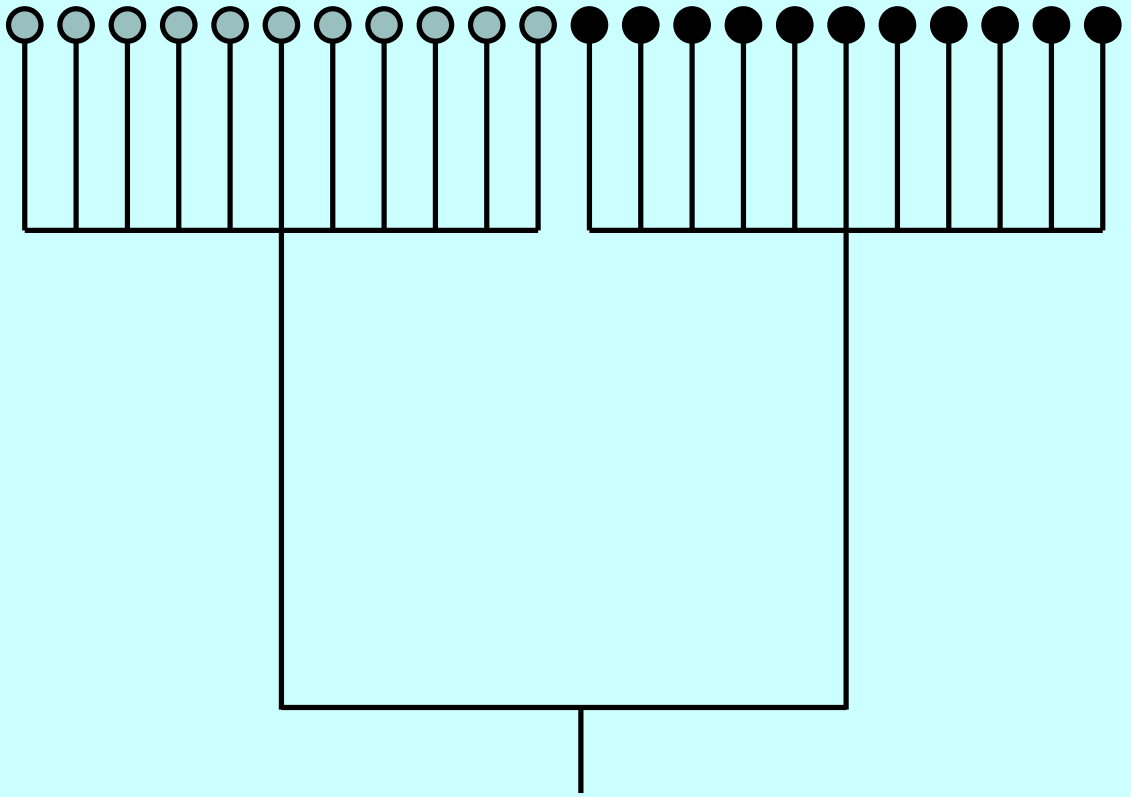
Comparative method and phylogenies

3 November 2016.

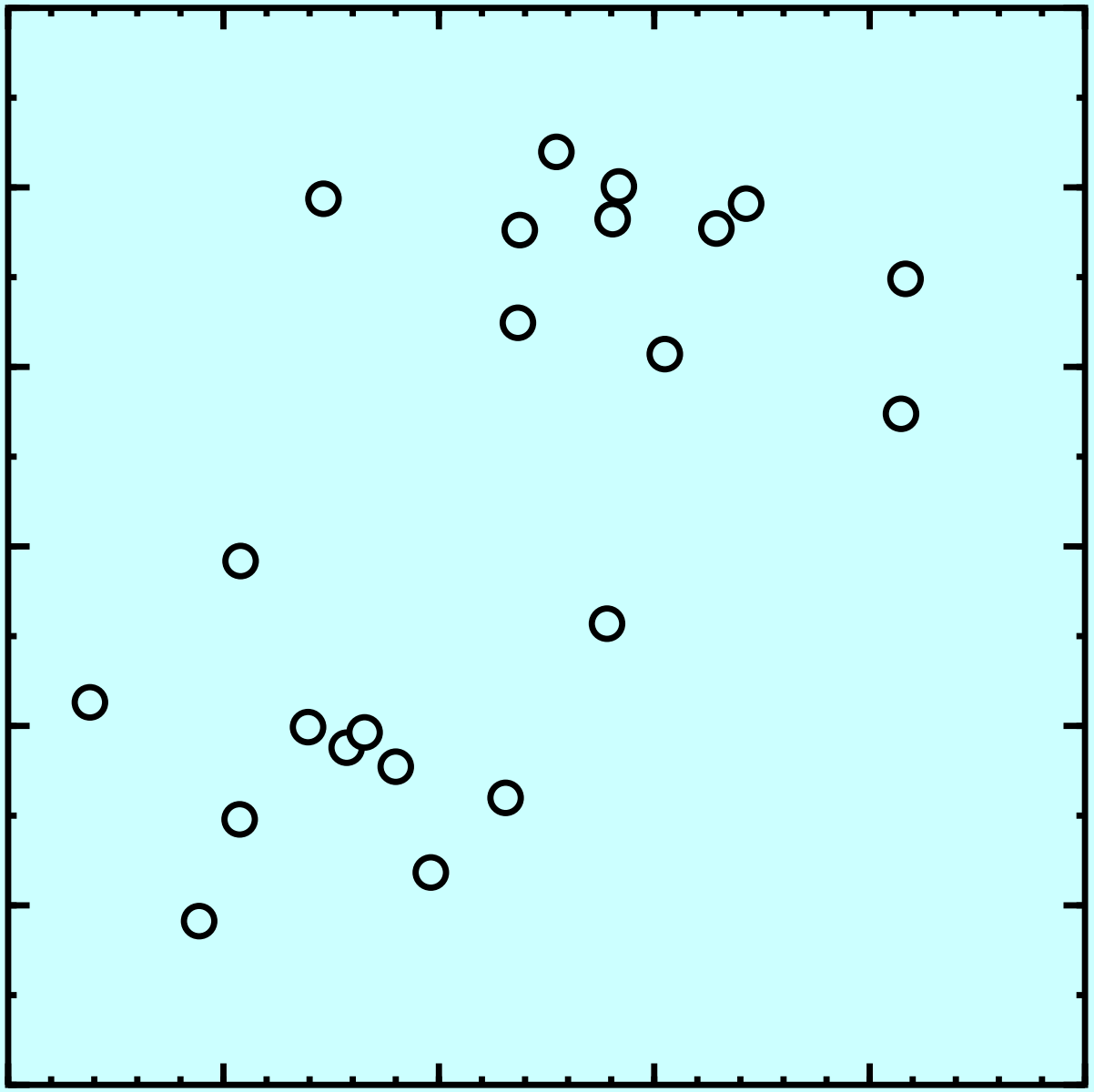
Joe Felsenstein

Biology 550D

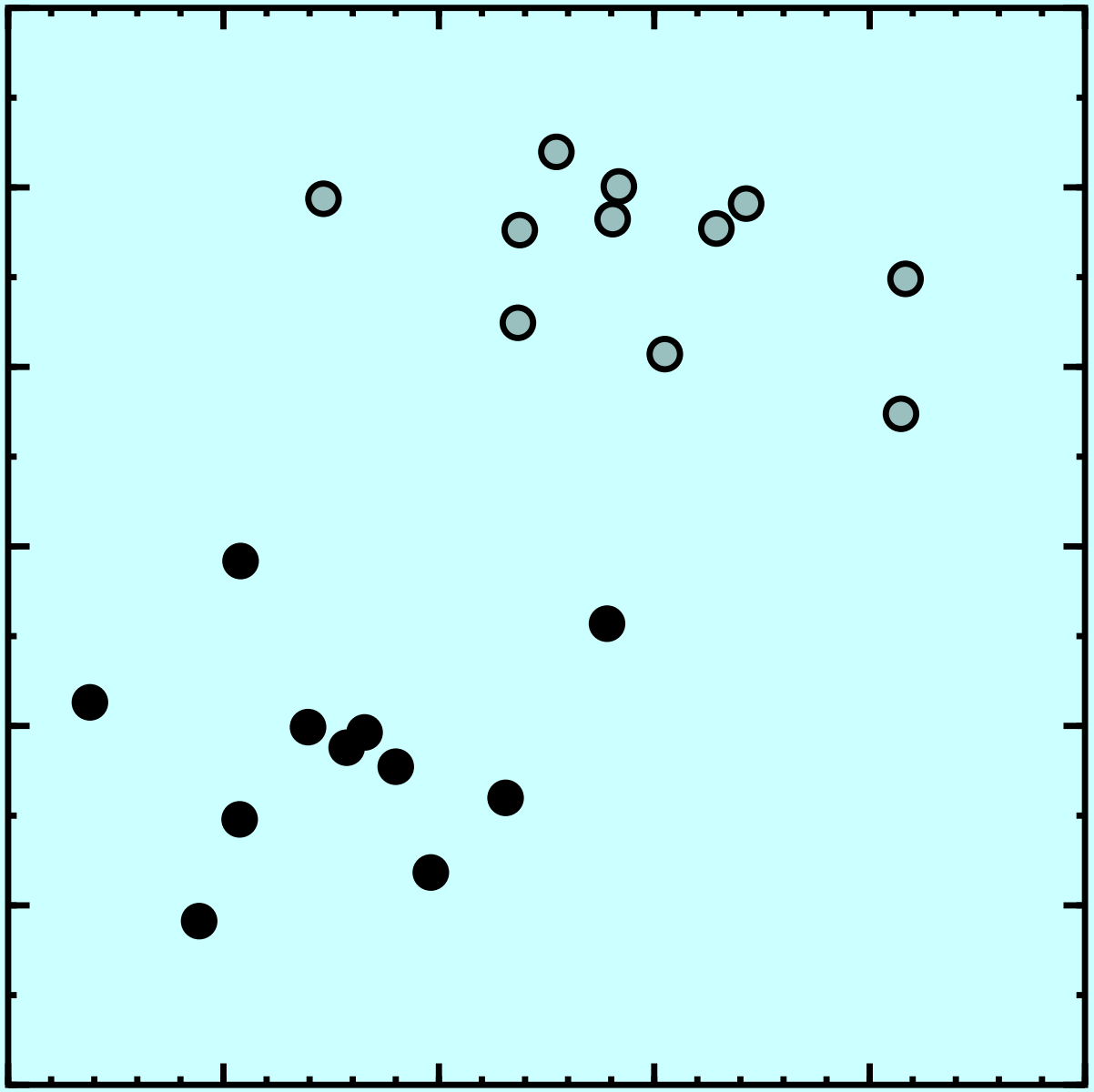
A simple case to show effects of phylogeny



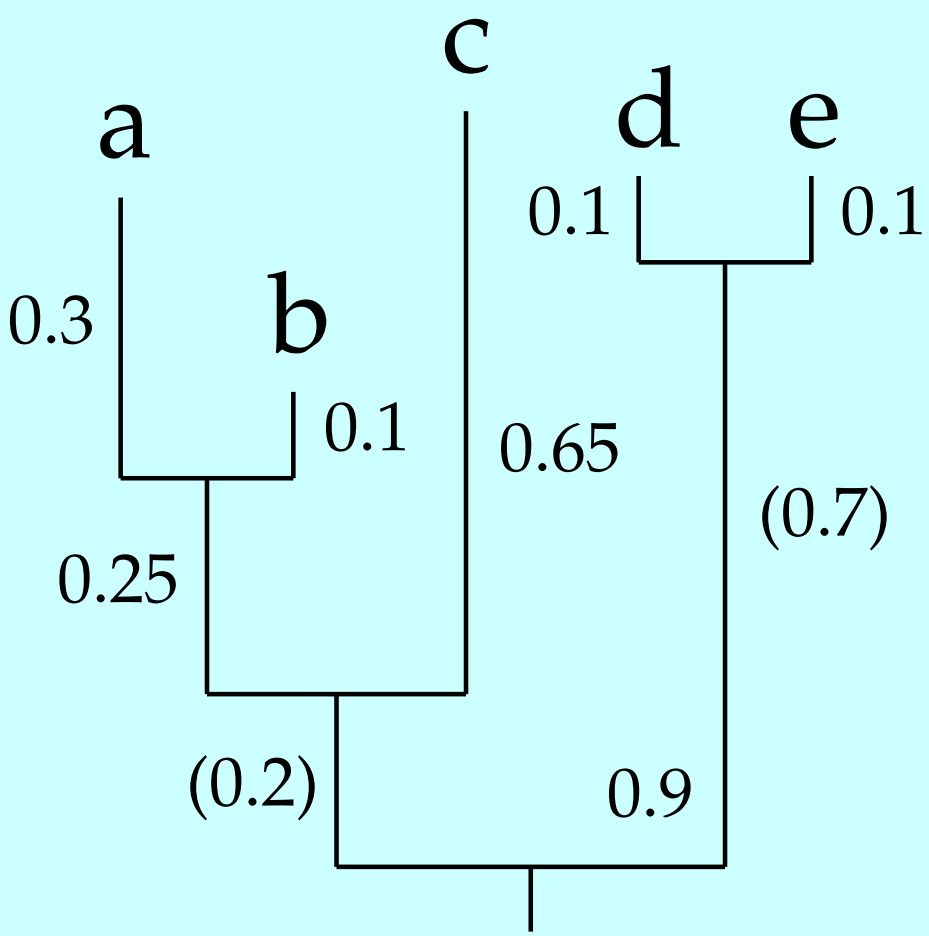
Two uncorrelated characters evolving on that tree



