An Evolutionary Model for Maximum Likelihood Alignment of DNA Sequences

MATH285K – Spring 2010 ***Presenter: Mandev Gill***

Reference: [1]

Introduction

We will look at a maximum likelihood method for the alignment of two DNA sequences which is based upon a statistical model of DNA sequence evolution. This approach allows only substitutions, single-base insertions, and single-base deletions. It is hoped to eventually replace this model with a more realistic version that can allow other events (such as inversions, large insertions, and large deletions).

The evolutionary model is a Markov process: the probability of a transition from the current state of a sequence is independent of the previous states of the sequences. The likelihood of a pair of modern sequences, *A* and *B*, separated from a common ancestral sequence *C* by divergence time *t* is

$$P_t(A,B) = \Sigma_C \; P_\infty(C)\, P_t(A|C) P_t(B|C)\ .$$

Here, $P_\infty(C)$ is the equilibrium probability of sequence *C*, and $P_t(A|C)$ is the transition probability from sequence *C* to sequence *A*. The values of these probabilities depend on parameters which are pertinent to the evolutionary process. We assume reversibility:

$$P_\infty(C)\, P_t(A|C) = P_\infty(A)\, P_t(C|A) \text{ for all } A \text{ and } C\text{, and all } t > 0\ .$$

Using this fact, we can write

$$P_t(A,B) = \Sigma_C \; P_\infty(A)\, P_t(B|C) P_t(C|A) = P_\infty(A)\, P_{2t}(B|A)\ .$$

This implies that it is unnecessary to sum over all possible ancestral sequences to compute the probability of two modern sequences arising from a common ancestral sequence. Instead, it is sufficient to treat one modern sequence as if it is the ancestor of the other.

The calculation of a transition probability can be separated into two components: a substitution process and an insertion-deletion process.

The Substitution Process

For the sake of simplicity, we adopt the Felsenstein (1981) substitution model (other reversible models could be used). Recall that a nucleotide can take on any of the values *A*, *G*, *C*, or *T*. Under this model, when a substation occurs, a base will be replaced by *A*, *G*, *C*, or *T* with respective probabilities π_A, π_G, π_C, or π_T (the equilibrium probabilities). Let *s* denote the rate of base substitution. The transition probability that a nucleotide which begins as type *i* is of type *j* at time *t* is

$$f_{ij}(t)= \pi_j(1- e^{-st}) \qquad \text{if } i \neq j$$

$$f_{ii}(t) = e^{-st} + \pi_i(1- e^{-st}) \,.$$

Note that it is possible under this model to, for instance, to substitute an *A* by another *A*.

The Insertion-Deletion Process

For simplicity, we think of the process in terms of imaginary links that separate the DNA bases of a sequence. In our model, a sequence of *N* bases has *N* normal links and one immortal link. There is a normal link (which we denote by a *) to the right of each base. The leftmost base in the sequence is considered to have an immortal link (denoted by $) to its left. For instance, the sequence *AGGGCCTA* can be depicted as

$$\$\, A * G * G * G * C * C * T * A * .$$

If the presence of nucleotides is considered without regard to the actual types of nucleotides, we can represent the sequence as:

$$\$ \; * \; * \; * \; * \; * \; * \; * \; * \, .$$

The insertion-deletion process is framed as a birth-death process of these links. A birth or death of one link does not affect the probability of a birth or death of another link. Both types of links can be associated with births, and they have the same birth rate (which we denote by λ). A newborn link is always a normal link. We adopt the convention that a new link appears immediately to the right of its parent link. The birth of a normal link is accompanied by the birth of a DNA base immediately to its left. The probabilities that the newborn DNA base will be *A*, *G*, *C*, or *T* are, respectively, π_A, π_G, π_C, and π_T. Normal links are subject to death (at rate μ), but not immortal links.

At any given instant, a sequence will either increase or decrease its length by a single nucleotide or stay the same length. The chance of more than one birth or death taking place within a sequence at the same instant is negligible. A sequence of *n* nucleotides will increase to length $n + 1$ at rate $(n + 1)\lambda$, and decrease to length *n* at rate $n\mu$. The presence of immortal links in this model is necessary for the existence of a realistic equilibrium distribution of sequence lengths. Without immortal links, sequences would tend over time to length 0 or infinity.

Likelihood of a Pair of DNA Sequences

The calculation of the likelihood requires the calculation of transition probabilities from ancestral sequence to descendant sequence, as well as calculation of the equilibrium probability of the ancestral sequence.

