CISC 889 Bioinformatics (Spring 2004)

Phylogenetic Trees (III)

Probabilistic methods

Bayes rule revisited.

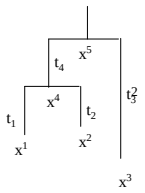
$$P(\text{model}|\text{data}) = P(\text{data}|\text{model})\, P(\text{model})/P(\text{data})$$

- model includes
  - our evolution theory,
  - a specific phylogenetic tree (topology and edge lengths), and
  - assignment of sequences to tree leaves.
- Data: a set of sequences that are used to infer phylogeny.

Let $P(x|y, t)$ be the probability in that sequence x is evolved from an ancestral sequence y over an edge of length t.

$P(x^1, x^2, x^3, x^4, x^5|T, t.)$
$= P(x^1| x^4, t_1)\, P(x^2| x^4, t_2)\, P(x^3| x^5, t_3)\, P(x^4| x^5, t_4)\, P(x^5)$

A shorthand notation $P(x^{\cdot}|T, t.)$ is used where $x^{\cdot}|$ stands for a set of sequences, and t. for edge lengths of the tree T.



In general, if we know

- P(x|y, t) for any x, y, and t
- A tree T, and assignment of sequences to tree nodes,

Then we can compute the likelihood for observing the sequences as they are assigned onto the tree leaves.

Q: Given n, the number of leaves, there are (2n-3)!! different trees (plus many different ways to assign length to tree edges), which tree can best interpret the data?

A: The tree that gives the maximum likelihood (ML).

In practice, to implement the ML method, two issues we need to address

1. A model of evolution, which gives the conditional probability P(x|y, t)
2. Method to find the maximum likelihood. For this, any optimization method may be utilized, such as
   - Descent gradient
   - Simulated annealing
   - Genetic algorithm

Models of evolution

Independence assumption: mutations occur independently at different positions.
Therefore,

$$P(x|y, t) = \prod_u P(x_u|y_u, t),$$

where $P(x_u|y_u, t)$ is the probability that residue x_u in sequence x mutates to residue y_u in sequence y over time t.

multiplicative assumption:

$$P(b|a, t + \Delta t) = \sum_c P(c|a, t) \cdot P(b|c, \Delta t).$$

For DNA sequences, probabilities for all possible mutations among four nucleotides during a given time period t form a 4 by 4 matrix

$$S(t) = \begin{pmatrix} P(A|A, t) & P(A|C, t) & P(A|G, t) & P(A|T, t) \\ P(C|A, t) & P(C|C, t) & P(C|G, t) & P(C|T, t) \\ P(G|A, t) & P(G|C, t) & P(G|G, t) & P(G|T, t) \\ P(T|A, t) & P(T|C, t) & P(T|G, t) & P(T|T, t) \end{pmatrix}$$

And, we have multiplicative property for these matrices

$$S(t) \cdot S(\Delta t) = S(t + \Delta t).$$

For each P(b|a, t) in the substitution matrix S(t), it is reasonable to assume that no mutation can occur at zero time interval:

- P(a|a, 0) = 1
- P(b|a, 0) = 0.

That is,

$$S(0) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Further, let's assume that P(b|a, Δt) over a infinitesimal interval Δt is proportional to Δt by a constant r, called mutation rate:

$$P(b|a, \Delta t) = r\,\Delta t.$$

Jukes-Cantor model for DNA sequences assumes that all nucleotide mutations have the same rate. That is,

$$S(\Delta t) = \begin{pmatrix} -3r & r & r & r \\ r & -3r & r & r \\ r & r & -3r & r \\ r & r & r & -3r \end{pmatrix} \quad \Delta t \equiv R\,\Delta t$$

Therefore,

$$S(t + \Delta t) = S(t) \cdot S(\Delta t) = S(t) \cdot [S(0) + S(\Delta t)]$$
$$= S(t) \cdot [I + R\Delta t]$$
$$[S(t + \Delta t) - S(t)] / \Delta t = S(t) \cdot R$$

$$S'(t) = S(t) \cdot R \quad \text{when } \Delta t \to 0.$$

Solving the differential equations, we have

- $P(a|a, t) = \frac{1}{4}(1 + 3e^{-4rt})$ for $a \in \{A, C, G, T\}$
- $P(b|a, t) = \frac{1}{4}(1 - e^{-4rt})$ for $a \neq b$, a and $b \in \{A, C, G, T\}$

In this model, when $t = \infty$, $P(b|a, \infty) = \frac{1}{4}$. That is, the nucleotide equilibrium frequencies are all equal.