Identifying the two clades



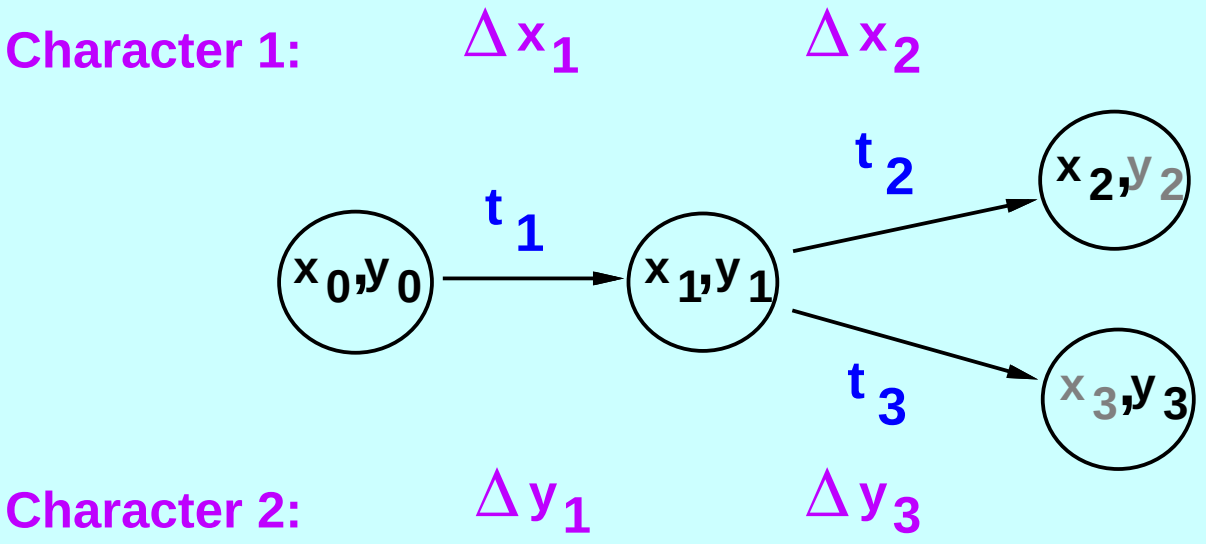
A tree on which we are to observe two characters



Contrasts on that tree

Contrast											Variance proportional to
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

Joint distribution for multiple species, characters



Consider change of two characters, each assessed in a different species. Say character x and character y, the first measured in species 2, the second in species 3. The result will give us the pattern for any two characters measured in any two species.

Seeing that covariances are zero in different branches ...

$$\mathrm{Cov}[\Delta x_1 + \Delta x_2,\ \ \Delta y_1 + \Delta y_3]$$

Given that changes in different branches are independent (whether changes of the same character or of different characters), the only nonzero covariance is between Δx_1 and Δy_1.

$$\mathrm{Cov}[\Delta x_1 + \Delta x_2,\ \ \Delta y_1 + \Delta y_3]$$

$$= \ \mathrm{Cov}[\Delta x_1,\ \ \Delta y_1]$$

So the covariance of different characters in different species is the product of the shared evolution to their common ancestor by the (infinitesimal) covariance of the character change per unit branch length.

Joint distribution for many species, many characters

The upshot is that if x_{ik} is character k in species i, and $x_{j\ell}$ is character ℓ in species j, the covariance between them is

$$\mathrm{Cov}[x_{ik}, x_{j\ell}] = t_{ij}\, v_{k\ell}$$

where t_{ij} is the time (branch length) to the latest common ancestor of species i and species j. $\mathbf{V}$ is the covariance matrix of evolutionary change for the characters.

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

$$\begin{bmatrix}
a & b & c & 0 & 0 & 0 & 0 \\
b & d & c & 0 & 0 & 0 & 0 \\
c & c & e & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & f & g & g & g \\
0 & 0 & 0 & g & h & i & i \\
0 & 0 & 0 & g & i & j & k \\
0 & 0 & 0 & g & i & k & l
\end{bmatrix}$$

“Pruning” a tree in the Brownian motion case

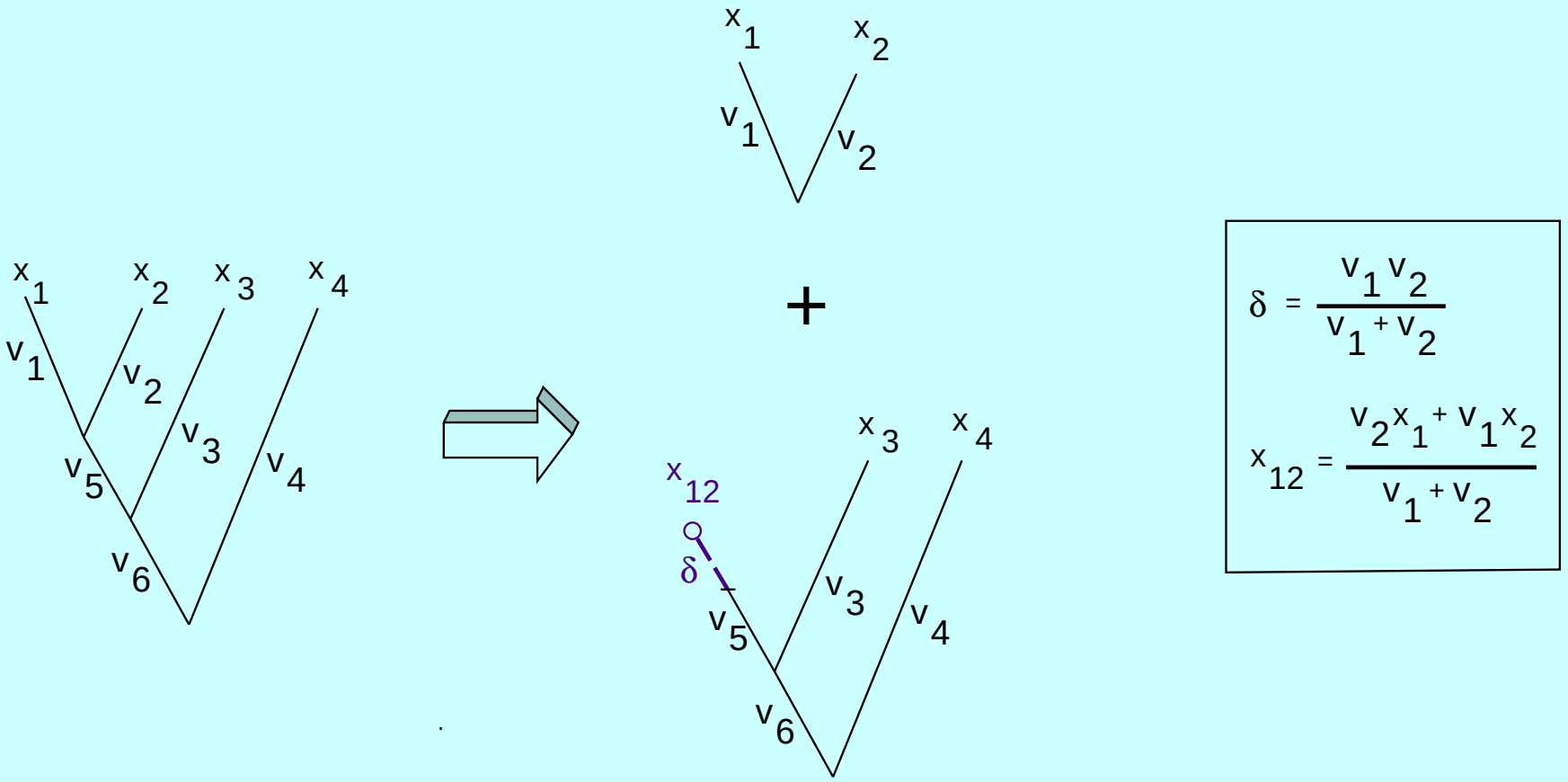
One can take two neighboring tips, and consider their difference $x_1 - x_2$ as well as a weighted average $ax_1 + (1-a)x_2$. Using weights $a : 1-a = 1/v_1 : 1/v_2$, the weighted average is independent of the difference, and the difference is also independent of the rest of the tree.

In fact, this weighted average behaves like a tip: Its covariances with the other species are the same as those of x_1 and x_2. It acts just as if the tree were pruned, cutting off species 1 and 2, leaving a single species whose variance is a bit bigger.

$$\mathrm{Var}[ax_1 + (1-a)x_2] = v_8 + v_9 + \frac{v_1 v_2}{v_1 + v_2}$$

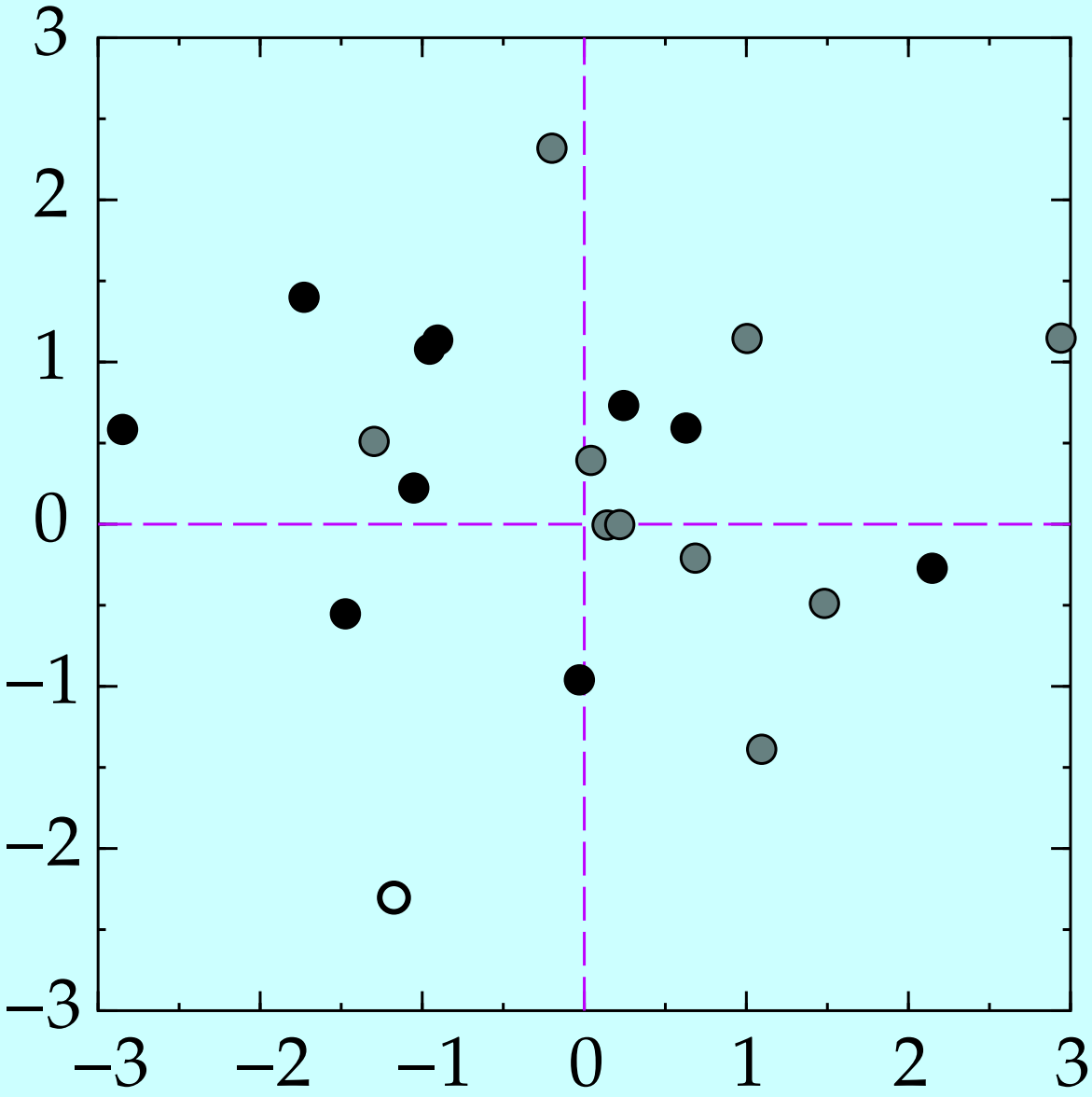
so in effect, a small extra amount of branch length is added.