If the ancestral sequence A has a length of n nucleotides, then

$$P_{\infty}(A) = \gamma_n \Pi_{i=1,\ldots,n} \pi_{n(i)}$$

where $n(i)$ denotes the nucleotide as position i, and γ_n is the equilibrium probability of sequences of n nucleotides in length. It can be shown that the distribution of γ_n obtained under the birth-death model is the geometric distribution

$$\gamma_n = (1 - \lambda/\mu)(\lambda/\mu)^n .$$

Now we turn to transition probabilities. Various paths are possible for a transition from one sequence to another. The transition probability from one sequence to another is the sum of the probabilities of all possible paths connecting the two sequences. The particular path of a transition from one sequence to another can be expressed well by alignment. For example, consider the following alignment, which we denote by α:

```
– T G T – C –
G – C – A C A
```

We can represent the information on presence and absence of bases in α by links and denote it α':

```
$ _ * * * _ * _
$ * _ * _ * * *
```

Let θ denote the collection of parameters μt, λt, st, π_A, π_G, π_C, and π_T. The probability of a specific transition path represented by α is

$$P(\alpha \mid \theta) = P(\alpha, \alpha' \mid \theta) = P(\alpha \mid \alpha', \theta) P(\alpha' \mid \theta) .$$

$P(\alpha' \mid \theta)$ is the product of γ_n and the appropriate transition probabilities for links (which we will describe shortly). $P(\alpha \mid \alpha', \theta)$ is a product of equilibrium and transition probabilities (determined by the way the descendent sequence evolves from the ancestral sequence through the births of new links and nucleotide substitutions).

Three types of transition probabilities are considered for links: $p_n(t)$ is the probability that after time t, n links are descended from a normal link, including the original link; $p_n'(t)$ is the probability that after time t, n links are descended from a normal link, but the original dies; $p_n''(t)$ is the probability that after time t, the immortal links has n descendants, including itself.

Returning to the example above, we have

$$P(\alpha' \mid \theta) = \gamma_4\, p_2''(t)\, p_0'(t)\, p_1(t)\, p_1'(t)\, p_2(t)$$

$$P(\alpha \mid \alpha', \theta) = \pi_G\, f_{GC}(t)\, \pi_A\, f_{CC}(t)\, \pi_A\, .$$

Explicit expressions can be obtained for $p_n''(t)$, $p_n''(t)$, and $p_n''(t)$ by solving the differential equations which govern the birth-death process. The solutions are:

$$p_0(t) = p_0''(t) = 0$$

$$p_n(t) = \mathrm{e}^{-\mu t}\, [1 - \lambda\beta(t)]\, [\lambda\beta(t)]^{\mathrm{n-1}}\, , \qquad n > 0$$

$$p_n'(t) = [1\text{- } \mathrm{e}^{-\mu t} - \mu\beta(t)]\, [1 - \lambda\beta(t)]\, [\lambda\beta(t)]^{\mathrm{n-1}}\, , \qquad n > 0$$

$$p_0'(t) = \mu\beta(t)$$

$$p_n''(t) = [1 - \lambda\beta(t)]\, [\lambda\beta(t)]^{\mathrm{n-1}}\, , \qquad n > 0$$

where, $\beta(t) = (1\text{- } \mathrm{e}^{(\lambda-\mu)t})/(\mu - \lambda \mathrm{e}^{(\lambda-\mu)t})$.   .

Alignment Algorithm

The alignment algorithm is a recursive algorithm that can produce maximum likelihood alignment between a sequence A and a sequence B and its likelihood for a given value of θ. The procedure consists of gradually filling in the entries of a matrix. Each matrix position corresponds to a subsequence of A and a subsequence of B. Each entry in the matrix is determined by considering its previously calculated neighboring entries.

