

Inferring the Past: Phylogenetic Trees (chapter 12)

- | The biological problem
- | Parsimony and distance methods
- | *Models for mutations and estimation of distances*
- | Maximum likelihood methods

Estimation of distances

- | Many alternative ways to derive the distances d_{ij} exist
- | We can construct a simple stochastic model for the evolution of a DNA sequence...
- | ...and then obtain the distances from the model
- | Key points:
 - mutations at sites are rare events in the course of time => poisson process
 - sites evolve individually and by an identical mechanism
 - number of mismatched bases is a sum of mutations at individual sites => binomial variable

A stochastic model for base substitutions

- | Consider a single homologous site in two sequences
- | Assume the sites diverged for time length t : the sites are separated by time $2t$
- | Suppose that the number of substitutions in any branch of length t has a Poisson distribution with mean λt
- | Probability that k substitutions occur is given by the Poisson probability $e^{-\lambda t}(\lambda t)^k/(k!)$, $k = 0, 1, 2, \dots$

Substitutions at one site

- | General model: $P(\text{substitution results in base } j \mid \text{site was base } i) = m_{ij}$
- | Felsenstein model: $m_{ij} = \pi_j$, with $\pi_j \geq 0$ and $\pi_1 + \pi_2 + \pi_3 + \pi_4 = 1$
- | Assume that the set of probabilities π_j is same at every position in the sequence

Substitutions at one site (2)

- | Probability $q_{ij}(t)$ that a base i at time 0 is substituted by a base j a time t later
- | $q_{ij}(t) = e^{-\lambda t} + (1 - e^{-\lambda t}) \pi_j$, if $i = j$
- | $q_{ij}(t) = (1 - e^{-\lambda t}) \pi_j$, otherwise

Substitutions at one site (3)

- | We assume stationarity: distribution of base frequencies is the same for every time t
- | In other words, we want that

$$P(\text{base } a \text{ time } t \text{ later} = j) = \pi_j^0$$

- | For our simple model, this can be shown to hold

Estimating distances

- | Distances should take into account the mutation mechanism
- | Average of λt substitutions occur at a particular site on a branch of length t
- | However, some of the substitutions do not change the base (A \rightarrow A or A \rightarrow G \rightarrow A, for example)

Mean number of substitutions in time t

- | What is the chance H that a substitution actually changes a base?
- | $H = \sum \pi_i(1 - \pi_i) = 1 - \sum \pi_i^2$
- | Average number of real substitutions is then $\lambda t H$
- | Distance K between two sequences is
 $K = 2\lambda t H$

Estimating distances from sequence data

- | We want to estimate $K = 2\lambda tH$ from sequence data
- | The chance $F_{ij}(t)$ that we observe a base i in one sequence and a base j in another is

$$F_{ij}(t) = \sum_l \pi_l q_{li}(t) q_{lj}(t)$$

by averaging over the possible ancestral nucleotides

Estimating distances from sequence data

- Expression $F_{ij}(t) = \sum_l \pi_l q_{li}(t) q_{lj}(t)$ can be simplified by assuming that the mutation process is reversible:

$$\pi_i m_{ij} = \pi_j m_{ji} \text{ for all } i \neq j$$

- From this it can be shown that

$$\pi_i q_{ij}(t) = \pi_j q_{ji}(t) \text{ for all } i, j \text{ and } t > 0$$

- Now the model simplifies into $F_{ij}(t) = \pi_i q_{ij}(2t)$

Estimating distances from sequence data

- | What is the probability $F = F(t)$ that the letters at a particular position in two immediate descendants from the same node are identical?

$$F = \sum_i \pi_i q_{ii}(2t) = e^{-2\lambda t} + (1 - e^{-2\lambda t})(1 - H)$$

Putting the sites together

- | Assume that
 - sites evolve independently of one other and
 - mutation process is identical at each site
 - The two sequences have been aligned against each other and gaps have been removed

- | Do the bases at site i in the sequences differ?

$X_i = 1$ if the i th pair of sites differ

$X_i = 0$ otherwise

Putting the sites together (2)

- | $P(X_i = 1) = 1 - F = (1 - e^{-2\lambda t})H$
- | Now $D = X_1 + \dots + X_s$ is the number of mismatched pairs of bases
- | D is a binomial random variable with parameters s and $1 - F$
- | Notice that D is the Hamming distance for the sequences

Putting the sites together (3)

- | F is unknown and has to be estimated from the sequence data
- | Recall that the observed proportion of successes is a good estimator of the binomial success probability: estimate $1 - F$ with D/s
- | $D/s = (1 - e^{-2\lambda t})H$
- | $2\lambda t = -\log(1 - D/(sH))$
- | Finally, we obtain $K = 2\lambda tH = -H \log(1 - D/(sH))$

Jukes-Cantor formula

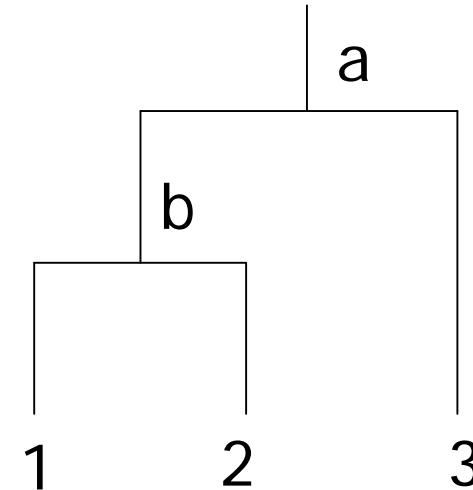
- | Estimate $2\lambda tH = -H \log(1 - D/(sH))$ of the distance K is known as the Jukes-Cantor formula
- | When H (chance that a substitution actually occurs) approaches 1, the estimate decreases and approaches the Poisson mean $2\lambda t$
- | H is usually not known and has to be estimated from the data as well

Inferring the Past: Phylogenetic Trees (chapter 12)

- | The biological problem
- | Parsimony and distance methods
- | Models for mutations and estimation of distances
- | *Maximum likelihood methods*

Maximum likelihood methods

- Consider the tree on the right with three sequences
- Probability $p(i_1, i_2, i_3)$ of observing bases i_1, i_2 and i_3 can be computed by summing over all possible ancestral bases,



$$p(i_1, i_2, i_3) = \sum_a \sum_b \pi_a q_{ai_3}(t_2) q_{ab}(t_2-t_1) q_{bi_2}(t_1) q_{bi_1}(t_1)$$

- Hard to compute for complex trees

Maximum likelihood estimation

- | We would like to calculate likelihood $p(i_1, i_2, \dots, i_n)$ in the general case
- | Calculations can be arranged using the peeling algorithm
- | Basic idea is to move all summation signs as far to the right as possible

Maximum likelihood estimation

- | Likelihood for the data is then obtained by multiplying the likelihoods of individual sites

- | General recipe for maximum likelihood estimation:
 - Maximize over all model parameters for a *given* tree
 - Maximize previous expression over *all* possible trees

Problems with tree-building

| Assumptions

- Sites evolve independently of one other
- Sites evolve according to the same stochastic model
- The tree is rooted
- The sequences are aligned
- Vertical inheritance

Additional material on phylogenetic trees

- | Durbin, Eddy, Krogh, Mitchison: Biological sequence analysis
- | Jones, Pevzner: An introduction to bioinformatics algorithms
- | Gusfield: Algorithms on strings, trees, and sequences