Kimura model for DNA sequences assumes different rates for transitions (A↔G and C ↔ T) and transversions (A ↔ T, G ↔ T, A ↔ C, and C ↔ G).

That is,

$$S(\Delta t)= \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} -2r-s & r & s & r \\ r & -2r-s & r & s \\ s & r & -2r-s & r \\ r & s & r & -2r-s \end{pmatrix} \end{array} \Delta t \equiv R\,\Delta t$$

Similar models are proposed for mutations among amino acids.

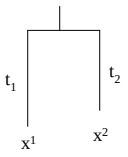
Should we be able to quantify the "time" as how many number mutations have occurred, the substitute matrices in those models would correspond to PAM matrices at respective times.

Maximum Likelihood

- The case of two sequences aligned with no gaps

$P(x^1, x^2|T, t_1, t_2) = \prod_{u=1 \text{ to } N} P(x^1_u, x^2_u|T, t_1, t_2)$

- Let x^1 = CCGGCCGCGCG
  x^2 = CGGGCCGCCCG



$$\begin{aligned} P(C,C|T, t_1, t_2) = {} & P(C|A, t_2)\,P(A|C, t_1)\,P(A) \\ & + P(C|C, t_2)\,P(C|C, t_1)\,P(C) \\ & + P(C|G, t_2)\,P(G|C, t_1)\,P(G) \\ & + P(C|T, t_2)\,P(T|C, t_1)\,P(T) \\ = {} & \tfrac{1}{4}\,[3 \times \tfrac{1}{4}\,(1- e^{-4rt1}) \times \tfrac{1}{4}\,(1- e^{-4rt2}) + \\ & \tfrac{1}{4}\,(1+3\, e^{-4rt1}) \times \tfrac{1}{4}\,(1+3\, e^{-4rt2})\,] \\ = {} & 1/16\,(1+3\, e^{-4r(t1\,+t2)}). \end{aligned}$$

$P(C,G|T, t_1, t_2) = 1/16\,(1- e^{-4r(t1\,+t2)})$.

Therefore, for an alignment that has n1 identical sites and n2 mutational sites, we have

$P(x^1, x^2|T, t_1, t_2) = 1/16^{n1+n2} \times (1+3\, e^{-4r(t1\,+t2)})^{n1} \times (1- e^{-4r(t1\,+t2)})^{n2}$,

which is a function of edge lengths in tree T.

In general, the probability can be computed by working up the tree from the leaves using post-order traversal. This is done by Felsenstein's algorithm (1981).

Once the probability is available, optimizing assignments of edge lengths t in the tree amounts to

$$\left.\frac{\partial P}{\partial t}\right|_{t_m} = 0$$

Where t_m is the tree length assignment that maximize the likelihood.

How to optimize tree topology?

- Discrete structure, therefore cannot take derivatives.

Basic strategy for searching the tree space

- A tree generation algorithm that can generate trees
- Assess the likelihood
  - Accept
  - Reject

A genetic algorithm implementation [Matsuda 1998]

Genetic algorithm

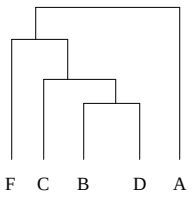
Input
- P, the population,
- r: the fraction of population to be replaced,
- f, a fitness,
- ft, the fitness_threshold,
- m: the rate for mutation.

Initialize population (randomly)
Evaluate: for each h in P, compute Fitness(h)
While $[Max_h f(h)] < ft$
do
1. Select
2. Crossover
3. Mutate
4. Update P with the new generation Ps
5. Evaluate: f(h) for all $h \in P$

Return the h in P that has the best fitness

Key components for implementing genetic algorithms

- Representing hypotheses (which are the trees here)
  - New Hampshire format
    (A,(F,(G,(B,D))))
- Genetic operators
  - Crossover:
    - Single
    - Two point
    - uniform
  - Mutation: point
- Fitness function: we use the likelihood computed for each tree using



More advanced topics in phylogenetic analysis

- Different heuristics for sampling the tree space
  - Monte Carlo
  - …
- More realistic evolutionary models
  - With gaps
  - Non-uniform: different rates at different sites
  - …
- Using different data sets and reconciliation
  - Sequences
  - Gene positions [Nadeau & Taylor 1984, PNAS 81:814-818, Pavzner, Sankoff, …]
  - …