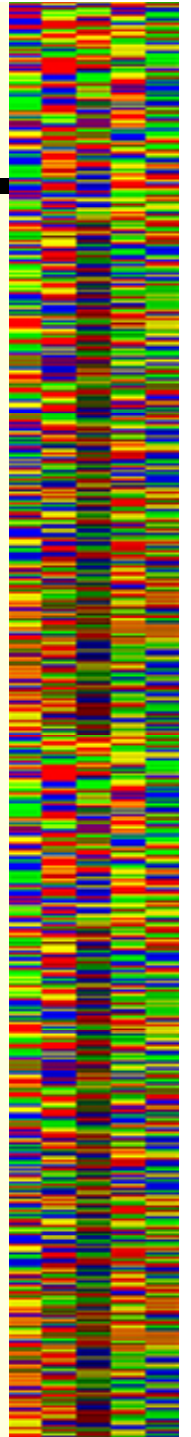
Multiple Alignment and Trees

Alignments to Leader

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

Human HBB 61 VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLGNVLVCVLAHHFG 120
Rana HBB 61 VHAHGKILGAIDNAIHNLDVKGTLHDLSEEHANELHVDPENFRRLGEVLIVVLGAKLG 120
Human HBB 62 KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLGNVLVCVLAHHFGK 121
Human HBA 57 KGHGKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPA 116
Human HBB 62 KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLGNVLVCVLAHHFGK 121
Bovine MY 63 KKHGNTVLTALGGILKKKGHHEAEVKHLAESHANKHKIPVKYLEFISDAI IHVLHAKHPS 122
Human HBB 62 KAHGKKVLGAFSDGLAHLDNLK-----TFATLSELHCDKLHVDPENFRLGNVLVCV 114
BrNapa HB 64 KAHAVKVFKMTCETAIQLRE-KGVVVADTLQYLGSVHFKSGVLDP-HFEVVKEALVRT 121
Human HBB 59 PKVKAHGKKVLGAFSDGLAHLDNLKGTFA----TLSELHCDKLHVDPENFRLGNVLVCV 114
CanLi LB 57 PKLKAHAEKVFKMTVDSAVQL-RAKGEVVLADPTLGSVHVQKGVLDP-HFLVVKEALLKT 114

Human HBB 121 KEFTPPVQAAYQKVVAGVANALAHKYH 147
Rana HBB 121 KAFSPQVQHVWEKFIAVLVDALSHSYH 147
Human HBB 122 EFTTPPVQAAYQKVVAGVANALAHKY 146
Human HBA 117 EFTPAVHASLDKFLASVSTVLTSKY 141
Human HBB 122 EFTTPPVQAAYQKVVAGVANALAHKY 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
```

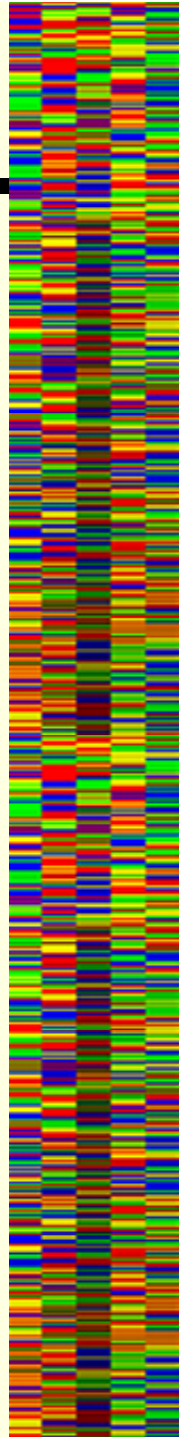


Multiple Alignment and Trees

Propagate Gaps vs Leader

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

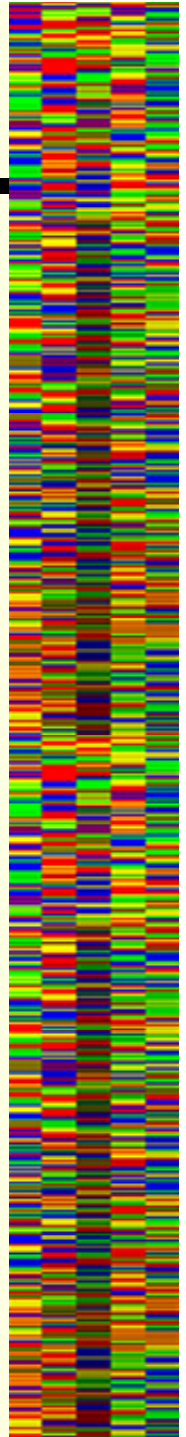
Human HBB	1	MVHLTPEEKSAVTALWG ++KV++ NVDEVG ++GE ALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