“Pruning” a tree in the Brownian motion case



(True in the sense that the log-likelihoods – which are a bit different than the usual likelihoods – add up, since the likelihoods multiply).

Contrasts for the 20-species two-clade example



The algebra

If $\mathbf{T}$ is the covariances of $\mathtt{n}$ tips on the tree, and $\mathbf{V}$ is the (unknown) covariances of the Brownian motion of the $\mathtt{p}$ characters, the log-likelihood of a set of characters (stacked as a vector) $\mathbf{x}$ is

$$\ln L = -(np/2)\ln(2\pi) - (1/2)\ln|\mathbf{T}\otimes\mathbf{V}| - (1/2)(\mathbf{x}-\mu)^t(\mathbf{T}\otimes\mathbf{V})^{-1}(\mathbf{x}-\mu)$$

If $\mathbf{C}$ is an $(\mathtt{n}-1)\times\mathtt{n}$ set of contrasts, each orthogonal to the grand mean, such that $\mathbf{CTC}^t$ is an $\mathtt{n}-1$-dimensional identity matrix, then taking the density of the transformed data $\mathbf{y} = \mathbf{C}\,\mathbf{x}$, this has expectation vector $\mathbf{0}$:

$$\ln L = K - (1/2)\ln|\mathbf{I}_{n-1}\otimes\mathbf{V}| - (1/2)\mathbf{y}^t(\mathbf{I}_{(n-1)}\otimes\mathbf{V})^{-1}\mathbf{y}$$

(where K collects the constant stuff, including the $\ln(\mathtt{v}_1+\mathtt{v}_2)$) Jacobian term.

... simplifying ...

This can also be expressed as

$$\ln L \;=\; K - ((n-1)/2)\ln|\mathbf{V}| - (1/2)\mathrm{tr}\left(\mathbf{SV}\right)^{-1})$$

where

$$\mathbf{S} \;=\; \sum_i \mathbf{y}^{(i)}\left(\mathbf{y}^{(i)}\right)^t$$

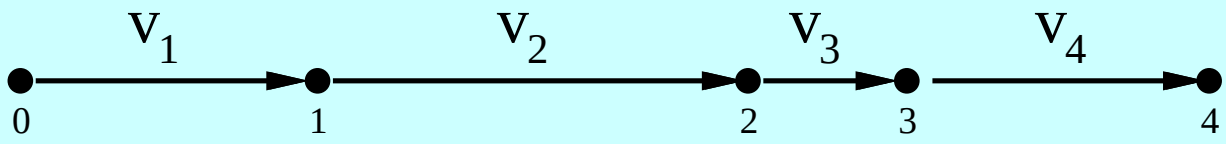
is the $\mathrm{p} \times \mathrm{p}$ sum of squares matrix of characters across contrasts. Inferring the Brownian motion phylogenetic covariances by maximum likelihood we find that

$$\widehat{\mathbf{V}} \;=\; \mathbf{S}/(n-1)$$

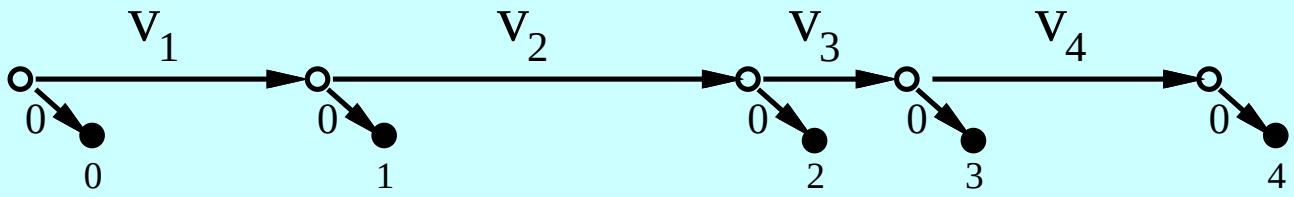
which leads to

$$\ln L \;=\; K' \;-\; ((n-1)/2)\ln|\widehat{\mathbf{V}}|$$

The case of observing ancestors



can be considered to be a tree with zero−length branches



This works out as one might hope. Computing the contrasts, they turn out to be simply

$$\frac{x_1 - x_0}{\sqrt{v_1}}, \quad \frac{x_2 - x_1}{\sqrt{v_2}}, \quad \frac{x_3 - x_2}{\sqrt{v_3}}, \quad \frac{x_4 - x_3}{\sqrt{v_4}}$$

which are obviously independent.

In the case where a finite sample is taken at each time, and there is within-species phenotypic variation, matters are more complicated but a comparative methods analysis allowing for sampling error works.

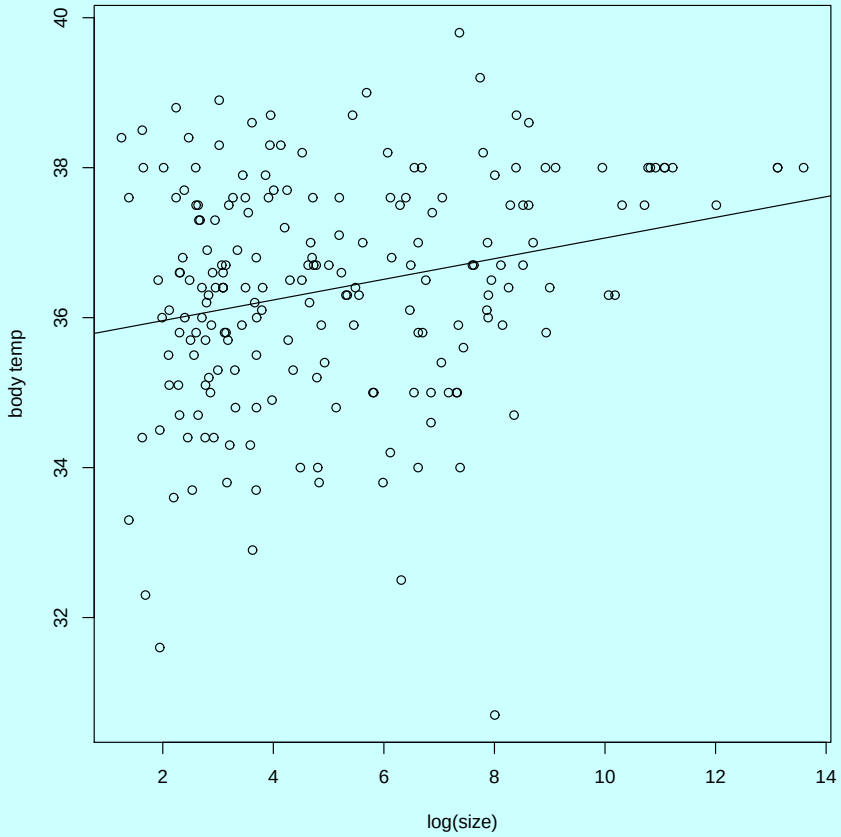
A research program?

What we could imagine doing is:

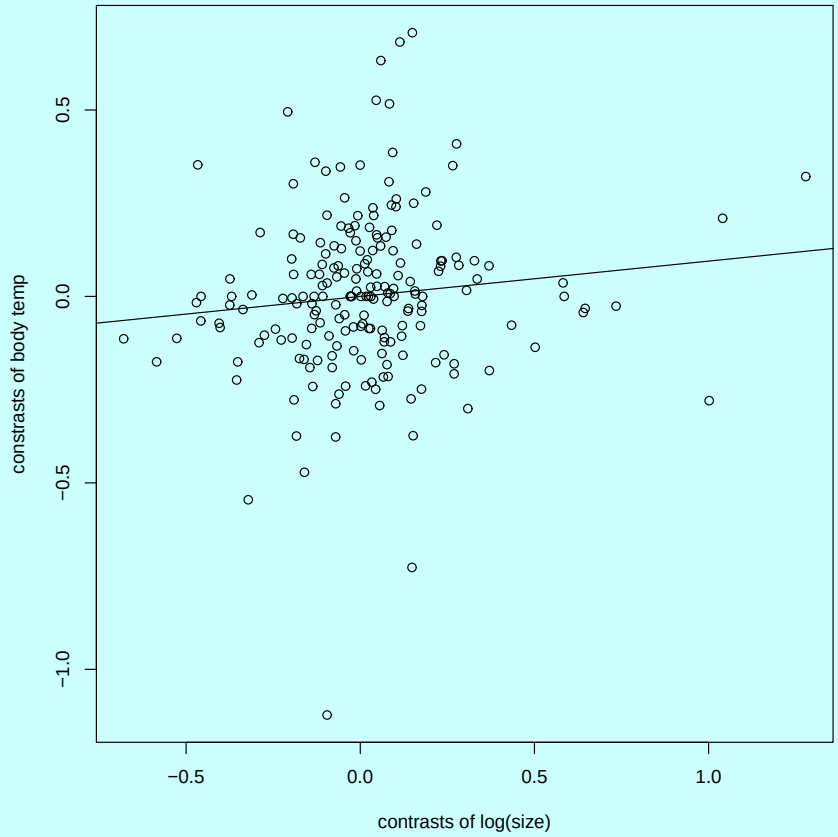
- We might hope to infer additive genetic covariances by doing quantitive genetics breeding experiments to infer them from covariances among relatives, perhaps even in multiple species.
- Infer the covariances of the changes along the phylogeny.
- From them, back-calculate the selective covariances.
- The genetic covariances may also be inferrable from differences between nearby tips on the tree if we do not have breeding experiments.
- There is little or no hope of inferring “selective correlations” more directly without a complete understanding of the functional ecology.

An example: Riek and Geiser, 2013

Alexander Riek and Fritz Geiser. 2013. Allometry of thermal variables in mammals: consequences of body size and phylogeny. *Biological Reviews* **88 (3):** 564-572.



