Biol 478/595 Intro to Bioinformatics

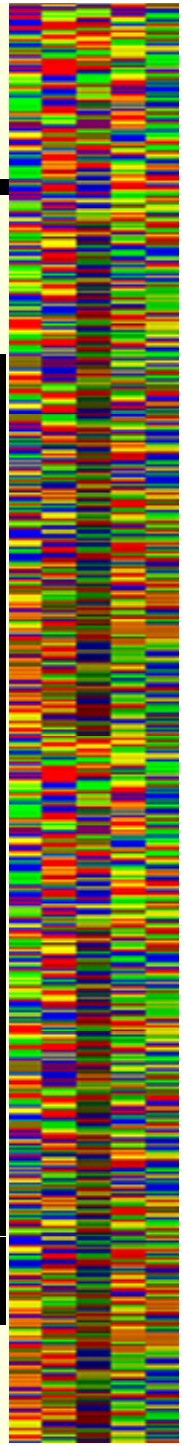
October					
15	W 1	MG	Evolution & Phylogeny		Ch 5.
16	F 3	MG	Evolution & Phylogeny	hw4 due	
17	M 6	MG	Evolution & Phylogeny		Handout
18	W8	MG	Phylogeny Statistics		
19	F 10	MG	Phylogeny Statistics	Mp1	
	M 13		October Break		
20	W 15	DK	Comparative Genomics		Ch 11
21	F 17	DK	Comparative Genomics	Hw5	
22	M 20	DK	Comparative Genomics Statistics		Ch 13

- ***Homework 4 due Friday***

Genomics - Gene Modeling

Limitation

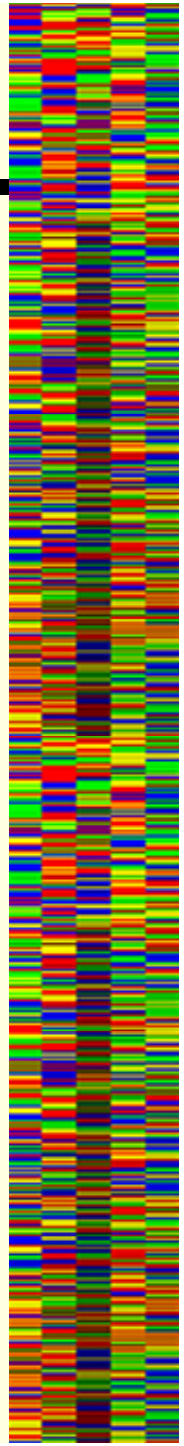
Gene prediction method	*Limitation*
Ab initio (Hidden Markov Model (HMM)-based) methods	**Poor sensitivity and specificity, leading to whole genes or exons being missed or wrongly predicted**
Similarity to existing expression sequence tags (ESTs)	**Contaminating ESTs derived from unspliced mRNA, genomic 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

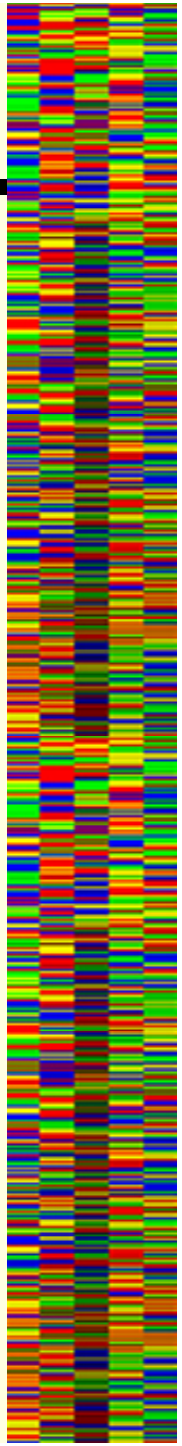
Goals

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

Multiple Alignments and Trees

Why make multiple alignments?

