# Molecular phylogenetic reconstruction

Pakorn Aiewsakun

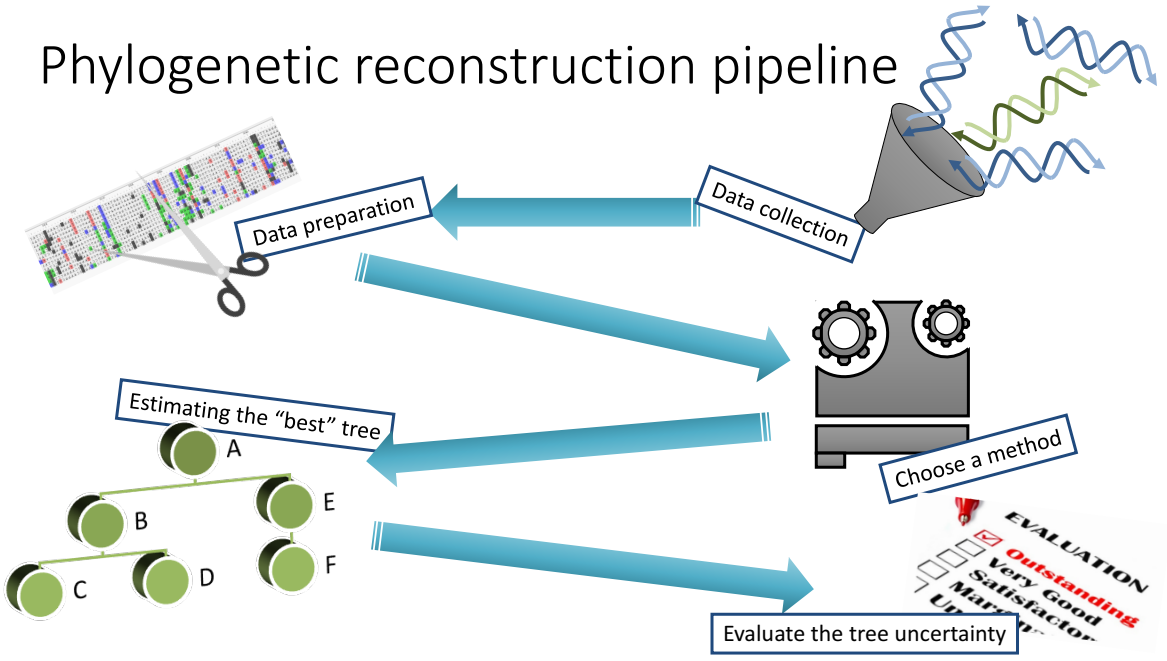
SCID 303 Bioinformatics

14/02/19

## Aims

- Able to describe the overview of the process of molecular phylogenetic reconstruction
- Understand the principles and philosophical concepts behind various phylogenetic reconstruction techniques
- Able to compare and contrast various phylogenetic reconstruction methods

## Phylogenetic reconstruction pipeline



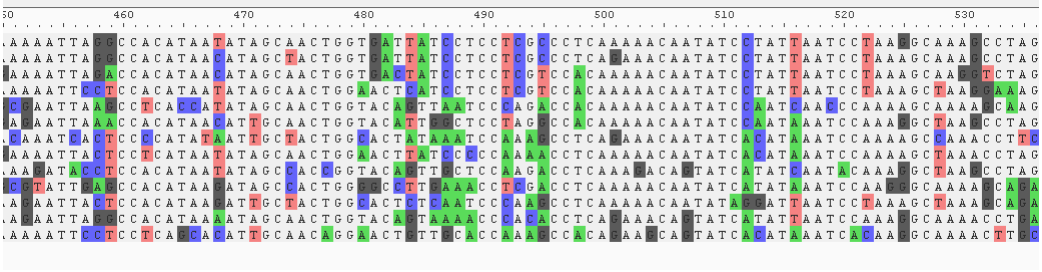

## Molecular phylogenetics

- Based on studies of gene/protein sequences
- First suggested by Zuckerkandl and Pauling (1962, Molecules as Documents of Evolutionary History, J. Theoret. Biol.)
- Most of the time, the results are comparable to those obtained from phenotypic-based methods, as phenotypes of an organism are largely correlated with its genome

