CMPT 830 - Bioinformatics and Computational Biology Chapter 5: Phylogenetic Trees

Ian McQuillan and Tony Kusalik
kusalik@cs.usask.ca

October 9, 2019

1

1 Phylogenetics

Phylogenetics

Definition 1. Phylogenetics is the study of evolutionary relationships.

- Taxonomists have been naming, grouping and classifying organisms for a long time.
- Now that we have lots of sequence data, we are using that more and more than just examining phenotype or morphology.
- Using phenotype alone had always been very difficult, especially for bacteria.

2

Phylogenetics

We would like to:

1. Infer genealogical relationships between different taxa.
2. Estimate the time of divergence between different taxa.

Definition 2. A *phylogenetic tree* is a graph in which the taxonomic units are nodes (or vertices) and their relationships are represented by branches (or edges).

3

Goals of Molecular Phylogeny

- Usually evolutionary relationships are depicted using a tree.
- All life shares a common origin, and are part of a single tree of life.
- This single tree of life should depict all species.

4

Tree of Life

- Tracing evolutionary history also involves extinct species.
- More than 99% of species that have existed are now extinct.[1]
- So, trying to determine the tree of life can be very tricky.

5

Homology and Phylogeny

- We've looked at homology through the lens of pairwise and multiple sequence alignment.
- If a group of sequences are homologous, then they share a common ancestor.
- That means they've all diverged from the common ancestor.
- The order of divergence can be described using a phylogenetic tree.

6

Homology and Phylogeny

- Any group of homologous proteins (or nucleic acid sequences) can be represented as a phylogenetic tree.
- Instead of viewing a set of homologous sequences with alignments, trees can be a useful depiction.
- The relationships between taxa are not obvious from a multiple sequence alignment, or a set of pairwise alignments.
- BLAST can give us back a bunch of orthologs in different organisms, but how are they related?

7

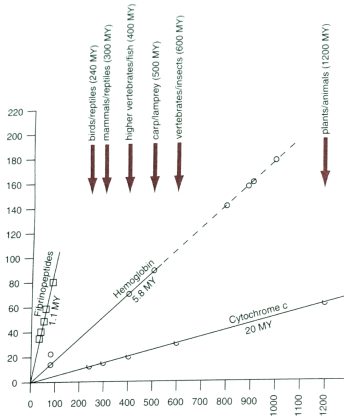
Molecular Clock Hypothesis

- In 1962, Zuckerkandl and Pauling and Margoliash proposed the notion of a "molecular clock".
- For any given gene or protein, the rate of molecular evolution is approximately constant.
- In 1971, Dickerson looked at three proteins with sequences available in a number of organisms.
- He plotted the relationship between the number of differences in amino acid sequences versus the divergence time for the organisms.
- The divergence times were all estimated using palaeontology techniques.

8

[1] Wilson, *The Diversity of Life*, 1992.

Molecular Clock Hypothesis



- The x-axis the in millions of years since divergence.
- The y-axis is in the amino acid changes per 100 residues.

9

Molecular Clock Hypothesis

- The amount of changes vs. time for each of those proteins was different.
- But each was very close to linear
- The slopes for each protein differed **significantly**.

10

Molecular Clock Hypothesis

- This means that the rate of change over time differs significantly for each protein.
- As an example, for a 1% change in amino acid sequence to occur requires 20 million years for cytochrome c protein.
- It requires only 1.1 million years for fibrinopeptides to collect the same number of changes.
- The rate of change is roughly constant for each individual protein tested.

11

Image from Pevsner, *Bioinformatics and Functional Genomics*.

Watch Out!

- We are measuring the number of substitutions per amino acid per year.
- This is different than the number of mutations.
- Mutation occurs when a biochemical process results in a change in sequence.
- But looking at observed substitutions is more complicated.

12

Selection

- The observed substitutions depends on mutation, but also on selection.
- For example, some substitutions are selected against.
- This doesn't mean that those mutations do not happen.
- It just means that these changes were deleterious to the organism.

13

Selection

- For example, substitutions in histones are almost never tolerated.
- If the mutation rate is relatively constant, then any changes in substation rates is due entirely to selection.
- Positive selection means change is encouraged.
- Negative selection means change is discouraged.

