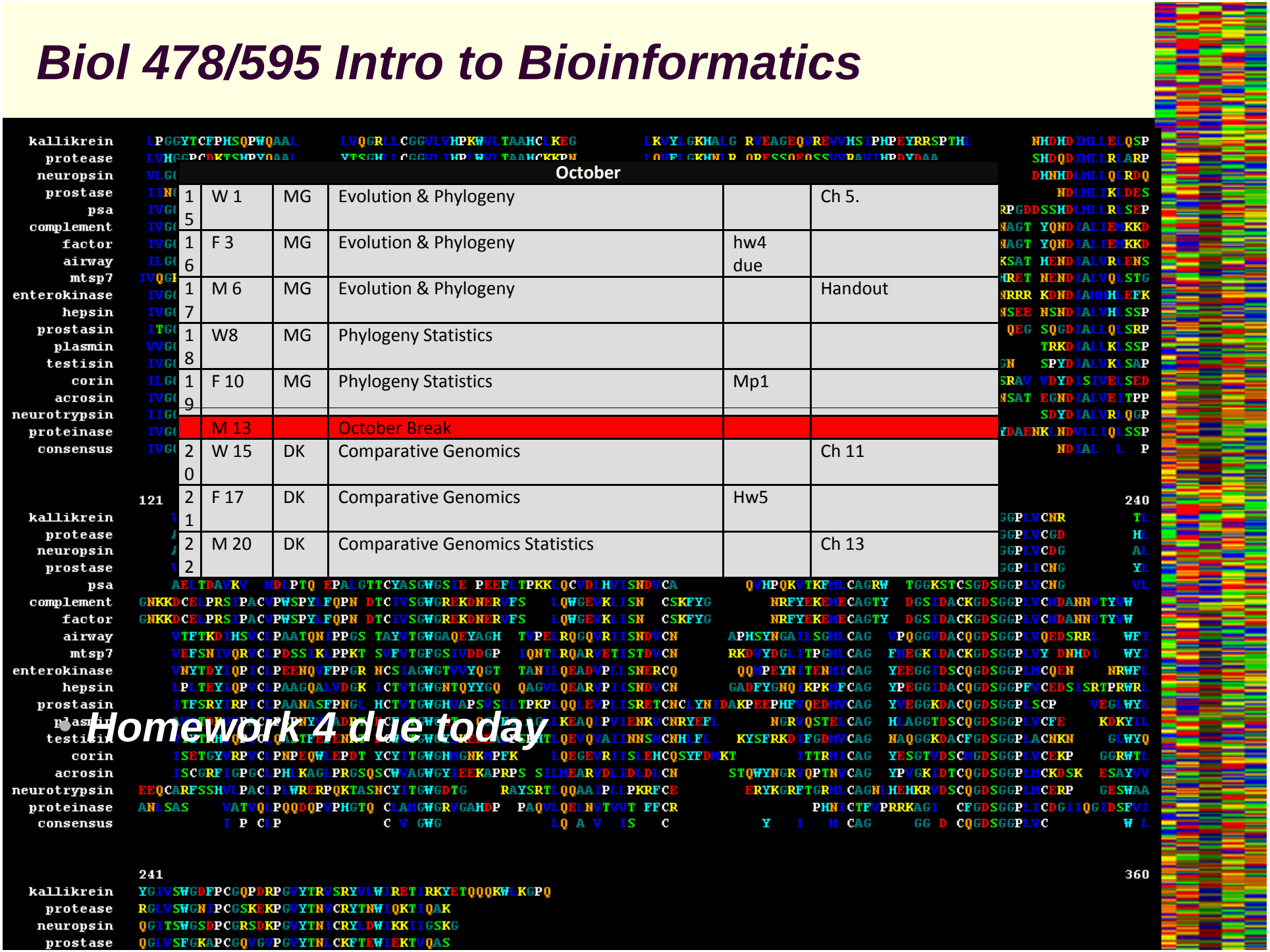
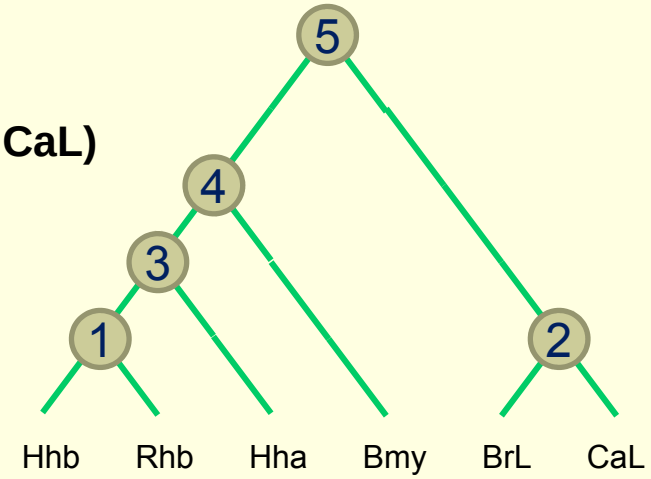
Biol 478/595 Intro to Bioinformatics



