Estimating Evolutionary Distances between Sequences

David Bryant

McGill Centre for Bioinformatics
Montréal, Québec

Abstract. These notes accompany a lecture at the summer school on mathematics for bioinformatics, Centre de Recherches Mathematiques, Montreal, August 2003. I develop basic Markov models for sequence evolution and show how these may be used to estimate evolutionary divergences between sequences. I then discuss extensions of the basic model to situations where the evolutionary rate varies over time and for different sites.
These notes are a sneak preview of a more detailed survey paper that is currently in preparation. Please respect this pre-publication status.

1 A discrete time model

1.1 Starting point - sequences

A *genetic sequence* is a string of finite length on a set of *states* which we number $1, 2, \ldots, r$ ($r = 4$ for nucleotide sequences, $r = 20$ for amino sequences). The positions in the sequence are called *sites*. We assume that, for each i, the states in site i in each sequence evolved from the same common ancestor: the sites are *homologous*.

To measure the distance between two sequences we could simply count the number of sites with differences. We would then lose *hidden mutations*. For example $A \to C \to G$ would be counted as one mutation when there were actually two, and $A \to C \to G \to A$ would be counted as zero mutations when there were actually three. To estimate how many hidden mutations there were we use a Markov model.

1.2 Markov chain model

We assume every site has the same probability distribution and this evolution is **independent** between different sites. Thus the probability of going from sequence A to sequence B (each with length m) over a certain time is given by

$$\mathbb{P}[A \to B] = \prod_{i=1}^{m} \mathbb{P}[A[i] \to B[i]].$$

Because we can separate out the probability we need only focus on the probability of mutation for one site.

We assume that time proceeds in discrete 'ticks' and study the evolution of an (ancestral) sequence A to the sequence B. We model the evolution of a site by a *Markov Chain*. Let R denote the *transition matrix*, so R_{ij} equals the probability of a site being in state j after one unit of time (tick) *conditional* on that site starting in state i. For each $k = 0, 1, 2, \ldots$, $(R^k)_{ij}$ is the probability of a site being in state j after k ticks *conditional* on that site starting in state i.

1.3 Stationary distributions

We've already made many significant assumptions about the evolutionary process. Here are some more assumptions.

1. The Markov chain is *irreducible* - every state can get to very other.
2. The chain is *aperiodic* - we never get into loops we can't get out of.
3. All of the states of the chain are *ergodic* - as we run the chain to the limit, the probability of being in each state j is non-zero and independent of the starting states. That is, there are $\pi_1, \ldots, \pi_r$, all positive, such that
$$\lim_{k\to\infty} (R^k)_{ij} = \pi_j$$

The values $\pi_1, \ldots, \pi_r$ comprise a *stationary distribution* (also called the *equilibrium distribution* or *equilibrium frequences*) for the states. These satisfy

$$\pi_j = \sum_{i=1}^{r} \pi_i R_{ij}. \tag{1}$$

If we sample the initial state from the stationary distribution, then run the chain for k steps, then the distribution of the final state will equal the stationary distribution. Equation (1) allows us to recover $\pi_1, \ldots, \pi_r$ just given the matrix R.

Note that if we sample the initial state from any distribution, then run the chain to infinity, the final state will always have the stationary distribution. To see this, consider non-negative $a_1, a_2, \ldots, a_r$ that sum to 1. Then

$$\begin{aligned}\lim_{k\to\infty} \sum_{i=1}^{r} a_i (R^k)_{ij} &= \sum_{i=1}^{r} a_i \pi_j \\ &= \pi_j.\end{aligned}$$

We use $\mathit{\Pi}$ to denote the $r \times r$ diagonal matrix with π_i's down the diagonal.

1.4 Divergence matrix

The matrix R^k gives the conditional probabilities

$$(R^k)_{ij} = \mathbb{P}[\text{in state } j \text{ after } k \text{ ticks} \mid \text{initial state } i].$$

Define the matrix $X(k)$ by

$$X(k)_{ij} = \mathbb{P}[\text{in state } j \text{ after } k \text{ ticks} \wedge \text{ initial state } i].$$

We assume that the initial state was sampled from the stationary distribution (that is, the process was already in equilibrium). Then

$$X(k)_{ij} = \pi_i (R^k)_{ij}$$

or in matrix notation

$$X(k) = \mathit{\Pi} R^k.$$

The matrix $X(k)$ is called the *divergence matrix*.

1.5 Time reversibility

We say that the Markov chain is *time reversible* if the divergence matrix $X(k)$ is symmetric for all k (that is, $X(k)_{ij} = X(k)_{ji}$ for all i, j, k)

Lemma 1. *The Markov chain is* time reversible *if and only if* $\Pi R = R^T \Pi$.