Rana HBB	1	MVHWTAEKAVINSVWQ ++KV++ DVEQDG ++ HEALTRLFIVYPWTQRYFSTFGDLSSPAAIAGNPK	60
Human HBB	4	LTPEEKSAVTALWG ++KV-- NVDEVG ++GE ALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV	61
Human HBA	3	LSPADKTNVKAAWG ++KVG AGAHAGEY ++ AEALERMFLSFPTTKTYFPHF-----DLSHGSAQV	56
Human HBB	4	LTPEEKSAVTALWG ++KV++ NVDEVG-- GE ALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV	61
Bovine MY	3	LSDGEWQLVLNAWG ++KV++ EADVAGHGQEVLRIRLFTGHPETLEKFDKFKHLKTEAEMKASEDL	62
Human HBB	4	LTPEEKSAVTALWG-- KV++ NVDEVG ++GE ALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV	61
BrNapa HB	6	FTEKQEALVKESWEILKQ ++ DIPKYS ++ LHFFSQILEIAPAADMFSFLRD--TDEVPHNNPKL	63
Human HBB	1	MVHLTPEEKSAVTALWG-- KV++ NVDEVG ++GE ALGRLLVVYPWTQRFFESFGDLSTPDAVMGN	58
CanLi LB	1	MGAFSEKQESLVKSSWEAFKQ ++ NVPHHS ++ AVFYTLILEKAPAAQNMFSLNSGVDPN----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++EADVAGHGQEV LIRLFTGHPETLEKFDKFKHLKTEAEMKASEDL	62
Human HBB	4	LTPEEKSAVTALWG--KV++NVDEVG++GEALGRLLVVYPWTQRRFFESFGDLSTPDAVMGNPKV	61
BrNapa HB	6	FTEKQEALVKESWEILKQ++DIPKYS++LHFFSQILEIAPAAKDMFSFLRD--TDEVPHNNPKL	63
Human HBB	1	MVHLTPEEKSAVTALWG--KV++NVDEVG++GEALGRLLVVYPWTQRRFFESFGDLSTPDAVMGN	58
CanLi LB	1	MGAFSEKQESLVKSSWEAFKQ++NVP HHS++AVFYTLILEKAPAAQNMF SFLSNGVDPN---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--EADVAGHGQEV LIRLFTGHPETLEKFDKFKHLKTEAEMKASEDL	62
BrNapa HB	6	FTEKQEALVKESWEILKQ--DIPKYS--LHFFSQILEIAPAAKDMFSFLRD--TDEVPHNNPKL	63
CanLi LB	1	MGAFSEKQESLVKSSWEAFKQ--NVP HHS--AVFYTLILEKAPAAQNMF SFLSNGVDPN---N	56



Multiple alignment and Trees

Do we need all the gaps?

