BIoinformatics 1-- lecture 8

Multiple sequence alignment

In class competition: Editing a multiple sequence alignment in Geneious

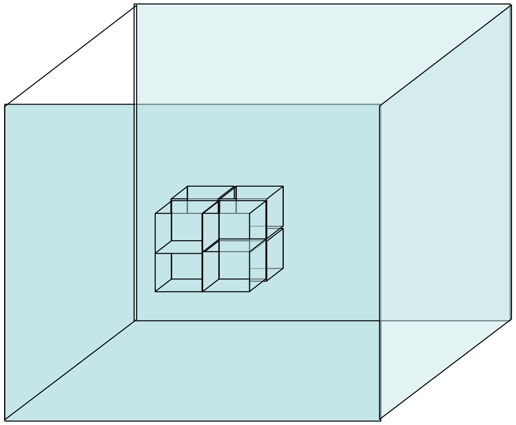
- Download and open “bad alignment” from the course web page
- --OR-- Open the Collaboration folder, find rpibioinfo. Look in Shared_sequences. Drag “bad alignment” to your in-class exercise folder

Fix the alignment

- Set *allow editing*
- Set *highlighting* (*agreements to consensus*)
- Move sequences around by adding/ removing gaps.
- Do not delete or change amino acids!
- Keep gaps together. (think global alignment, end gaps count)
- Maximize **pairwise % identity**
- Minimize alignment **length**.
- The winner has the highest % identity, with shortest Length.

All alignment methods are fundamentally pairwise

- Can we do Dynamic Program for three sequences*?

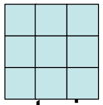


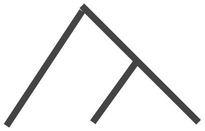
S(i,j,k) = MAX {
A(i-1,j-1,k-1)+S(i,j,k),
A(i-1,j,k)-gap,
A(i,j-1,k)-gap,
A(i,j,k-1)-gap,
A(i-1,j-1,k)-gap,
A(i-1,j,k-1)-gap,
A(i,j-1,k-1)-gap }

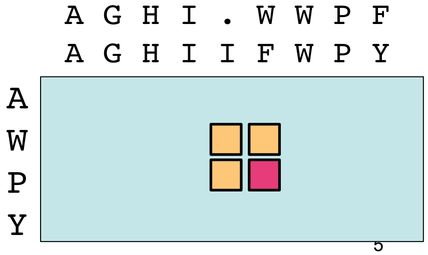
*or more?

Progressive alignment

1. align all pairs
2. pairwise align two most similar first
3. align next most similar to previous alignment
4. repeat until all sequences are aligned



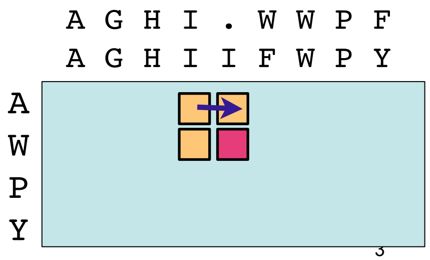




```
A G H I . W W P F
A G H I I F W P Y
```

S(P,[W,F]) =(1/2)(S(P,W) + S(P,F))

No gap penalty for aligning a gap to a gap



A(i,j) = A(i-1,j) - gap(i)

If *i* is already a gap position in any sequence, set gap(i)=0.

Progressive, using Pairwise distances

- Guide tree
- progressive alignment

	A	B	C	D	E	F
A		97	81	82	59	32
B			77	80	55	31
C				90	65	40
D					61	42
E						33
F						

A
B
C
D
E
F

Draw guide tree here

Fill in J-C distances on whiteboard.

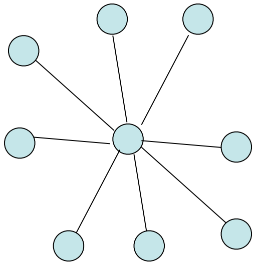
Sequence distance versus similarity

Maximizing similarity and Minimizing distance are equivalent if

- $d(i,j) + s(i,j) = s_{max}$,
  where s_{max} is the maximum possible similarity, and the minimum distance is d=0. For each position in the alignment.
- Distance based on identity score (p-distance)
  $d = 100 - \%identity$
- Distance using empirical J-C correction
  $d = -\ln((S_{real}-S_{rand})/(S_{ident}-S_{rand}))$
  where S_{ident} = score of an identity alignment, and S_{rand} = typical score of a false alignment.
- For proteins, S_{rand} = 25%. “Twilight zone” (R. Doolittle, 1986)