Proof
Suppose that $\Pi R = R^T \Pi$. Then

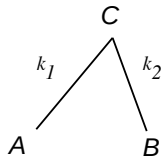
$$X(k) = \Pi R^k = R^T \Pi R^{k-1} = (R^k)^T \Pi = (X(k))^T$$

so $X(k)$ is symmetric. Conversely, if $X(1)$ is symmetric then

$$\Pi R = X(1) = (X(1))^T = R^T \Pi$$

□

We will assume the Markov chain is time reversible. This means that the direction of evolution does not affect the probabilities of different sites in two sequences. Consider the simple two leaf tree



where there are k_1 time units down the left branch and k_2 down the right. Let A, B, C denote the states at a particular site in three sequences, where the first two sequences evolved from the common ancestor. We assume that evolution is independent down the different branches. Then

$$\begin{aligned}
\mathbb{P}[A = a \wedge B = b] &= \sum_{c=1}^{r} \mathbb{P}[A = a \wedge B = b | C = c]\mathbb{P}[C = c] \\
&= \sum_{c=1}^{r} \mathbb{P}[A = a | C = c]\mathbb{P}[B = b | C = c]\pi_c \\
&= \sum_{c=1}^{r} \pi_c (R^{k_1})_{ca} (R^{k_2})_{cb} \\
&= \sum_{c=1}^{r} \pi_a (R^{k_1})_{ac} (R^{k_2})_{cb} \qquad \text{since } X(k_1) \text{ is symmetric} \\
&= \pi_a R^{k_1+k_2}_{ab}
\end{aligned}$$

Thus $\mathbb{P}[A = a \wedge B = b]$ depends only on the sum $k_1 + k_2$ and we can consider either sequence to be the root without affecting the probability.

2 From Discrete to Continuous

2.1 Making a discrete time model continuous

Suppose now that the ticks occur according to a Poisson process, where the probability of observing k ticks over time t equals

$$\mathbb{P}[k \text{ ticks}] = e^{-\mu t}\frac{(\mu t)^k}{k!}.$$

The expected number of ticks equals μt, and μ is the expected number of ticks per unit time (this is the *mean instantaneous substitution rate*, see below).

Let $P(t)_{ij}$ denote the probability of being in state j at time t conditional on being in state i at time 0. Then

$$\begin{aligned} P(t) &= \sum_{k=0}^{\infty} R^k \mathbb{P}[k \text{ ticks}] \\ &= \sum_{k=0}^{\infty} R^k e^{-\mu t}\frac{(\mu t)^k}{k!} \\ &= e^{-\mu t}\sum_{k=0}^{\infty}\frac{(\mu t R)^k}{k!} \\ &= e^{-\mu t}e^{\mu t R}. \end{aligned}$$

Note that $e^{\mu t R}$ is a *matrix* exponential. The exponential of a matrix A is defined by

$$e^A = \sum_{k=0}^{\infty}\frac{A^k}{k!}.$$

Carrying on, we have

$$\begin{aligned} P(t) &= e^{-\mu t}e^{R\mu t} \\ &= e^{-\mu t I}e^{R\mu t} && I \text{ is the identity matrix} \\ &= e^{(R-I)\mu t} \\ &= e^{Q\mu t} \end{aligned}$$

where $Q = R - I$. The matrix Q is called the *(instantaneous) rate matrix*.

The continuous time divergence matrix is defined

$$X(t) = \Pi P(t) = \Pi e^{Q\mu t}$$

Clearly, X is symmetric when the original chain is time reversible.

2.2 Rate of mutation

One aspect of evolutionary models that often causes confusion is the notion of evolutionary rate. Typically, the evolutionary rate of a model is not made explicit, causing problems if we compare results from different models, or reconstruct branch lengths from simulated sequences.

The key observation is that there are two measures of rates in use. The first is called the *mean instantaneous substitution rate* and is equal to the parameter μ defined above. The parameter μ equals the expected number of ticks per unit time. The problem with this measure is that multiplying the rate matrix Q by a constant k, then dividing μ by k, gives an identical Markov process with a different mean instantaneous substitution rate.

The second measure is the *mutation rate*. This equals the expected number of mutations, *including hidden mutations*, per unit time. We now show how to compute the mutation rate for a given model. We consider once again the discrete case. After one step, the probability of a mutation equals the sum of the off-diagonal elements in $X(1) = \Pi R$. As the entries of ΠR sum to one, this probability equals 1 minus the trace of ΠR. [Recall that the *trace* of a matrix A, denoted here by $\mathrm{tr}(A)$, equals the sum of the diagonal entries]. Thus the expected number of mutations after one tick equals

$$1 - \mathrm{tr}(\Pi R).$$

The expected number of mutations after k ticks is therefore given by

$$\mathbb{E}[\text{number of mutations}|k \text{ ticks}] = k(1 - \mathrm{tr}(\Pi R)).$$