# Molecular phylogenetics

- Why molecular phylogenetics?
  - Organisms can be compared even if they are morphologically very different – a common measure for evolutionary divergence
  - Clear character states
  - Many characters to be compared, diluting the effects of homoplasy
  - There are lots of data available

# Molecular data



## Molecular data

- Things to be considered when collecting the data
  - Are the genes/proteins present in all organisms under the investigation?
  - Are they homologous genes?
  - Is there enough phylogenetic signal in the data? Nucleotide VS protein data?
  - Can you align them with confident?
  - What are appropriate outgroups? How many taxa should I include in my tree?

## Phylogenetic reconstruction methods

- There are no uniquely correct methods for inferring phylogenies
- Different methods may give different answers, and perhaps you might want to try several methods
- The results may or may not reflect the true phylogenetic tree. Difficult to verify that your trees are correct

## Phylogenetic reconstruction methods

- Maximum parsimony: identify phylogenetic tree(s) that results in the smallest total number of evolutionary changes to explain the data
- Distance-matrix methods: construct an all-to-all distance matrix between each taxon pair and hierarchically grouping them using a clustering algorithm, e.g. NJ and UPGMA
- Maximum likelihood: identify the tree that most likely gives rise to the observed data under a specific model of evolution
- Bayesian inference: identify the landscape of possible trees and models based on your data and prior knowledge of their evolutionary processes

## Phylogenetic reconstruction methods – maximum parsimony

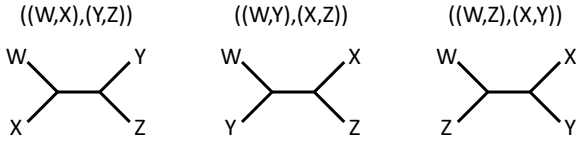
- A maximum parsimony tree is the tree that requires the smallest number of total evolutionary changes (tree length) in order to explain the evolutionary relationships of the organisms under study
- Grouping them based on synapomorphic traits/characters
- The focus of this method is the tree topology… but it can return "branch lengths" nonetheless, if you want

# Phylogenetic reconstruction methods – maximum parsimony

- Example

There are 3 possible (unrooted) trees connecting 4 taxa

```
W: GGTGTTAACGCCTC
X: AGTGTTAACACCCC
Y: TGTATTAGTACAAC
Z: AGTTTTGATACAAC
```

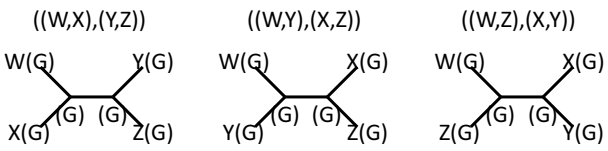



# Phylogenetic reconstruction methods – maximum parsimony

- Example

Invariant sites are uninformative as they don't distinguish the 3 possibilities

```
W: GGTGTTAACGCCTC
X: AGTGTTAACACCCC
Y: TGTATTAGTACAAC
Z: AGTTTTGATACAAC
Invariant site:  ** **     *   *
```



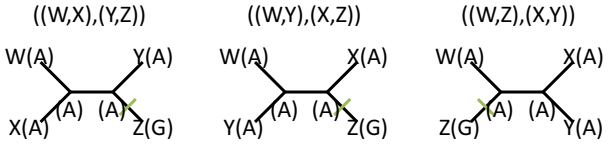

0 changes for all trees, under the parsimony principle

# Phylogenetic reconstruction methods – maximum parsimony

- Example

Some variant sites are also uninformative

```
W: GGTGTTAACGCCTC
X: AGTGTTAACACCCC
Y: TGTATTAGTACAAC
Z: AGTTTTGATACAAC
Variant site: *  *  **** **
```



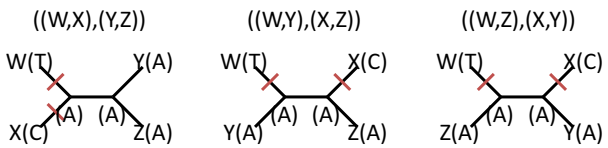

1 change for all trees, under the parsimony principle

# Phylogenetic reconstruction methods – maximum parsimony

- Example

Some variant sites are also uninformative

```
W: GGTGTTAACGCCTC
X: AGTGTTAACACCCC
Y: TGTATTAGTACAAC
Z: AGTTTTGATACAAC
Variant site: *  *  **** **
```