Star, using all-to-one distances

- no guide tree
- star alignment, all sequences are aligned one.
- BLAST does this

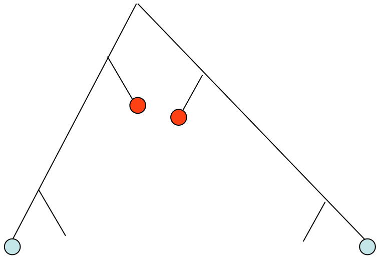


Multiple sequence alignment

- The power of many....
- A is not detectably similar to B, but C is similar to A and C is similar to B. Therefore A is homologous to B.
- Transitive-BLAST = using the hits of a BLAST search to do additional BLAST searches.

Evolution is a random walk through sequence space.... at various speeds

- A phylogram showing long distances (bacteria) and short distances (plants, animals, subsurface bacteria)
- Slowly evolving sequences serve as bridges in a transitive-BLAST search.



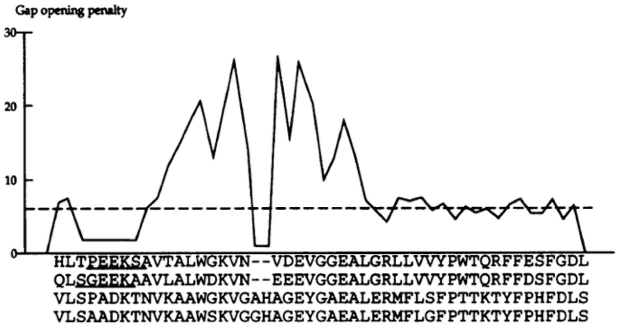
CLUSTALW

JD Thompson, DG Higgins, TJ Gibson - Nucleic acids research, 1994

- Start with unrooted tree, using Neighbor joining.
- choose root to get guide tree
- progressive alignment
  - matches are scored using sequence weights
  - gaps are position dependent
    - GOP lower for polar residues
    - GOP zero where there is already a gap

Install CLUSTALW

CLUSTALW Position specific gap penalty



```
HLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDL
QLSGEEKAAVLALWDKVN--EEEVGGEALGRLLVVYPWTQRFFDSFGDL
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLS
VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLS
```

Figure 3. The variation in local gap opening penalty is plotted for a section of alignment. The inital gap opening penalty is indicated by a dotted line. Two hydrophilic stretches are underlined. The lowest penalties correspond to the ends of the alignment, the hydrophilic stretches and the two positions with gaps. The highest values are within 8 residues of the two gap positions. The rest of the variation is caused by the residue specific gap penalties (12).

MUSCLE

RC Edgar - Nucleic acids research, 2004

- Iterative MSA
  - k-mer distance matrix
  - UPGMA tree
  - **progressive alignment**--> MSA1
  - Kimura distances from MSA1
  - UPGMA tree
  - **progressive alignment** -->MSA2
  - For all tree branches:
    - split tree into two
    - calculate profiles
    - align profiles
    - accept or reject the alignment.
    - Repeat

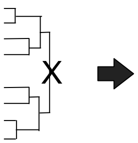
based on short identical matches

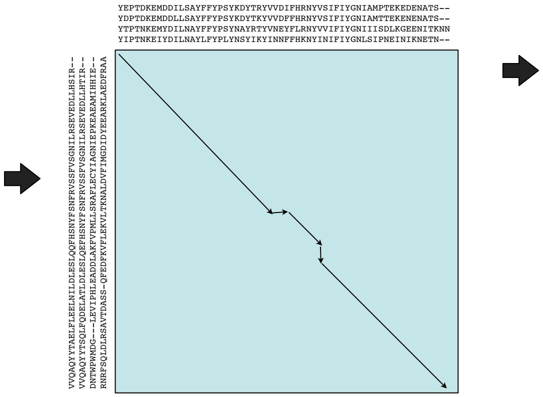
Z&B p174

Install MUSCLE. Try it.