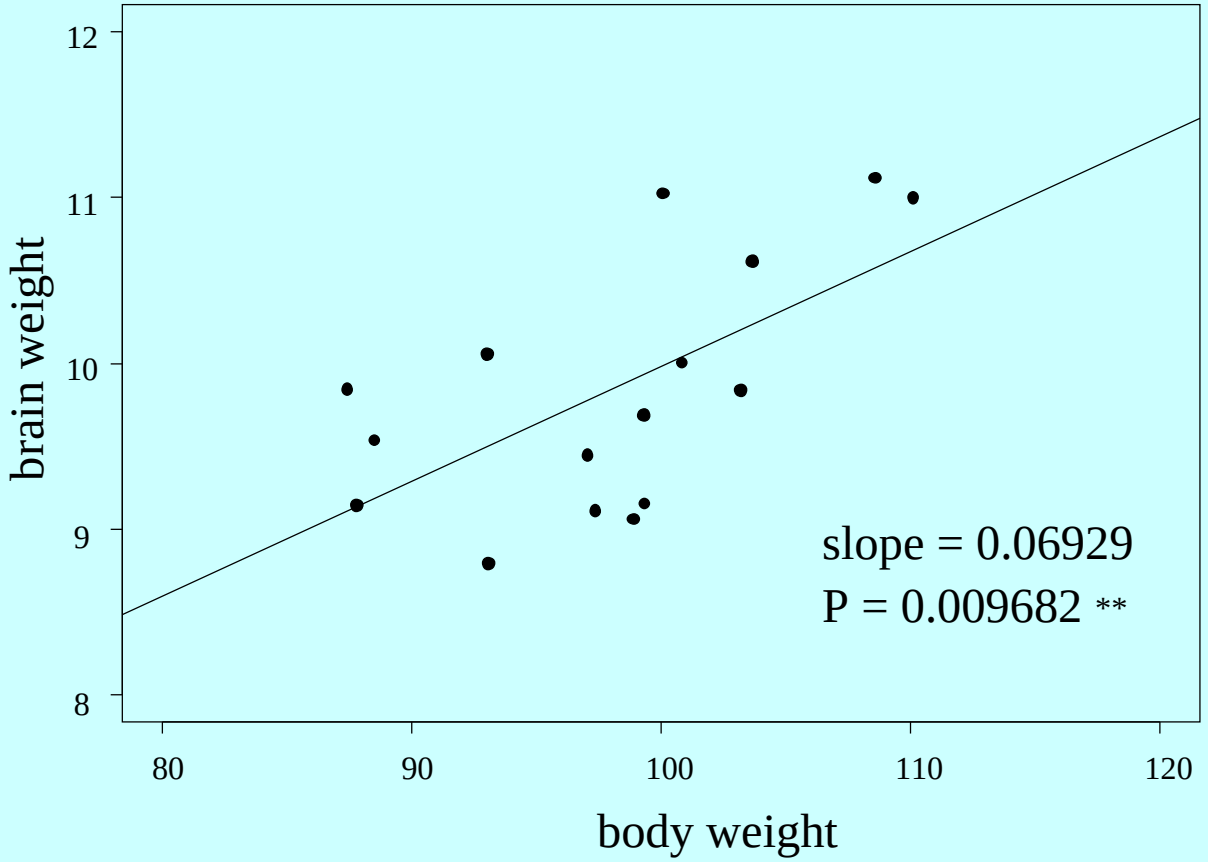
body temperature vs. log(body size)
(P for slope $\neq 0$ is 0.000375)



contrasts vs. contrasts
(P for slope $\neq 0$ is 0.116)

A simulated example

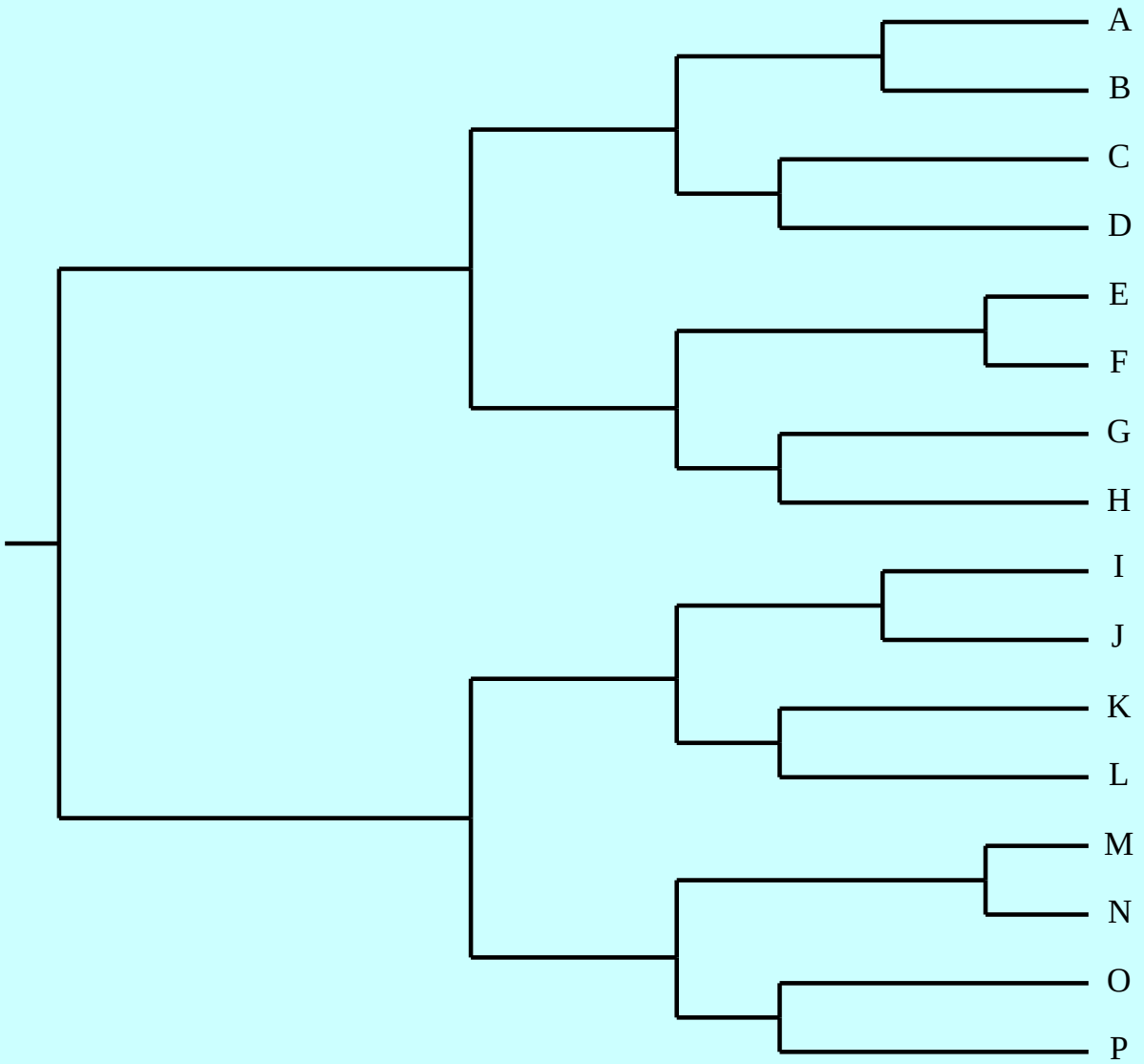
Using an ordinary regression with the species as points, we see a significant relationship between brain weight and body weight:



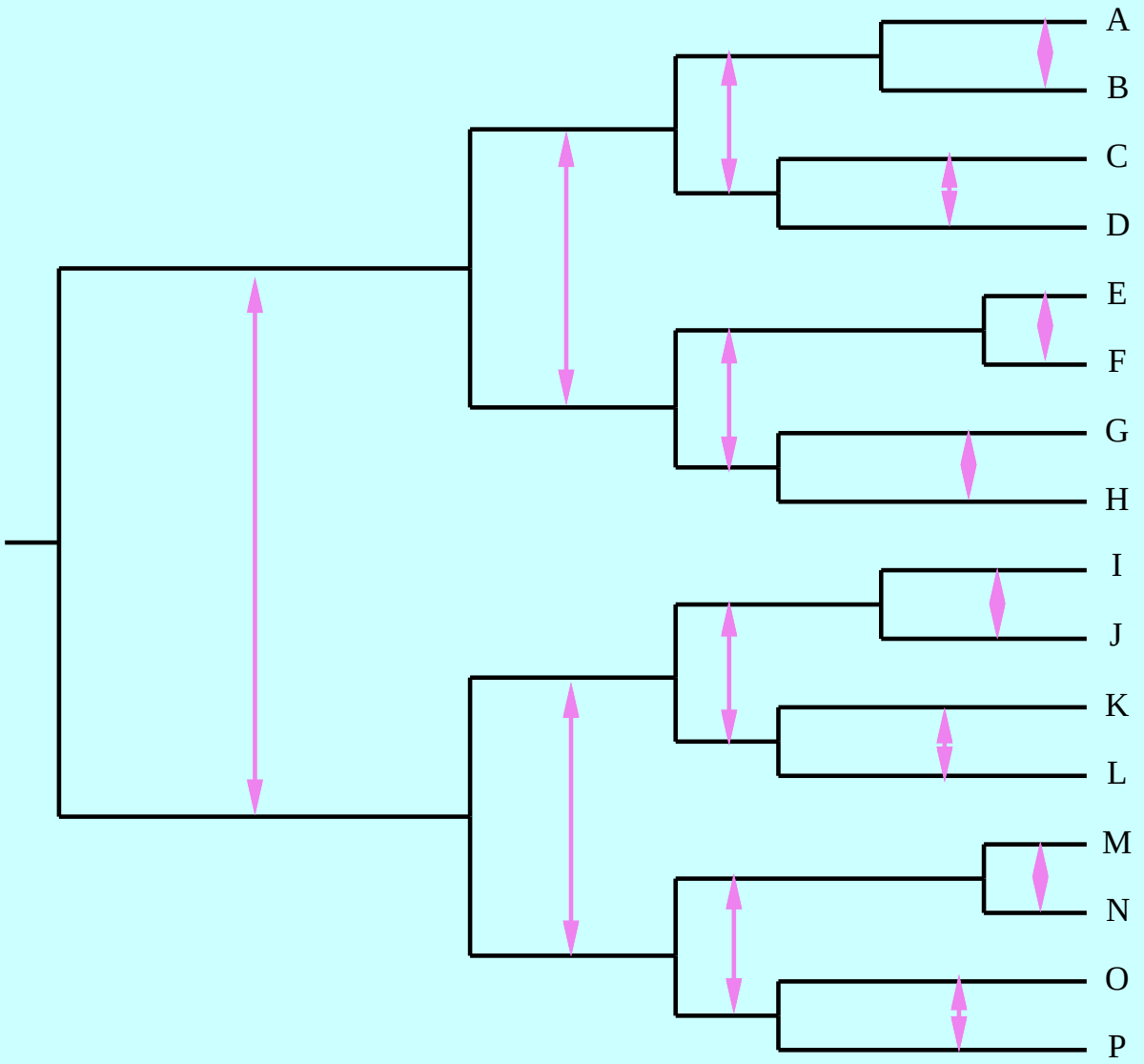
It looks as if we have 16 independent data points and a positive correlation between brain weight and body weight across species.

But the points are not independent

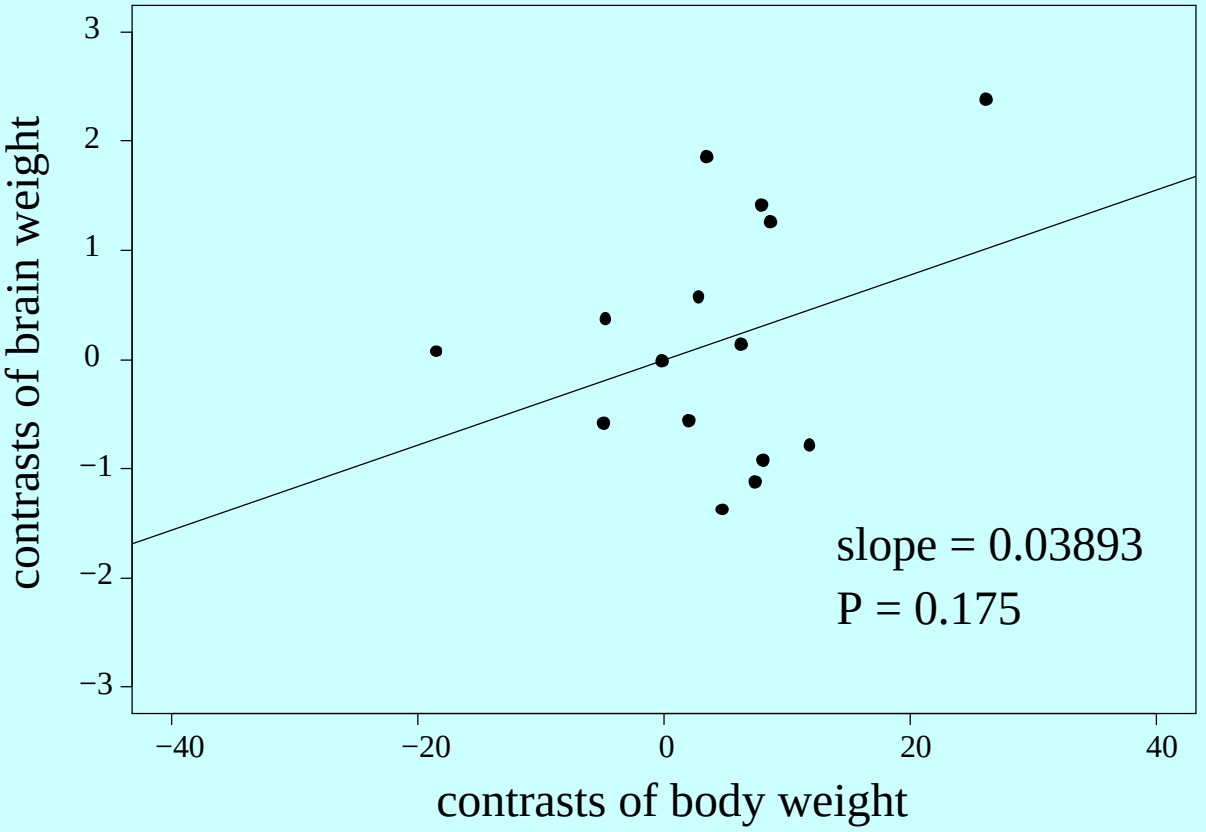
They evolved on a phylogeny. More closely related points are similar.