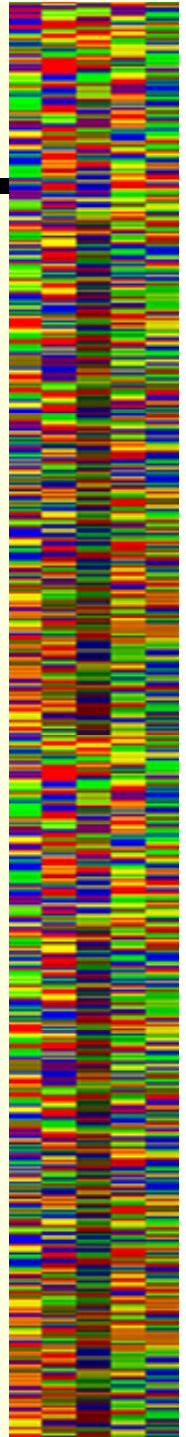
Human HBB	1	MVHLTPEEKSAVTALWG--KV--NVDEVG--GEALGRLLVVY	PWTQRFFESFGDLSTPDAVMGNPK	60
Rana HBB	1	MVHWTAEKAVINSVWQ--KV--DVEQDG--HEALTRLFIVY	PWTQRYFSTFGDLSSPAAIAGNPK	60
Human HBA	3	LSPADKTNVKAAWG--KVGAHAGEYGA	AEALERMFLSFPTTKTYFPHF-----DLSHGSAQV	56
Bovine MY	3	LSDGEWQLVLNAWG--KVEADVAGHG	QEVLI RLFTHGPELLEKFDKFKHLKTEAEMKASEDL	62
BrNapa HB	6	FTEKQEALVKESWEILKQ--DIPKYS--LHFFSQILEIA	PAAKDMFSFLRD--TDEVPHNNPKL	63
CanLi LB	1	MGAFSEKQESLVKSSWEAFKQ--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--KVGAHAGEYGA	AEALERMFLSFPTTKTYFPHF-----DLSHGSAQV	56	
Bovine MY	3	LSDGEWQLVLNAWG--KVEADVAGHG	QEVLI RLFTHGPELLEKFDKFKHLKTEAEMKASEDL	62	
BrNapa HB	6	FTEKQEALVKESWEILKQ--DIPKYS	LHFFSQILEIA	PAAKDMFSFLRD--TDEVPHNNPKL	63
CanLi LB	1	MGAFSEKQESLVKSSWEAFKQ--NVPHHS	AVFYTLILEKA	PAAQNMFSFLSNGVDPN----N	56



Human HBB	1	MVHLTPEEKSAVTALWGKV--NVDEVG	GEALGRLLVVY	PWTQRFFESFGDLSTPDAVMGNPK	60
Rana HBB	1	MVHWTAEKAVINSVWQKV--DVEQDG	HEALTRLFIVY	PWTQRYFSTFGDLSSPAAIAGNPK	60
Human HBA	3	LSPADKTNVKAAWGKVGAHAGEYGA	AEALERMFLSFPTTKTYFPHF-----DLSHGSAQV	56	
Bovine MY	3	LSDGEWQLVLNAWGKVEADVAGHG	QEVLI RLFTHGPELLEKFDKFKHLKTEAEMKASEDL	62	
BrNapa HB	6	FTEKQEALVKESWEILKQDIPKYS	LHFFSQILEIA	PAAKDMFSFLRD--TDEVPHNNPKL	63