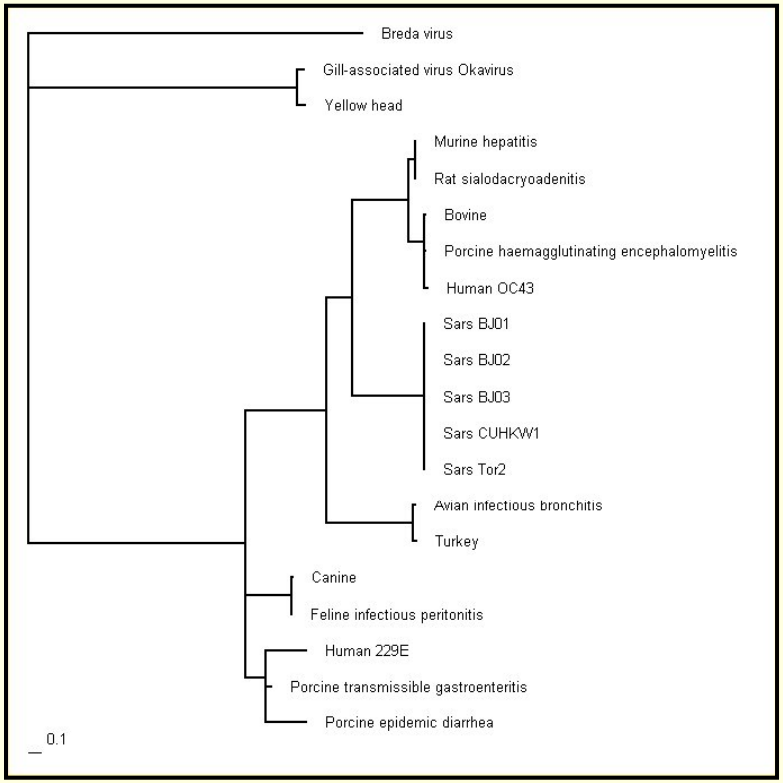
- **Find evolutionarily constrained regions of proteins**
  - **Structural cores**
  - **Active sites**
  - **Binding surfaces**
- **Understand evolution of proteins**
- **Find origins/causes of disease**
- **Understand evolution of species (phylogenetics)**
  - **DNA used more often for species comparison**