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)**

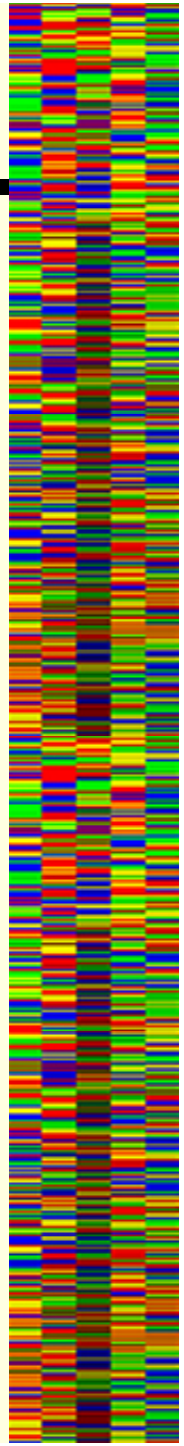


Multiple Alignment and Trees

Progressive Alignments

- **All multiple alignment methods, including multi-dimensional dynamic programming, assume some kind of tree underlying the data!**

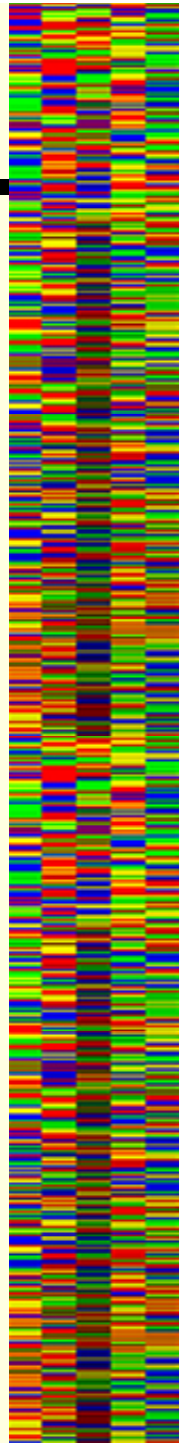
- **Tree construction methods use multiple alignments as data. The realized tree will always tend to look like the guide tree or implicit tree assumed when constructing the multiple alignment**



Multiple Alignment and Trees

Progressive Alignment

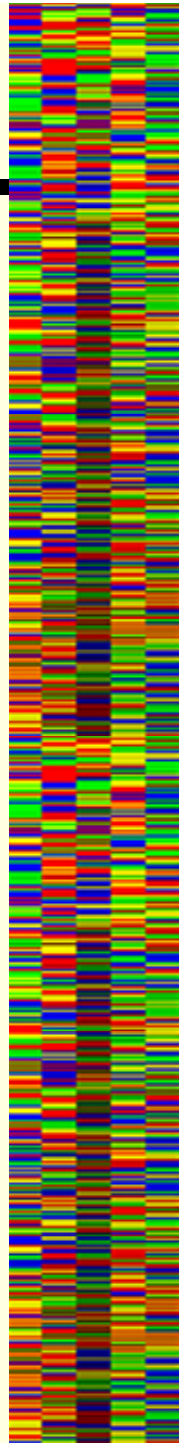
- **Progressive alignments are fast enough to allow hundreds or even thousands of sequences to be aligned**
- **Basic progressive alignment procedure**
  - **Determine distances between sequences**
  - **Use a distance-based method to construct a guide tree for the sequences (UPGMA, Neighbor Joining)**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



Multiple Alignment and Trees

Progressive Alignment

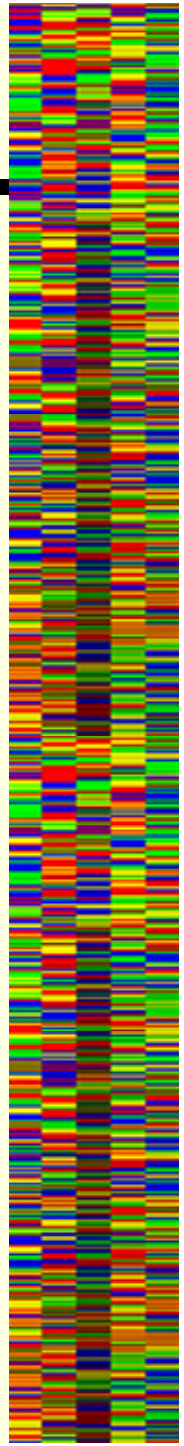
- **General problems with progressive alignments**
  - **Solution is not guaranteed to be optimal. The greedy strategy used in progressive alignment is highly like to be "trapped" in a local optimum.**
  - **Error in early stages propagate through to final stage - there is no error correction possible**
  - **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**