Let A_m denote the subsequence consisting of the first m bases of the sequence A. Define B_n analogously. Because the model is reversible, we can, without loss of generality, consider A to be an ancestor of B. Let $S(A_m,B_n)$ denote the set of all possible alignments between A_m and B_n. Each possible alignment $\alpha(A_m,B_n)$ between A_m and B_n is a member of exactly one of the following three subsets of $S(A_m,B_n)$:

$S_0(A_m,B_n) = \{\ \alpha(A_m,B_n)$ s.t. rightmost link of A_m has no descendant links in $B_n\ \}$

$S_1(A_m,B_n) = \{\ \alpha(A_m,B_n)$ s.t. rightmost link of A_m has exactly one descendant link in $B_n\ \}$

$S_2(A_m,B_n)$ = { $\alpha(A_m,B_n)$ s.t. rightmost link of A_m has at least two descendant links in B_n }

The likelihood of a specific subsequence alignment $\alpha(A_m,B_n)$ for a certain value of θ will be written $l_\theta\ [\alpha^i\ (A_m,B_n)]$ where i = 0,1,or 2, depending on the subset of $S(A_m,B_n)$ to which $\alpha(A_m,B_n)$ belongs. For a certain value of θ, the alignment of highest likelihood in $S_i(A_m,B_n)$ is written $\alpha_{max}{}^i(A_m,B_n)$. Additionally, define

$$l_\theta\ [\alpha_{max}(A_m,B_n)] = max\ _i\ \{\ l_\theta\ [\alpha_{max}{}^i\ (A_m,B_n)]\ \}.$$

The maximum likelihood alignment between A and B for a particular value of θ can be determined by a recursive procedure that updates each $l_\theta\ [\alpha_{max}{}^i\ (A_m,B_n)]$.

Let $a(m)$ denote the type of nucleotide at the mth position of sequence A (define $b(n)$ analogously). The recursive procedure starts with the boundary conditions

$$l_\theta\ [\alpha_{max}{}^0\ (A_0,B_0)] = l_\theta\ [\alpha_{max}{}^2\ (A_0,B_0)] = 0$$

$$l_\theta\ [\alpha_{max}{}^1\ (A_0,B_0)] = \gamma_0\ p_1{}''(t)$$

$$l_\theta\ [\alpha_{max}{}^1\ (A_m,B_0)] = l_\theta\ [\alpha_{max}{}^2\ (A_m,B_0)] = 0\ ,\ \text{where } 1 \le m \le \text{length}(A)$$

$$l_\theta\ [\alpha_{max}{}^0\ (A_m,B_0)] = \gamma_m\ p_1{}''(t)\ \Pi_{\ i=1,\dots,m}\ \ \pi_{a(i)}\ ,\ \text{where } 1 \le m \le \text{length}(A)$$

$$l_\theta\ [\alpha_{max}{}^0\ (A_0,B_n)] = l_\theta\ [\alpha_{max}{}^1\ (A_0,B_n)] = 0\ , \qquad \text{where } 1 \le n \le \text{length}(B)$$

$$l_\theta\ [\alpha_{max}{}^2\ (A_0,B_n)] = \gamma_0\ p_{n+1}{}''(t)\ \Pi_{\ i=1,\dots,n}\ \ \pi_{b(i)}\ ,\text{where } 1 \le n \le \text{length}(B)\ .$$

For $1 \le m \le$ length(A) and $1 \le n \le$ length(B), the recursive procedure follows these rules:

$$l_\theta\ [\alpha_{max}{}^0\ (A_m,B_n)] = (\lambda/\mu)\ \pi_{a(m)}\ p_0{}'(t)\ l_\theta\ [\alpha_{max}(A_{m-1},B_n)]$$

$$l_\theta\ [\alpha_{max}{}^1\ (A_m,B_n)] = (\lambda/\mu)\ \pi_{a(m)}\ max[f_{\ a(m)b(n)}(t)\ p_1(t)\ ,\ \pi_{b(n)}\ p_1{}'(t)\ ]\ l_\theta\ [\alpha_{max}(A_{m-1},B_{n-1})]$$

$$l_\theta\ [\alpha_{max}{}^2\ (A_m,B_n)] = \pi_{b(n)}\ \lambda\ \beta(t)\ max\{\ l_\theta\ [\alpha_{max}{}^1\ (A_m,B_{n-1})]\ ,\ l_\theta\ [\alpha_{max}{}^2\ (A_m,B_{n-1})]\ \}$$

The maximum likelihood alignment between A and B has likelihood

$$l_\theta\ [\alpha_{max}(A,B)] = max\ _i\ \{\ l_\theta\ [\alpha_{max}{}^i\ (A,B)]\ \}.$$

Recovery of the actual maximum likelihood alignment is obtained by tracing back through the likelihood matrix on the path that led to the maximum likelihood value.

Reference

[1] Thorne JL, Kishino H, Felsenstein J (1991) An evolutionary model for the maximum likelihood alignment of DNA sequences. J Mol Evol 33:114-124