2 changes for all trees, under the parsimony principle

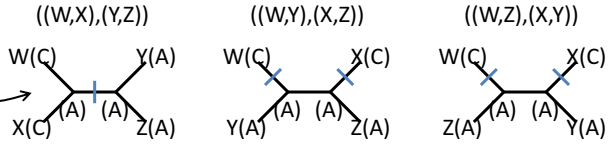
# Phylogenetic reconstruction methods – maximum parsimony

- Example

Only the variant informative sites are considered in the tree determination

```
W: GGTGTTAACGCCTC
X: AGTGTTAACACCCC
Y: TGTATTAGTACAAC
Z: AGTTTTGATACAAC
Variant site: *  *  **** **
```

The maximum parsimonious tree topology!

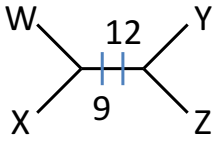



The 1st tree gives the fewest changes (1 VS 2); the tree is most preferable under the maximum parsimony principle

# Phylogenetic reconstruction methods – maximum parsimony

- Example

((W,X),(Y,Z))

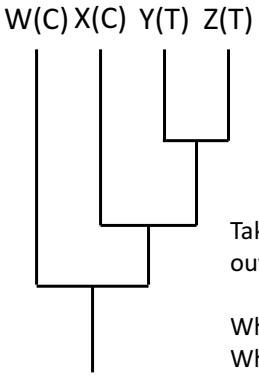



Only 2 sites are used to determine the best tree topology

```
W: GGTGTTAACGCCTC
X: AGTGTTAACACCCC
Y: TGTATTAGTACAAC
Z: AGTTTTGATACAAC
Invariant site:  ** **     *   *
  Variant site: *  *  **** **
Informative site:        *   *
Uninformative site: *  *  ** *  *
```

# Phylogenetic reconstruction methods – maximum parsimony

- Example




```
W: GGTGTTAACGCCTC
X: AGTGTTAACACCCC
Y: TGTATTAGTACAAC
Z: AGTTTTGATACAAC
```

Take site 9 for example. Let's say we root the tree using W as an outgroup:

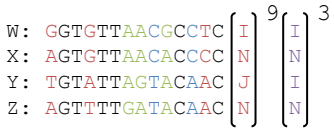
When did the change occur?
What is the derived trait?
What is the ancestral trait?
Are the answers still the same when you change the root?

# Phylogenetic reconstruction methods – maximum parsimony

- The basic (implicit) assumption of the maximum parsimony method is that evolutionary changes are improbable
- However, molecular states are finite, and homoplasy can occur, especially in independently rapidly evolving entities
- When a tree contains both fast and slow evolving organisms, this method tends to (falsely) group fast evolving entities together – this is called "long branch attraction"

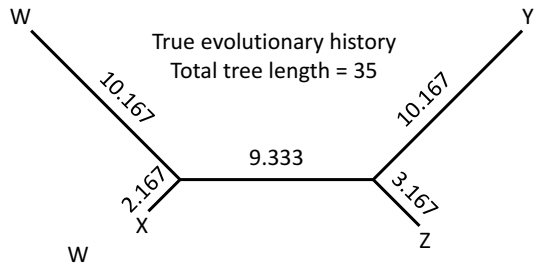
# Phylogenetic reconstruction methods – maximum parsimony

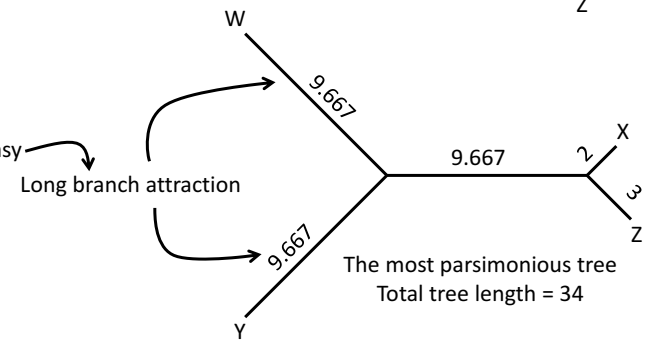
- Long branch attraction




Homoplasy

This is because this method doesn't take into account rate variation among lineages




