Bioinformatics II:

Building Phylogenies using Maximum likelihood

Dr Manaf A Guma
University of Anbar- college of applied science-Heet.
Department of chemistry

1

What is a maximum likelihood method?

- Maximum likelihood is the third method used to build trees.
- Likelihood provides probabilities of the sequences given a model of their evolution on a particular tree.
- A parameter is some descriptor of the model.
- By (Felsenstein 1981).

2

Concepts

- What are the statics that are provided by this method?
- Molecular phylogenetic methods use a given set of aligned sequences to construct a phylogenetic Tree.
- For example: sequence 1, 2 , 3 and 4.
- There are several ways to construct phylogenetic trees.
- The Maximum Likelihood method will pick out the tree that most represents the true tree.
- So, the more probable the sequences given the tree, the more the tree is preferred.

3

ML is based on a Markov model of evolution

- **Observed:** The species labeling the leaves
- **Hidden:** The ancestral states
- **Transition probabilities:** The mutation probabilities
- **Assumptions**:
  - Only mutations are allowed
  - Sites are independent

4

Models of evolution at a specific site

- Transition probability matrix:

$$M = [m_{ij}], \qquad i, j \in \{\mathrm{A}, \mathrm{C}, \mathrm{T}, \mathrm{G}\}$$

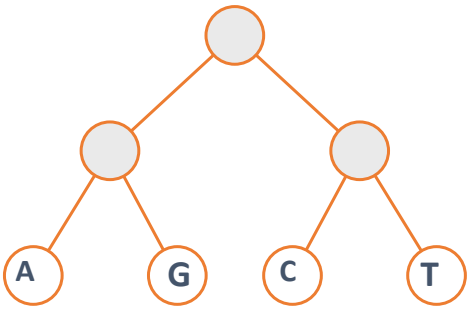
where

$m_{ij} = \mathrm{Prob}(i \to j \text{ mutation in 1 time unit})$

- Branches may have different lengths!

5

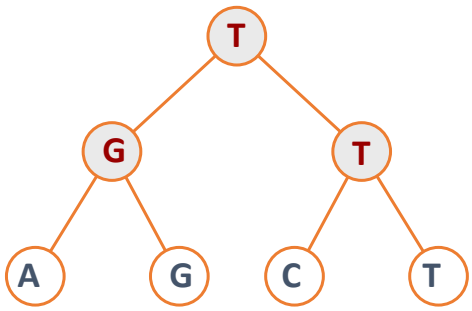
The probability of an assignment



Probability = ?

6

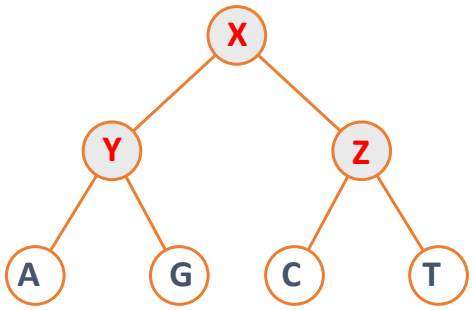
The probability of an assignment



Probability = $m_{TG} \cdot m_{GA} \cdot m_{GG} \cdot m_{TT} \cdot m_{TC} \cdot m_{TT}$

7

Ancestral reconstruction: most likely assignment



$L^* = \max_{X,Y,Z} \{m_{XY} \cdot m_{YA} \cdot m_{YG} \cdot m_{XZ} \cdot m_{ZC} \cdot m_{ZT}\}$

Compute using Viterbi algorithm

8

Do you remember?

- Phylogenetic tree is a data structure, characterized by:

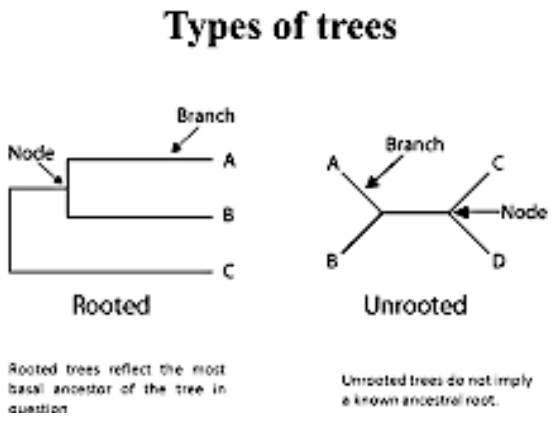
1. topology (form). Like rooted or unrooted trees.
2. its branch lengths.

