Statistical nonmolecular phylogenetics: can molecular phylogenies illuminate morphological evolution?

2 August 2014.

Joe Felsenstein

Workshop on Molecular Evolution, MBL, Woods Hole
Where this lecture fits in

Lately, more integration of

- work on molecular evolution
- work on between-species differences of measurable characters
- work on within-species differences of measurable characters

How can they fit together?
A standard quantitative genetics model

\[ P = \mu + \left\{ \begin{array}{lll} 0.6 & \text{AA} & 2 \\ 0.1 & \text{Aa} & 4 \\ -0.2 & \text{aa} & 7 \end{array} \right\} + \left\{ \begin{array}{lll} 6 & \text{BB} & \text{Bb} \\ 6 & \text{Cc} & \text{cc} \end{array} \right\} + \left\{ \begin{array}{lll} 0.3 & \text{DD} & \text{Dd} \\ 0.7 & \text{EE} & \text{ee} \end{array} \right\} + \text{environmental effect} \]
A model of quantitative characters on a phylogeny

- Brownian motion with multiple characters with different variances and with covariation as well.
A model of quantitative characters on a phylogeny

- Brownian motion with multiple characters with different variances and with covariation as well.
- This started with approximating gene frequencies in the 1960s by Anthony Edwards and Luca Cavalli-Sforza.
A model of quantitative characters on a phylogeny

- Brownian motion with multiple characters with different variances and with covariation as well.
- This started with approximating gene frequencies in the 1960s by Anthony Edwards and Luca Cavalli-Sforza.
- I expanded it to model quantitative characters determined by these genes.
Models for long-term evolution

The use of quantitative genetics approximations to model long-term evolution in lineages was largely introduced by Russ Lande in the 1980s.

Russell Lande, from his website at Imperial College, U.K., where he has been in recent years.
Where do the covariances come from?

- **Genetic covariances** (the same loci affect two or more traits). Genetic drift or natural selection can change the gene frequencies at these loci, and thus make correlated changes in the two traits.
Where do the covariances come from?

- **Genetic covariances** (the same loci affect two or more traits). Genetic drift or natural selection can change the gene frequencies at these loci, and thus make correlated changes in the two traits.

- **Selective covariances** (Olof Tedin, 1926; G. Ledyard Stebbins 1950) The same environmental conditions can select changes in two or more traits – even though they may have no genetic covariance. This source of evolutionary covariance is widely ignored.
Part 1: Morphometrics and phylogenies

Fred Bookstein is a co-author on this part of the talk

Fred Bookstein

me

“J. F. L. Bookenstein”
(Our reconstructed common ancestor)
How to use morphometric coordinates on phylogenies?

Is it possible to simply use the coordinates of landmarks $(x_1, y_1), (x_2, y_2), \ldots, (x_p, y_p)$ as continuous phenotypes $x_1, y_1, \ldots, x_p, y_p$ using Brownian motion along a phylogeny?

Yes, but ...

We must do proper morphometrics (correct for translation? rotation?)
Can we superpose specimens?

Consider two cases:

Are these different?
Why superposition is in principle impossible

Consider two cases:

Are these different?  No!
Dealing with translation

In effect one is centering each specimen so that the mean of its points is at \((0, 0)\). (The assumption is that the horizontal and vertical placement of the specimen on the digitizer is not useful information).

This has the effect of dropping two degrees of freedom so that each specimen now has \(2p - 2\) coordinates. It now “lives” in a \((2p - 2)\)-dimensional space.
The annoying issue of rotation

Sadly, there is no corresponding transform that tosses out rotation, as there is for translation.
Degrees of freedom and other transforms

In the morphometrics literature this is dealt with by a Procrustes Transform.

What Fred and I do is: choose the angles of rotation of all but the first specimen \((\theta_2, \theta_3, \ldots, \theta_p)\) to maximize the resulting likelihood.

All of these reduce the degrees of freedom of each specimen by 3, to \(2p - 3\).

But does this mean that the multivariate density function does not exist? No, it does exist, just in a \((2p - 3)\)-dimensional subspace.