Since $\mathbb{E}[k] = \mu t$ we have $\mathbb{E}[k(1 - \mathrm{tr}(\Pi R))] = \mu t \cdot (1 - \mathrm{tr}(\Pi R))$ and

$$\begin{aligned}\mathbb{E}[\text{number of mutations}] &= \mathbb{E}[\text{number of mutations}|k \text{ ticks}] \\ &= \mathbb{E}[k(1 - \mathrm{tr}(\Pi R))] \\ &= \mu t \cdot (1 - \mathrm{tr}(\Pi R)).\end{aligned}$$

Note that here we have used the *law of total expectation*, which says that for random variables X, Y we have $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$. Since $1 - \mathrm{tr}(\Pi R) = -\mathrm{tr}(\Pi Q)$ we obtain

Theorem 1. *The expected number of mutations per unit time equals* $-\mu \cdot \mathrm{tr}(\Pi Q)$.

You can also derive this result using calculus.

Often texts on molecular evolution only give the rate matrix Q to describe the model. It is not always the case that $R = Q + I$ is a valid transition matrix, since the diagonal entries of Q may be less than -1. However, if you scale set $\mu = -\mathrm{tr}(\Pi Q)$ and $R = \frac{-1}{\mathrm{tr}(\Pi Q)} + I$ then R will be a valid transition matrix, and the original Markov process will correspond to a Markov chain with ticks distributed according to a Poisson process with mean μt. Thus we have not lost any generality by building the continuous model from the discrete case.

3 Distance estimates from inverting probabilities

3.1 Divergence

The *divergence* d between two sequences equals the expected number of mutations on the path connecting them (that is, from one sequence, to their common ancestor, to the other sequence). For a particular Markov model, we have

$$d = -\mu t \cdot \mathrm{tr}(\Pi Q)$$

where μ is the instantaneous substitution rate, t is time, Π the matrix with equilibrium frequencies down the diagonal, Q the rate matrix for the process. Recall that $-\mu \cdot \mathrm{tr}(\Pi Q)$ equals the expected number of mutations per unit time.

3.2 Example: Cavander-Farris model

Recall that, under the Cavander-Farris two state model (with rate equal to one), the conditional probabilities are

$$P(t) = \begin{bmatrix} \frac{1}{2} + \frac{1}{2}e^{-2*t} & \frac{1}{2} - \frac{1}{2}e^{-2*t} \\ \frac{1}{2} - \frac{1}{2}e^{-2*t} & \frac{1}{2} + \frac{1}{2}e^{-2*t} \end{bmatrix}.$$

The divergence matrix is

$$X(t) = \Pi P(t) = \begin{bmatrix} \frac{1}{4} + \frac{1}{4}e^{-2*t} & \frac{1}{4} - \frac{1}{4}e^{-2*t} \\ \frac{1}{4} - \frac{1}{4}e^{-2*t} & \frac{1}{4} + \frac{1}{4}e^{-2*t} \end{bmatrix},$$

so the probability of observing a change is

$$p = X(t)_{12} + X(t)_{21} = \frac{1}{2} - \frac{1}{2}e^{-2*t}.$$

We can invert this formula, to give

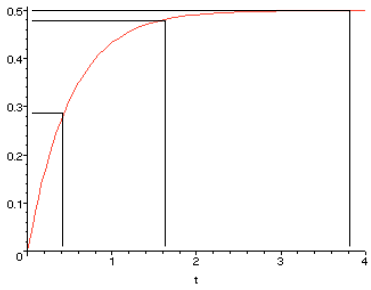
$$t = -\frac{1}{2}\log(1 - 2p)$$

Let $\hat{p}$ denote the proportion of sites where there is a change between the two sequences. As the number of sites gets large, this proportion will get closer and closer to p. Hence

$$\hat{t} = -\frac{1}{2}\log(1 - 2\hat{p})$$

is a *consistent* estimate for t. (An estimate is consistent if it approaches the true value as the amount of data gets larger.)

The curve $p = \frac{1}{2} - \frac{1}{2}e^{-2*t}$ has the graph



Notice two things: as t gets larger the curve flattens out, so small changes in p correspond to large changes in t. If $p \geq 0.5$ then there is no corresponding value of t and the distance correction is undefined.

When $\hat{p}$ approaches (or exceeds) the limit (0.5) we say that the data is *saturated*. Either our sequences are not long enough, in which case $\hat{p}$ is large because of sampling error, or we have the wrong model.

3.3 Example: Jukes-Cantor model

From Maple, and the examples last week, we have that the probability of observing a change is

$$p = 1 - \mathrm{tr}(X(t)) = \frac{3}{4} - \frac{3}{4}e^{-\mu t}$$

The mutation rate of the JC model equals $\frac{3}{4}\mu$.

