Model-based Distances

Recall that 1 weakness of the Hamming distance = proportion of sites that differ is that it fails to account for back substitutions or multiple substitutions on an edge

and unless sequences are very closely related, the Hamming distance tends to under estimate the amount of evolutionary distance between 2 taxa.

To address this, we use our models JC, K2P, K3ST, GTR, GM to "correct" the Hamming distance and to account for unseen changes.

Example: Jukes-Cantor model and Jukes-Cantor distance

Parameters: single edge thought of as path from taxon a to b in tree

$\vec{p}_0 = (.25 \;\; .25 \;\; .25 \;\; .25)$

$Q = \begin{pmatrix} -\alpha & \alpha/3 & \alpha/3 & \alpha/3 \\ & -\alpha & & \\ & & -\alpha & \\ & & & -\alpha \end{pmatrix}$, branch length t, and Markov matrix $M(t) =$

$M = \begin{pmatrix} 1-a & a/3 & a/3 & a/3 \\ & 1-a & & \\ & & 1-a & \\ & & & 1-a \end{pmatrix}$ with $a(t) = \frac{3}{4}\left(1 - \frac{4}{3}\alpha t\right)$

Assuming the JC model, the expected pattern frequency array is

$P = \text{diag}([.25 \;\; .25 \;\; .25 \;\; .25])\, M(t) = \begin{pmatrix} \frac{1}{4}(1-a) & a/12 & a/12 & a/12 \\ a/12 & \frac{1}{4}(1-a) & a/12 & a/12 \\ a/12 & & \ddots & \\ a/12 & & & \end{pmatrix}$ with a as above

[No data here. Yet.]

If we interpret α = rate in units $\frac{\#\text{ substitutions}}{\text{time}}$, then

$$\alpha t = \left(\frac{\#\text{ substitutions per site}}{\text{time } t}\right)(\text{time } t)$$

(model at single site)

$$\boxed{\alpha t = \#\text{ of substitutions over time } t}$$

including those hidden ones

This is (will be) the Jukes Cantor distance once we solve for it:

$$a(t) = a = \frac{3}{4}\left(1 - e^{-\frac{4}{3}\alpha t}\right)$$

$$\Rightarrow \boxed{\alpha t = -\frac{3}{4}\ln\left(1 - \frac{4}{3}a\right)} \quad = \#\text{ of subst. per site over elapsed time } t$$

Since we don't have a in hand, we must estimate it from data,

i.e from the empirical pattern freq. array

In theory, S_0 and S_1 disagree with probability $12\left(\frac{a}{12}\right) = a$

(Sum of off-diagonal entries)

Thus, we estimate a with $\hat{a} = \frac{\#\text{ of sites with non-constant pattern}}{\#\text{ of sites } n}$

$$= \text{Hamming distance!}$$

Defn: The Jukes-Cantor distance $d_{JC}(S_0, S_1)$ between aligned sequences is

$$d_{JC} = -\frac{3}{4}\ln\left(1 - \frac{4}{3}\hat{a}\right) \qquad \hat{a} = d_{\text{Hamming}}(S_0, S_1)$$

Ex.

S_0: A A A [C] G G C [A] [T] G

S_1: A A A [T] G G C [T] [A] G

$d_{Hamming}(S_0, S_1) = 3/10 = .3$

$d_{JC}(S_0, S_1) = -\frac{3}{4} \ln\left(1 - \frac{4}{3}(.3)\right) \approx .38$

Several comments:

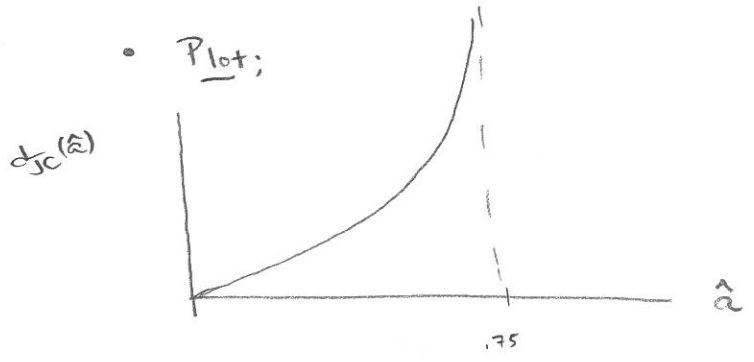
- $d_{JC} = .38 > .3 = d_{Hamming}$ to account for hidden mutations
- $d_{JC}(S_0, S_1) = -\frac{3}{4} \ln\left(1 - \frac{4}{3}\hat{a}\right) \Rightarrow 0 \le \hat{a} < \frac{3}{4}$ for the log to make sense

This makes sense: if 2 sequences are generated at random using the root distribution (1/4 1/4 1/4 1/4) then S_0 and S_1 will agree roughly 1/4 of the time and disagree in roughly $\hat{a} = 3/4$ of the sites $d(S_0, S_1) >>> 0$

S_0: _ _ _

S_1: _ _ _

- Plot:



Thus as $\hat{a}$ gets close to .75, $d_{JC}(S_0, S_1) \to \infty$, huge distances

