Convergence of the TKF91 model

Bhalchandra D. Thatte

Rényi Alfréd Matematikai Kutató Intézet, Budapest.

(Bayesian Phylogenetics Meeting, June 2008)

Overview

- The problem
- The TKF91 process of sequence evolution
- Setting up the differential equations for the TKF91 process
- Solutions to the differential equations
- Results of Hakimi & Patrinos (1972), Zaretskii (1965) and Buneman (1971)
- Applying the results of Hakimi et al.
- Reconstruction from sequence lengths.

The problem

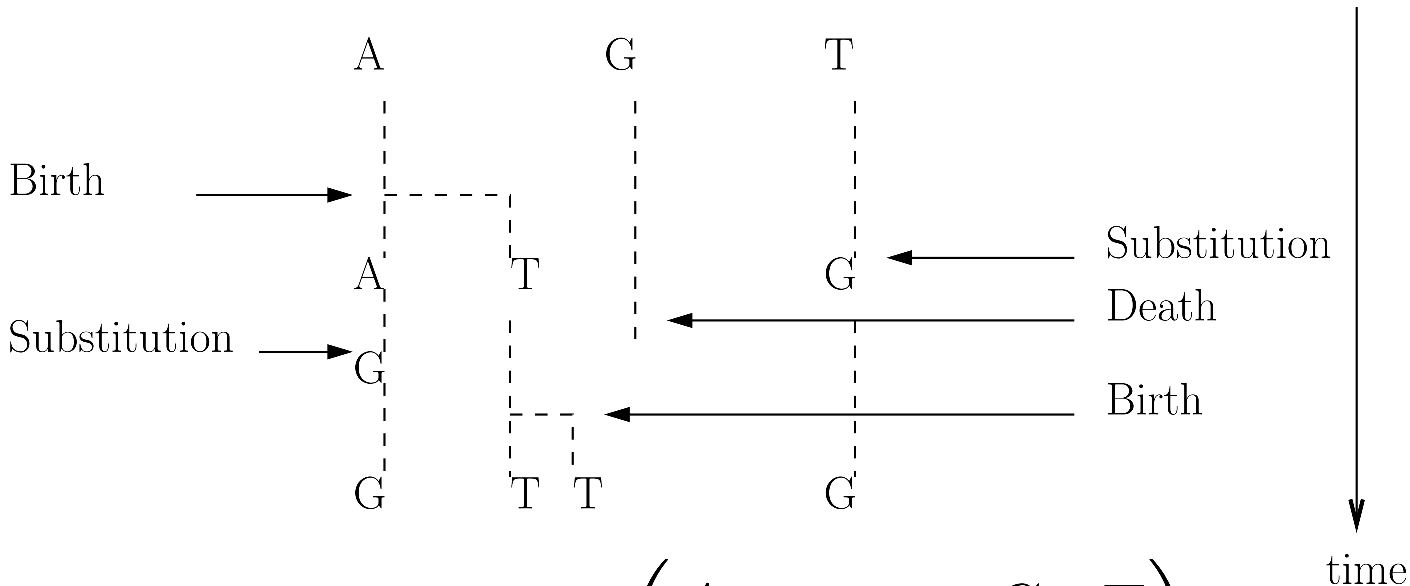
Given a multiple alignment of sufficiently long sequences, can we construct uniquely the phylogenetic tree topology, edge lengths and other model parameters?

Examples of pure substitution models

1. Hadamard conjugation method
   (Hendy, Penny, Steel, ...)
2. Chang's results on very general Markov transition matrices

The TKF91 process

Thorne, Kishino and Felsenstein (1991) proposed a process to model substitutions and single character insertions and deletions.



Alignment $\begin{pmatrix} A & - & - & G & T \\ G & T & T & - & G \end{pmatrix}$

Specification of the TKF91 process

- λ = Poisson birth rate
- μ = Poisson death rate, where $\mu > \lambda$.
- s = Poisson substitution rate
- Each character in the sequence, and each newly born character evolves under the three processes independently of all other characters.
- When a substitution or a birth event occurs, the substituted or newly born character is chosen from a uniform distribution on the set of possible characters.

Setting up the differential equations - I

Let $p_n^H(t)$ be the probability that a character survives for time t, and at time t, it has n descendants.

$$\frac{dp_n^H}{dt} = \lambda(n-1)p_{n-1}^H + \mu n p_{n+1}^H - (\lambda+\mu)np_n^H$$

with the initial conditions

$$\begin{aligned} p_1^H(t=0) &= 1 \\ p_n^H(t=0) &= 0 \text{ for } n > 1 \end{aligned}$$

Setting up the differential equations - II

Let $p_n^N(t)$ be the probability that the character dies before time t, but leaves behind n descendents at time t.

By convention, $p_n^N(t) = 0$ for $n < 0$.