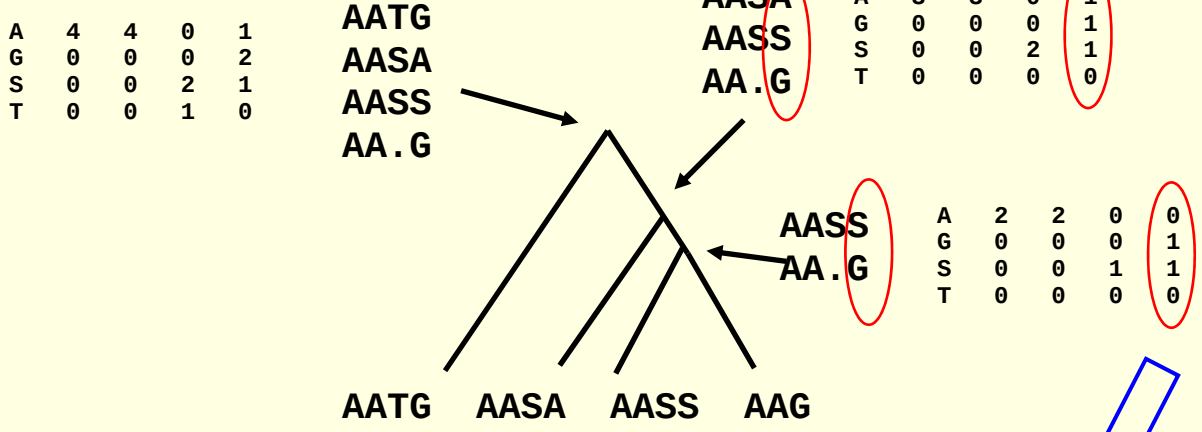


Multiple Alignment and Trees

Progressive alignment

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

Scoring table

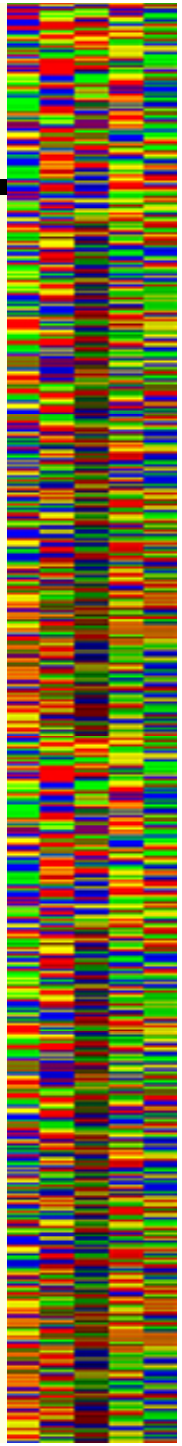


$$S_{ij} = 1 \times G{:}A + 1 \times S{:}A$$

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  MVHLTPEEKSAVTALWGKVNV--DEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK   60
Rana HBB     1  MVHWTAEEKAVINSVWQKVDV--EQDGHEALTRLFIVYPWTQRYFSTFGDLSSPAAIAGNPK   60
Human HBA    3     LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSH------GSAQ   55
Bovine MY    3     LSDGEWQLVLNAWGKVEADVAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASED   61
BrNapa HB    6     FTEKQEALVKESWEILKQDIPKYSLHFFSQILEIAPAAKDMFSFLRD--TDEVHNPNPK   62
CanLi LB     1  MGAFSEKQESLVKSSWEAFKQNVPHHSAVFYTLILEKAPAAQNMFSFL----SNGVDPNNPK   54
                                *                   *

Human HBB   61  VKAHGKKVLGAFSDGLAHLDNLKG------TFATLSELHCDKLHVDPENFRLLGNVLVCV   114
Rana HBB    61  VHAHGKKILGAIDNAIHNLDDVKG------TLHDLSEEHANELHVDPENFRRLGEVLIVV   114
Human HBA   56  VKGHGKKVADALTNAVAHVDDMPN------ALSALSDLHAHKLRVDPVNFKLLSHCLLVT   109
Bovine MY   62  LKKHGNTVLTALGGILKKKGHHEA------EVKHLAESHANKHKIPVKYLEFISDAIIHV   115
BrNapa HB   63  LKAHAVKVFKMTCETAIQLREKGKVVVADTTLQYLGSVHFKSGVLDP-HFEVVKEALVRT   121
CanLi LB    55  LKAHAEKVFKMTVDSAVQLRAKGEVVLADPTL---GSVHVQKGVLDP-HFLVVKEALLKT   114
                  *                                 *