Inverting this equation gives an estimator of

$$\tfrac{3}{4}\mu t = -\frac{3}{4}\log(1-\frac{4}{3}p)$$

so the sequences become saturated as p approaches $\frac{3}{4}$.

3.4 General case

The divergence matrix for a given model can be written

$$X(t) = \Pi e^{Q\mu t}$$

which we can invert to give

$$Q\mu t = \log(\Pi^{-1}X(t)).$$

Note that log is the *matrix logarithm*, which is defined by the series

$$\log(I+A) = -\sum_{k=1}^{\infty}(-1)^k\frac{A^k}{k}$$

The log of a matrix exists whenever the eigenvalues of the matrix are all positive.

Multiplying by $-\Pi$ and taking the trace of both sides gives

$$d = -\mathrm{tr}(\mu\Pi Q)t = -\mathrm{tr}(\Pi\log(\Pi^{-1}X(t))). \tag{2}$$

For each pair of states i, j let F_{ij} denote the observed proportion of sites with an i in sequence A and a j in sequence B. Then as the sequences get longer and longer, $F \to X(t)$. We therefore have a general estimator

$$\hat{d} = -\mathrm{tr}(\Pi\log(\Pi^{-1}F)). \tag{3}$$

of the expected number of mutations between A and B. This works for any model.

If we have a general time reversible model, then $X(t)$ is symmetric, so we can replace $X(t)$ in (2) by $\frac{1}{2}(X(t)+X(t)^T)$ and F in (3) by $\frac{1}{2}(F+F^T)$. This reduces the variance of the estimator.

3.5 Obtaining specific distance corrections from the general distance correction

All of the standard distance corrections can be derived from this general formula. or example, under the Jukes-Cantor model, suppose that we have observed that the proportion of changed sites equals p. Our estimate for $X(t)$ would then be the matrix

$$F = \begin{bmatrix} \frac{1-p}{4} & \frac{p}{12} & \frac{p}{12} & \frac{p}{12} \\ \frac{p}{12} & \frac{1-p}{4} & \frac{p}{12} & \frac{p}{12} \\ \frac{p}{12} & \frac{p}{12} & \frac{1-p}{4} & \frac{p}{12} \\ \frac{p}{12} & \frac{p}{12} & \frac{p}{12} & \frac{1-p}{4} \end{bmatrix}$$

Substituting this matrix into (3) gives the standard Jukes-Cantor correction

$$\hat{d} = -\frac{3}{4}\log(1-\frac{4}{3}p).$$

3.6 LogDet

Lockhart et al. propose the LogDet distance correction

$$\hat{d} = -1/r \log\left(\frac{\det F}{\sqrt{\det \Pi_x \Pi_y}}\right)$$

where Π_x is the character state frequencies for one sequence and Π_y is the character state frequencies for the other.

The LogDet is closely related to the general GTR correction, in fact it is far more closely related than is widely acknowledged. By studying the eigenvalues, one can easily show that for a matrix A,

$$\log(\det A) = \mathrm{tr}(\log(A)).$$

Hence if $\Pi = \sqrt{\Pi_x \Pi_y}$ then

$$\begin{aligned} -1/r \log\left(\frac{\det F}{\sqrt{\det \Pi_x \Pi_y}}\right) &= -1/r \log(\det \Pi^{-1} F) \\ &= -1/r \mathrm{tr}(\log(\Pi^{-1} F)) \end{aligned}$$

whereas the GTR correction equals

$$-\mathrm{tr}(\Pi \log(\Pi^{-1} F))$$

Both estimators are equivalent if $\pi_i = 1/r$ for all r.

4 ML estimates

The general inversion formulae implicitly estimate the parameter models separately for each pair of sequences. For ML estimates we fix the model parameters first, then $d = -\mu \cdot \mathrm{tr}(\Pi Q) \cdot t$ for each pair of sequences.

In general, the maximum likelihood estimate for t is the t that maximizes

$$L(t) = \prod_{ij} (X(t)_{ij})^{F_{ij}}$$

or equivalently, that minimizes

$$-\log L(t) = -\sum_{ij} F_{ij} \log(X(t)_{ij})$$

With different models, various types of mutations are equivalent. For example, under the K2P model, we count only the number of transitions, transversions and matches. The probability for a a transition equals

$$P_{trans}(t) = X(t)_{13} + X(t)_{24} + X(t)_{31} + X(t)_{42} = \tfrac{1}{4} + \tfrac{1}{4} e^{-\mu t} - \tfrac{1}{2} e^{-\frac{1}{2}\mu(\kappa+1)t}$$