Using contrasts on the phylogeny ...



Is evolution of brain and body weight correlated?



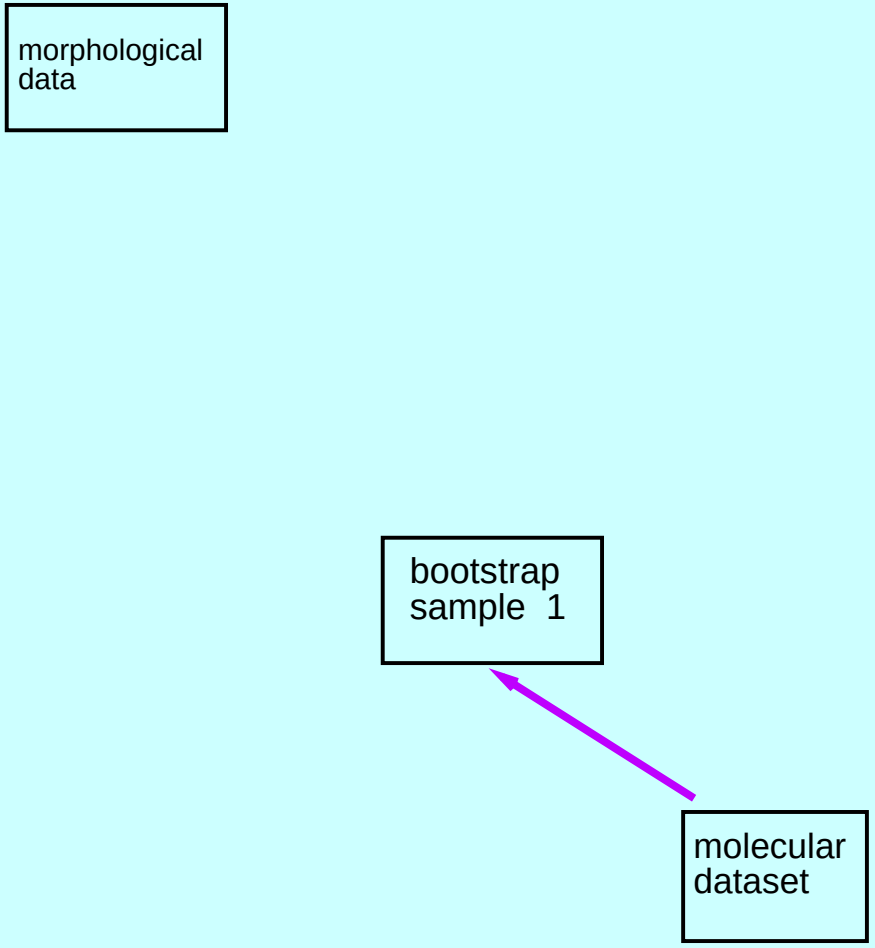
Using the contrasts method we see no significant relationship.