In that space, all the usual machinery of the phylogenetic comparative method is available: contrasts to evaluate covariation of characters, reconstruction of ancestors, etc.
A simulation test

1. Generate 50 100-species trees by a pure birth process
2. For each evolve 100 forms by (covarying) Brownian Motion up the tree
3. These are the true covariances:

- All 10 landmarks move by independent and equal Brownian Motion of the coordinates with variance (per unit branch length) of 0.001, \textit{plus}
- the vertical coordinate of the pectoral fin and the two coordinates of the top of the tail move in a perfectly correlated change with variance 0.003.
20 of the 100 fishes from data set #2, centered and rotated.
20 of the 100 fishes from data set #2, also rescaled
First PC 1 for data set #2

This principal component shows both size changes and the fin extensions, and it is not easy to see which is which.
First shape PC 1 for data set #2

Now we’ve inferred a scale (size) component and removed it from the covariances, and then taken the first PC of the residual on size. We can see the fin component more clearly.
Making the first shape PC sparser by “medianizing”

To make PC1 be sparser we can add in a little location (not forcing the changes to maintain the centroid superposition). This is done by subtracting from the $x$ components, their median, and similarly for the $y$ components. So it minimizes the $L^1$ norm of the PC coefficients. The result is very clear.
What do we get from the Morphometric Consensus?

Using a Procrustes superposition and assuming the forms are i.i.d. and then computing principal components:

... we get a not-as-clear result with some size still there – we have ignored the tree and taken out size by standardizing centroid size, which is affected more by the fin component in the MMC methods.
Part 2: Fossils and phylogenies

(see also Tracy Heath’s talk later today – her method and this one are quite compatible)

This is also quite similar (I’m not sure how much as neither is yet published) to a recent project by Liam Revell, Luke Mahler, Graham Reynolds, and Graham Slater (reported on by Jonathan Losos at his anole research blog).
Present methods for calibration

Can take a fossil to indicate a bound on how recently a common ancestor was present. Use various priors on how much earlier or how much more recently.

But there is another way, which is being explored by me and (independently) by Alexander Pyron (2011) and by Fredrik Ronquist et al. (2012)
Another way of using fossils

Infer tree of present-day species from molecular sequences
Using fossils

Infer covariances of morphology using it, present-day species
Using fossils

Infer placement of fossil species using their data

Statistical nonmolecular phylogenetics: can molecular phylogenies illuminate morphological evolution? – p.25/85
Using fossils

Use fossil and present-day morphology, covariances, tree, also stratigraphic models
Using fossils

Use fossil and present-day morphology, covariances, tree, also stratigraphic models
Using fossils

Use fossil and present-day morphology, covariances, tree, also stratigraphic models

Statistical nonmolecular phylogenetics: can molecular phylogenies illuminate morphological evolution? – p.28/85
Using fossils

Use fossil and present-day morphology, covariances, tree, also stratigraphic models
Using fossils

Use fossil and present-day morphology, covariances, tree, also stratigraphic models
Using fossils

Use fossil and present-day morphology, covariances, tree, also stratigraphic models
Using fossils

Use fossil and present-day morphology, covariances, tree, also stratigraphic models
Using fossils

Use fossil and present-day morphology, covariances, tree, also stratigraphic models

Statistical nonmolecular phylogenetics: can molecular phylogenies illuminate morphological evolution? – p.33/85
Using fossils

Use fossil and present-day morphology, covariances, tree, also stratigraphic models
Using fossils

Use fossil and present-day morphology, covariances, tree, also stratigraphic models
Using fossils

Use fossil and present-day morphology, covariances, tree, also stratigraphic models
Using fossils

Use fossil and present-day morphology, covariances, tree, also stratigraphic models.
Using fossils

Use fossil and present-day morphology, covariances, tree, also stratigraphic models
Using fossils

Use fossil and present-day morphology, covariances, tree, also stratigraphic models
Using fossils

Use fossil and present-day morphology, covariances, tree, also stratigraphic models
A simple result

The upshot is that to find the maximum likelihood placement of a fossil lineage, we

- Hook it up somewhere
- Obtain the contrasts for that tree
- Infer the phylogenetic covariances of the characters from the contrasts
- The log-likelihood for this placement is (a constant plus) $-(n-1)/2$ times the log of the determinant of the covariance matrix, minus a penalty which depends on the sum of the logs of the standard deviations of the contrasts.