CanLi LB	1	MGAFSEKQESLVKSSWEAFKQNVPHHS	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**

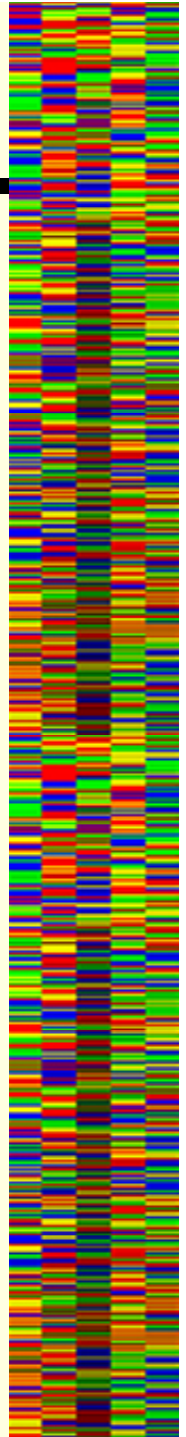
$$\sum_{\text{columns}} \sum_{i=1}^n \sum_{j=i+1}^n S_{ij}$$

for the letters (s) is sequence *i* and *j*

- **Number of mutations (as in PAM calculation)**
 - *doesn't overcount due to where we get 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 KVG AHAGEYG
Bovine MY 3 KV EADVAGHG
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)

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

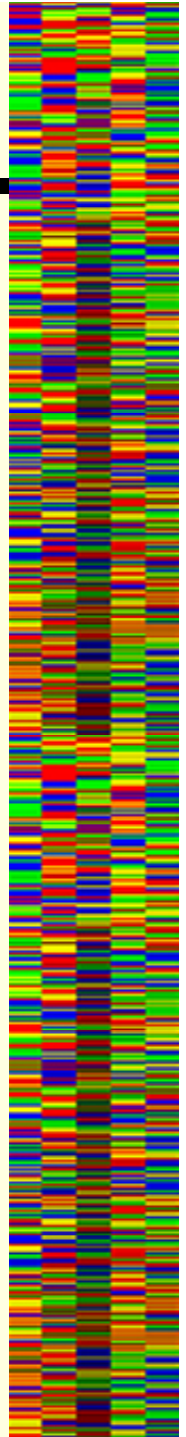
```

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 KVGAGEYGA
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

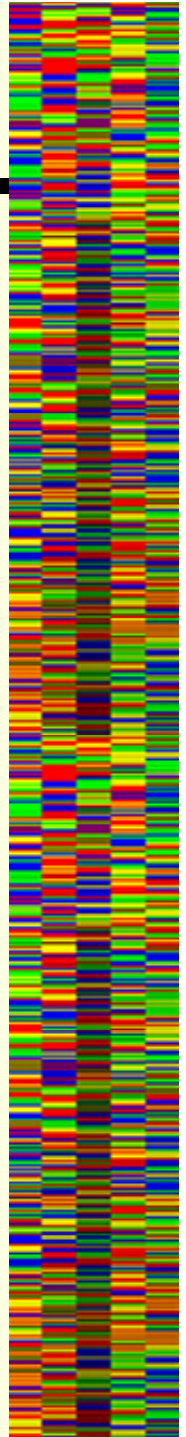
- More gaps gives higher SOP score
 - Lower alignment is not clearly better

```
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)

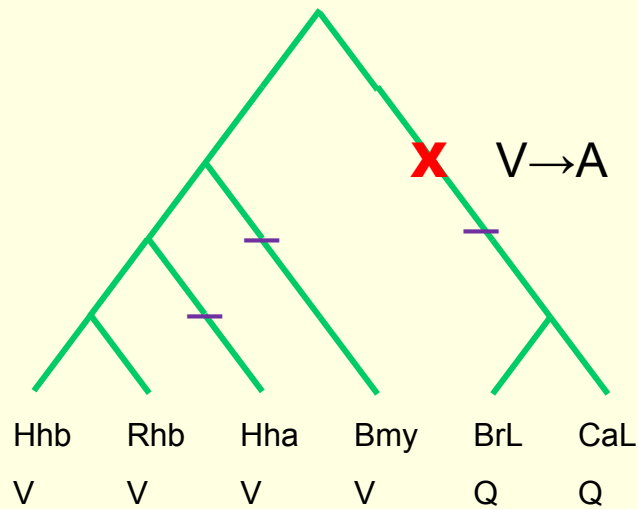


Multiple Alignments and Trees

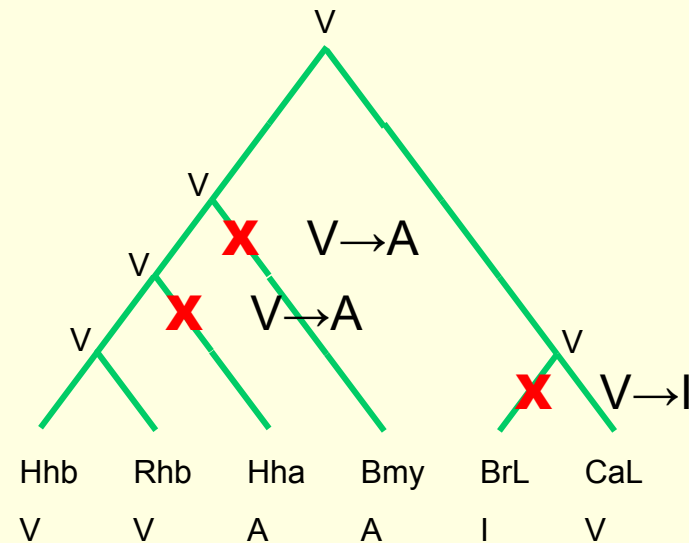
Mutational Transitions

- 3 indels, each in one branch (-)

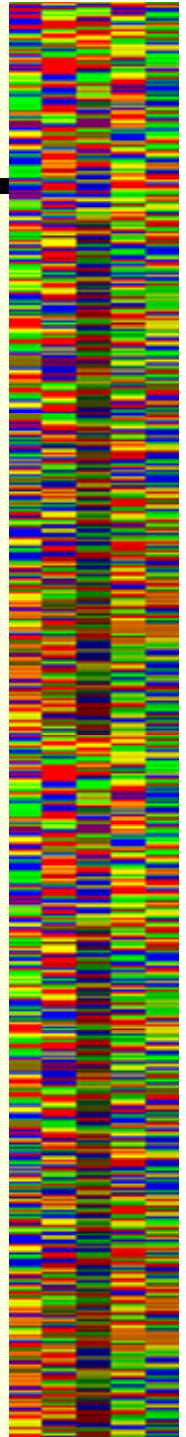
Human HBB	1	--KV--NVDEVG--
Rana HBB	1	--KV--DVEQDG--