- Stores information regarding the relationship of several species or sequences.

9

What are the difference ?

- Rooted tree: assumed ancestral state "d" is the root species.
- Unrooted tree... no implicit "directionality", but is a measure of similarity between species.



10

How to approach the maximum likelihood tree?

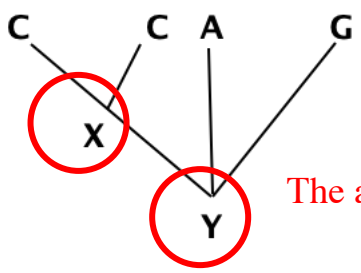
1. Assumes that all sequences at each site are considered independent.

	1	2		J					n
1	A	G	G	C	T	C	C	A	A..A
2	A	G	G	T	T	C	G	A	A..A
3	A	G	C	C	C	A	G	A	A..A
4	A	T	T	T	C	G	G	A	A..C

11

2.The log-likelihood is computed for a given topology by using a particular probability model.

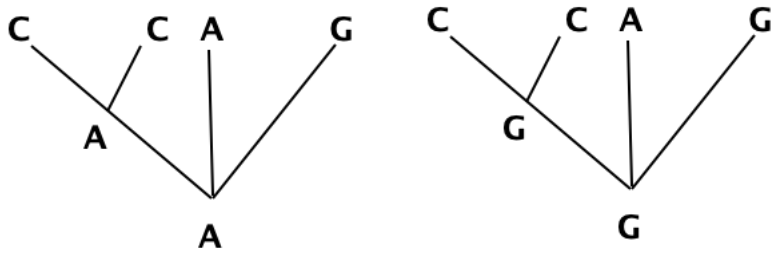
- a-



The ancestors are Hidden

12

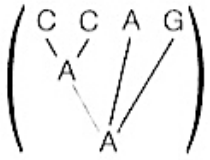
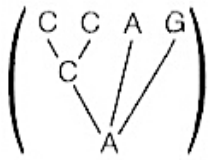
- b- L (j) = Prob + ……..+ Prob N

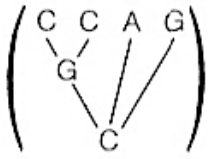


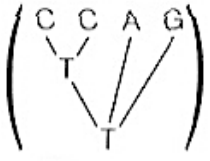
- c- ln L= ln L(1) + ln L(2)..+ ln L(j)+… + ln L(N) = $\sum_{i=1}^{n}$ ln L(i)

13

Analysis for site *j*

$L_{(j)}$ = Prob  + Prob 

\+ … + Prob 

\+ … + Prob 

14

Then,

- 3. After procedure is done for, the topology that shows the highest likelihood is chosen as the true (realistic) tree.
- #Rooted trees = $\frac{(2n-3)!}{2^{n-2}(n-2)!}$
- #Unrooted trees = $\frac{(2n-5)!}{2^{n-3}(n-3)!}$
- How many topologies do we have to go through for?
- How many topologies do we have to go through for n sequences?

15

- For i=2, I ≤ 10, i++,
- Print i
- Then seq will have # of topologies

```
3 seq. will have # of topologies: 1.
4 seq. will have # of topologies: 3.
5 seq. will have # of topologies: 15.
6 seq. will have # of topologies: 105.
7 seq. will have # of topologies: 945.
8 seq. will have # of topologies: 10395.
9 seq. will have # of topologies: 135135.
10 seq. will have # of topologies: 2.02703×10^6
```

16

What does the result tell?

1. The results is consistent.
2. But the time consuming. ML can be slow.
3. ML converges to correct answer as more data is added.
4. Can put in a Bayesian statistical framework, to obtain a distribution of possible phylogenies.

17