# Phylogenetic reconstruction methods – maximum parsimony

- Pros
  - Many evolutionary scenarios are compared and evaluated, but this feature can make the method very slow
  - The results shed some light on the past evolutionary history of organisms. However, it is important to remember that the true evolutionary pathway does not have to be the simplest one

# Phylogenetic reconstruction methods – maximum parsimony

- Cons
  - Typically only returns a cladogram without branch lengths, but this can be overcome by mapping changes onto the "best" tree. These branch lengths however do not reflect the true numbers of evolutionary changes
  - When there are > 4 taxa, there might be more than one equally maximum parsimonious trees
  - Lots of data are disregarded as uninformative sites

# Phylogenetic reconstruction methods – maximum parsimony

- Cons
  - Unable to account for rate variation among lineages, and therefore is prone to long branch attraction (and thus isn't too popular in molecular phylogenetic analysis)
  - The classic method attributes the same cost to every type of changes. However, this can be modified by assigning different costs to different types of changes.

## Phylogenetic reconstruction methods – distance-based methods

- Distance-based method estimates the (mean) number of changes among pairs of taxa
- A tree is then estimated from the pairwise (dis)similarity matrix, using a hierarchical clustering method, typically, the neighbour joining (NJ) method, or unweighted pair group method with arithmetic mean (UPGMA)
- Both the estimated topology and branch lengths are meaningful

## Phylogenetic reconstruction methods – distance-based methods

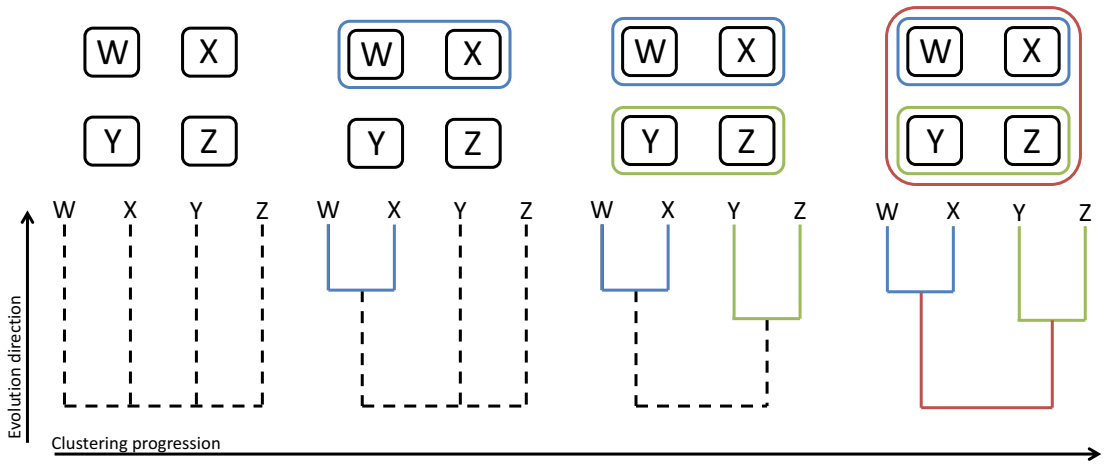
- To compute distances between pairs of sequences, there are a number of evolutionary models that you can use, all of which can take into account:
  - Different state (e.g. nucleotide, or amino acid) frequencies
  - Different probabilities of changing from one state to another
  - Multiple hits (typically assuming a time-continuous Markov process)
  - Rate variation among sites

# Phylogenetic reconstruction methods – distance-based methods

- Once you obtain the distance matrix, the distances are gathered into a tree using the NJ or UPGMA method
- Both methods are greedy heuristic methods which joins the two closest taxa/sub-trees in each step into a higher-level cluster
- While UPGMA focuses on taxa, NJ keeps track of nodes