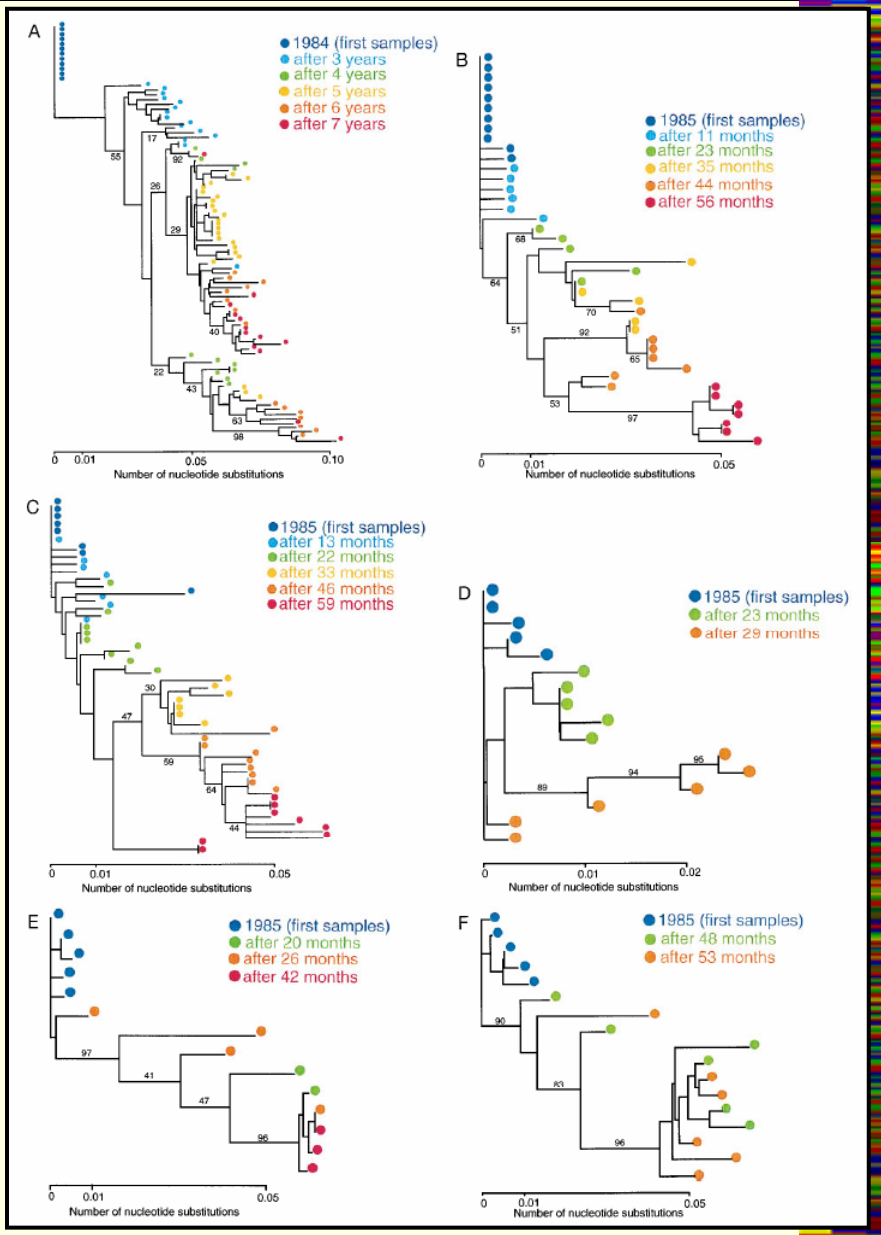


Multiple Alignments and Trees

Examples of Trees

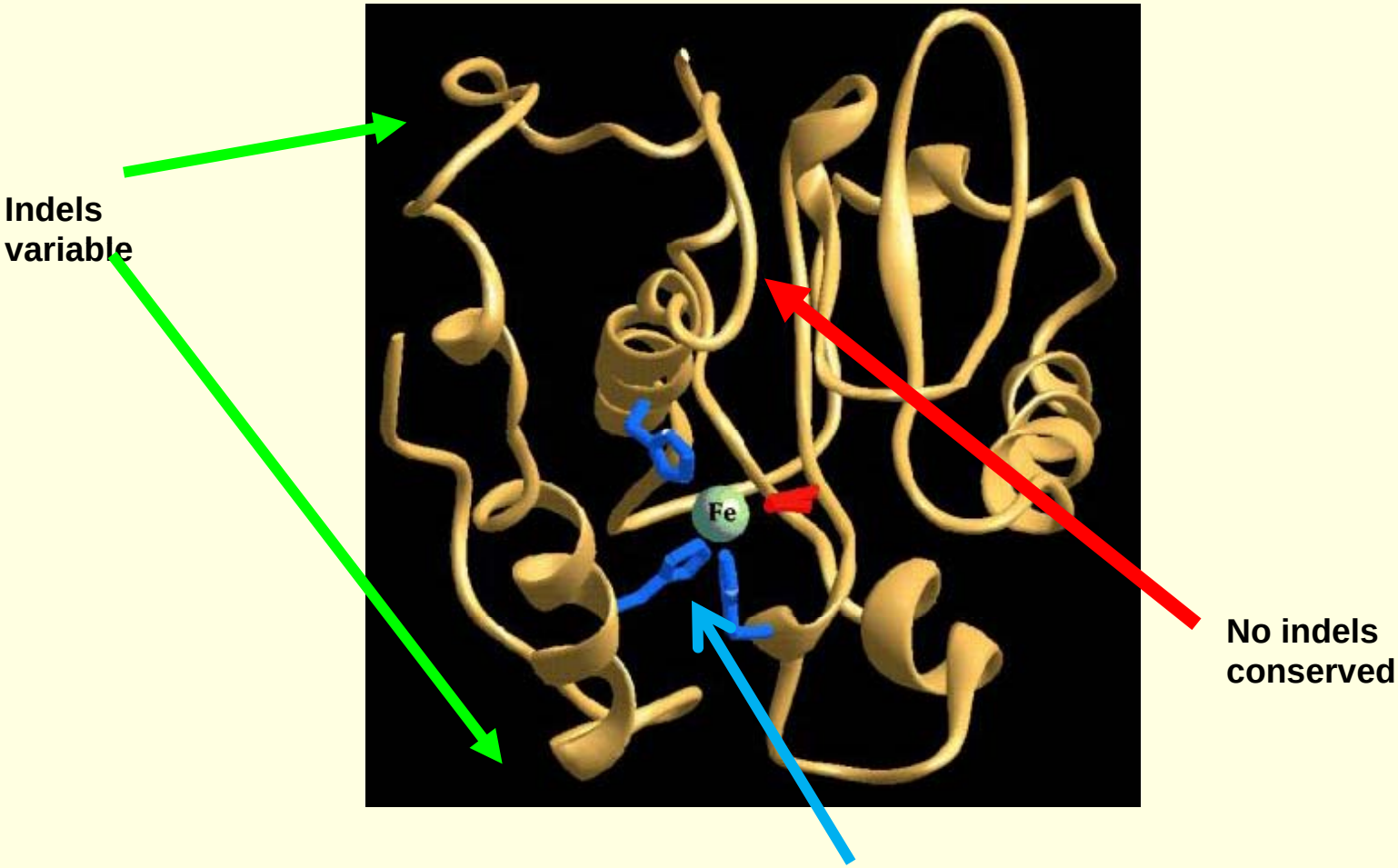
- **More than just phylogeny**





Multiple Alignment and Trees

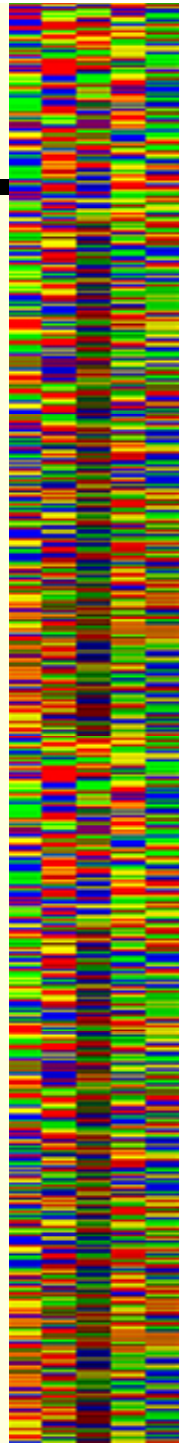
Multiple alignment and protein structure



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  MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK  60
Rana HBB    1  MVHWTAEEKAVINSVWQKVDVEQDGHEALTRLFIVYPWTQRYFSTFGDLSSPAAIAGNPK  60

Human HBB 61   VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG  120
Rana HBB  61   VHAHGKKILGAIDNAIHNLDDVKGTLHDLSEEHANELHVDPENFRRLGEVLIVVLGAKLG  120

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

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

Human HBB 62   KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK  121
Human HBA 57   KGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPA  116

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

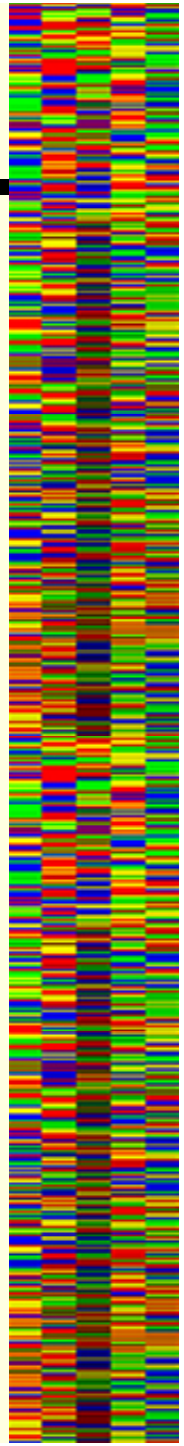
```
Human HBB 4    LTPEEKSAVTALWGKVNVDEVG--GEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV  61