So we minimize the determinant plus penalty to find the best placement. We can consider whether we can do likelihood ratio tests, too, at least for placement within a single branch.
An example: the true tree with F a fossil species
Traffic-light colors shows where fossil can be placed

Green = within 1 log-likelihood unit, Orange = within 2 units, Red = lower than that. Green arrow is the ML placement. Gray placements are ruled out by date of the fossil.
Calibrating the molecular clock

Molecular trees don’t usually have branch lengths on a time scale, and we need that. How to infer the calibration of the clock?
Calibrating the molecular clock

For example if (not a real example) the placement of F turned out to be as shown, with the branch length shown in red, that in turn scales the whole molecular tree, as we know the time of F.
Calibrating the molecular clock

There will be two quantities to infer, the scaling of the molecular tree on the time scale, and the placement of the connection to the fossil. We make an ML estimate and accept other values that are not rejected by a Likelihood Ratio Test with 2 degrees of freedom.
A qualification

- The present method takes the molecular tree as known.
A qualification

- The present method takes the molecular tree as known.
- Uncertainty in it could be modelled by doing the analysis multiple times on bootstrap samples (or Bayesian posterior samples) of the tree estimates.
A qualification

- The present method takes the molecular tree as known.
- Uncertainty in it could be modelled by doing the analysis multiple times on bootstrap samples (or Bayesian posterior samples) of the tree estimates.
- Pyron and Ronquist both use a more comprehensive “total evidence” approach of allowing the morphological data to influence Bayesian inference of the tree.
A qualification

- The present method takes the molecular tree as known.
- Uncertainty in it could be modelled by doing the analysis multiple times on bootstrap samples (or Bayesian posterior samples) of the tree estimates.
- Pyron and Ronquist both use a more comprehensive “total evidence” approach of allowing the morphological data to influence Bayesian inference of the tree.
- I suspect this will have little effect if there is a lot of molecular data, so I am sticking with this approach.
Part 3: A threshold model for 0/1 characters

(This was published in *American Naturalist* in 2012)
Current methods for statistical treatment of 0/1 characters

Pagel (1994) and Lewis (2001) treat such data with

\[
\begin{array}{cccc}
A & C & B & D \\
\end{array}
\]

Pagel allows inference of whether change is correlated, on a known tree. Lewis infers the tree, but does not allow for correlations among characters. Neither takes into account contributions to a 0/1 character from multiple underlying loci.
The threshold model

A relevant model was invented in 1934 by Sewall Wright (1889-1988) shown here in 1954.
The threshold model

A relevant model was invented in 1934 by Sewall Wright (1889-1988)

rumor has it he then absent-mindedly erased the board with the guinea pig
Sewall Wright, at the University of Chicago, 1928

this guy
(seen in 1928,

Same place,
May, 2013

Statistical nonmolecular phylogenetics: can molecular phylogenies illuminate morphological evolution? – p.52/85
The threshold model

Sewall Wright (1934), guinea pig digit number (from Wright’s follow-up 1934 second paper)
The threshold model on a tree
Computing the likelihood

With two species, one character:

$$(1,0)$$

Disadvantages:

Quite hard to compute likelihoods: need to compute area in a corner of a correlated multivariate normal distribution.

With 5 species, one character:

$$L = \text{Prob} (1,1,0,1,1)$$

$$= \int_0^\infty \int_0^\infty \int_{-\infty}^0 \int_0^\infty \int_0^\infty \phi(x_1,x_2,x_3,x_4,x_5 \mid \text{Tree}) \, dx_1 \, dx_2 \, dx_3 \, dx_4 \, dx_5$$
Likelihoods under the threshold model on a tree

To compute the likelihood for a tree under the threshold model with \( p \) characters, want to compute:

\[
L = \int_{c}^{\infty} \int_{-\infty}^{c} \cdots \int_{c}^{\infty} |V|^{-1} (2\pi)^{-np/2} \\
\times \exp \left(-\frac{1}{2} (x - \mu)^t V^{-1} (x - \mu) \right) \, dx_{11} \, dx_{12} \cdots dx_{np}
\]

where \( \mu \) is the appropriate vector of means, and

\[
V = A \otimes T
\]

involves the tree and the “evolutionary" covariance matrix of the characters.

In other words, the probability density of the (unknown) liabilities gets integrated over the region of their values that corresponds to the observed discrete characters.
MCMC on liabilities