When the tree is noisy: Propagating bootstrap sampling

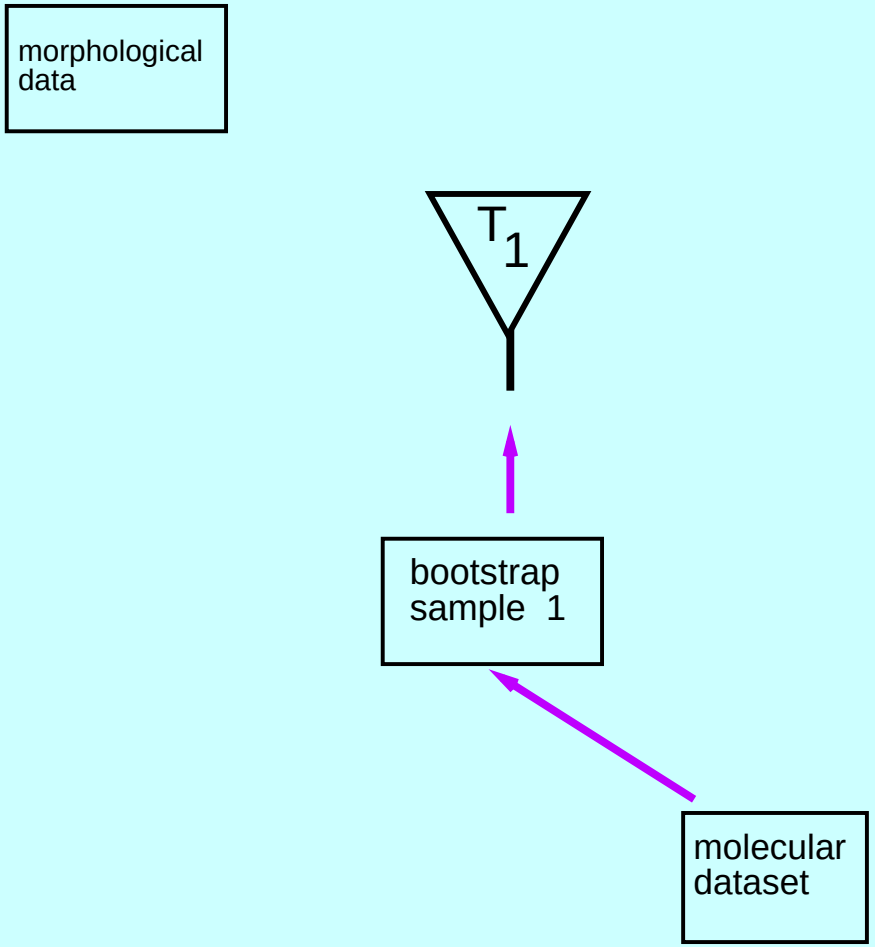
morphological data

molecular dataset

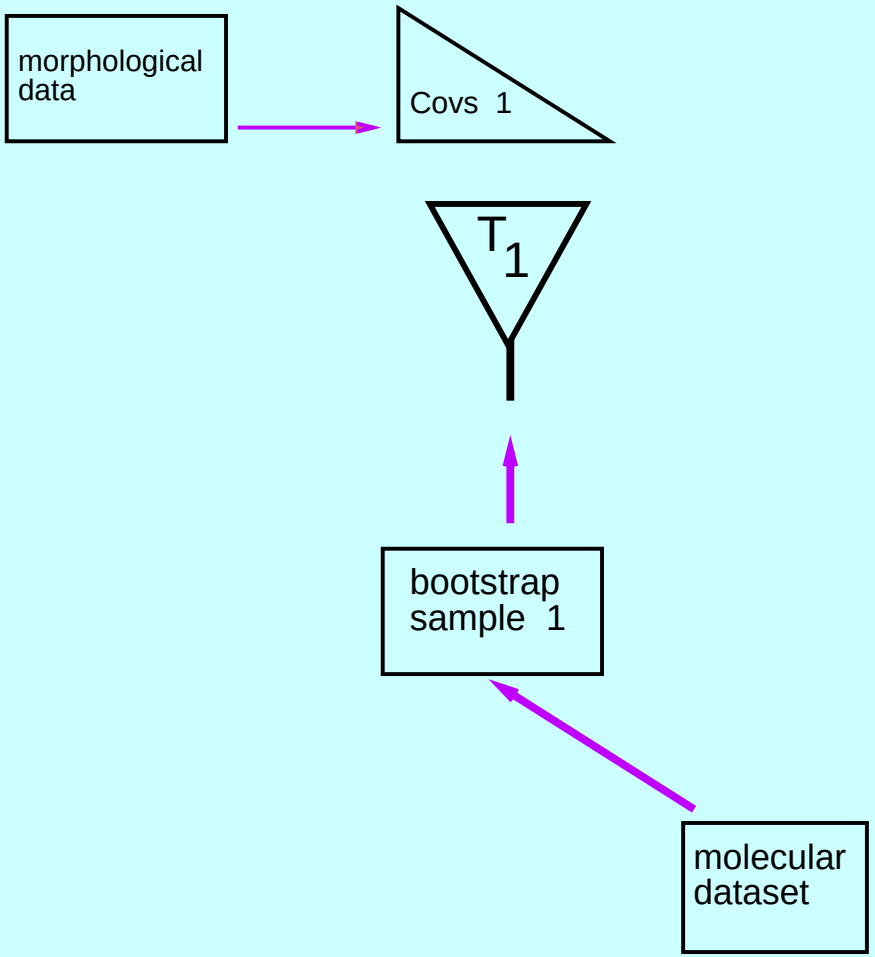
Propagating bootstrap sampling



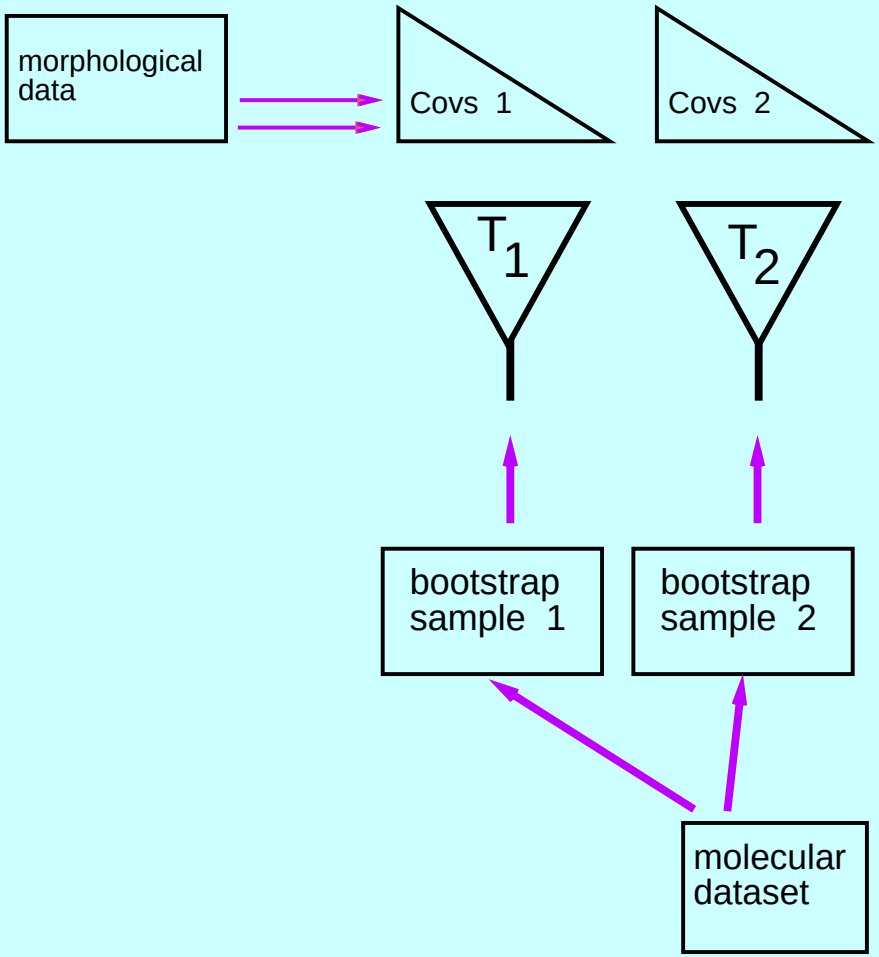
Propagating bootstrap sampling



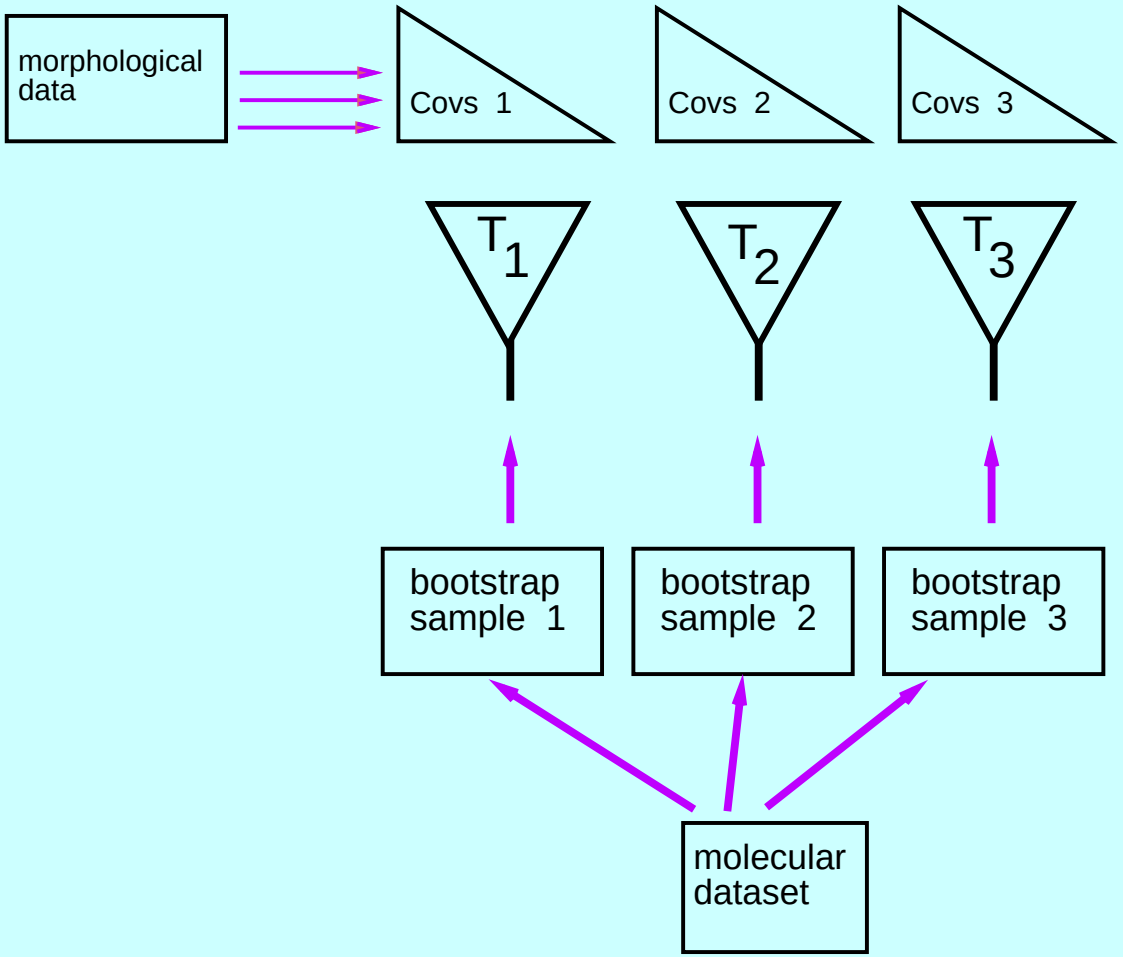
Propagating bootstrap sampling



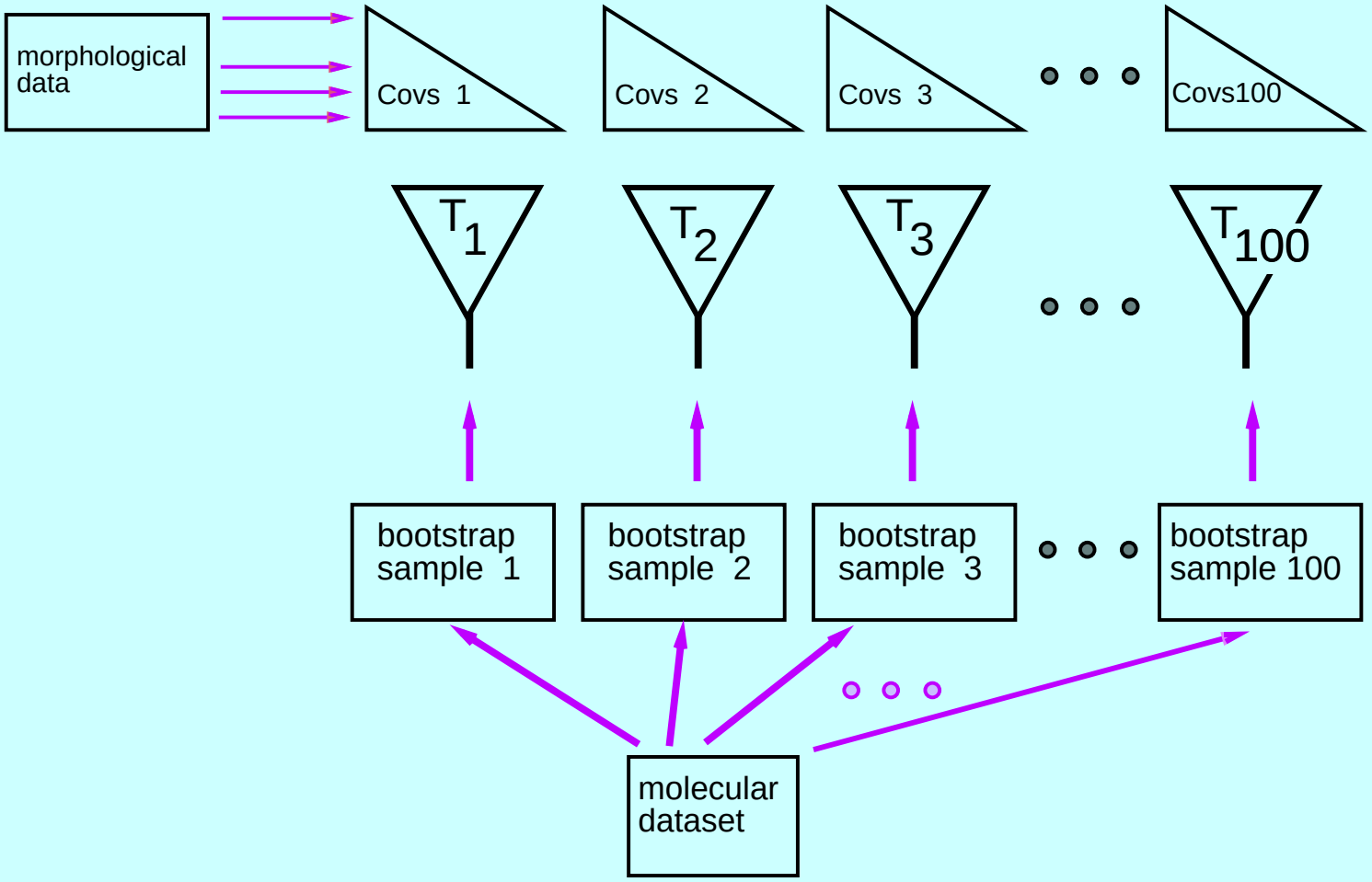
Propagating bootstrap sampling



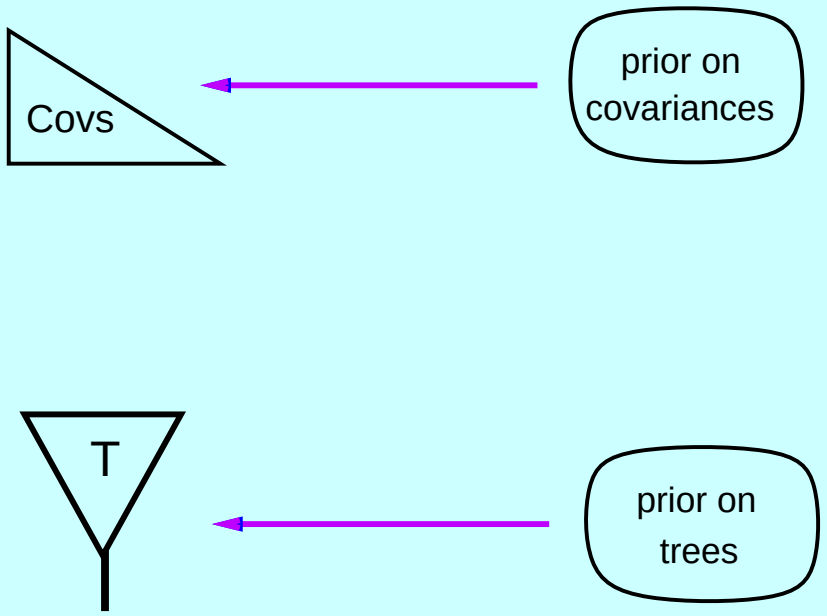
Propagating bootstrap sampling



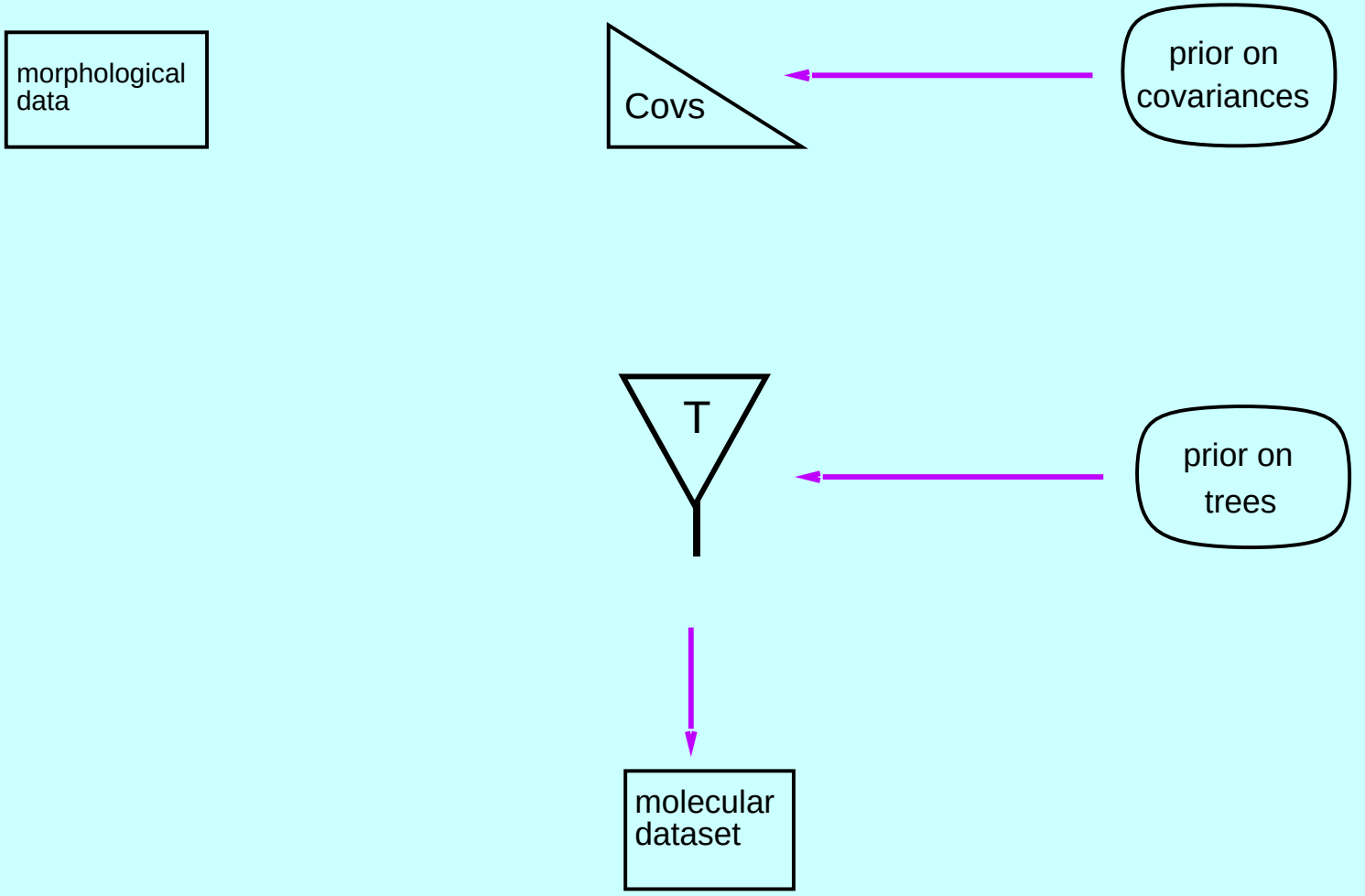
Propagating bootstrap sampling



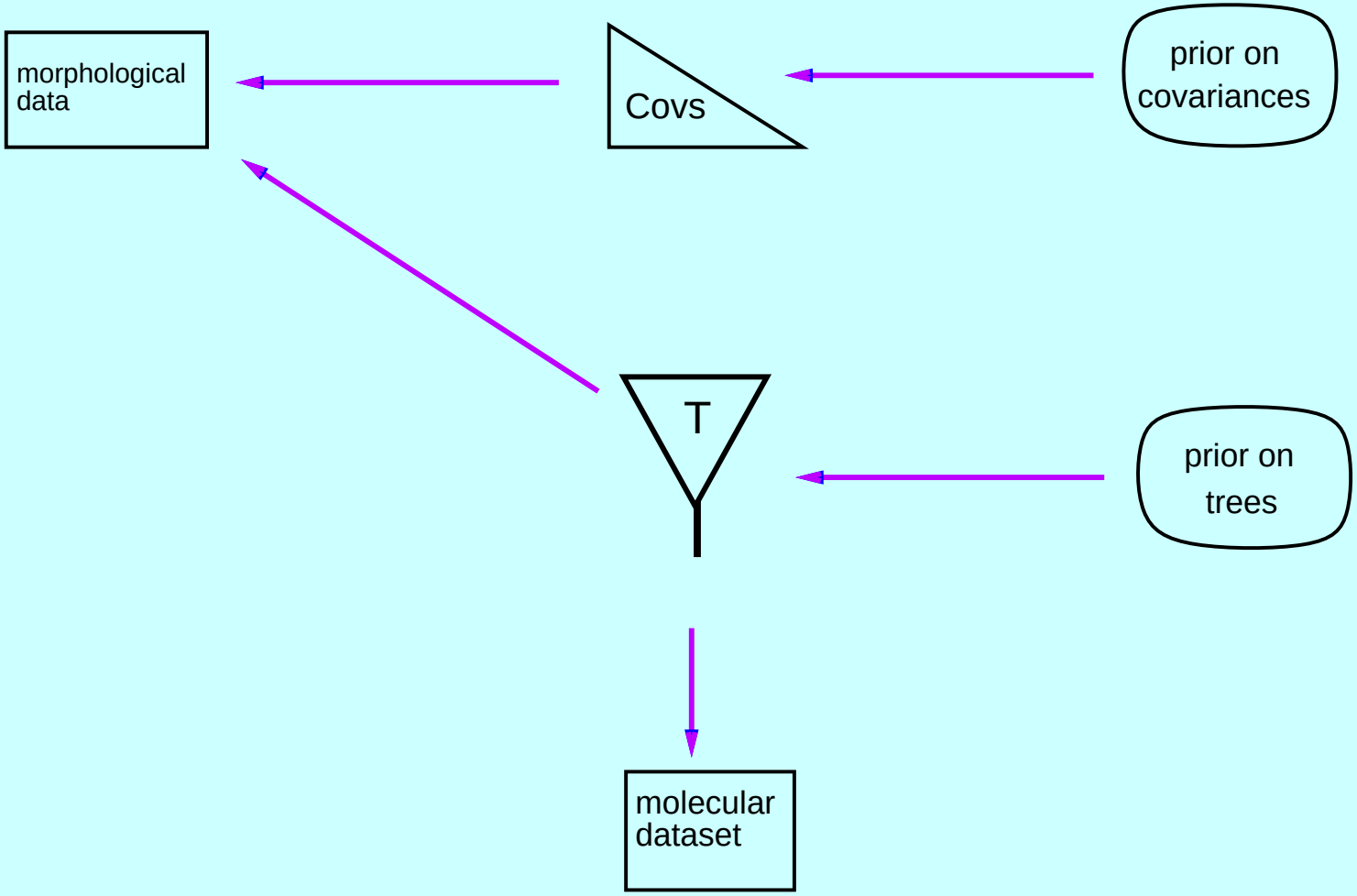
A Bayesian model



A Bayesian model

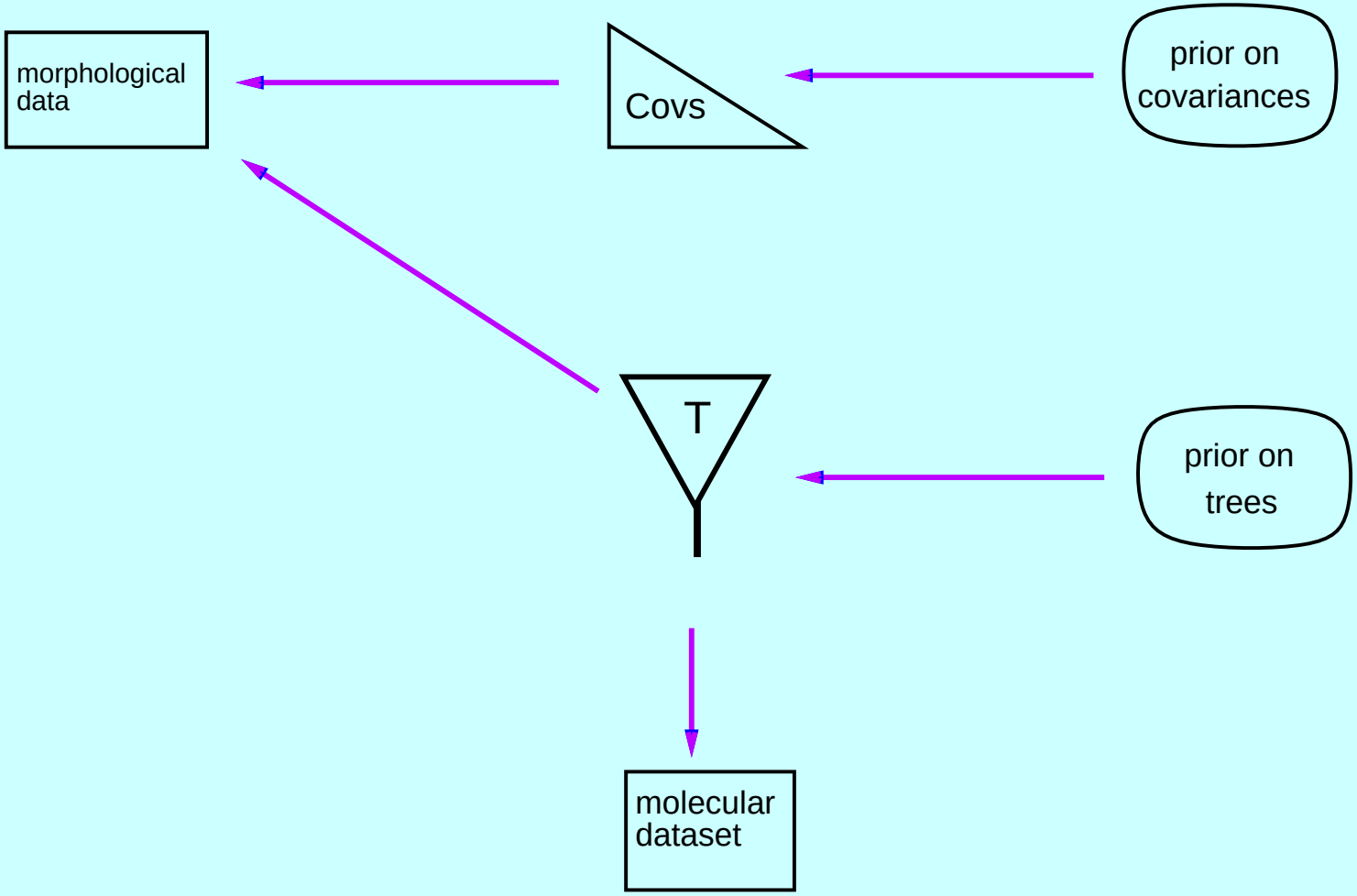


A Bayesian model

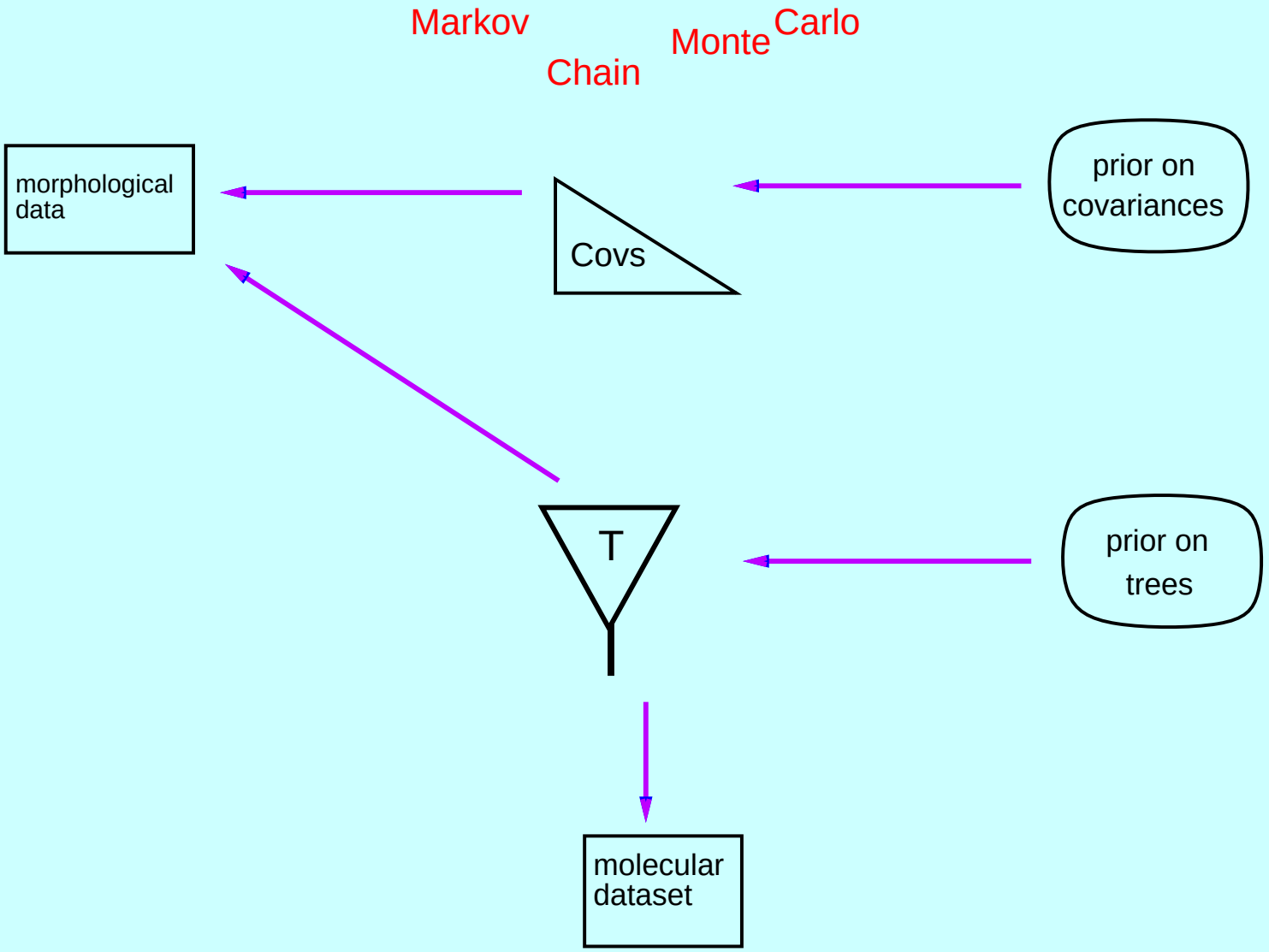


Bayesian MCMC

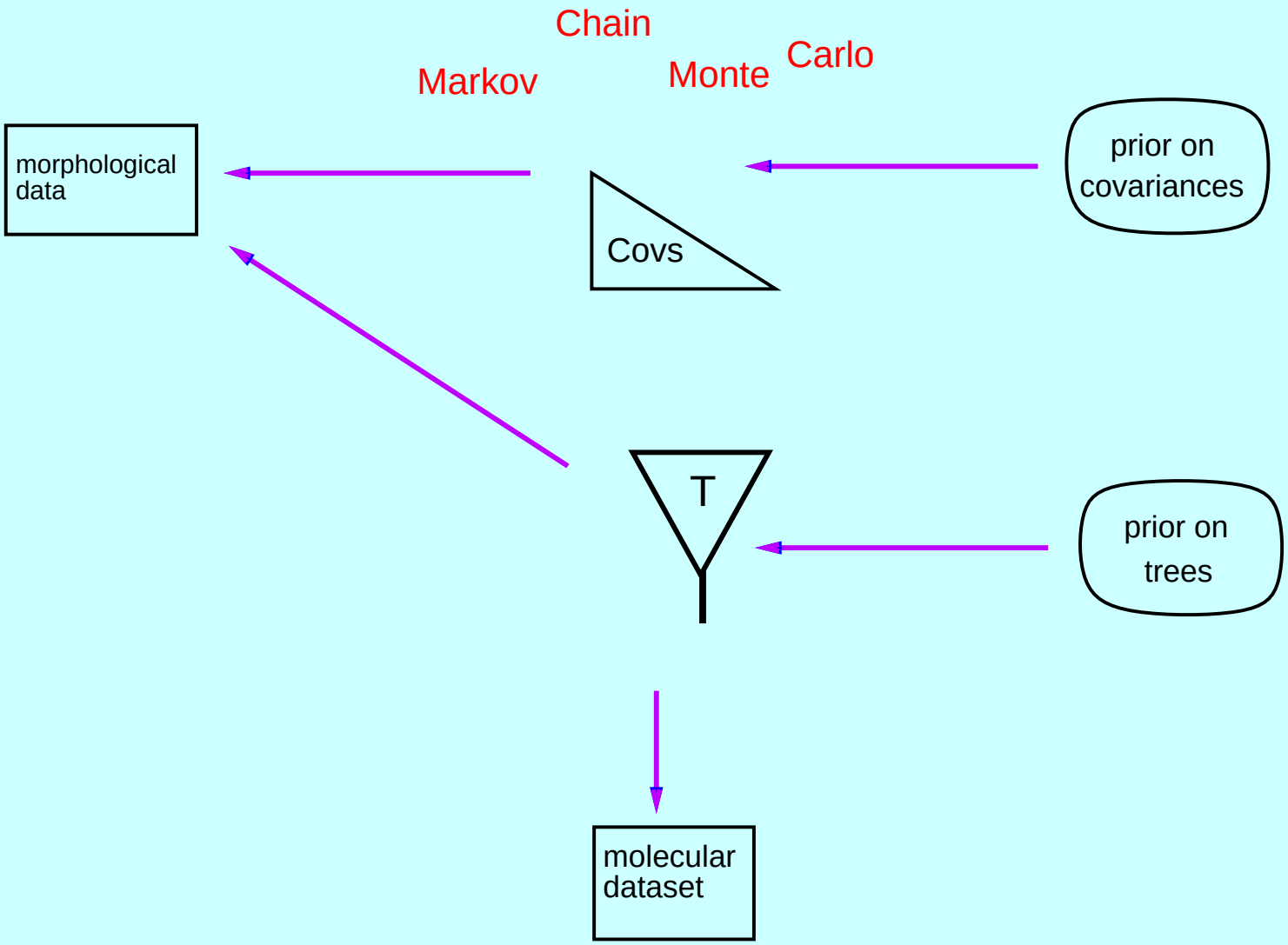
Markov Chain Monte Carlo



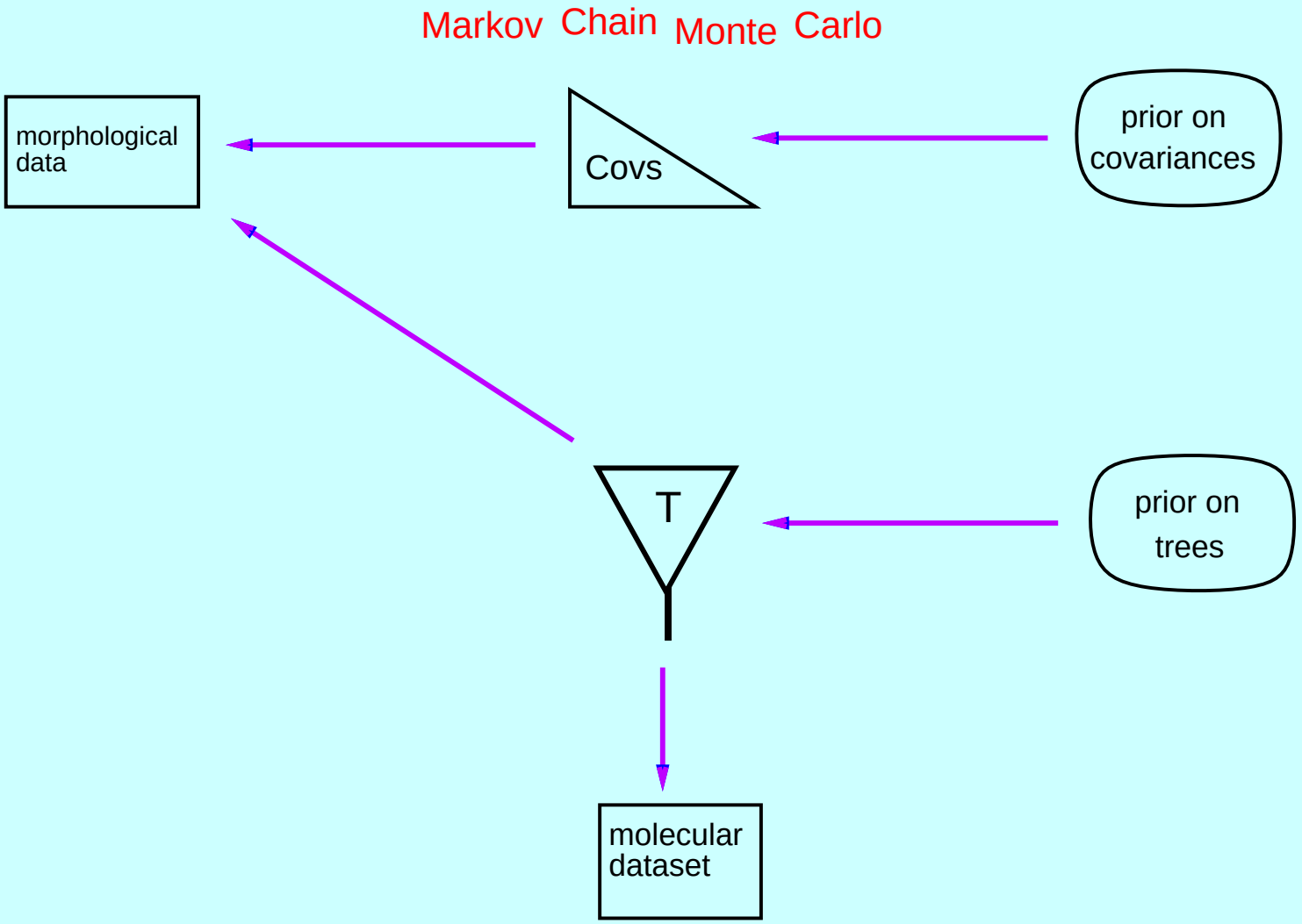
Bayesian MCMC



Bayesian MCMC



Bayesian MCMC



Some complications

- (As noted above) dealing with uncertainty about the phylogeny

Some complications

- (As noted above) dealing with uncertainty about the phylogeny
- Small sample size from species means their species means are uncertain. Must use a model with another level of variation – within-species phenotypic variation (Ricklefs and Starck, 1996; Ives et al., 2007; Felsenstein, 2008)

Some complications

- (As noted above) dealing with uncertainty about the phylogeny
- Small sample size from species means their species means are uncertain. Must use a model with another level of variation – within-species phenotypic variation (Ricklefs and Starck, 1996; Ives et al., 2007; Felsenstein, 2008)
- Rate of change of morphological characters need not be constant on the molecular tree branch lengths.

Some complications

- (As noted above) dealing with uncertainty about the phylogeny
- Small sample size from species means their species means are uncertain. Must use a model with another level of variation – within-species phenotypic variation (Ricklefs and Starck, 1996; Ives et al., 2007; Felsenstein, 2008)
- Rate of change of morphological characters need not be constant on the molecular tree branch lengths.
- Note – regressions involving contrasts should assume that they all have expectation zero. (They do because we don't know which lineage at a fork will move further to the right on the phenotype scale).

Some complications

- (As noted above) dealing with uncertainty about the phylogeny
- Small sample size from species means their species means are uncertain. Must use a model with another level of variation – within-species phenotypic variation (Ricklefs and Starck, 1996; Ives et al., 2007; Felsenstein, 2008)
- Rate of change of morphological characters need not be constant on the molecular tree branch lengths.