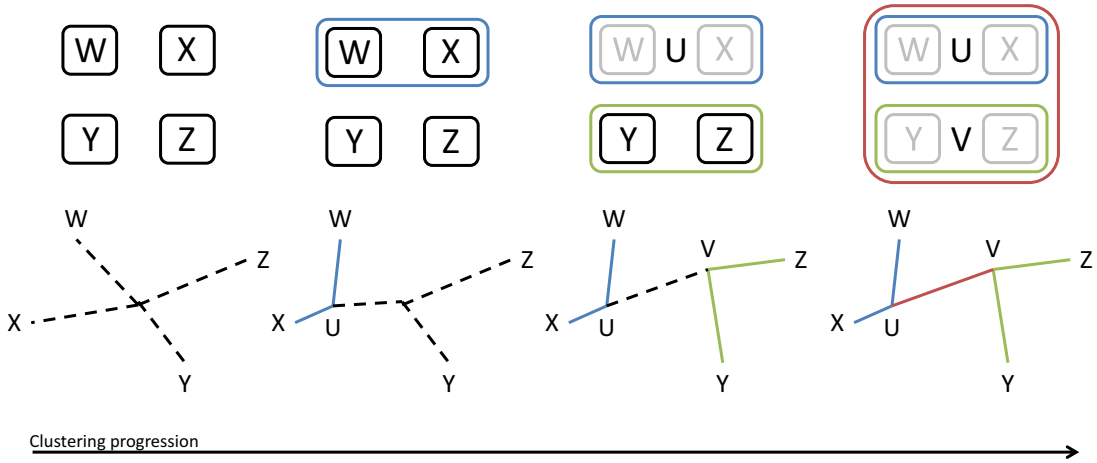
# Phylogenetic reconstruction methods – distance-based methods

- UPGMA



# Phylogenetic reconstruction methods – distance-based methods

- NJ




# Phylogenetic reconstruction methods – distance-based methods

**UPGMA**

- Focuses on the taxa
- Has a built-in "strict molecular clock" assumption, thus producing a rooted tree
- Always give a tree with non-negative branch lengths

**NJ**

- Focuses on the nodes. In fact, it adds a new node every time two taxa are joined
- Produces an unrooted tree
- Accommodate rate variation among lineages
- It can produce a tree with negative branches

## Phylogenetic reconstruction methods – distance-based methods

- Pros
  - Simple and super fast
  - No data got thrown away
  - Allowing for molecular distance correction using various models
- Cons
  - Only returns one tree, and no other evolutionary scenarios are explored
  - Doesn't tell anything about the past evolutionary states of the organisms
  - Sensitive to the assumptions in the evolutionary model, and you can't quite say exactly which model is the best one.

## Phylogenetic reconstruction methods – maximum likelihood method

- A maximum likelihood tree is a tree that most likely gives rise to the observed data under a specific model of evolution
  - State (e.g. nucleotide, amino acid) frequencies
  - Probabilities of changing from one state to another
  - Rate variation among sites
  - Substitution dynamics (typically assume a time-continuous Markov process)
- Evaluates various tree topologies and branch length values, and chooses the best one that can account for the conversion of one sequence into another best

# Phylogenetic reconstruction methods – maximum likelihood method

- Given a tree of your interest, this method computes its "likelihood", which is the plausibility (but not a probability) of that tree and its branch lengths (under a specific evolutionary model) that will produce the observed data
- The tree with the highest likelihood the best tree under the maximum likelihood criterion

# Phylogenetic reconstruction methods – maximum likelihood method

- Pros
  - This method evaluates both the relationships and branch lengths at the same time, and it considers all possible evolutionary scenarios, just like in the maximum parsimony method
  - Uses all the data, just like the distance-matrix based method
  - You can quantify exactly how much more probable is one model over the other by comparing their likelihood values, accommodating model selection
  - Robust to violations of the assumptions in the evolutionary model

## Phylogenetic reconstruction methods – maximum likelihood method

- Pros
  - The framework also accommodates ancestral state reconstruction, allowing us to learn more about the past evolutionary processes of the gene/protein
- Cons
  - Computationally intensive. Might not be possible to examine all model parameter values (e.g. tree topologies, branch lengths, transition rate values, etc.)