MUSCLE iterative alignment

```
XP_001615335   YEPTDKEMDDILSAYFFYPSYKDYTRYVVDIFHRNYVSIFIYGNIAMPTEKEDENATS--
XP_002259219   YDPTDKEMDDLLSAYFFYPSYKDYTKYVVDFFHRNYVSIFIYGNIAMTTEKENENATS--
XP_001347897   YTPTNKEMYDILNAYFFYPSYNAYRTYVNEYFLRNYVVIFIYGNIIISDLKGEENITKNN
   XP_726635   YIPTNKEIYDILNAYLFYPLYNSYIKYINNFFHKNYINIFIYGNLSIPNEINIKNETN--
   XP_671449   ------------------------------------------------------------
XP_001458064   VVQAQYYTAELFLEELNILDLESLQQFHSNYFSNFRVSSFVSGNILRSEVEDLLHSIR--
XP_001347129   VVQAQYYTSQLFQDELATLDLESLQEFHSNYFSNFRVSSFVSGNILRSEVEDLLHTIR--
XP_002283970   DNTWPWMDG---LEVIPHLEADDLAKFVPMLLSRAFLECYIAGNIEPKEAEAMIHHIE--
XP_002367832   RNRFSQLDLRSAVTDASS-QFEDFKVFLEKVLTKNALDVFIMGDIDYEEARKLAEDFRAA
```




YEPTDKEMDDILSAYFFYPSYKDYTRYVVDIFHRNYVSIFIYGNIAMPTEKEDENATS--
YDPTDKEMDDLLSAYFFYPSYKDYTKYVVDFFHRNYVSIFIYGNIAMTTEKENENATS--
YTPTNKEMYDILNAYFFYPSYNAYRTYVNEYFLRNYVVIFIYGNIIISDLKGEENITKNN
YIPTNKEIYDILNAYLFYPLYNSYIKYINNFFHKNYINIFIYGNLSIPNEINIKNETN--
VVQAQYYTAELFLEELNILDLESLQQFHSNYFSNFRVSSFVSGNILRSEVEDLLHSIR--
VVQAQYYTSQLFQDELATLDLESLQEFHSNYFSNFRVSSFVSGNILRSEVEDLLHTIR--
DNTWPWMDG---LEVIPHLEADDLAKFVPMLLSRAFLECYIAGNIEPKEAEAMIHHIE--
RNRFSQLDLRSAVTDASS-QFEDFKVFLEKVLTKNALDVFIMGDIDYEEARKLAEDFRAA

```
YEPTDKEMDDILSAYFFYPSYKDYTRYVVDIFHRNYV..SIFIYGNIAMPTEKEDENATS--
YDPTDKEMDDLLSAYFFYPSYKDYTKYVVDFFHRNYV..SIFIYGNIAMTTEKENENATS--
YTPTNKEMYDILNAYFFYPSYNAYRTYVNEYFLRNYV..FIYGNIIISDLKGEENITKNN
YIPTNKEIYDILNAYLFYPLYNSYIKYINNFFHKNYI..NIFIYGNLSIPNEINIKNETN--
VVQAQYYTAELFLEELNILDLESLQQFHS..NYFSNFRVSSFVSGNILRSEVEDLLHSIR--
VVQAQYYTSQLFQDELATLDLESLQEFHS..NYFSNFRVSSFVSGNILRSEVEDLLHTIR--
DNTWPWMDG---LEVIPHLEADDLAKFVP..MLLSRAFLECYIAGNIEPKEAEAMIHHIE--
RNRFSQLDLRSAVTDASS-QFEDFKVFLE..KVLTKNALDVFIMGDIDYEEARKLAEDFRAA
```

In one iteration:
The phylogenetic tree is cut at a random branch, the two subtrees are converted to profiles, and aligned. The new alignment is either accepted or rejected

Building and pruning multiple sequence alignments

- Steps in making a good MSA
  - – Database search
  - – Automatic multiple sequence alignment
  - – Removing N and C-terminal extensions, if necessary
  - – Removing redundant sequences, if necessary
  - – Removing false hits, if necessary
  - – Manual re-alignment, if necessary

Phyogenetic trees

What is a phylogenetic tree?

A model of evolutionary relationships -- common ancestors and speciation events.

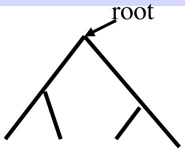
Why build phylogenetic trees?

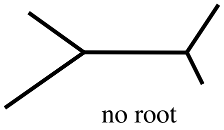
To trace the branch order of "taxa" (taxon = a gene, a species, a population, etc.)