Bovine MY 3    LSDGEWQLVLNAWGKVEADVAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASEDL  62

Human HBB 62   KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK  121
Bovine MY 63   KKHGNTVLTALGGILKKKGHHEAEVKHLAESHANKHKIPVKYLEFISDAIIHVLHAKHPS  122

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

Multiple Alignment and Trees

Pairwise alignments to leader (Human HBB)

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

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

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

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

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

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



Multiple Alignment and Trees

Alignments to Leader

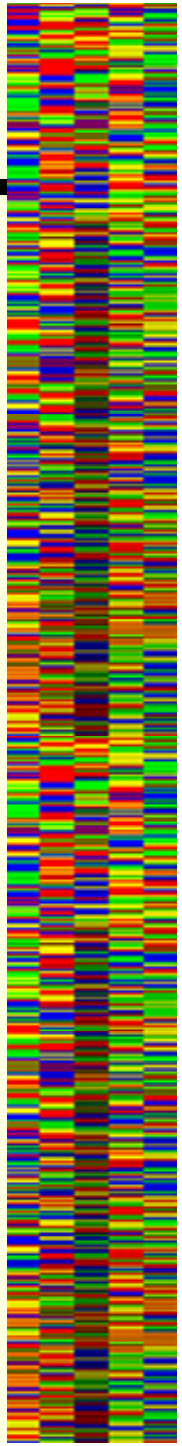
```
Human HBB    1  MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK  60
Rana HBB     1  MVHWTAEEKAVINSVWQKVDVEQDGHEALTRLFIVYPWTQRYFSTFGDLSSPAAIAGNPK  60
Human HBB    4     LTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV  61
Human HBA    3     LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF------DLSHGSAQV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