Human HBA	3	--KVGAAHAGEYG--
Bovine MY	3	--KV--EADVAGHG
BrNapa HB	6	ILKQ--DIPKYS--
CanLi LB	1	AFKQ--NVPHHS--



SOP=7
mutations=1



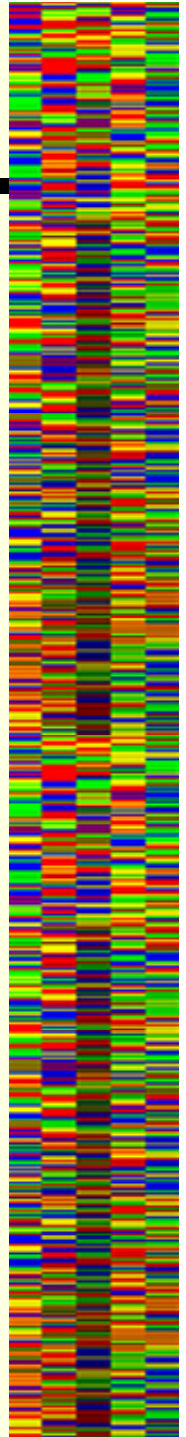
SOP=4
mutations=3



Multiple Alignment and Trees

Final leader-follower alignment

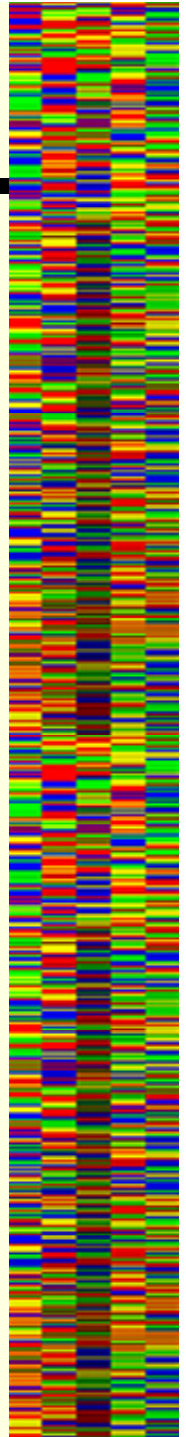
Human HBB	1	MVHLTPEEKSAVTALWG--KV--NVDEVG--GEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60
Rana HBB	1	MVHWTAEKAVINSVWQ--KV--DVEQDG--HEALTRLFIVYPWTQRYFSTFGDLSSPAAIAGNPK	60
Human HBA	3	LSPADKTNVKAAGW--KVG AHAGEYG--AEALERMFLSFPTTKTYFPHF-----DLSHGSAQ	55
Bovine MY	3	LSDGEWQLVLNAWG--KV--EADVAGHGQEV LIRLFTGH PETLEKFDKFKHLKTEAEMKASED	61
BrNapa HB	6	FTEKQEALVKESWEILKQ--DIPKYS--LHFFSQILEIAPA AKDMFSFLRD--TDEVPHN NPK	62
CanLi LB	1	MGAFSEKQESLVKSSWEAFKQ--NVPHHS--AVFYTLILEKAPAAQNMFSFLSNGVDPN----NPK	54
Human HBB	61	VKAHGKKVLGAFSDGLAHL DNLK G-----TFA----T LSELHCDKLHVDPENFRLLGNVLCV	114
Rana HBB	61	VHAHGKKILGAIDNAIHN LDDVKG-----TLH----DLSEEHANELHVDPENFRRLGEVLIVV	114
Human HBA	56	VKGHGKKVADALTN AVAHVDDMPN-----ALS----ALSDLHAHKL RVDPVNFKLLSHC LLVT	109
Bovine MY	62	LKKHGNTVLTALGGILK KKGHHEA-----EVK----HLAESHANKHKIPVKYLEFISDAI IHV	115
BrNapa HB	63	LKAHAVKVFKMT CETAIQLRE-KGKV VVADTTLQ----YLG SVHFKSGVLD P-HFEVVKEALVRT	121