the probability of a match is

$$P_{mat}(t) = X(t)_{11} + X(t)_{22} + X(t)_{33} + X(t)_{44} = \tfrac{1}{4} + \tfrac{1}{4} e^{-\mu t} + \tfrac{1}{2} e^{-\frac{1}{2}\mu(\kappa+1)t}$$

and the probabiity of a transversion equals

$$P_{tranv}(t) = \tfrac{1}{2} - \tfrac{1}{2}e^{-\mu t}.$$

Let $\hat{p}_{trans}$, $\hat{p}_{mat}$ and $\hat{p}_{tranv}$ be the observed proportions of transitions, matches, and transversions. The maximum likelihood estimate under the K2P model equals μt, where t maximizes

$$\hat{p}_{trans}\log(P_{trans}(t)) + \hat{p}_{mat}\log(P_{mat}(t)) + \hat{p}_{tranv}\log(P_{tranv}(t))$$

5 Coping with different rates

So far we have assumed that all of the sites evolved at the same rate. There are a number of models, and associated distance corrections, where this assumption is relaxed.

5.1 Review

So far, we have studied Markov process models of sequence evolution. Given two sequences separated by a branch (or path) of length t the probability of a site being in state j in the second sequence *given* that the site has state i initially equals $P(t)_{ij}$, where

$$P(t) = e^{Q\mu t}$$

Let $\pi_1, \ldots, \pi_r$ denote the stationary distribution, and let Π be the diagonal matrix with π_i's down the diagonal. The divergence matrix is

$$X(t) = \Pi P(t)$$

So $X(t)_{ij}$ is the probability of observing state i in the first sequence and j in the other. The process is time reversible iff $X(t)$ is symmetric. The mutation rate is the expected number of mutations per unit time, and equals $-\mu\mathrm{tr}(\Pi Q)$. If F_{ij} denotes the proportion of sites with an i in sequence one and a j in sequence two then the general distance correction is

$$-\mathrm{tr}(\Pi\log(\Pi^{-1}F)).$$

5.2 Invariant sites model

Suppose that a proportion ϕ of the sites are not free to vary at all. These sites will give 'false matches' and should be removed from the analysis. Let $\hat{P}$ be the transition matrix for this modified process. Then for a randomly chosen site,

$$\hat{P}(t)_{ij} = (1-\phi)P(t)_{ij}$$

and

$$\hat{P}(t)_{ii} = \phi + (1-\phi)P(t)_{ii}.$$

Hence

$$P(t) = \phi I + (1-\phi)e^{Q\mu t}$$

and the divergence matrix then becomes

$$X_{inv}(t) = \Pi(\phi I + (1-\phi)e^{Q\mu t}).$$

Inverting this formula gives an estimator

$$-\mathrm{tr}(Q)\mu t = -\mathrm{tr}(\Pi\log(\frac{1}{1-\phi}(\Pi^{-1}F - \phi I))).$$

See Swofford et al. (1996) and references therein for a discussion of applying the invariant sites model and the estimation of ϕ.

5.3 Gamma distribution

The *gamma distribution* with parameters $\alpha, \beta > 0$ has probability density function

$$f(x|\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, \qquad 0 < x < \infty$$

and mean $\alpha\beta$. Note

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt.$$

Gamma distributions provide a mathematically convenient model for the distribution of a rate at a particular site. All the sites are still i.i.d., we have simply allowed the rate at a site to be a random variable. Typically, $\alpha = 1/\beta$, so the rate distribution has mean 1.

Once we have fixed a rate, the conditional probabilities are given by

$$P(t|x) = e^{Q\mu x t}$$

so

$$P(t) = \int_0^\infty P(t|x) f(x|\alpha,\beta) dx$$

where the integral operates on each element of the matrix separately. We will see later how to evaluate this integral.

5.4 Covarion and variable rate models

Suppose that each site in the sequence has an associated 'on-off' switch, that changes position according to a Markov chain. When the switch is off, the site cannot change (e.g. it is under functional constraint), when the switch is on, the site can change. In fact, we can extend the model to have switches with not just 'on-off' but different speeds.

The transition probabilities between two sequences depend only on the total amount of time that the switch for that site is on (or off), or, equivalently, the integral of the rates for that site. This 'integrated rate' will have some distribution $f(x)$ that we can determine from the model of rate evolution. Once we have computed $f(x)$, we can calculate $P(t)$, and the divergence matrix, using

$$P(t) = \int_0^\infty P(t|x) f(x) dx$$

5.5 General case

We now develop the tools required to compute $P(t)$, and the distance corrections, for a general model of rate evolution. Things get a bit technical, but the goal is worthwhile. At the end we will have an analytical formula for $P(t)$ and distance corrections for almost any reasonable model of site by site rate evolution.