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

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

Multiple Alignment and Trees

Propagate Gaps vs Leader

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

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

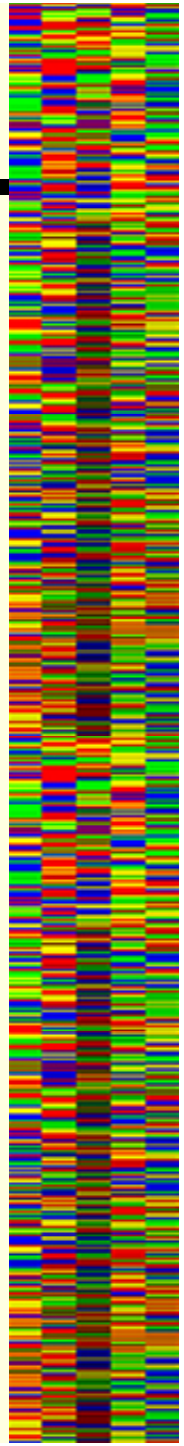



Multiple Alignment and Trees

Remove redundant leader sequences

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

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



Multiple alignment and Trees

Do we need all the gaps?

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

?

```
Human HBB    1  MVHLTPEEKSAVTALWG--KV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK   60
Rana HBB     1  MVHWTAEEKAVINSVWQ--KV--DVEQDGHEALTRLFIVYPWTQRYFSTFGDLSSPAAIAGNPK   60
Human HBA    3     LSPADKTNVKAAWG--KVGAHAGEYGAEALERMFLSFPTTKTYFPHF------DLSHGSAQV   56
Bovine MY    3     LSDGEWQLVLNAWG--KVEADVAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASEDL   62
BrNapa HB    6     FTEKQEALVKESWEILKQ--DIPKYSLHFFSQILEIAPAAKDMFSFLRD--TDEVPHNNPKL   63
CanLi LB     1  MGAFSEKQESLVKSSWEAFKQ--NVPHHSAVFYTLILEKAPAAQNMFSFLSNGVDPN----N     56
```

?

```
Human HBB    1  MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK   60
Rana HBB     1  MVHWTAEEKAVINSVWQKV--DVEQDGHEALTRLFIVYPWTQRYFSTFGDLSSPAAIAGNPK   60
Human HBA    3     LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF------DLSHGSAQV   56
Bovine MY    3     LSDGEWQLVLNAWGKVEADVAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASEDL   62
BrNapa HB    6     FTEKQEALVKESWEILKQDIPKYSLHFFSQILEIAPAAKDMFSFLRD--TDEVPHNNPKL   63
CanLi LB     1  MGAFSEKQESLVKSSWEAFKQNVPHHSAVFYTLILEKAPAAQNMFSFLSNGVDPN----N     56
```

Multiple Alignment and Trees

Which is best?

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

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

- ***Qualitative***
  - **I'm an expert, the bottom one is better, trust me**
- **Quantitative**
  - **Sum of pairs score**

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

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

  - **Number of mutations (as in PAM calculation)**
    - *– doesn't overcount due to where we get the sequences*

Multiple Alignment and Trees

Which is best?