14

Molecular Clock Hypothesis

- But in principle, the molecular clock hypothesis is useful.
- It allows us to date phylogenetic relationships.
- If proteins are evolving at constant rates, that can be used to estimate time since the sequences diverged.

15

Calculation

It is possible to estimate the rate of nucleotide substitution over time.

Rates

The rate can be calculated by

$$r = \frac{K}{2T}$$

where

- T is the time of divergence between two sequences from a common ancestor,
- $2T$ accounts time of divergence across two separate lineages,
- K is the number of substitutions per site.

16

Calculation

- Of the variables, T can sometimes be calculated by using fossil data.
- For example, humans and rodents diverged approximately 80 million years ago.
- Alpha globins from rat and human differ by 0.093 nonsynonymous[3] substitutions per site.
- So,

$$r = \frac{0.093 \text{ substitutions/site}}{2(8 \times 10^7 \text{years})}.$$

17

Solve for T

- By the same token, we can solve for another variable.
- Of particular interest is solving for T,

$$T = \frac{K}{2r}.$$

- This gives a technique of predicting the time of divergence without new fossil data.

18

Solve for Time

- Let's say we've already figured out the rate r by looking at the same protein in a few species.
- Now let's say we sequence the same protein in a different species.
- Then we can estimate the time of divergence of the new species to the others.
- Solve for T.

19

[3] nucleotide substitutions that cause a change in amino acids

Caveats

- There are some caveats.
- The molecular clock hypothesis does not apply exactly to all proteins.
- Rate of evolution at the molecular level can vary between organisms.
  - Viruses often change rapidly.
  - It is known that rodents tend to have a faster molecular clock than primates.

20

Caveats

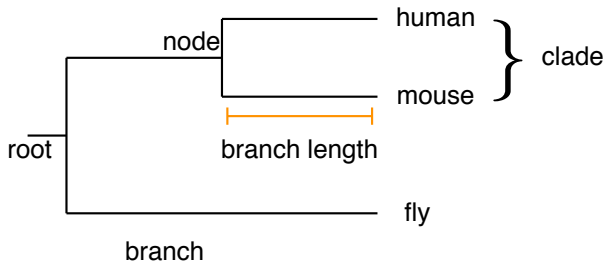
- Rate of evolution can be affected by environmental factors; e.g. geographical location, changes in climate, shifts in composition of ecological communities
- Also, if a gene changes its function, it can be under different selective pressures.
- This can cause substitution rates to change between similar proteins.
- Genes that become non-functional often mutate far more rapidly.

21

2 Phylogenetic Trees

Phylogenetic Trees

- Nodes can be species, populations, individuals or molecules. Referred to as OTU, operational taxonomic unit.
- The root is the common ancestor of all the taxa.
- Usually, the external nodes (or leaves) represent existing organisms and the internal nodes represent hypothetical, or perhaps extinct species.
- A *clade* is a subtree at a given node.



22

Phylogenetic Trees

There are:

1. Scaled trees: branch lengths are proportional to the differences between pairs of neighbouring nodes.
2. Unscaled trees: the branching is only representing the ancestral information, without involving distances between neighbours.

23

Phylogenetic Trees

- If a phylogenetic tree is based on a single gene, it is usually referred to as a *gene tree*, rather than a *species tree*.
- One must be careful, as there are situations when an individual from one species can have a gene more similar to other individuals from other species rather than his own.
- The situation is further complicated with *horizontal gene transfer* - the process of transferring genetic material to another cell which is not its offspring.

24

Genetic Transfer

- Homologs: sequences that have common ancestral origins.
  - Orthologs: homologs produced by speciation.
  - Paralogs: homologs produced by gene duplication.
  - Xenologs: homologs resulting from horizontal gene transfer.
- If four genes in a tree are related to each other in the same way in which the organisms are related, then likely the genes are othologs.
- If the trees are much different, they could be xenologs.

25

Phylogenetic Analysis

Usually phylogenetic analysis is based on either

1. Distance-based methods,
2. Character-based methods.

26

3 Distance-Based Methods

Distance-Based Methods

- Compute pairwise distances (using some metric), and then use only this distance to derive the trees, rather than using the original data.

