Week 8: Paired sites tests, gene frequencies, continuous characters

Genome 570

May, 2008
An example – two trees

Tree I

Mouse

Bovine

Gibbon

Orang

Gorilla

Chimp

Human

Tree II

Mouse

Bovine

Gibbon

Orang

Gorilla

Chimp

Human
The differences of log likelihoods

<table>
<thead>
<tr>
<th>Tree</th>
<th>site</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>231</th>
<th>232</th>
<th>ln L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diff</td>
<td></td>
<td>+0.012</td>
<td>+0.111</td>
<td>+0.013</td>
<td>+0.015</td>
<td>+0.010</td>
<td>-0.431</td>
<td></td>
<td>+0.012</td>
<td>+0.010</td>
</tr>
</tbody>
</table>

Week 8: Paired sites tests, gene frequencies, continuous characters – p.3/44
The histogram of differences of log-likelihoods

Difference in log likelihood at site
Paired sites tests

- Winning sites test (Prager and Wilson, 1988). Do a sign test on the signs of the differences.

- z test (me, 1993 in PHYLIP documentation). Assume differences are normal, do z test of whether mean (hence sum) difference is significant.

- t test. Swofford et. al., 1996: do a t test (paired)

- Wilcoxon ranked sums test (Templeton, 1983).

- RELL test (Kishino and Hasegawa, 1989 per my suggestion). Bootstrap resample sites, get distribution of difference of totals.
In our example ...

- Winning sites test. 160 of 232 sites favor tree I. $P < 3.279 \times 10^{-9}$

- $z$ test. Difference of log-likelihood totals is $0.948104$ standard deviations from 0, $P = 0.343077$. Not significant.

- $t$ test. Same as $z$ test for this large a number of sites.

- Wilcoxon ranked sums test. Rank sum is $4.82805$ standard deviations below its expected value, $P = 0.000001378765$

- RELL test. 8,326 out of 10,000 samples have a positive sum, $P = 0.3348$ (two-sided)
The Shimodaira-Hasegawa test

- Starts with a set of user-specified trees
- Gets the sitewise log-likelihoods
- adjusts each trees’ log-likelihoods to add up to same value
- then resamples columns (sites) from these
- asks how often a tree will get more than X worse then the best
- for each X (the log-likelihood difference between one tree and the best one) can get a P value.
Brownian motion along a tree

Week 8: Paired sites tests, gene frequencies, continuous characters – p.8/44
Covariances of species on the tree

\[
\begin{bmatrix}
    v_1 + v_8 + v_9 & v_8 + v_9 & v_9 & 0 & 0 & 0 & 0 \\
    v_8 + v_9 & v_2 + v_8 + v_9 & v_9 & 0 & 0 & 0 & 0 \\
    v_9 & v_9 & v_3 + v_9 & 0 & 0 & 0 & 0 \\
    0 & 0 & 0 & v_4 + v_{12} & v_{12} & v_{12} & v_{12} \\
    0 & 0 & 0 & v_{12} & v_5 + v_{11} + v_{12} & v_{11} + v_{12} & v_{11} + v_{12} \\
    0 & 0 & 0 & v_{12} & v_{11} + v_{12} & v_6 + v_{10} + v_{11} + v_{12} & v_{10} + v_{11} + v_{12} \\
    0 & 0 & 0 & v_{12} & v_{11} + v_{12} & v_{10} + v_{11} + v_{12} & v_7 + v_{10} + v_{11} + v_{12}
\end{bmatrix}
\]
Covariances are of form

\[
\begin{pmatrix}
a & b & c & 0 & 0 & 0 & 0 \\
b & d & c & 0 & 0 & 0 & 0 \\
c & c & e & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & f & g & g & g \\
0 & 0 & 0 & g & h & i & i \\
0 & 0 & 0 & g & i & j & k \\
0 & 0 & 0 & g & i & k & l \\
\end{pmatrix}
\]
An outcome of Brownian motion on a 5-species tree
An outcome of Brownian motion on a 5-species tree
An outcome of Brownian motion on a 5-species tree
An outcome of Brownian motion on a 5-species tree
Likelihood under Brownian motion with two species

\[ f(x; \mu, \sigma^2) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left( -\frac{(x - \mu)^2}{2\sigma^2} \right) \]