First some statistics. The *moment generating function* (m.g.f.) for a random variable X with probability density function f is defined

$$M_f(z) = \mathbb{E}[e^{Xz}]$$

The m.g.f.'s for standard distributions can all be found in a decent statistics textbook. The m.g.f. for the gamma distribution is $\left(\frac{1}{1-z\beta}\right)^{\alpha}$.

We can extend the definition of moment generating functions to matrices:

Lemma 2. *Let A be a square matrix (with real eigenvalues). Suppose that f has m.g.f. $g(z)$. Then*

$$\int_0^\infty e^{Ax} f(x)dx = g(A)$$

where the integral is evaluated entry by entry.

Proof. First diagonalize A, giving $A = VDV^{-1}$ where D is a diagonal matrix with eigenvalues $\lambda_1, \ldots, \lambda_n$ on the diagonal. Then $e^{Ax} = Ve^{Dx}V^{-1}$. For each i, j there are coefficients $a_1, a_2, \ldots, a_n$ such that

$$e_{ij}^{Ax} = \sum_{k=1}^{n} a_k e^{\lambda_k x}.$$

Hence

$$\begin{aligned}\int_0^\infty e_{ij}^{Ax} f(x)dx &= \sum_{k=1}^{n} a_k \int_0^\infty e_{ij}^{\lambda_k x} f(x)dx \\ &= \sum_{k=1}^{n} a_k g(\lambda_k)\end{aligned}$$

and

$$\begin{aligned}\int_0^\infty e^{Ax} f(x)dx &= Vg(D)V^{-1} \\ &= g(A).\end{aligned}$$

We can now derive the general distance correction

Theorem 2. *Suppose that*

$$P(t) = \int_0^\infty e^{Q\mu x} f(x)dx$$

where x is a random variable with density function $f(x)$ and moment generating function $g(z)$ (both dependent on t). Then

$$X(t) = \Pi g(Q\mu t)$$

and

$$-\mathrm{tr}(\Pi Q)\mu t \approx -\mathrm{tr}(\Pi g^{-1}(\Pi^{-1}F))$$

as $F \to X(t)$.

5.6 Example: gamma distance correction

The moment generating function for the gamma distribution with parameters α, β is

$$g(z) = \left(\frac{1}{1-z\beta}\right)^{\alpha}.$$

The inverse $g^{-1}(y)$ is

$$g^{-1}(y) = (1 - y^{\frac{1}{\alpha}})/\beta$$

so the general correction is

$$-\mathrm{tr}(\Pi Q)\mu t = -\frac{1}{\beta}\mathrm{tr}(\Pi(I - (\Pi^{-1}F)^{\frac{1}{\alpha}})).$$

(See file DistCorrect.txt for the maple code needed to derive K2P+gamma correction)

5.7 Example: Markov rate models and the covarion model

Suppose that the rate x evolves according to a Markov process with rate matrix $\hat{Q}$, rate classes $r_1, r_2, \ldots, r_s$, and initial distribution $\hat{\pi_1}, \ldots, \hat{\pi_s}$. If $\tau_i(t)$ denotes the time spent at rate i (a random variable), then the integrated rate equals

$$\tau(t) = \sum_{i=1}^{s} r_i \tau_i(t).$$

Let R be the diagonal matrix with r_i's on the diagonal, and let $\hat{\Pi}$ be the diagonal matrix with π_i's on the diagonal. Let u denote the matrix of ones. Darroch and Morris (1968) prove that $\tau(t)$ has moment generating function

$$\mathbb{E}[e^{-z\tau(t)}] = u^T \Pi e^{t(\hat{Q} - zR)} u$$

From this we can use Theorem 2 to derive a general distance correction.

6 Bounding the sampling error in an estimated distance

We now switch track a bit to start looking at the construction of phylogenetic trees. Rather than survey all the methods (there are many excellent references for this) we instead concentrate on some analytical results bounding the probability of obtaining the correct tree.

6.1 Bounding $\mathbb{P}[|\hat{\boldsymbol{p}} - \boldsymbol{p}| \leq \boldsymbol{\delta}]$

Once again consider a two leaf tree, where the leaves are separated by a branch of length t. Suppose that sequences of length n are evolved on this tree, according to the Cavender Farris model with rate 1.

For a particular site, the probability that the two sequences have different states is

$$p = \tfrac{1}{2} - \tfrac{1}{2}e^{-2t} < \tfrac{1}{2} \tag{4}$$

Each site is independent, and has the same probability of change. The number $n\hat{p}$ of observed sites that have changed therefore has a binomial distribution with probability p.

We can bound the probability that $\hat{p}$ is different from p using a result of Chernoff.