  Example 3. Let's say that we are trying to construct a phylogenetic tree for OTUs (e.g. species) a, b, c, d, e. We get our pairwise distance matrix.

	a	b	c	d	e
a	-	-	-	-	-
b	$d_{a,b}$	-	-	-	-
c	$d_{a,c}$	$d_{b,c}$	-	-	-
d	$d_{a,d}$	$d_{b,d}$	$d_{c,d}$	-	-
e	$d_{a,e}$	$d_{b,e}$	$d_{c,e}$	$d_{d,e}$	-

27

UPGMA

- One method is the *unweighted-pair-group method with arithmetic mean* (or UPGMA, for short).
- After the pairwise distances have been calculated, UPGMA groups together the two OTUs that have the least distance.

28

UPGMA

- If OTUs (e.g. species) a and b are the most similar, then we group them into (ab) and we construct a new matrix with rows and columns $(ab), c, d, e$.
- The entries involving only c, d, e remain the same; however, for those involving a or b, we calculate the entries as follows:

  procedure

  For each species $x \in \{c, d, e\}$, we let

$$d_{(ab),x} = 1/2(d_{a,x} + d_{b,x}).$$

29

UPGMA

- We then continue the process over and over again until all the species have been grouped.
- As we "move up" the tree, the smallest distance could be the result of merging two clades with more than one taxon (or one taxon and one clade).

30

UPGMA

- At each step, we compute the average distance of one clade Y to another clade Z as follows:

$$d_{Y,Z} = \frac{1}{|Y||Z|} \sum_{y \in Y} \sum_{z \in Z} d_{y,z}.$$

- This is a more general version of our previous formula:

  procedure
  For each species $x \in \{c, d, e\}$, we let

$$d_{(ab),x} = 1/2(d_{a,x} + d_{b,x}).$$

31

UPGMA example

Example 4. Consider the following sequences that exist for species a, b, c, d.

```
a: ATGCTAAGTTGCCAGGCGTGTGAACCCTGT
b: ATGTTGAGTTCCGAGGCGTTTGCACCTTGT
c: ATGCCAAGTTCCGAGGCGTTTGAACCTTGT
d: ATGGGCCGTTGGGAGGCGTTTGCACCGTGT
```

If we are just calculating our distance based on the number of mismatches, then this would be our distance matrix:

	a	b	c	d
a	-	-	-	-
b	6	-	-	-
c	5	4	-	-
d	9	7	8	-

32

Example Continued

Example 5.

- The smallest distance is between b and c, so we group those together and calculate our new distance matrix:

- General formula:

$$d_{Y,Z} = \frac{1}{|Y||Z|} \sum_{y \in Y} \sum_{z \in Z} d_{y,z}$$

- $d_{(bc),a} = \frac{1}{2} \sum_{y \in \{b,c\}} \sum_{z \in \{a\}} d_{y,z} = \frac{1}{2}(d_{b,a} + d_{c,a}) = \frac{1}{2}(6 + 5) = 5.5,$
- $d_{(bc),d} = \frac{1}{2} \sum_{y \in \{b,c\}} \sum_{z \in \{d\}} d_{y,z} = \frac{1}{2}(d_{b,d} + d_{c,d}) = \frac{1}{2}(7 + 8) = 7.5$

	a	(bc)	d
a	-	-	-
(bc)	5.5	-	-
d	9	7.5	-

33

Example Continued

Example 6.

- The smallest is between (bc) and a.
- General formula:

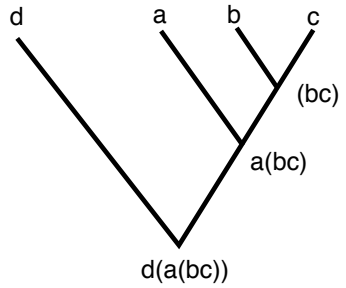
$$d_{Y,Z} = \frac{1}{|Y||Z|} \sum_{y \in Y} \sum_{z \in Z} d_{y,z}$$

- $d_{a(bc),d} = \frac{1}{3} \sum_{y \in \{a,b,c\}} \sum_{z \in \{d\}} d_{y,z} = \frac{1}{3}(d_{a,d} + d_{b,d} + d_{c,d}) = \frac{1}{3}(9 + 7 + 8) = \frac{1}{3}(24) = 8$

	(a(bc))	d
(a(bc))	-	-
d	8	-

Final Tree

Example 7. That gives us our final tree:

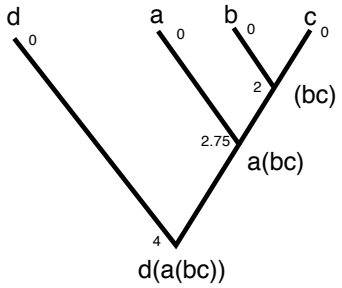


Scaled Tree

- In the construction of our tree, we did not make the tree scaled.
- We define something called the "height" of each node of our tree.
- The height of each leaf (taxa) is 0.
- If C is the parent of X and Y, then the height of C is $\frac{d_{X,Y}}{2}$.