CanLi LB	55	LKAHAEKVFKMTVDS AVQL-RAKG-----EVVLADPTLGSVHVQKGVLD P-HFLVVKEALLKT	114
Human HBB	115	LAH HFGKEFTPPV----QAAYQKV VAGVANALAHKYH	147
Rana HBB	115	LGAKLGKAFSPQV----QH VWEKFI AVLVDALSHSYH	147
Human HBA	110	LAAHLPAEFTPAV----HASL DKFLASVSTVLT SKY	141
Bovine MY	116	LHAKHPSDFGADA----QAAMSKALELFRNDMAAQY	147
BrNapa HB	122	LKEGLGEKYNEEV----EGAW SKAYDHLALAI	149
CanLi LB	115	FKEAVGDKWNDELGN AWEVAYDELA AAIKKAMG	147



Multiple Alignment

Final leader-follower alignment

Human HBB	1	MVHLTPEEKSAVTALWG--KV--NVDEVG--GEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK	60	
Rana HBB	1	MVHWTAEKAVINSVWQ--KV--DVEQDG--HEALTRLFIVYPWTQRYFSTFGDLSPPAAIAGNPK	60	
Human HBA	3	LSPADKTNVKAAG--KVGAHAGEYG--AEALERMFLSFPTTKTYFPHF-----DLSHGSAQ	55	
Bovine MY	3	LSDGEWQLVLNAWG--KV--EADVAGHGQEVLIIRLFTGHPETLEKFDKFKHLKTEAEMKASED	61	
BrNapa HB	6	FTEKQEALVKESWEILKQ--DIPKYS--LHFFSQILEIAPAAKDMFSFLRD--TDEVHNPKNPK	62	
CanLi LB	1	MGAFSEKQESLVKSSWEAFKQ--NVPHHS--AVFYTLILEKAPAAQNMFSFLSNGVDPN----NPK	54	
		*		*
Human HBB	61	VKAHGKKVLGAFSDGLAHLNLRK-----TFA----TLSELHCDKLHVDPENFRLLGNVLCV	114	
Rana HBB	61	VHAHGKKILGAIDNAIHNLDLVK-----TLH----DLSEEHANLHVDPENFRRLGEVLIVV	114	