Theorem 3. *Let n be the number of sites. For any $\delta > 0$ we have*

$$\mathbb{P}[\hat{p} - p \geq \delta] \leq e^{-2n\delta^2}$$
$$\mathbb{P}[p - \hat{p} \geq \delta] \leq e^{-2n\delta^2}.$$

6.2 Bounding $\mathbb{P}[|d - \hat{d}| < \epsilon]$

We now determine which δ we need in order to force $|d - \hat{d}| < \epsilon$.

$$\begin{aligned}\hat{d} - d &= -\tfrac{1}{2}\log(1-2\hat{p}) - \tfrac{-1}{2}\log(1-2p) \\ &= -\tfrac{1}{2}\log\left(\frac{1-2\hat{p}}{1-2p}\right) \\ &= -\tfrac{1}{2}\log\left(\frac{1-2p-2(\hat{p}-p)}{1-2p}\right) \\ &= -\tfrac{1}{2}\log(1-\frac{2(\hat{p}-p)}{1-2p})\end{aligned}$$

It follows (after some manipulation) that $\hat{d} - d < \epsilon$ if and only if

$$(\hat{p} - p) < (\tfrac{1}{2} - p)(1 - e^{-2\epsilon})$$

and that $d - \hat{d} < \epsilon$ if and only if

$$(p - \hat{p}) < (\tfrac{1}{2} - p)(e^{2\epsilon} - 1)$$

Note that $(1 - e^{-2\epsilon}) - (e^{2\epsilon} - 1) = 2 - \cosh(2\epsilon) < 0$ for all ϵ, so $(\frac{1}{2} - p)(1 - e^{-2\epsilon}) < (\frac{1}{2} - p)(e^{2\epsilon} - 1)$.

Lemma 3. *If* $|p - \hat{p})| < (\frac{1}{2} - p)(1 - e^{-2\epsilon})$ *then* $|d - \hat{d}| < \epsilon$. *Hence*

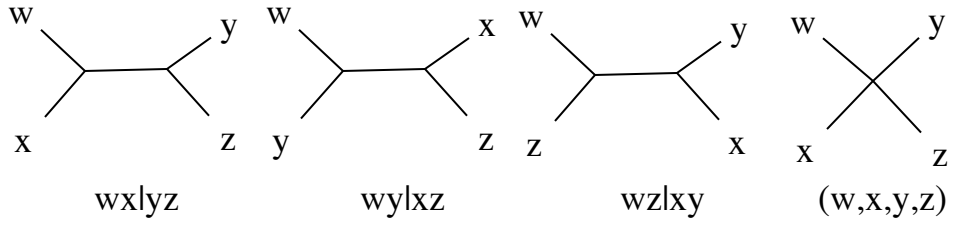
$$\mathbb{P}[|d - \hat{d}| < \epsilon] \leq 2\exp(-2n(\tfrac{1}{2} - p)^2(1 - e^{-2\epsilon})^2)$$

7 Reconstructing the four leaf tree

7.1 Four point method

The question we are working towards is 'what is the probability of reconstructing the correct tree'. The answer will, of course, depend on the tree, the branch lengths, the model of evolution, and the method used. We will start with a very simple tree, a very simple model, and a very simple method.

There are four different trees with four leaves.



We can easily characterise the distance matrices that come from each of these trees. (The distance between two leaves in a tree is the sum of the lengths of the edges along the path between them). Suppose that D is an arbitrary distance matrix. Define the four summations

$$\begin{aligned} S_{wx|yz} &= D_{wx} + D_{yz} \\ S_{wy|xz} &= D_{wy} + D_{xz} \\ S_{wz|xy} &= D_{wz} + D_{xy}. \end{aligned}$$

Then D is a distance matrix on $wx|yz$ if and only if

$$S_{wx|yz} < S_{wy|xz} = S_{wz|xy},$$

D is a distance matrix on $wy|xz$ if and only if

$$S_{wy|xz} < S_{wx|yz} = S_{wz|xy},$$

D is a distance matrix on $wz|xy$ if and only if

$$S_{wz|xy} < S_{wx|yz} = S_{wy|xz},$$

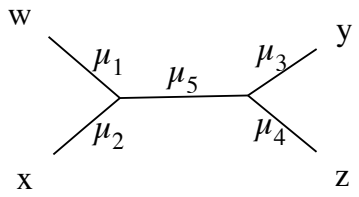
D is a distance matrix on $wz|xy$ if and only if

$$S_{wz|xy} = S_{wx|yz} = S_{wy|xz}$$

This gives a consistent method for an arbitrary distance matrix (on four sequences): choose the tree $wx|yz$ for which $S_{wx|yz}$ is minimum. If there is a tie, choose (w, x, y, z). This is called the *four-point method*.