- Note – regressions involving contrasts should assume that they all have expectation zero. (They do because we don't know which lineage at a fork will move further to the right on the phenotype scale).
- How to infer the effect of an environmental variable when only its present-day values are known but not its values when the past changes were occurring? (note: regressing on the present-day values is generally **wrong**, see paper by Hansen and Bartoszek, *Systematic Biology*, 2012).

Some complications

- (As noted above) dealing with uncertainty about the phylogeny
- Small sample size from species means their species means are uncertain. Must use a model with another level of variation – within-species phenotypic variation (Ricklefs and Starck, 1996; Ives et al., 2007; Felsenstein, 2008)
- Rate of change of morphological characters need not be constant on the molecular tree branch lengths.
- Note – regressions involving contrasts should assume that they all have expectation zero. (They do because we don't know which lineage at a fork will move further to the right on the phenotype scale).
- How to infer the effect of an environmental variable when only its present-day values are known but not its values when the past changes were occurring? (note: regressing on the present-day values is generally **wrong**, see paper by Hansen and Bartoszek, *Systematic Biology*, 2012).
- Might be able to assume environment does Brownian motion and infer covariances. But this itself is a somewhat arbitrary assumption.

Poor inference of covariation – what to do with that?

- Covariances are hard to infer with only (say) 50 species sampled

Poor inference of covariation – what to do with that?

- Covariances are hard to infer with only (say) 50 species sampled
- ... particularly if they samples are not independent but on a tree

Poor inference of covariation – what to do with that?

- Covariances are hard to infer with only (say) 50 species sampled
- ... particularly if they samples are not independent but on a tree
- ... particularly if the quantitative characters are thresholded

Poor inference of covariation – what to do with that?

- Covariances are hard to infer with only (say) 50 species sampled
- ... particularly if they samples are not independent but on a tree
- ... particularly if the quantitative characters are thresholded
- How do we propagate the resulting uncertainty when biologists want “fly on the wall” certainty?

Poor inference of covariation – what to do with that?

- Covariances are hard to infer with only (say) 50 species sampled
- ... particularly if they samples are not independent but on a tree
- ... particularly if the quantitative characters are thresholded
- How do we propagate the resulting uncertainty when biologists want “fly on the wall” certainty?