Human HBB  115  LAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH          147   Structural alignment
Rana HBB   115  LGAKLGKAFSPQVQHVWEKFIAVLVDALSHSYH          147   shaded regions show alpha
Human HBA  110  LAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR          141   helix
Bovine MY  116  LHAKHPSDFGADAQAAMSKALELFRNDMAAQYKVLGFHG    147
BrNapa HB  122  LKEGLGEKYNEEVEGAWSKAYDHLALAIKAEMKQEDSQKP   149
CanLi  LB  115  FKEAVGDKWNDELGNAWEVAYDELAAAIKKAMGSA        147
```

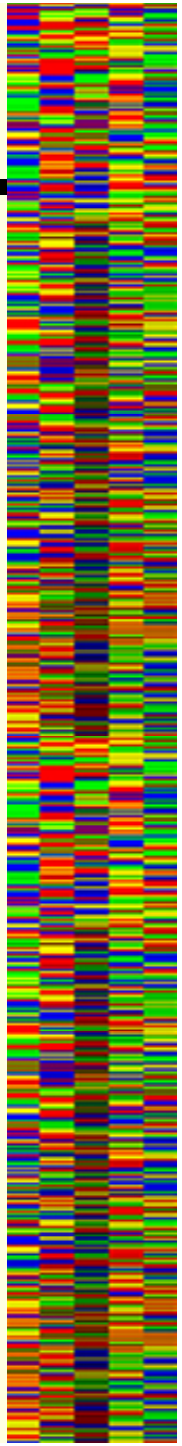
```
Human HBB  1    MVHLTPEEKSAVTALWG--KVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN   58
CANLI LB   1    MGAFSEKQESLVKSSWEAFKQNVPHHSAVFYTLILEKAPAAQNMFSFLSNGVDPN----N   56

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

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

Multiple Alignments and Trees

Progressive alignments

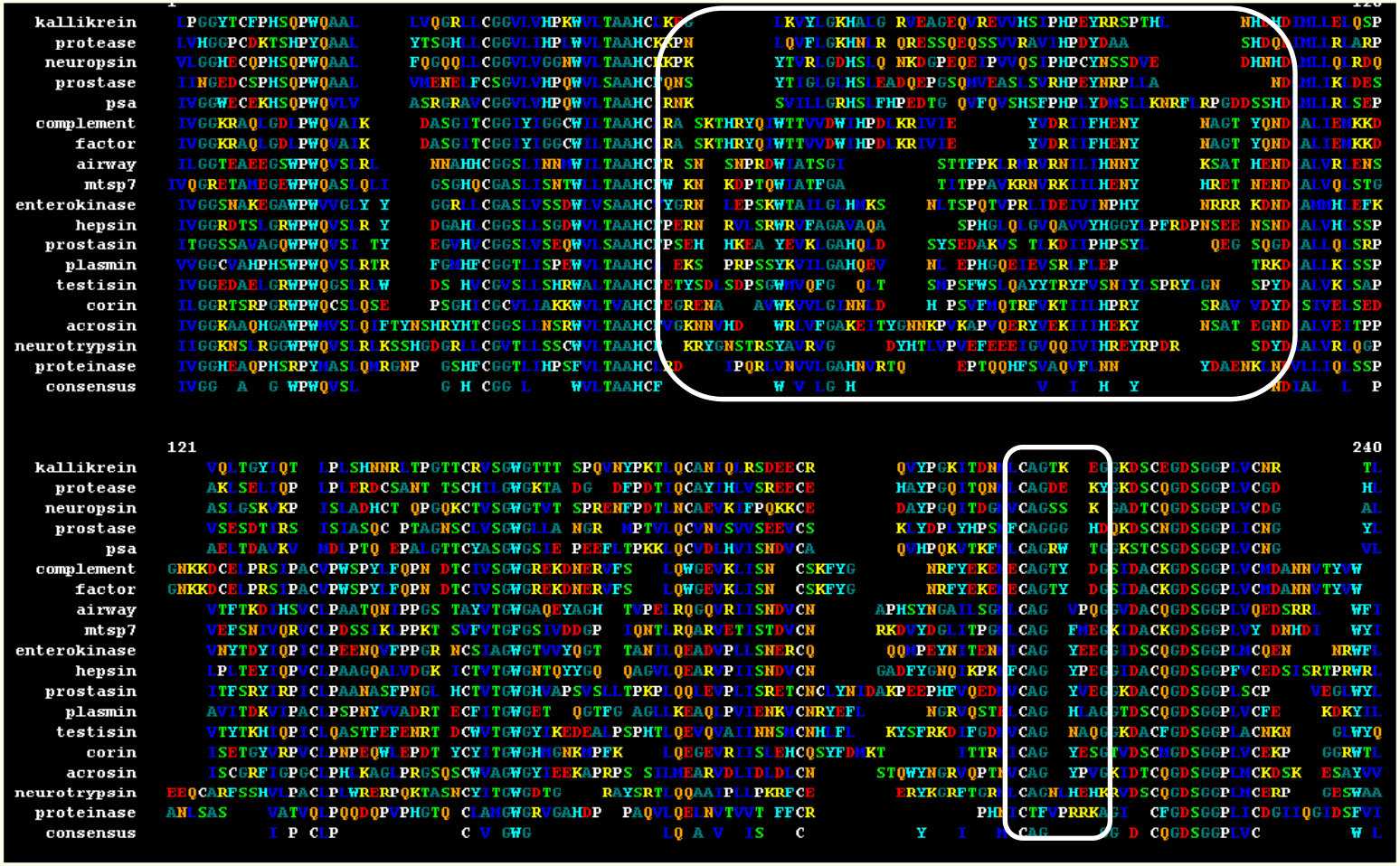
- **No computational alignment is error free**
  - **Inferences based on alignments are only as good as the alignment**
  - **Using automated alignments to build trees guarantees disaster – the guide tree tends to be reproduced because the alignment has been optimized to fit that tree**