\[ L = \prod_{i=1}^{p} \frac{1}{(2\pi)^{\frac{1}{2}} \sigma_i \sqrt{v_1 v_2}} \exp \left( -\frac{1}{2\sigma_i^2} \left[ \frac{(x_{1i} - x_{0i})^2}{v_1} + \frac{(x_{2i} - x_{0i})^2}{v_2} \right] \right) \]
Minimizing for each character $i$

\[ Q = \frac{(x_{1i} - x_{0i})^2}{v_1} + \frac{(x_{2i} - x_{0i})^2}{v_2} \]

so:

\[ \frac{dQ}{dx_{0i}} = -2\frac{(x_{1i} - x_{0i})}{v_1} - 2\frac{(x_{2i} - x_{0i})}{v_2} = 0 \]

and then:

\[ \hat{x}_{0i} = \frac{\frac{1}{v_1}x_{1i} + \frac{1}{v_2}x_{2i}}{\frac{1}{v_1} + \frac{1}{v_2}} \]

The result is that

\[ Q = \frac{(x_{1i} - x_{2i})^2}{v_1 + v_2} \]
Likelihood after estimating initial coordinates

Substituting this in we end up with

\[
L = \frac{1}{(2\pi)^p (\nu_1 \nu_2)^{\frac{1}{2}p} \prod_i \sigma_i^2} \exp \left( -\frac{1}{2} \sum_{i=1}^{p} \frac{(x_{1i} - x_{2i})^2}{\sigma_i^2 (v_1 + v_2)} \right)
\]

and this finally turns into:

\[
\ln L = -p \ln (2\pi) - 2 \sum_{i=1}^{p} \ln \sigma_i - \frac{1}{2} p \ln (\nu_1 \nu_2) - \frac{1}{2} \sum_{i=1}^{p} \frac{(x_{1i} - x_{2i})^2}{\sigma_i^2 (v_1 + v_2)}
\]
If there is a clock ...

If instead we constrain $v_1 = v_2$ because assume a clock:

$$\ln L = K' - p \ln(v_1 + v_2) - \frac{1}{2} \frac{D^2}{(v_1 + v_2)}$$

which leads to

$$\hat{v}_1 = \hat{v}_2 = \frac{D^2}{4p}$$

(which is half as big as it should be!)
Using only differences between populations (REML)

\[
L = \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi \sigma_i} \sqrt{v_1 + v_2}} \exp \left( -\frac{1}{2\sigma_i^2} \frac{(x_{1i} - x_{2i})^2}{v_1 + v_2} \right)
\]

\[
\ln L = K - \frac{p}{2} \ln (v_1 + v_2) + \frac{1}{2 (v_1 + v_2)} \sum_{i=1}^{n} \left( \frac{x_{i1} - x_{i2}}{\sigma_i} \right)^2
\]
Likelihood with two species using REML

\[
\ln L = K - \frac{p}{2} \ln (v_1 + v_2) + \frac{D^2}{2(v_1 + v_2)}
\]

\[
\ln L = K - \frac{p}{2} \ln (v_T) + \frac{D^2}{2v_T}
\]

\[
\hat{v}_T = \frac{D^2}{p}
\]
“Pruning” a tree in the Brownian motion case

\[ \delta = \frac{v_1 v_2}{v_1 + v_2} \]

\[ x_{12} = \frac{v_2 x_1 + v_1 x_2}{v_1 + v_2} \]
What about quantitative characters?

For neutral mutation and genetic drift, can show that for a quantitative character with additive genetic variance $V_A$ and population size $N$ the genetic (additive) value of the population mean is:

$$\text{Var}(\Delta \bar{g}) = \frac{V_A}{N}$$

If mutation and drift are at equilibrium:

$$E\left[V_A^{(t+1)}\right] = V_A^{(t)} \left(1 - \frac{1}{2N}\right) + V_M$$
In neutral traits additive genetic variance rules

so that

$$E[V_A] = 2NV_M$$

whereby

$$Var[\Delta \bar{g}] = (2NV_M) / N = 2V_M$$

an analogue of Kimura’s result for neutral mutation. There is a precise analogue of this for multiple characters.