## Phylogenetic reconstruction methods – Bayesian inference

- This method computes something call “posterior probability” of a tree and its associated parameter values of your (pre-specified) evolutionary model, given the observed molecular data and your prior belief in their evolutionary processes
- It does not give you just one tree, but “a landscape” or “a population” of trees and parameter values
- You then summarise the entire landscape to get a consensus tree

# Phylogenetic reconstruction methods – Bayesian inference

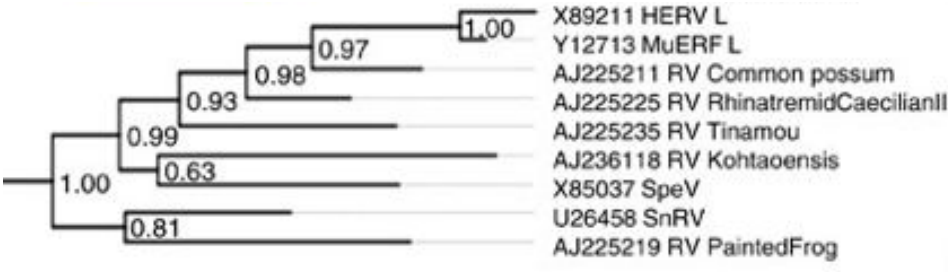
- Pros
  - Everything that the maximum likelihood method can do, Bayesian inference can, and more!
  - It doesn't gives just one best tree (and its associated parameter values), but a distribution of trees. It can tell you about the uncertainty of the grouping!
  - Allows you to integrate your prior knowledge of the model into the calculation!

# Phylogenetic reconstruction methods – Bayesian inference

- Cons
  - Very computationally intensive. Only possible recently
  - Prior knowledge can influence the results quite significantly

# Evaluate tree uncertainty

- The Bayesian framework gives you a population of trees, so you can evaluate how robust your tree topology is, based on "the Bayesian posterior probability clade support" values



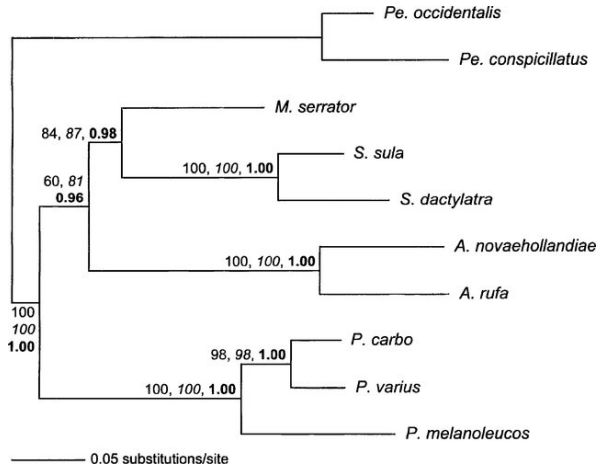

> 0.85 is typically considered good enough

# Evaluate tree uncertainty – bootstrapping

- For other methods, you need to built your own population of trees.
- "Bootstrapping" is a common technique used to generate, i.e. resample, new datasets, and thus new trees

```
W: GGTGTTAACGCCTC                W: CGCTGCTCCTTGCC
X: AGTGTTAACACCCC      Bootstrapping X: CGCCGCTCCTTGCC
Y: TGTATTAGTACAAC                Y: CGCAATTCCTTACC
Z: AGTTTTGATACAAC                Z: CGCATTTCCTTTCC
   1 2 3 4 5 6 7 8 9 10 11 12 13 14    14 2 11 13 4 9 5 11 11 6 6 4 14 11
```

# Evaluate tree uncertainty – bootstrapping




The maximum-likelihood tree (the branch lengths represent the expected number of substitution per site) with A. anhinga excluded. The percentage of bootstrap replicates (out of 1000) that supported each node (for parsimony and, in italics, maximum likelihood) and the Bayesian posterior probabilities (in bold) are shown.

Kennedy et al., 2005 Syst Biol.

> 75% is typically considered good enough

# Evaluate tree uncertainty – bootstrapping

- These clade support, however, DOES NOT determine how accurate a tree is
- It only tells you how well a tree reflects the underlying data, or how the data supports particular groupings
- If the data are biased, clade support values can be erroneous, supporting false groupings