To understand the evolution of traits

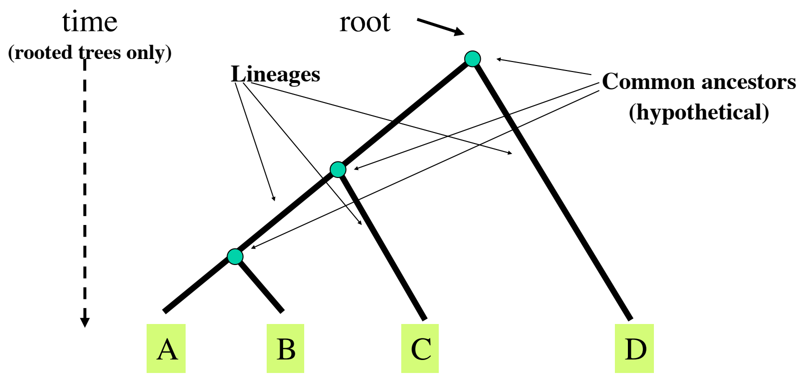
As part of a multiple sequence alignment algorithm

Trees can be "rooted" or "unrooted"



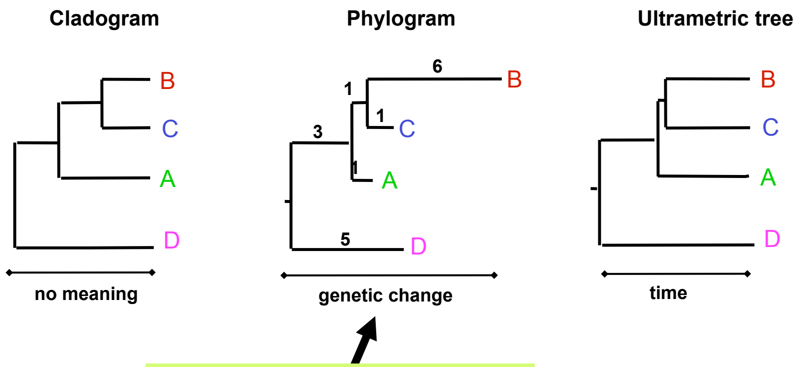


Tree Terminology



Terminal nodes (leaves) represent taxa, which are observed species/genes/populations.

Evolutionary time

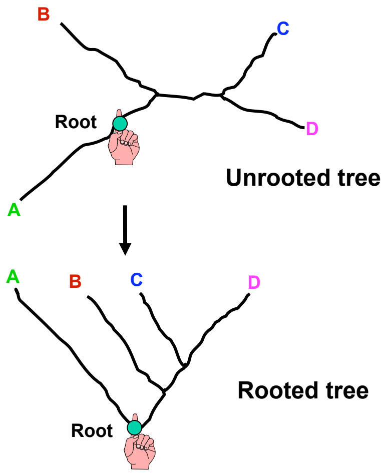


(D:5,(A:1,(C:1,B:6):1):3)

parenthesis (notation can have both labels and distances.

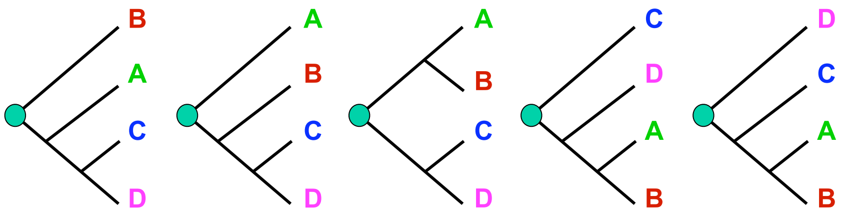
Inferring evolutionary *relationships* between the taxa requires rooting the tree:

To root a tree mentally, imagine that the tree is made of string. Grab the string at the root ● and tug on it until the ends of the string (the taxa) fall opposite the root:



Where the tree is rooted changes its meaning.

Each of these trees is possible by choosing a different root.