- **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**

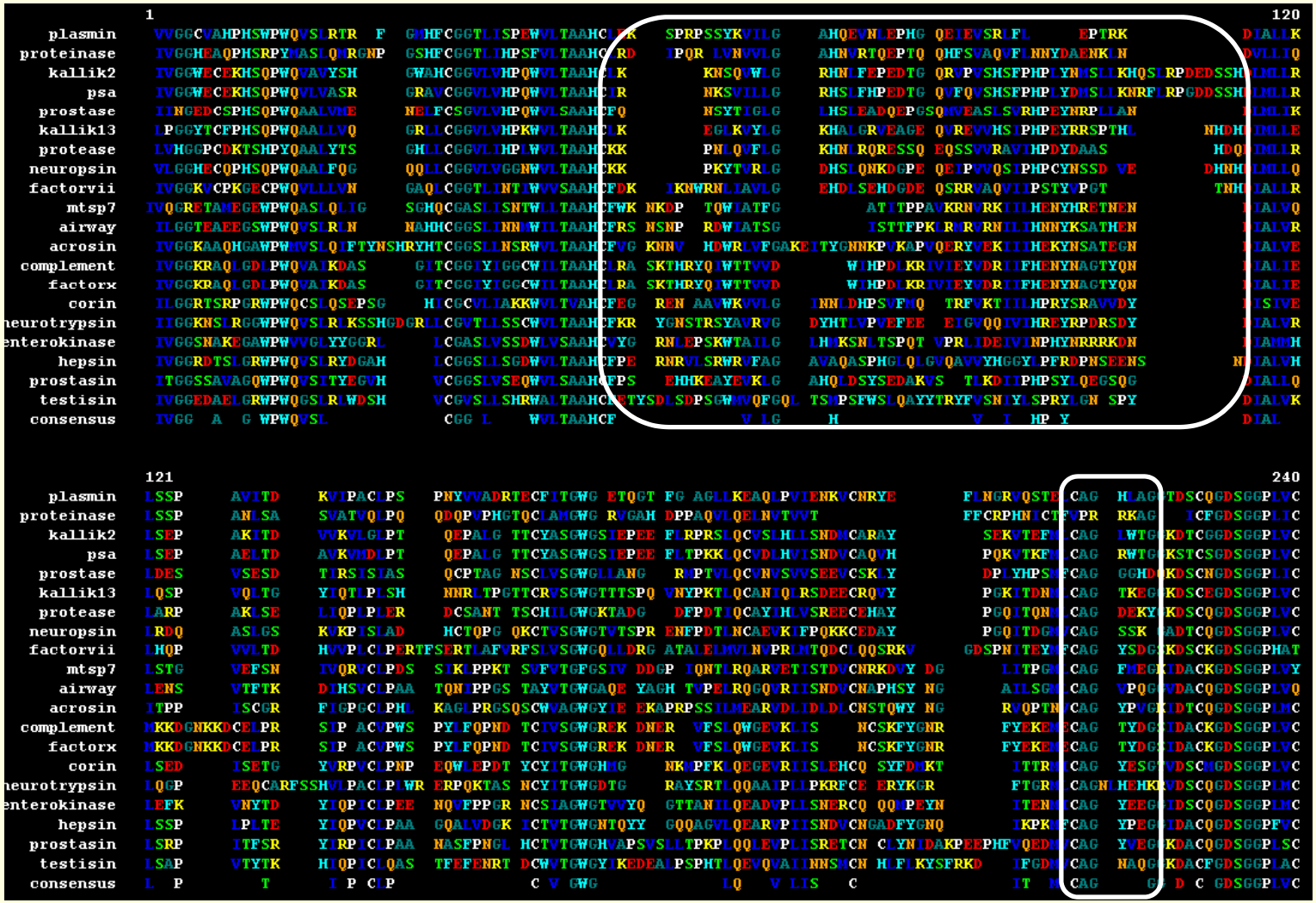


Multiple Alignments and Trees

Manual Correction



Multiple Alignment and Trees

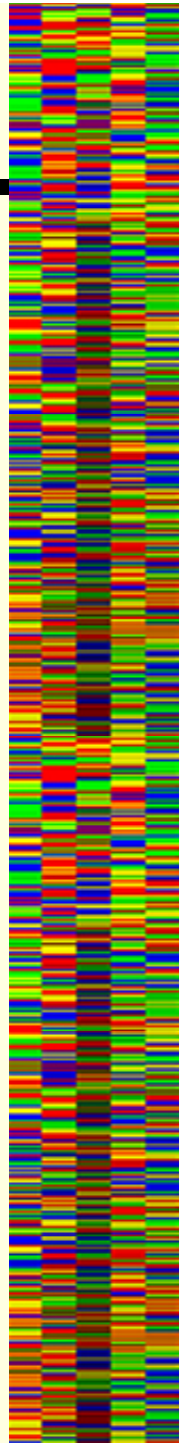


Note: this alignment only partly "corrected"