Tree with Heights

Example 8. Our tree with heights included is:



37

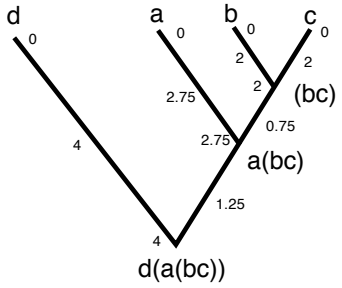
Scaled Tree

- We still need to compute the length of each branch and draw it to scale.
- If C is the parent of X and Y, the length of the branch between C and X is the height of C minus the height of X (same with Y).
- Thus, the branches connecting the ancestor (bc) to b and c should each be $2 - 0 = 2$ units long.
- The branch connecting $a(bc)$ to (bc) should be 0.75, etc..

38

Tree with Heights

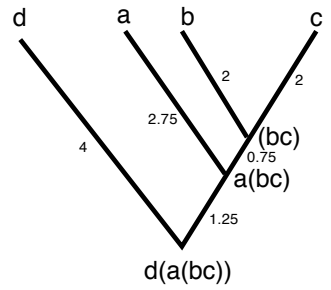
Example 9. Our tree with lengths of branches included is:



39

Scaled Tree

Example 10. Drawing our tree with proportional branch lengths gives us:



- Notice that the sum of lengths along any path from root to a leaf is the same.

40

UPGMA

- It is usually better not to use number of mismatches as a distance metric.
- This is often used as a starting point, but then they are adjusted.
- Two common formulae are p-distance and Poisson correction. The proportion of mismatches can be converted to these.

41

Mismatches

- It is common to start with a multiple sequence alignment when generating a phylogenetic tree.
- It is also common to chop out parts of the alignments that have gaps so only matches and mismatches remain.
- It is also common to restrict to parts of sequences that are available for all taxa.

42

Assumptions

- One must be sure that sequences are homologous before constructing a tree – otherwise the non-homologs are harming the tree.
- It might be beneficial to perform pairwise alignments first to confirm sequences are homologs.

43

Assumptions

- UPGMA is built on assumption that the rate of nucleotide or amino acid substitution is constant for all branches of the tree.
- The molecular clock hypothesis needs to apply.
- If this is true, branch lengths can be used to estimate dates of divergence.

44

Distance-Based Methods

- UPGMA and *neighbour-joining* are two distance based methods.
- Neighbour joining is more commonly used.
- We will describe neighbour joining intuitively.

45

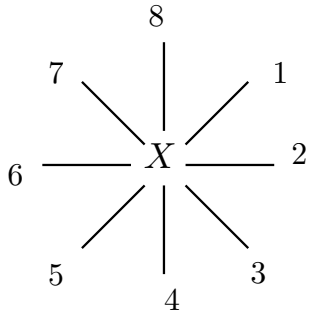
Neighbour-Joining

- Neighbour-joining operates on unrooted trees instead of rooted.
- Neighbour-joining does not require the molecular clock hypothesis.

46

Neighbour-Joining

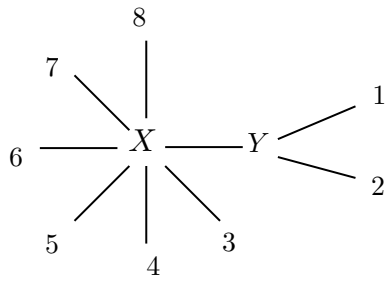
- Then it starts by putting all taxa in a star-like topology.



47

Neighbour-Joining

- It calculates the pairwise distances as before.
- It finds the two most similar and groups them together. Say it's 1 and 2. Then it makes them "neighbours".