- Expanding to more species may put the model at risk

Poor inference of covariation – what to do with that?

- Covariances are hard to infer with only (say) 50 species sampled
- ... particularly if they samples are not independent but on a tree
- ... particularly if the quantitative characters are thresholded
- How do we propagate the resulting uncertainty when biologists want “fly on the wall” certainty?
- Expanding to more species may put the model at risk
- Expanding to more characters just adds new parameters to estimate

References for phylogenetic comparative methods

Felsenstein, J. 1985. Phylogenies and the comparative method. *American Naturalist* **125:** 1–5. **[Introduces the contrasts method]**

Felsenstein, J. 1988. Phylogenies and quantitative characters. *Annual Review of Ecology and Systematics* **[Suggests using bootstrapping to correct comparative methods for uncertainty about the phylogeny** **19:** 445–471.

Harvey, P. H. and M. D. Pagel. 1991. *The Comparative Method in Evolutionary Biology.* Oxford University Press, Oxford. **[The major book introducing statistical phylogenetic comparative methods]**

Grafen, A. 1989. The phylogenetic regression. *Philosophical Transactions of the Royal Society of London, Series B* **326:** 119–157. **[Using generalized least squares to evaluate the likelihood for Brownian Motion phylogenies and do comparative methods analysis, without the contrasts methods. In the simplest case, is exactly equivalent to the contrasts method. Discusses ways of coping with unresolved parts of the phylogeny and with varying evolutionary rates.]**

References, continued

Ricklefs, R. E. and J. M. Starck. 1996. Applications of phylogenetically independent contrasts: A mixed progress report. *Oikos* **77:** 167–172. **[Pointing put that small sample size within species is a problem for comparative methods]**

Ives, A. R., P. E. Midford, and T. Garland. 2007. Within-species variation and measurement error in phylogenetic comparative methods. *Systematic Biology* **56:** 252-270. **[Taking small sample size into account when we know the within-species phenotypic covariances]**

Hansen, T. F., and K. Bartoszek. 2012. Interpreting the evolutionary regression: the interplay between observational and biological errors in phylogenetic comparative studies. *Systematic Biology* **61(3):** 413 âĂŞ- 425.

Felsenstein, J. 2008 Comparative methods with sampling error and within-species variation: contrasts revisited and revised. *American Naturalist* **171:** 713–725. **[Inferring both between=species evolutionary covariances and within-species phenotypic variation]**

Felsenstein, J. 2004. *Inferring Phylogenies.* Sinauer Associates, Sunderland, Massachusetts. **Mentions this model and also sample size issues in contrasts method.**