Multiple Alignment and Trees

Progressive Alignment - ClustalW

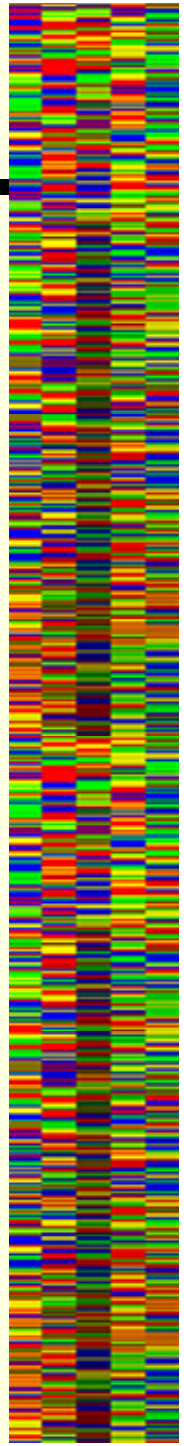
- **Most widely used program for multiple alignment**
- **Incorporates many heuristics in an attempt to improve alignments**
- **Deals well with length differences**
- **Progressive alignment using neighbor joining method for *guide tree***
- **W stands for weighting due to position specific weighting 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 - D2/5) D = 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

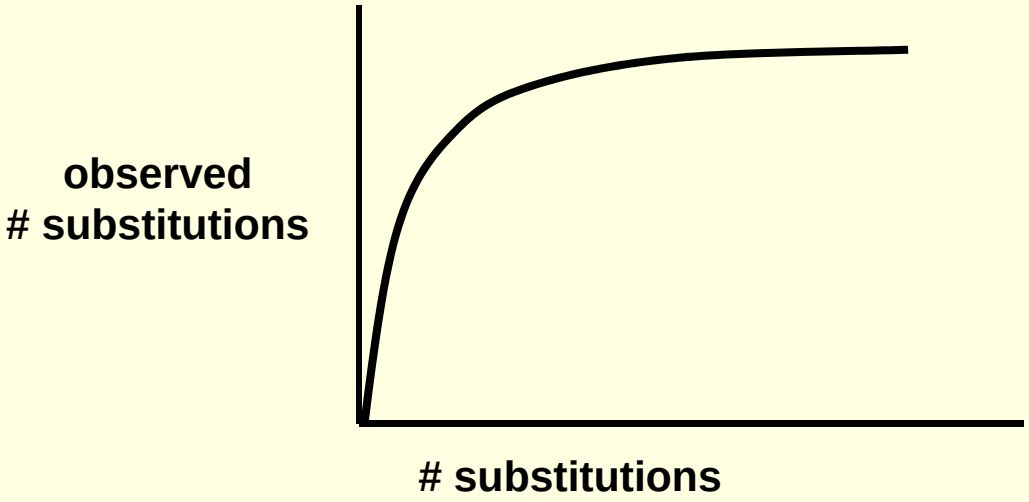
Progressive Alignment – ClustalW

- **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**
  - **Bootstrappingto evaluate final tree**
- **Drawbacks**
  - **Tree/alignment is not global optimum**
  - **Errors introduced early in alignment procedure can't be fixed**
  - **Many (heuristic) parameters**



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

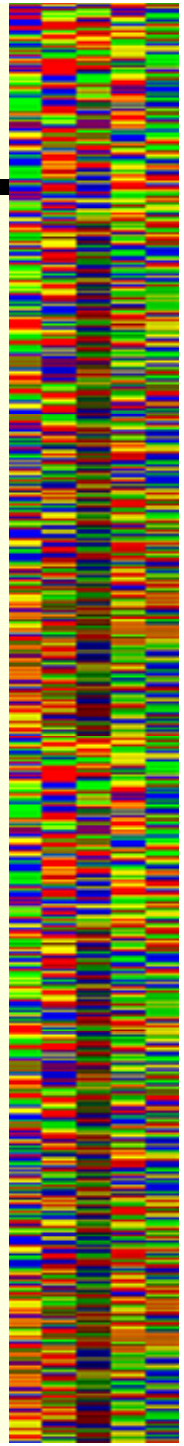
Progressive Alignment – ClustalW

- **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**

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

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

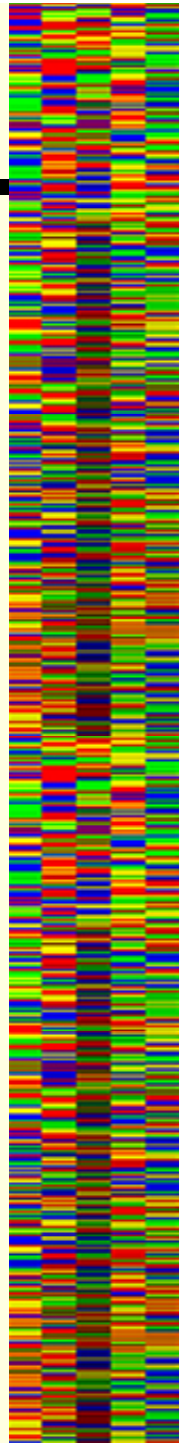
- **where *N* and *M* are the lengths of the sequences, *A* is the average value of a mismatch in the scoring matrix, and *B* is the percent similarity between the two sequences.**