Human HBA	56	VKGHGKKVADALTNVAHVDDMPN-----ALS----ALSDLHAHKLKRVDPVNFKLLSHCLLVV	109	
Bovine MY	62	LKKHGNTVLTALGGILKKGHEA-----EVK----HLAESHANKHKIPVKYLEFISDAIIHV	115	
BrNapa HB	63	LKAHAVKVFKMTCTETAIQLRE-KGKVVVADTTLQ----YLGSVHFKSGVLDP-HFEVVKEALVRT	121	
CanLi LB	55	LKAHAEKVFKMTVDSAVQL-RAKG-----EVVLADPTLGSVHVQKGVLDP-HFLVVKEALLKT	114	
		*		*
Human HBB	115	LAHFFGKEFTPPV----QAAYQKVVAGVANALAHKYH	147	agrees with structural
Rana HBB	115	LGAKLGKAFSPQV----QHVWEKFI AVLVDALSHSYH	147	unclear in structure
Human HBA	110	LAHLLPAEFTPAV----HASLDKFLASVSTVLTISKY	141	disagrees with structural
Bovine MY	116	LHAKHPSDFGADA----QAAMSKALELFRNDMAAQY	147	
BrNapa HB	122	LKEGLGEKYNEEV----EGAWSKAYDHLALAI----	149	
CanLi LB	115	FKEA VGDKWNDELGN AWEVAYDELA AAIKKAMG	147	



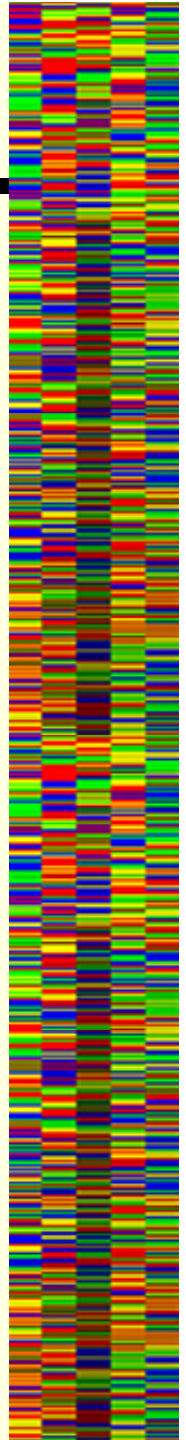
Multiple Alignment and Trees

Leader makes a big difference

- Human HBB vs CanLi as Leader

Human HBB	1	MVHLTPEEKSAVTALWG-- KV --NVDEVG--GEALGRLLVVY PWT QRFFESFGDLSTPDAVMGNPK	60
Rana HBB	1	MVHWTAEKAVINSVWQ-- KV --DVEQDG--HEALTRLFIVY PWT QRYFSTFGDLSSPAAIAGNPK	60
Human HBA	3	LSPADKTNVKAAWG-- KV GAHAGEYG--AEALERMFLSF PTT KTYFPHF-----DLSHGSAQ	55
Bovine MY	3	LSDGEWQLVLNAWG-- KV --EADVAGHGQEV LIRLFTGH PET LEKFDKFKHLKTEAEMKASED	61