Thus to transform characters to independent Brownian motions of equal evolutionary variance, we could use the additive genetic variance $V_A$. 
With selection ... life is harder

There is the quantitative genetics formula of Wright and Fisher (1920’s)

\[ \Delta z = h^2 S \]

and Russ Lande’s (1976) recasting of that in terms of slopes of mean fitness surfaces:

\[ S = V_P \frac{d \log (\bar{w})}{dx} \]

\[ \Delta z = (V_A/V_P) V_P \frac{d \log (\bar{w})}{dx} = V_A \frac{d \log (\bar{w})}{dx} \]
Selection towards an optimum

If fitness as a function of phenotype is:

\[ w(x) = \exp\left[-\frac{(x - p)^2}{2V_s}\right] \]

Then the change of mean phenotype “chases” the optimum:

\[ m' - m = \frac{V_A}{V_s + V_P} (p - m) \]
A character changing by “chasing” an adaptive peak

The course of change of the population mean is expected to be somewhat smoother than the changes of the peak of the fitness surface.
Sources of evolutionary correlation among characters

Variation (and covariation) in change of characters occurs for two reasons:

1. Genetic drift, with the covariances being proportional to the additive genetic covariances
2. Selection, with the covariances being affected by both the additive genetic covariances and the covariation of the selection pressures.
A simple example of selective covariance

covariation due not to genetic correlation but to covariation of the selection pressure

These are Bergmann's, Allen's and Glogler's Rules
They are presumably not the result of genetic correlations but result from patterns of selection

a simple example:
(temperate) (arctic) (temperate) (arctic) (temperate)

size

color

limb length

A simulated example with two characters

After 100 generations:

Genetic covariances are negative, but the wanderings of the adaptive peak in the two characters is positively correlated.
A simulated example with two characters

After 1000 generations:

Genetic covariances are negative, but the wanderings of the adaptive peak in the two characters is positively correlated.
A simulated example with two characters

After 10,000 generations:

Genetic covariances are negative, but the wanderings of the adaptive peak in the two characters is positively correlated.
Correcting for correlations among characters

Can we transform the set of characters to remove their correlations and thus end up with independent Brownian motions of equal variance?

- We might hope to infer additive genetic covariances by doing quantitative genetics breeding experiments to infer them from covariances among relatives.
- There is little or no hope of inferring “selective correlations” without a complete understanding of the functional ecology.
A tree with punctuated equilibrium

Week 8: Paired sites tests, gene frequencies, continuous characters – p.33/44
The punctuated tree when we sample 10 species
Correlation of states in a discrete-state model

Week 8: Paired sites tests, gene frequencies, continuous characters – p.35/44
A simple case two show effects of phylogeny
Two uncorrelated characters evolving on that tree
Identifying the two clades
A tree on which we are to observe two characters
Contrasts on that tree

Variance proportional to

\[
\begin{align*}
    y_1 &= x_a - x_b & 0.4 \\
    y_2 &= \frac{1}{4} x_a + \frac{3}{4} x_b - x_c & 0.975 \\
    y_3 &= x_d - x_e & 0.2 \\
    y_4 &= \frac{1}{6} x_a + \frac{1}{2} x_b + \frac{1}{3} x_c - \frac{1}{2} x_d - \frac{1}{2} x_e & 1.11666
\end{align*}
\]
Contrasts for the 20-species two-clade example
An example: Smith and Cheverud 2002


![Graph showing interspecific allometric equation](image)

Fig. 1. The interspecific allometric equation (specific regression, identified as IA) and the independent contrasts equation (identified as IC) plotted for 105 primate species in raw data space, transformed to natural logarithms. The interspecific allometric equation is \( \ln y = 0.139 + 0.080(\ln x) \), with \( r = 0.53 \). The phylogenetically corrected form of this equation, taken from the independent contrasts analysis, is \( \ln y = 0.160 + 0.056(\ln x) \), with \( r = 0.26 \). The two equations are not significantly different from each other. The identified species are Mandrillus sphinx (M), Pongo pygmaeus (O), Gorilla gorilla (G), Pan troglodytes (P), and Homo sapiens (H).
Treating within-species variation

Week 8: Paired sites tests, gene frequencies, continuous characters – p.43/44
How it was done

This freeware-friendly presentation prepared with

- Linux (operating system)
- LaTeX (mathematical typesetting)
- ps2pdf (turning Postscript into PDF)
- Idraw (drawing program to modify plots and draw figures)
- Adobe Acrobat Reader (to display the PDF in full-screen mode)