Phylogenetics:
Likelihood

COMP 571
Luay Nakhleh, Rice University

1

The Problem

- Input: Multiple alignment of a set S of sequences
- Output: Tree T leaf-labeled with S

2

Assumptions

- Characters are mutually independent
- Following a speciation event, characters continue to evolve independently

3

4

- The likelihood of model M given data D, denoted by L(M|D), is p(D|M).
- For example, consider the following data D that result from tossing a coin 10 times:
  - HTTTTHTTTT

5

- Model M1:
  - A fair coin (p(H)=p(T)=0.5)
  - $L(M1|D)=p(D|M1)=0.5^{10}$

6

- Model M2:
  - A biased coin (p(H)=0.8,p(T)=0.2)
  - $L(M2|D)=p(D|M2)=0.8^{2}0.2^{8}$

- Model M3:
  - A biased coin (p(H)=0.1,p(T)=0.9)
  - L(M3|D)=p(D|M3)=$0.1^2 0.9^8$

7

- The problem of interest is to infer the model M from the (observed) data D.

8

- The maximum likelihood estimate, or MLE, is:

$$\hat{M} \leftarrow \text{argmax}_M p(D|M)$$

9

10

- D=HTTTTHTTTT
- M1: p(H)=p(T)=0.5
- M2: p(H)=0.8, p(T)=0.2
- M3: p(H)=0.1, p(T)=0.9
- MLE (among the three models) is M3.

11

- A more complex example:
  - The model M is an HMM
  - The data D is a sequence of observations
  - Baum-Welch is an algorithm for obtaining the MLE M from the data D

12

- The model parameters that we seek to learn can vary for the same data and model.
- For example, in the case of HMMs:
  - The parameters are the states, the transition and emission probabilities (no parameter values in the model are known)
  - The parameters are the transition and emission probabilities (the states are known)
  - The parameters are the transition probabilities (the states and emission probabilities are known)

Back to Phylogenetic Trees

13

- What are the data D?
  - A multiple sequence alignment
  - (or, a matrix of taxa/characters)

Back to Phylogenetic Trees

14

- What is the (generative) model M?
  - The tree topology
  - The branch lengths
  - The model of evolution (JC, ..)

Back to Phylogenetic Trees

15

- What is the (generative) model M?
  - The tree topology, T
  - The branch lengths, λ
  - The model of evolution (JC, ..), E

Back to Phylogenetic Trees

16

- The likelihood is p(D|T,λ,E).
- The MLE is

$$(\hat{T}, \hat{\lambda}, \hat{E}) \leftarrow \text{argmax}_{(T,\lambda,E)} p(D|T, \lambda, E)$$

Back to Phylogenetic Trees

17

- In practice, the model of evolution is estimated from the data first, and in the phylogenetic inference it is assumed to be known.
- In this case, given D and E, the MLE is

$$(\hat{T}, \hat{\lambda}) \leftarrow \text{argmax}_{(T,\lambda)} p(D|T, \lambda)$$

Assumptions

18

- Characters are independent
- Markov process: probability of a node having a given label depends only on the label of the parent node and branch length between them t

Maximum Likelihood

19

- Input: a matrix D of taxa-characters
- Output: tree T leaf-labeled by the set of taxa, and with branch lengths λ so as to maximize the likelihood P(D|T,λ)

P(D|T,λ)

20

$$
\begin{aligned}
P(D|T,\lambda) &= \prod_{site\ j} p(D_j|T,\lambda) \\
&= \prod_{site\ j} \left(\sum_R p(D_j, R|T,\lambda)\right) \\
&= \prod_{site\ j} \left(\sum_R \left[p(root) \cdot \prod_{edge\ u\to v} p_{u\to v}(t_{uv})\right]\right)
\end{aligned}
$$

21

- What is $p_{i\to j}(t_{uv})$ for a branch u→v in the tree, where i and j are the states of the site at nodes u and v, respectively?

22

* For the Jukes-Cantor model with the parameter μ (the overall substitution rate), we have

$$p_{i\to j}(t) = \begin{cases} \frac{1}{4}(1+3e^{-t\mu}) & i=j \\ \frac{1}{4}(1-e^{-t\mu}) & i\neq j \end{cases}$$

23

* If branch lengths are measured in expected number of mutations per site, v (for JC: v=(μ/4+μ/4+μ/4)t=(3/4)μt)

$$p_{i\to j}(\nu) = \begin{cases} \frac{1}{4}(1+3e^{-4\nu/3}) & i=j \\ \frac{1}{4}(1-e^{-4\nu/3}) & i\neq j \end{cases}$$

24

* The ML problem is NP-hard (that is, finding the MLE (T,λ) is very hard computationally)
* Heuristics involve searching the tree space, while computing the likelihood of trees
* Computing the likelihood of a leaf-labeled tree T with branch lengths can be done efficiently using dynamic programming

P(D|T,λ)

25

Let $C_j(x,v)$ = P(subtree whose root is v | v_j=x)

Initialization: leaf v and state x $\quad C_j(x,v) = \begin{cases} 1 & v_j = x \\ 0 & \text{otherwise} \end{cases}$

Recursion: node v with children u,w

$$C_j(x,v) = \left[\sum_y C_j(y,u) \cdot P_{x\to y}(t_{vu})\right] \cdot \left[\sum_y C_j(y,w) \cdot P_{x\to y}(t_{vw})\right]$$

Termination:

$$L = \prod_{j=1}^{m} \left[\sum_x C_j(x, root) \cdot \mathrm{P}(x)\right]$$

Running Time

26

* Takes time $O(nk^2m)$, where n is the number of leaves in the tree, m is the number of sites, and k is the maximum number of states per site (for DNA, k=4)

Unidentifiability of the Root

27

* If the base substitution model is reversible (most of them are!), then rooting the same tree differently doesn't change the likelihood.

Questions?