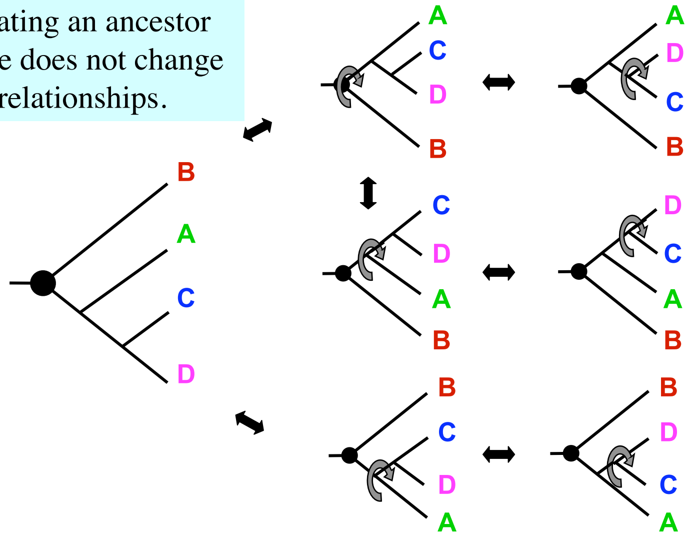


This one says
C and D
branched *late*.

This one says C
and D branched
early.

On the other hand, taxon order doesn't matter.

Rotating an ancestor node does not change the relationships.

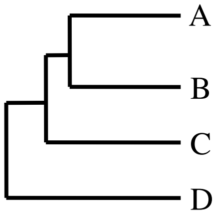


(, (, (,)))

Trees can be represented in "parenthesis notation".

Each set of parentheses represents a branch-point (bifurcation), the comma separates left and right lineages.

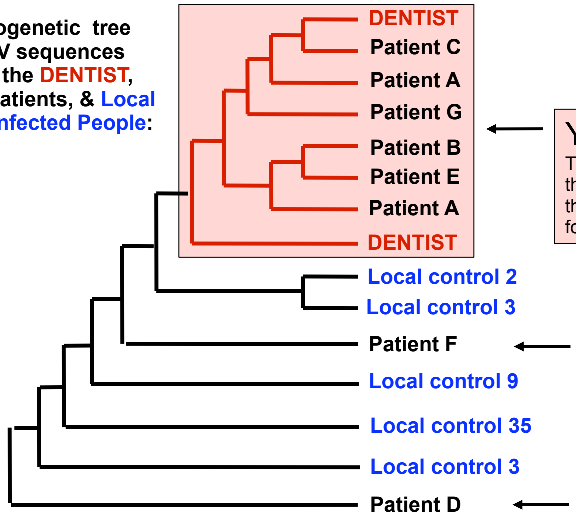
(D,(C,(A,B))) =



Parenthesis notation can contain sequence labels too.

Did the *Florida Dentist* infect his patients with HIV?

Phylogenetic tree of HIV sequences from the DENTIST, his Patients, & Local HIV-infected People:



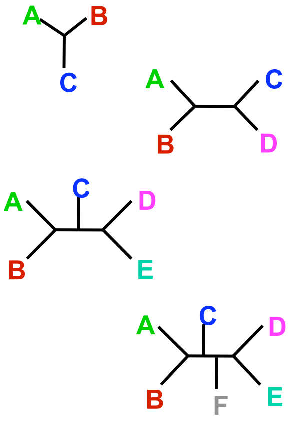
Yes:
The HIV sequences from these patients fall within the clade of HIV sequences found in the dentist.

No

No

From Ou et *al*. (1992) and Page & Holmes (1998)

Explosive tree growth



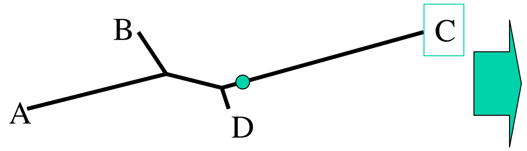
# Taxa (N)	# Unrooted trees
3	1
4	3
5	15
6	105
7	945
8	10,935
9	135,135
10	2,027,025
.	.
.	.
.	.
.	.
30	≈3.58 x 10^{36}

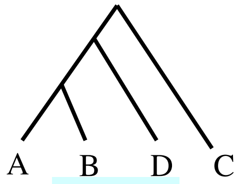
Methods that try all possible trees are possible only for small numbers of taxa.

$(2N-5)!/[2^{N-3}*(N-3)!]$ = # unrooted trees for N taxa

Two strategies for rooting a tree:

1. Choose the **midpoint** between the two most distant branches.

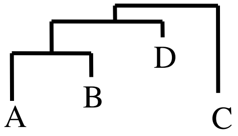




cladogram

2. Choose one taxon as the "**out group**." (it branches first.)

A good outgroup is not too distant from the rest of the tree.



phylogram