BrNapa HB	6	FTEKQEALVKESWEIL KQ --DIPKYS--LHFFSQILEIAP AAK DMFSFLRD--TDEVPHNNPK	62
CanLi LB	1	MGAFSEKQESLVKSSWEAF KQ --NVPHHS--AVFYTLILEKAP AA QNMFSFLSNGVDPN----NPK	54

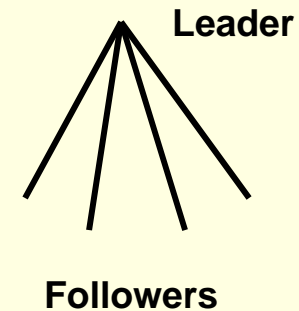
CanLi LB	7	KQESLVKSSWEAFKQ NVP-HHSAVFYTLILEKAP AAQ --NMFS-F-L--SNGVDP----N----NPK
Human HBB	7	EEKSAVTALWG-- KVN VD-EVGGEALGRLLVVY PWT Q--RFFE-S-F--GDLSTP----DAVMGNPK
Rana HBB	4	WTAEKAVINSVWQ KV DVEQD-GHEAL--TRLFIVY PWT Q--RYFS-T-F--GDLSSP----AAIAGNPK
Human HBA	4	SPADKTNVKAAWG KV GAHAG-EYGAEALERMFLSF PTT K--TYFPHFDL--SHG-----SAQ
Bovine MY	1	MG---LSDGEWQLVLNAWG KV EADVAGHGQEV LIRLFTGH PET LEKFDKFK-H-L--KTEAEM----K---ASED
BrNapa HB	1	MGEIVFTEKQEALVKESWEIL KQ DIP-KYSLHFFSQILEIAP AAK --DMFS-F-LRDTDEVPH----N----NPK



Multiple Alignment and Trees

Problems revealed by leader-follower alignments

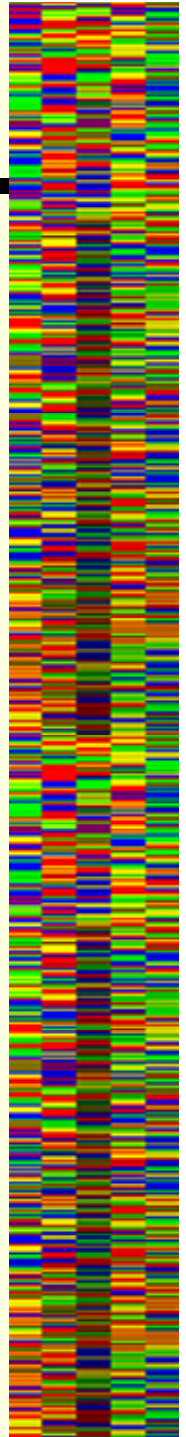
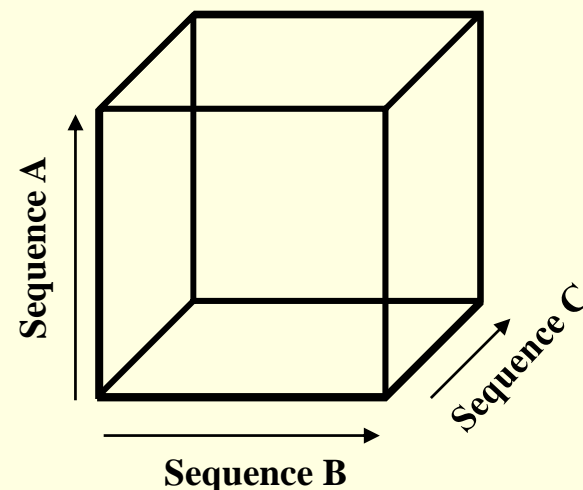
- **Alignment depends on leader**
 - **Model is not biologically realistic**
- **Inaccuracy in pairwise alignments compromises multiple alignment**
- **Some alignments are more informative (better) than others**



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>
 - download: <ftp://fastlink.nih.gov/pub/msa/>