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

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

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

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

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

Multiple Alignments and Trees

SOP score

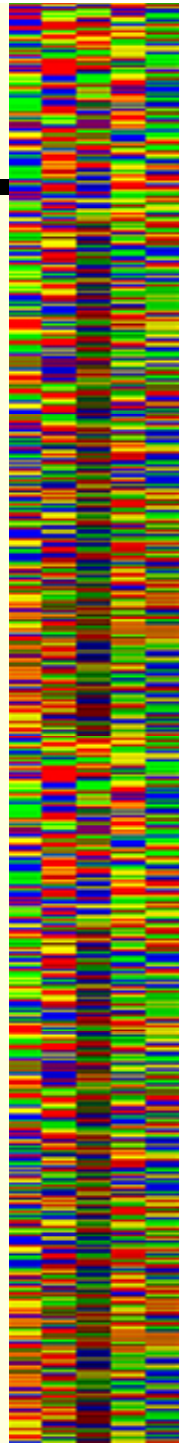
- **More gaps gives higher SOP score**
  - **Lower alignment is not clearly better**

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

SOP=42
gaps: 6 (52)

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

SOP=49
gaps: 13 (57)

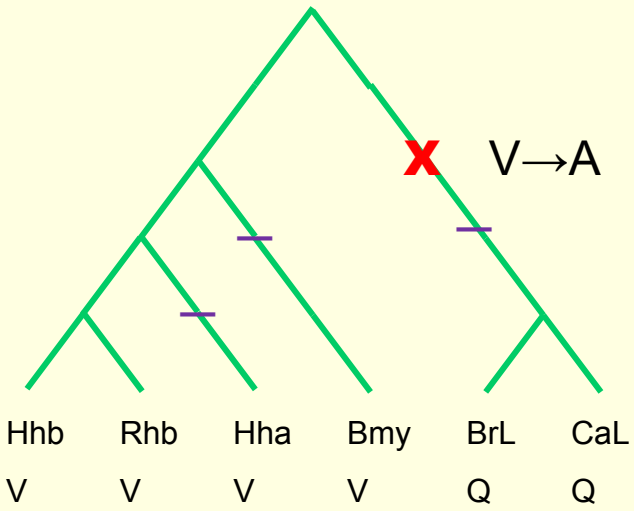


Multiple Alignments and Trees

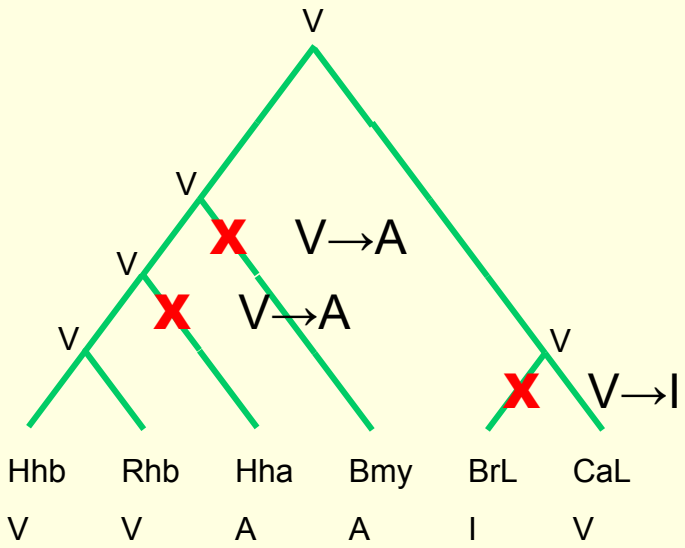
Mutational Transitions

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

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



SOP=7
mutations=1

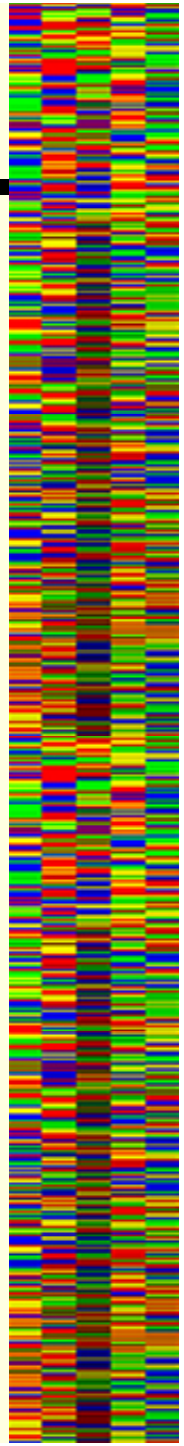


SOP=4
mutations=3

Multiple Alignment and Trees

Final leader-follower alignment

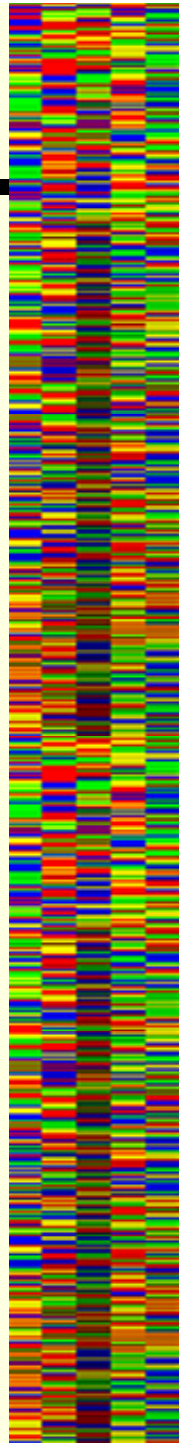
```
Human HBB     1  MVHLTPEEKSAVTALWG--KV--NVDEVG--GEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK   60