The differential equation for $p_n^N(t)$ is

$$\frac{dp_n^N}{dt} = \lambda(n-1)p_{n-1}^N + \mu(n+1)p_{n+1}^N + \mu p_{n+1}^H - (\lambda+\mu)np_n^N$$

with the initial condition

$$p_n^N(t=0) = 0 \text{ for all } n$$

Solving the differential equations

Lemma 1. *Solutions to the above differential equations are given by*

$$\begin{aligned} p_n^H(t) &= e^{-\mu t}(1-\lambda\beta(t))(\lambda\beta(t))^{n-1} \text{ for } n>0 \\ p_n^N &= \mu\beta(t) \text{ for } n=0 \\ &= (1-e^{-\mu t}-\mu\beta(t))(1-\lambda\beta(t))(\lambda\beta(t))^{n-1} \text{ for } n>0 \end{aligned}$$

where

$$\beta(t) = \frac{1-e^{(\lambda-\mu)t}}{\mu-\lambda e^{(\lambda-\mu)t}}$$

Blocks of an alignment

A part of the alignment of two sequences S_1 and S_2:

$$\begin{pmatrix} S_1 & \# & \# & \# & \# & \# & - & \# & - & \# & \# \\ S_2 & \# & \# & - & \# & - & \# & \# & \# & - & \# \end{pmatrix}$$

The alignment shows four types of blocks.

$$A \equiv \begin{pmatrix} \# & \# \\ \# & \# \end{pmatrix} \quad B \equiv \begin{pmatrix} \# & \# & \# \\ \# & - & \# \end{pmatrix}$$

$$C \equiv \begin{pmatrix} \# & \# & - & \# \\ \# & - & \# & \# \end{pmatrix} \quad D \equiv \begin{pmatrix} \# & - & \# & \# \\ \# & \# & - & \# \end{pmatrix}$$

Computing $\mathbb{P}(A)$, $\mathbb{P}(B)$ and $\mathbb{P}(C \vee D)$

- Probabilities of observing blocks of type A, B, and $C \vee D$ can be computed using the transition matrix for the Markov chain on the columns of an alignment of two sequences.
- The states of a Markov chain are the three types of columns $\binom{\#}{\#}$, $\binom{\#}{-}$ and $\binom{-}{\#}$, and the end state, and the matrix of transition probabilities is given by Hein, Jensen, Pedersen (2003).
- The probabilities $\mathbb{P}(A)$, $\mathbb{P}(B)$ and $\mathbb{P}(C \vee D)$ do not depend on the root.
- $e^{-\mu t_{ij}}$ can be estimated for each pair (S_i, S_j) of sequences in a multiple alignment.

A result of Hakimi, et al.

Lemma 2. *Let T be a tree on the vertex set V. Let f be a non-zero real valued function defined on the set of subsets of V of cardinality 2, satisfying the additivity condition*

$$f(\{x, y\}) = \sum_{i=0}^{r-1} f(\{x_i, x_{i+1}\})$$

where $x_0, x_1, \ldots, x_r$ is the unique path in T connecting $x = x_0$ and $y = x_r$. Then the value of f on all pairs of leaf nodes of T determines uniquely the tree T and the function f.

In other words ...

Suppose $T_1(U,E)$ and $T_2(V,F)$ are two trees having the same leaf set X. Let there be non-zero real-valued functions $f : U^{(2)} \longrightarrow \mathbb{R}$ and $g : V^{(2)} \longrightarrow \mathbb{R}$ that satisfy the additivity condition. If f and g agree on $X^{(2)}$, then there is an isomorphism π from T_1 to T_2 such that $\pi(x) = x$ for each $x \in X$, and $f(\{u,v\}) = g(\{\pi(u), \pi(v)\})$.

From a multiple alignment to the tree

Combining $e^{-\mu t_{ij}}$ that were calculated for each pair (S_i, S_j) of sequences, and the result of Hakimi, et al., a unique tree is constructed for sufficiently long sequences.

From the sequence lengths to the tree

Suppose a sequence has length X_0 at time $t = 0$.
Let $\mathbb{P}(X, t)$ be the probability that the sequence has length X at time t. Then $\mathbb{P}(X, t)$ satisfies

$$
\begin{aligned}
\frac{dP(X,t)}{dt} &= \lambda X P(X-1,t) + \mu(X+1)P(X+1,t) \\
&\quad -\lambda(X+1)P(X,t) - \mu X P(X,t)
\end{aligned}
$$

with the initial conditions

$$
\begin{aligned}
P(X = X_0, 0) &= 1 \\
P(X \neq X_0, 0) &= 0
\end{aligned}
$$

Solving for the moments of X

The differential equation for the first moment is

$$\frac{dM_1}{dt} = (\lambda - \mu)M_1 + \lambda$$

with the initial condition

$$M_1(0) = X_0$$

$$\frac{dM_2}{dt} = 2(\lambda - \mu)M_2 + (3\lambda + \mu)M_1 + \lambda$$

with the initial condition

$$M_2(0) = X_0^2$$

Köszönöm

Allan Wilson Centre for Molecular Ecology and Evolution

Mike Steel

Rényi Institute and the HUBI project (Istvan Miklos and Peter Erdős)

organizers of the conference,

all of you (and myself) for turning up so early in the morning.