Bioinformatics – Lecture Notes

Announcements

Next Thursday (3/28) I will be out of town. Isabel Darcy will teach the class. She will talk about DNA structure and how it relates to its function. She will demonstrate how the mathematical field of topology can aid in studying these topics.

Tuesday – NS&M Colloquium
Reception at 2:00 pm
Seminar at 2:30 pm
Galaxy Room – Student Union
Dr. Jacques Banchereau from Baylor Health Care System will be talking about immunology and the need for good biostatistical analysis.

Class 18 – March 21, 2002 -

A. Phylogenetic Trees – Maximum Likelihood Approaches

The basic approach initially proposed by Felsenstein is to estimate the branch lengths of a tree of given topology using the maximum likelihood method. Then local modifications are made to the topology and branch lengths are optimized again in a recursive fashion. New taxa are added one by one causing local modification of the tree topology.

Felsenstein observed that parsimony methods yields incorrect tree topologies when the amount of evolutionary change is sufficiently divergent in different branches. We will set up the "Pulley Principle" which is necessary for efficient computation.

Assumptions of this method:

1) n nucleotide sequences are given, each of the same length m
2) no insertions or deletions (only substitutions) have occurred in the construction of the phylogenetic tree
3) The evolutionary process is a reversible Markov process P(t) whose substitution rate matrix P'(0) is given by nucleotide specific substitution rates u_1, u_2, u_3, and u_4, as described earlier

$$\begin{matrix} -(u_2+u_3+u_4) & u_2 & u_3 & u_3 \\ u_1 & -(u_1+u_3+u_4) & u_3 & u_3 \\ u_1 & u_2 & -(u_1+u_2+u_4) & u_3 \\ u_1 & u_2 & u_3 & -(u_1+u_2+u_3) \end{matrix}$$

P(t) is reversible means that

$\pi_i\, p_{i,j}(t) = \pi_j\, p_{j,i}(t)$ for all states i, j, and all times t.

4) $u = u_2 + u_2 + u_3 + u_4 = 1$ so that t is the expected number of substitutions
5) the base changes at different sites of the length m oligonucleotides are independent events (not true biologically)

Assume that $u = u_2 + u_2 + u_3 + u_4$ and $\pi_i = u_i/u$ (this can be considered to be a nucleotide frequency).

From before we have

$p_{i,j}(t) = e^{-ut}\delta_{i,j} + (1 - e^{-ut})\,\pi_j$ where $\delta_{i,j} = 1$ if i = j and 0 otherwise. We must show that the frequencies in π are stationary that is,

$(\pi_1, \pi_2, \ldots \pi_n)P(t) = (\pi_1, \pi_2, \ldots \pi_n)$ for all t.

To do this we must show that for each j, $\sum_i \pi_i\, p_{i,j}(t) = \pi_j$

$$\sum_i \pi_i\, p_{i,j}(t) = \left[\sum_i \pi_i \{e^{-ut}\delta_{i,j} + (1-e^{-ut})\pi_j\}\right]$$
$$= \left[\sum_i \pi_i (1-e^{-ut})\pi_j\right] + \pi_j e^{-ut}$$
$$= \left[(1-e^{-ut})\pi_j \sum_i \pi_i\right] + \pi_j e^{-ut}$$
$$= \left[(1-e^{-ut})\pi_j\right] + \pi_j e^{-ut}$$
$$= \pi_j$$

Hence, the nucleotide frequencies are stationary.

To check reversibility, we need to show that $\pi_i\, p_{i,j}(t) = \pi_j\, p_{j,i}(t)$
for all states i, j, and all times t.

if $i = j$

$$\pi_i p_{i,j}(t) = \pi_i\{e^{-ut}\delta_{i,j} + (1-e^{-ut})\pi_j\}$$
$$= \pi_i\{e^{-ut} + (1-e^{-ut})\pi_j\}$$
$$= \pi_j\{e^{-ut} + (1-e^{-ut})\pi_i\}$$
$$= \pi_j\{e^{-ut}\delta_{i,j} + (1-e^{-ut})\pi_i\}$$
$$= \pi_j p_{i,j}(t)$$

if $i \neq j$

$$\pi_i p_{i,j}(t) = \pi_i\{e^{-ut}\delta_{i,j} + (1-e^{-ut})\pi_j\}$$
$$= \pi_i\{(1-e^{-ut})\pi_j\}$$
$$= \pi_j\{(1-e^{-ut})\pi_i\}$$
$$= \pi_j\{e^{-ut}\delta_{i,j} + (1-e^{-ut})\pi_i\}$$
$$= \pi_j p_{i,j}(t)$$

I. Likelihood of a Tree

The likelihood of a tree L(tree) = Pr(data|tree)

Consider the tree with m specific sites. If we had to consider all internal nodes of the tree, it would become a huge combinatorial problem. Instead, to find the likelihood for this tree, one has to multiply the likelihoods of m site specific trees.

(see section 4.4.1 for notes)

II. Recursive Definition of Likelihood

Calculation of the likehood is just one part of the problem of finding the maximum-likelihood tree. A great savings in computation can be obtained by using a *recursive* definition of likelihood. To do this we pick the *l*th site among the 1,…, m sites of all the n nucleotides of length m for which we want to construct a phylogenetic tree.

(see section 4.4.2 for notes)