Multiple Alignment and Trees

Progressive Alignment - ClustalW

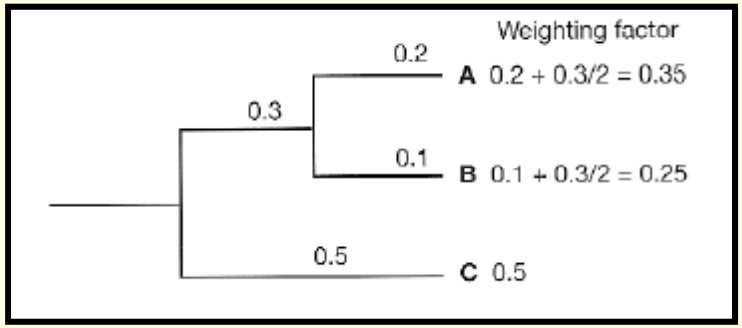
- **Variable gap penalties – accounts for preferences of gaps for surface regions of proteins**
- **Try to guess surface loops**
  - **Use lower gap penalties where gaps occur in one or more sequences**
    - *$GOP = GOP_{default} * 0.3 * (W/N)$*
    - *where W is the number of sequences without gaps and N is the number of sequences (range 0.3→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 →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

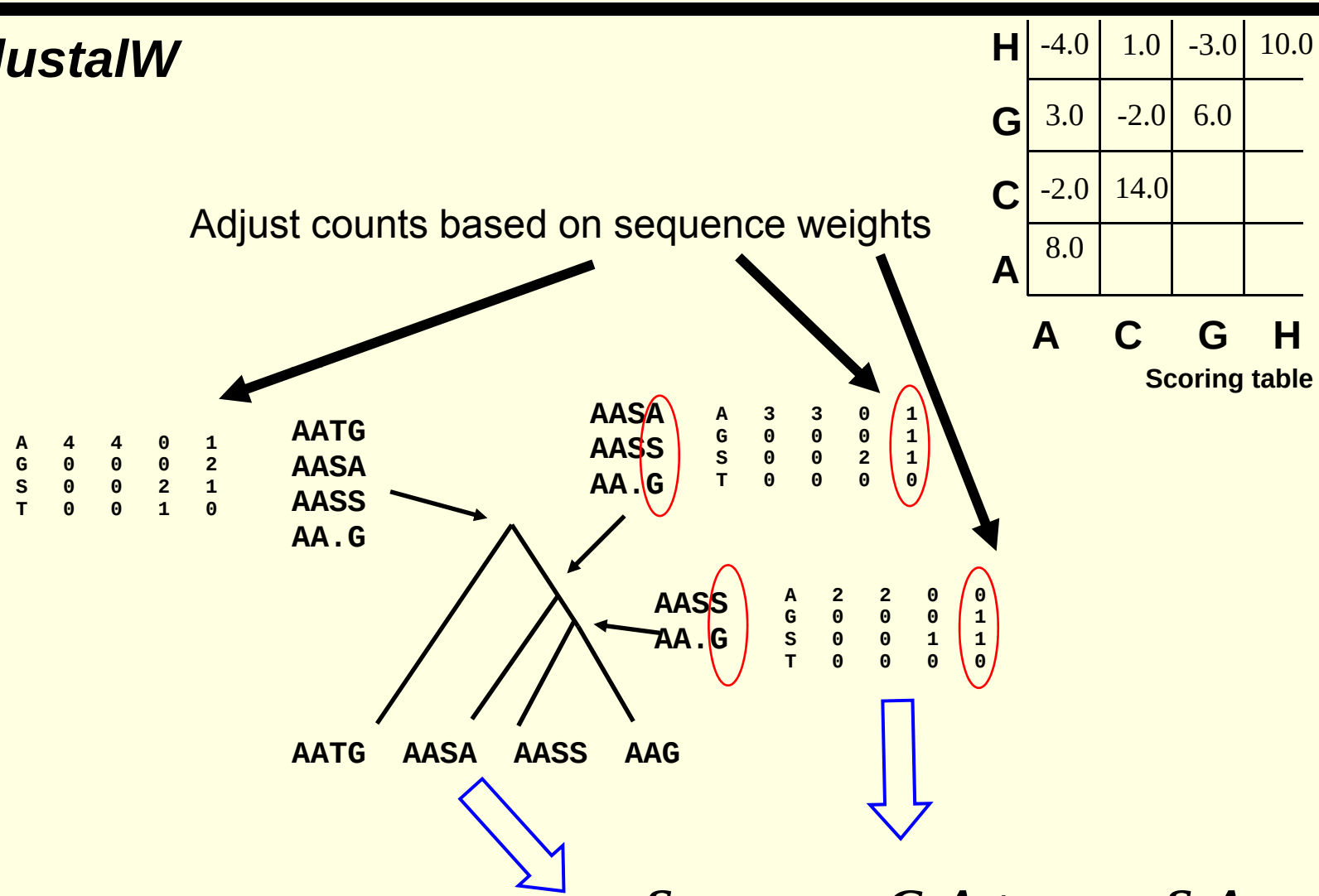
Progressive Alignment – ClustalW

- **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



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

Multiple Alignment and Trees

ClustalW

```
CLUSTAL W (1.8) multiple sequence alignment

HBB3_RANCA      MVH--WTAEEKAVINSVWQKVDVE--QDGHEALTRLFIVYPWTQRYFSTFGDLSSPAAIA
HBB_HUMAN       MVH--LTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVM
HBA_HUMAN       MV---LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS-----H
HBL2_BRANA      MGEIVFTEKQEALVKESWEILKQDIPKYSLHFFSQILEIAPAAKDMFSFLRDTD--EVPH
LGB_CANLI       MG--AFSEKQESLVKSSWEAFKQNVPHHSAVFYTLILEKAPAAQNMFSFLSN----GVDP
MYG_BOVIN       MG---LSDGEWQLVLNAWGKVEADVAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMK
                *     :  :   :   *  .  .    .      ::   * :   *  : .

HBB3_RANCA      GNPKVHAHGKKILGAIDNAIHNLDDVKGTLHD------LSEEHANELHVDPENFRRLGEV
HBB_HUMAN       GNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAT------LSELHCDKLHVDPENFRLLGNV
HBA_HUMAN       GSAQVKGHGKKVADALTNAVAHVDDMPNALSA------LSDLHAHKLRVDPVNFKLLSHC
HBL2_BRANA      NNPKLKAHAVKVFKMTCETAIQLREKGKVVVADTTLQYLGSVHFKSGVLDP-HFEVVKEA
LGB_CANLI       NNPKLKAHAEKVFKMTVDSAVQLRAKGEVVLADPT---LGSVHVQKGVLDP-HFLVVKEA
MYG_BOVIN       ASEDLKKHGNTVLTALGGILKKKGHHEAEVKH------LAESHANKHKIPVKYLEFISDA
                 . .:: *. .:         :       .       *.. * ..  :    :  : .

HBB3_RANCA      LIVVLGAKLGKAFSPQVQHVV
HBB_HUMAN       LVCVLAHHFGKEFTPPVQAAY
HBA_HUMAN       LLVTLAAHLPAEFTPAVHASL
HBL2_BRANA      LVRTLKEGLGEKYNEEVEGAW
LGB_CANLI       LLKTFKEAVGDKWNDELGNAW
MYG_BOVIN       IIHVLHAKHPSDFGADAQAAM
                :: .:       :
```

```
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
```

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:
```

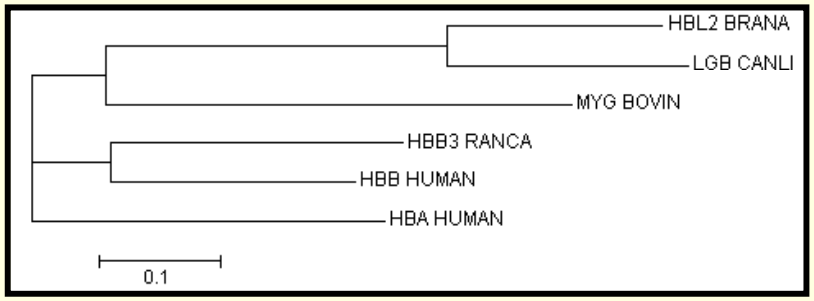
Multiple Alignment and Trees

ClustalW

- **Trees - Newick format**

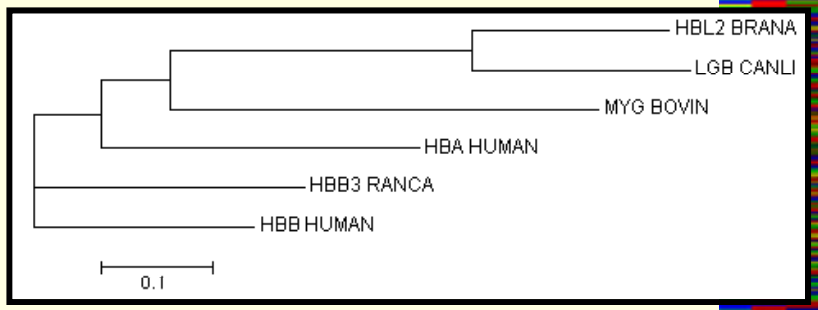
```
(
(
(
HBL2_BRANA:0.17743,
LGB_CANLI:0.19841)
:0.28051,
MYG_BOVIN:0.38318)
:0.05975,
(
HBB3_RANCA:0.24072,
HBB_HUMAN:0.20145)
:0.06376,
HBA_HUMAN:0.28909);
```

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)**