(1) (1) (1) (0) (0)
A  2.03  B  1.64  C  0.57  D  −1.4  E  −0.3

1.69

0.81

−0.29

0.23
MCMC on liabilities: Gibbs sampling in the interior

Gibbs sampler for internal node values

\[
\begin{align*}
&\begin{array}{cccccc}
(1) & (1) & (1) & (0) & (0) \\
A & B & C & D & E \\
2.03 & 1.64 & 0.57 & -1.4 & -0.3 \\
\end{array}
\end{align*}
\]

\[
\begin{align*}
v_1 & \quad v_2 \\
x & \quad v_6
\end{align*}
\]

\[
\begin{align*}
x & \quad \text{drawn from normal distribution,} \\
\text{mean} & \quad = \quad \frac{(1/v_1) \times 2.03 + (1/v_2) \times 1.64 + (1/v_6) \times 0.81}{(1/v_1) + (1/v_2) + (1/v_6)} \\
\text{var} & \quad = \quad \frac{1}{(1/v_1) + (1/v_2) + (1/v_6)}
\end{align*}
\]
MCMC on liabilities: result of Gibbs sampling

Gibbs sampler for internal node values

(1) (1) (1) (0) (0)
A B C D E
2.03 1.64 0.57 -1.4 -0.3

1.48

0.81

-0.29

0.23
MCMC on liabilities: rejection at tips

How to update the liability at a tip?
(must condition on ancestor and observed phenotype)

Statistical nonmolecular phylogenetics: can molecular phylogenies illuminate morphological evolution? – p.60/85
An example
An example
An example
An example
An example
An example
An example
An example
An example
An example
An example
An example
An example
An example
A 3-character simulation
A 3-character simulation

For these true covariances:

\[
\begin{bmatrix}
1.64 & 0.8 & 0 \\
0.8 & 1.36 & -0.6 \\
0 & -0.6 & 1 \\
\end{bmatrix}
\]

100 data sets with 100-species trees were analyzed.
Inferred correlation coefficients

<table>
<thead>
<tr>
<th>character 1</th>
<th>character 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>−1.0</td>
<td>−1.0</td>
</tr>
<tr>
<td>−0.5</td>
<td>−0.5</td>
</tr>
<tr>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

character 1

character 2

character 3

Statistical nonmolecular phylogenetics: can molecular phylogenies illuminate morphological evolution? – p.77/85
What about QTLs?

- We can integrate this work with QTL inference.
What about QTLs?

- We can integrate this work with QTL inference.
- Not only identify QTLs, but to see them change across species, including some QTLs causing variation within some species, some within others.
What about QTLs?

- We can integrate this work with QTL inference.
- Not only identify QTLs, but to see them change across species, including some QTLs causing variation within some species, some within others.
- Could even allow us to infer on which of two correlated characters the selection really acted.
The Reunion

- For the last 40-50 years population-genetic work within species has been (mostly) isolated from work on molecular evolution between species.
- Now we are in a gradual Reunion of these two lines of work (not a New Synthesis, though) as observations can be made that connect them (coalescents across species boundaries, Ds/Dn inferences, etc.)
- As this happens, Russ Lande’s vision will become more and more of a reality – quantitative genetics will become directly relevant to multi-species evolutionary biology.

More generally we are seeing increased connections between
- Within- and between-species work
- Morphological and genomic studies
- Paleontological and neontological studies
What we can ... and cannot ... infer

- BUT ... we have limited power from any one sample of species. Biologists must learn to accept that, and find ways to propagate that uncertainty through the analysis that flow from these inferences. We cannot (ever!) have a Fly-On-The-Wall account of evolution.
What we can ... and cannot ... infer

- BUT ... we have limited power from any one sample of species. Biologists must learn to accept that, and find ways to propagate that uncertainty through the analysis that flow from these inferences. We cannot (ever!) have a Fly-On-The-Wall account of evolution.

- Furthermore we must always be sensitive to the limits of our models – as we expand the tree to less related groups, the models are called severely into question.
An advertisement

Steve Arnold and I are running a course on Evolutionary Quantitative Genetics this summer (as we have done the previous three years at NESCENT), this time at the National Institute for Mathematical and Biological Synthesis (NIMBioS) in Knoxville. It starts Monday.

Other faculty include Patrick Carter, Adam Jones, Paul Hohenlohe, Marguerite Butler, Liam Revell, Brian O’Meara and Josef Uyeda.

Next year we hope to have it occur at either NESCENT or NIMBioS.
Thanks to...

NSF for several grants to me

NIH NIGMS for several past grants to me and Mary Kuhner

Felsenstein / Rudd family funds (more and more necessary)
References


Pyron, R. A. 2011. Divergence time estimation using fossils as terminal taxa and the origins of Lissamphibia. *Systematic Biology* **60**: 466-481. [Pioneering paper on using morphology and molecules to place fossils and date divergences]


Lewis, P. O. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology* **50**: 913-925. [Uses 0/1 stochastic process to infer morphological phylogenies]
References