48

Neighbour-Joining

- It then continues making neighbours based on the next smallest distance.
- This could either be between a taxon and a taxon, or a clade and a taxon, or a clade and a clade.
- It iterates distances using 'sum of branch lengths'.
- See 'Saitou and Nei, The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 1987', for procedure.

49

4 Character-Based Methods

Character-Based Methods

- The distance-based methods "look at the big picture", assign numerical scores, then create trees.
- Character-based are concerned with specific well-defined features that can exist in a limited number of states.
- A character could be a binary value for the presence or absense of some feature.

50

Character-Based Methods

Definition 11. Parsimony (in the context of phylogenetic trees): the process of giving preference to one evolutionary pathway over another on the basis of which requires the smallest number of changes.

- Changes could be mismatches in sequence.
- Intuition: mutations are very rare, and the model with the fewest is perhaps most likely to be correct.

- Typically, character-based methods work on unrooted trees.

51

Maximum Parsimony

- Maximum parsimony dictates that the best tree is the one with the shortest branch lengths possible.
- It can equally well be applied to morphological characters, as was described by Hennig in 1966.
- The goal is to find the most parsimonious explanation for the observed data.

52

An Example

The concept is more easily described with an example. We will use a multiple sequence alignment.

Example 12. Consider the following multiple sequence alignment:

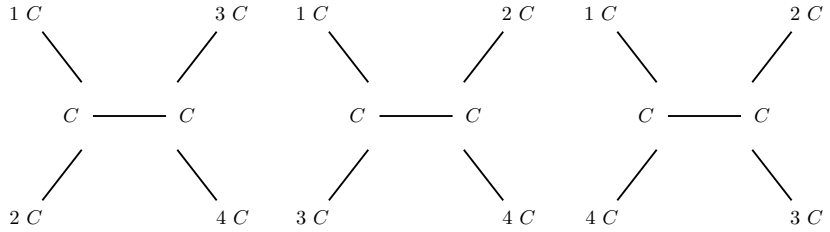
Sequence	1	2	3
1	C	A	T
2	C	C	C
3	C	A	T
4	C	G	C

At each site i of the multiple sequence alignment (i runs from 1 to 3), we examine the set of all possible unrooted phylogenetic trees with the leaves being the four letters at site i, from sequences 1 through 4.

53

Site 1

Example 13. For site 1, we get the following three possible trees.

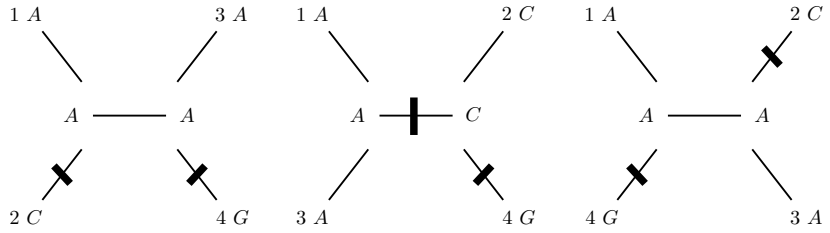


Thus, this site is said to be *uninformative*, because all possible trees give the same number of mutations.

54

Site 2

Example 14. For site 2, we get the following three possible trees.

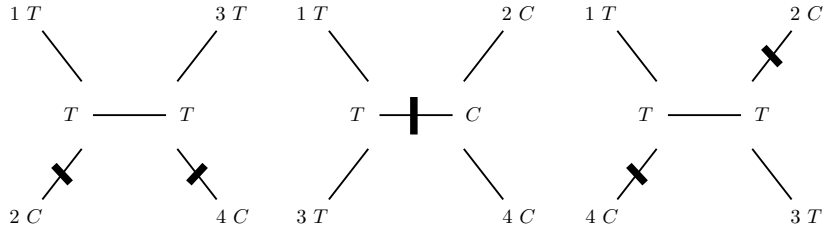


Each require two permutations, thus we have another uninformative site.

55

Site 3

Example 15. For site 3, we get the following three possible trees.



The second tree requires fewer mutations than the others, thus this site is informative.

56

Example continued

Example 16.