7.2 When is the four point method correct?

Now suppose that $D = \hat{d}$, the estimated distance. Furthermore, suppose that we evolved the sequences down $wx|yz$, where the branch lengths are



wx|yz

For each i, j let $\delta_{ij} = \hat{d}_{ij} - d_{ij}$, where d is the true tree distance. Then from looking at the tree we have

$$\begin{aligned} S_{wy|xz} - S_{wx|yz} &= \hat{d}_{wy} + \hat{d}_{xz} - \hat{d}_{wx} - \hat{d}_{yz} \\ &= (\delta_{wy} + \delta_{xz} - \delta_{wx} - \delta_{yz}) + (d_{wy} + d_{xz} - d_{wx} - d_{yz}) \\ &= (\delta_{wy} + \delta_{xz} - \delta_{wx} - \delta_{yz}) + 2\mu_5 \end{aligned}$$

which is definitely positive when

$$|\delta_{wy}| + |\delta_{xz}| + |\delta_{wx}| + |\delta_{yz}| < 2\mu_5.$$

The same holds for $S_{wz|xy} - S_{wx|yz}$.

Thus if $|\delta_{ij}| < \frac{1}{2}\mu_5$ for all i, j we know that the four point method will return the correct tree $wx|yz$.

Let p_{ij} denote the probability of a change between sequences i and j.

$$\begin{aligned}
\mathbb{P}[\text{incorrect tree}] &\leq \mathbb{P}[\delta_{ij} \geq \tfrac{1}{2}\mu_5 \text{ for some } i, j\ ] \\
&\leq \mathbb{P}[\delta_{wx} \geq \tfrac{1}{2}\mu_5] + \mathbb{P}[\delta_{wy} \geq \tfrac{1}{2}\mu_5] + \cdots + \mathbb{P}[\delta_{yz} \geq \tfrac{1}{2}\mu_5] \\
&\leq \sum_{i<j} \exp(-2k(\tfrac{1}{2} - p_{ij})^2(1 - e^{-\mu_5})^2) \\
&\leq 12\exp(-8k(\tfrac{1}{2} - \Delta)^2 z^2)
\end{aligned}$$

where Δ is the maximum of p_{ij} and $z = \frac{1}{2}(1 - 2e^{-\mu_5})$ is the probability of a mutation along the central edge.

We now have

Theorem 4. *The probability that the four point condition will obtain the best tree is bounded below*

$$1 - 12\exp(-8k(\tfrac{1}{2} - \Delta)^2 z^2)$$

7.3 Extension to N taxa

We can show that if you correctly reconstruct the restriction of a tree to all subsets of four taxa than you have correctly reconstructed the tree.

Hence for N taxa we can think of the simple method that uses the four point method to construct every four taxa "subtree" then checks to see if these can be combined into a single tree. Let μ be the expected number of mutations along the shortest edge of the tree. We know that this method will work if all of the estimated distances are within $\mu/2$ of the true distances. Let $z = \frac{1}{2}(1 - 2e^{-\mu})$, the probability of a mutation along the central edge. From Lemma 3 , the probability of getting the wrong tree is bounded above by

$$N^2\exp(-n(\tfrac{1}{2} - \Delta)^2 z^2)$$

Rearranging we obtain

Theorem 5. *If*

$$n > \frac{-\log(\epsilon) + 2\log(N)}{4(1 - 2\Delta)^2 z^2}$$

then

$$\mathbb{P}[\textit{reconstructing the wrong tree}] < \epsilon]$$

8 Further reading

An excellent general introduction to models in phylogenetics appears in the chapter

Swofford, D., Olsen, G., Waddell, and P., Hillis, D. (1996) Phylogenetic inference. *in* Hillis, D., Moritz, C., and Mable, B. *Molecular Systematics* (2nd edition)

A concise introduction to the models used (suitable for mathematicians) and material on general corrections is

Rodríguez, F., Marín, J.L., and Medina, J.R. (1990) The general stochastic model of nucleotide subsitution. *Journal of Theoretical Biology* 142, 485–501.

For Markov chains and markov processes I still depend on the two classic texts:

Feller, W. *An introduction to probability theory and its applications* Vol. I. Third edition John Wiley and Sons, Inc., New York-London-Sydney.

and

Karlin, S. and Taylor, H. M. *A first course in stochastic processes* Second edition. Academic Press, New York-London, 1975.

A covarion (two state) model for rate evolution was developed and analysed by

Tuffley, C. and Steel, M.A. (1997) Modelling the covarion hypothesis of nucleotide substitution, Mathematical Biosciences 147: 63-91.

A Markov rate evolution model with an arbitrary number of rate classes was developed by

Galtier, N. (2001) Maximum Likelihood phylogenetics analysis under a covarion-type model. Molecular Biology and Evolution 18:866–873.

The analysis on bounds for the error on distance corrections is based on

Erds, P.L., Steel, M.A., Szkely, L.A. and Warnow, T. (1999) A few logs suffice to build (almost) all trees (Part 1). Random Structures and Algorithms 14(2): 153-184.