"Saturated sequences"

informally, you can not differentiate them from 2 randomly selected sequences

- Since JC is time-reversible, on a tree

or more generally we can consider paths in trees

Similar methods can be used to derive distance formulas for K2P, K3ST, GTR.

$$d_{K2P}(S_1,S_2) = -\frac{1}{2}\ln\left(1-2\hat{b}-\hat{e}\right) - \frac{1}{4}\ln\left(1-2\hat{e}\right)$$

$\hat{b}$ = proportion of observed transitions

$\hat{e}$ = proportion of observed transv.

$$d_{K3}(S_1,S_2) = -\frac{1}{4}\left(\ln(1-2\hat{b}-2\hat{c}) + \ln(1-2\hat{b}-2\hat{d}) + \ln(1-2\hat{e}-2\hat{d})\right)$$

$\hat{b}$, $\hat{c}$, $\hat{d}$ are best estimates for $M_{K3P} = \begin{pmatrix} * & b & c & d \\ & & & \end{pmatrix}$

OR $P = \mathrm{diag}([1/4\ \ 1/4\ \ 1/4\ \ 1/4])\, M_{K3P}$

If $\hat{e}=\hat{d}$, then this simplifies to K2P.

There is also a more general GTR distance ⇝ see book. Requires

1) normalization 2) Knowledge of Trace of a matrix.

Why normalize?

Eg. JC model

Probability (Model)

$$d_{JC}(S_1,S_2) = -\frac{3}{4}\ln\left(1-\frac{4}{3}a\right)$$

$$\| \quad \alpha t$$

Estimate from data.

$$d_{JC}(S_0,S_1) = -\frac{3}{4}\ln\left(1-\frac{4}{3}\hat{a}\right)$$

intertwined

$$\alpha t = 2\alpha\left(\frac{t}{2}\right) = 4\alpha\left(\frac{t}{4}\right) = c\alpha\left(\frac{t}{c}\right)$$

We can compute the product αt,

but neither α or t separately

However, α represents a total mutation rate in JC

$$\frac{1}{4}\alpha + \frac{1}{4}\alpha + \frac{1}{4}\alpha + \frac{1}{4}\alpha = \alpha$$

rate leaving state A; rate leaving state G, C, T

Normalize so that $\alpha = 1$

$$Q = \begin{pmatrix} -1 & 1/3 & 1/3 & 1/3 \end{pmatrix} \text{ etc.}$$

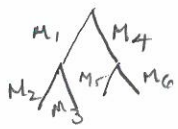
$$\alpha t = l(t) = \frac{\#\text{ of subst.}}{t} \cdot t$$

$$= t \cdot \#\text{ of substitutions}$$

i.e. t measures the expected number of substitutions over the elapsed time.

MATLAB eg's

The log-det distance: Essentially a distance for the general Markov model.



First the definition:

Let F_{ab} be the expected frequency array (or use $\hat{F}_{ab}$ if generated from data. Let $\vec{p}_a, \vec{p}_b$ be the base distribution at a, b respectively and g_a, g_b the product of entries in $\vec{p}_a, \vec{p}_b$, then the LOG-DET distance is

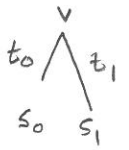
$$d_{\text{log-det}}(a,b) = -\frac{1}{4} \ln \left(|\det F_{ab}| - \frac{1}{2} \ln g_a g_b \right) = -\frac{1}{4}\left(\ln \det(F_{ab}) - \frac{1}{2}\ln g_a g_b\right)$$

On data, use $\hat{F}_{ab}$

Exercise for math students: You can use either the expected frequency array F_{ab} or NF_{ab} → expected count data and get the same answer.

log-det has several good properties

- very general model: $\in M$
- additive



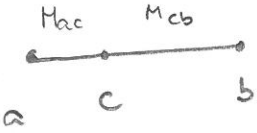
$$d(s_0,s_1) = d(s_0,v) + d(v,s_1)$$
$$t_0 + t_1$$

- used in proofs ...

Its units are a bit mysterious, but somehow quantifies the amount of mutation between s_0 and s_1 since it's additive.

To understand why $d_{LD}(a,b)$ is additive, ignore $\frac{-1}{4}$ and $-\ln(g_a g_b)$

i.e. $d_{LD}(a,b) = \frac{-1}{4}\ln\left(|\det(\bar{F}_{ab})| - \frac{1}{2}\overset{\ln}{\downarrow} g_a g_b\right)$



$M_{ac} = M_{ab} M_{bc}$

Ignore $-1/4$:

$$d_{LD}(a,b) = \ln\left(\det(\mathrm{diag}(p_a))\det(M_{ac}M_{cb}) - \tfrac{1}{2}\ln g_a g_b\right)$$

$$= \ln(\sqrt{g_a}) - \ln(\sqrt{g_b}) + \ln(\det(M_{ac})) + \ln(\det(M_{cb}))$$

$$= \underbrace{\ln(\sqrt{g_a}) - \ln(\sqrt{g_c}) + \ln(\det(M_{ac}))}_{d_{LD}(a,c)} + \underbrace{\ln(\sqrt{g_c}) - \ln(\sqrt{g_b}) + \ln(\det(M_{cb}))}_{d_{LD}(c,b)}$$

□

Parting thoughts:

Model-based distances ... good bad

etc.