Rana HBB      1  MVHWTAEEKAVINSVWQ--KV--DVEQDG--HEALTRLFIVYPWTQRYFSTFGDLSSPAAIAGNPK   60
Human HBA     3     LSPADKTNVKAAWG--KVGAHAGEYG--AEALERMFLSFPTTKTYFPHF------DLSHGSAQ   55
Bovine MY     3     LSDGEWQLVLNAWG--KV--EADVAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASED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

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

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



Multiple Alignment

Final leader-follower alignment

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

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

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



Multiple Alignment and Trees

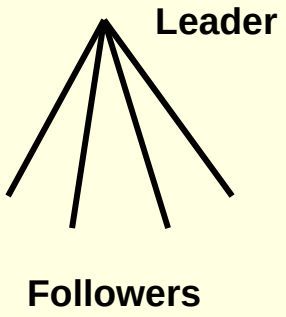
Leader makes a big difference

- **Human HBB vs CanLi as Leader**

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

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

Multiple Alignment and Trees

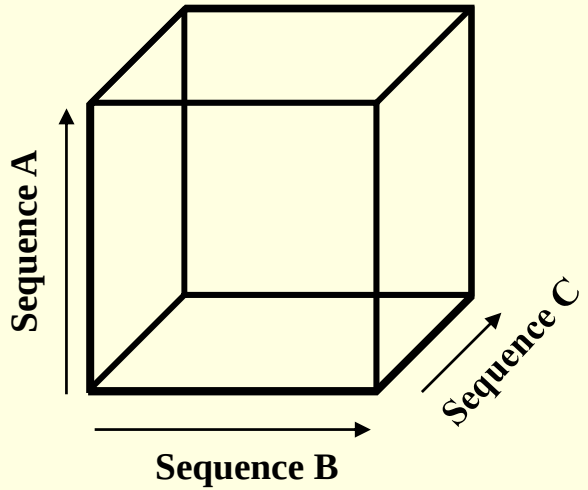
Problems revealed by leader-follower alignments



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

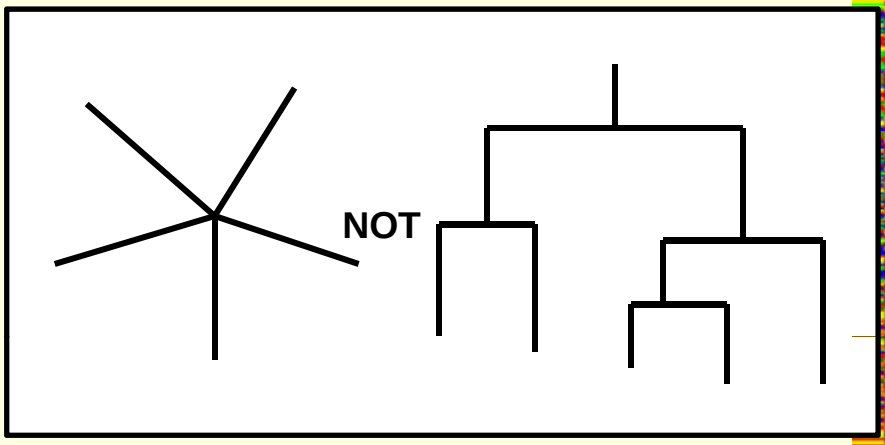
Dynamic Programming in Multiple Dimensions

- 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/*