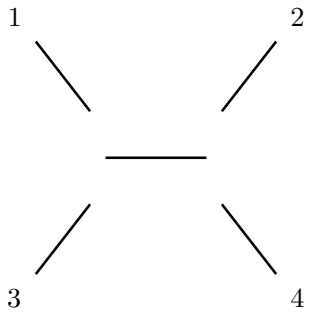
- The tree that invokes the smallest number of mutations overall is the winner.
- The first tree invokes 4 mutations (0 at site 1, 2 at site 2 and 2 at site 3).
- The second tree invokes 3 mutations (0 at site 1, 2 at site 2 and 1 at site 3).
- The third tree invokes 4 mutations (0 at site 1, 2 at site 2 and 2 at site 3).

57

Example continued

Example 17.

- Overall, tree number 2 is the winner.
- That is represented in Newick format description by $((1,3))(2,4))$.
- The tree would look like:



58

More character-based methods

- The previous example is an instance of so-called "unweighted parsimony" because all types of mutations are weighted equally.
- There is also weighted parsimony, where we weight different types of mutations differently.
- Informative sites that support (or are consistent with) the final or "winner" tree are called *synapomorphies* (a derived state shared by several taxa).
- Other informative sites are called *homoplasies* (a character trait that has arisen independently, or through horizontal gene transfer, in several taxa).

59

Procedure

- Only if there are few taxa is it practical to go through every possible tree.
- After that, there needs to be heuristics to cut down on search space.

60

Phylogenetics

- There is no clear winner between distance-based and character-based methods.
- Typically, we use distance based-methods and character-based methods in tandem, and we only trust results if they work in a variety of circumstances.

61

Phylogenetic trees

This "tree of life" was determined by both character and distance based methods from 16S rRNA genes (nice and short, only 1,542 bases).

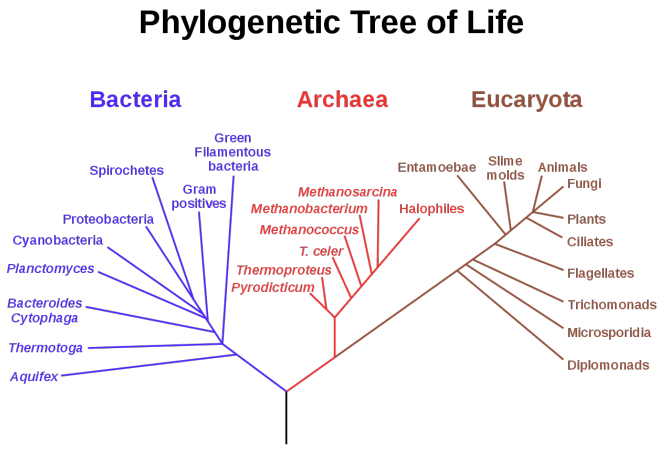


image from http://en.wikipedia.org/wiki/File:Phylogenetic_tree.svg

62

Evaluating Trees

- After making a tree, it is beneficial to assess its accuracy.
- Good criteria to assess accuracy are
  - consistency,
  - efficiency,
  - robustness.

63

Evaluating Trees

- The most common approach is to use *bootstrapping*.
- It provides a technique to assess robustness of the phylogenetic tree.
- It randomly permutes the original data set many times.
- It checks how often the trees created from permuted data match generated tree.

64

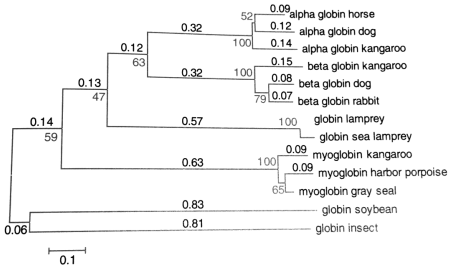
Bootstrapping

- Assume the original tree was built from a multiple sequence alignment.
- Fake data is made by randomly picking columns.
- Then a tree is generated from random data.
- Do this 100 or 1000 times and compare to the original inferred tree.

65

Bootstrapping

- Different software will give you the percentage of time each clade in the original tree ends up in the bootstrapped trees.
- The tree below was generated with the software Mega, and bootstrap values are listed beside each clade.
- 70% is often considered good support for the clade[4].



[4]Hillis, Bull, An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biol.*, 1993.

66

ClustalW

- To get a little familiarity with distance-based approaches online, we can use ClustalW.
- ClustalW uses either Neighbour-Joining or UPGMA to build its guide tree.

  website:

  http://clustalw.ddbj.nig.ac.jp or http://www.genome.jp/tools/clustalw/ or http://www.clustal.org

- You can paste in a FASTA file (which can contain more than one sequence, one after the other).

67

ClustalW

- The output is not a visual representation of a tree, but an abstract formulation of a tree (there are a couple of different formats, like Phylip).
- You will then probably want a program which displays the abstract formulation in a visually pleasing manner.

  Newick tree viewers

  http://www.trex.uqam.ca/index.php?action=newick&project=trex or http://etetoolkit.org/treeview/

- Start with the FASTA file sample.fasta, run it through ClustalW, then through a tree viewer.

68

ClustalW

- Just like the other distance based approaches, ClustalW
  1. calculates pairwise distances (between every two sequences).
  2. creates a phylogenetic tree (the guide tree), at each step merging together the sequences of smallest distance.
  3. using the guide tree, calculates a multiple sequence alignment
- For step 3, the alignment is created using the guide tree, by at each step, aligning the two most closely related sequences into an alignment.
- In this way, one can see how this is using a "once a gap, always a gap" approach.

69

picture from Pevsner, *Bioinformatics and Functional Genomics*, 2009.

Phylogeny Programs

- For the purposes of multiple sequence alignment, ClustalW is good.
- But for the purposes of building accurate trees, try Mega or Phylip (http://evolution.genetics.washington.edu/phylip.html).
- Both are free (at least for education purposes) for major platforms.

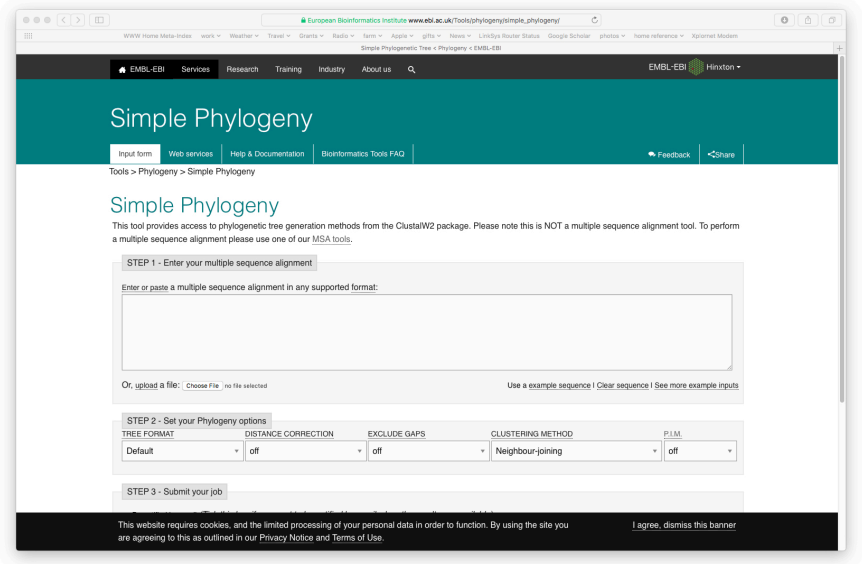
70

ClustalW

- The multiple sequence alignment can be used as the basis for a phylogenetic tree.
- Save the alignment file (.aln file).
- Visit EBI's (ClustalW2) "Simple Phylogeny" site. https://www.ebi.ac.uk/Tools/phylogeny/simple_phylogeny/

71

ClustalW Phylogeny



- This looks like the regular ClustalW program, but it is not the same program.
- It takes a multiple sequence alignment as input. It's assumed it's already been aligned. Use the .aln file and click 'submit'.

72

ClustalW Phylogeny

Parameters of Interest

- 'Exclude gaps': When set to 'on', sections of gaps in the multiple sequence alignment are removed.
- 'Clustering method': This can be set to Neighbour-Joining or UPGMA.
- 'Distance correction': This corrects for multiple substitutions at the same site. It is set to 'off' but is recommended to be 'on' for distant sequences.
- The various 'Tree Formats' give additional information in the 'Result Summary' tab, such as the distance matrix used to construct tree.

_____ 73

ClustalW

- Try UPGMA with 'distance matrix', 'exclude gaps' on and 'distance correction' on.
- Compare to neighbour-joining.
- A *cladogram* is an unscaled tree.

_____ 74