BI694

Bioinformatics & Phylogenetics

Phylogenetic Analysis using PAUP

The Nexus file format and PAUP command block

→ example using a small dataset

Phylogenetic Analysis using PAUP

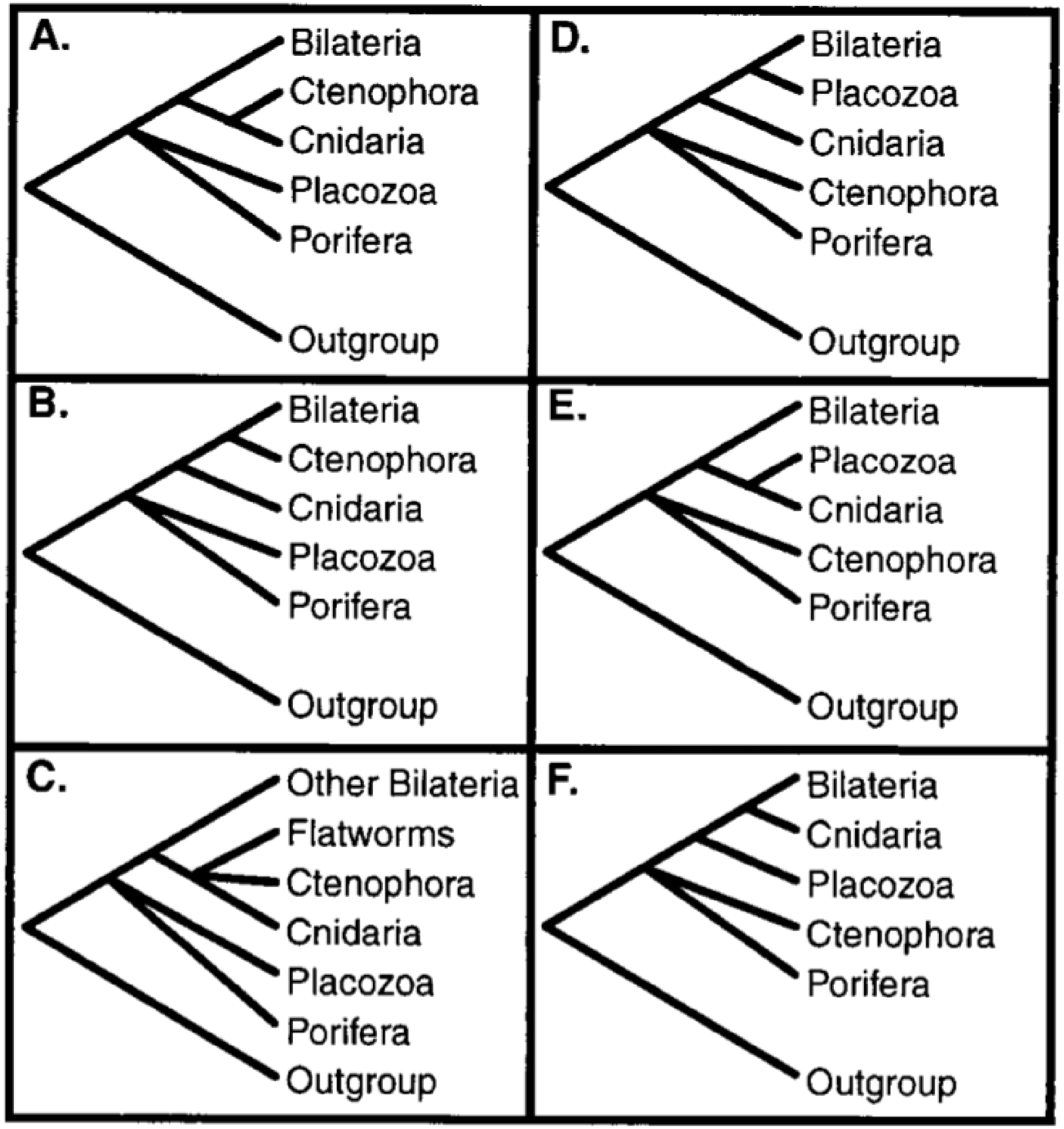


FIG. 1. Six alternative hypotheses for the origin of the Bilateria.

Collins (1998) PNAS

Phylogenetic Analysis using PAUP

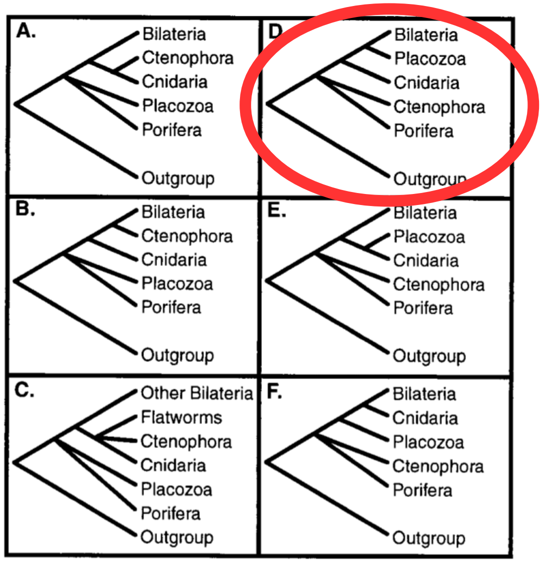
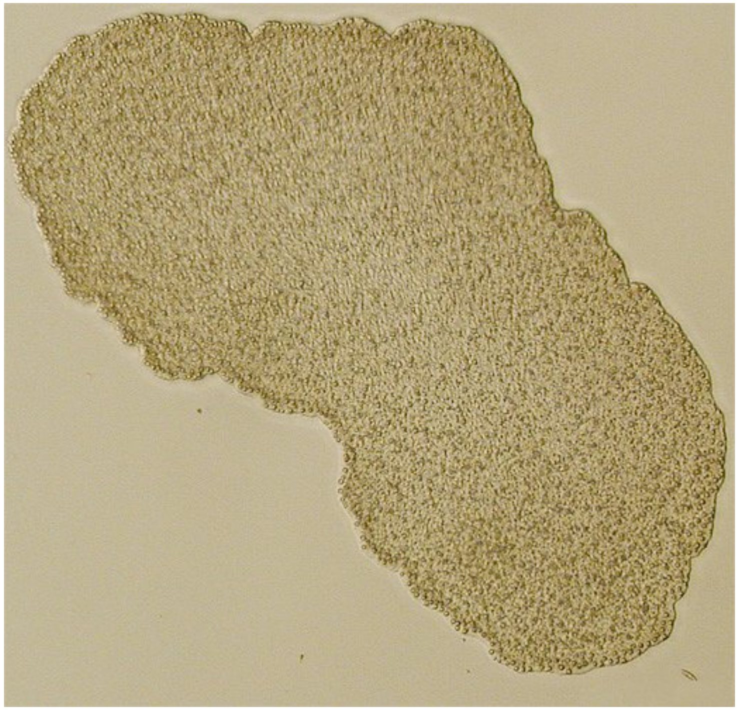
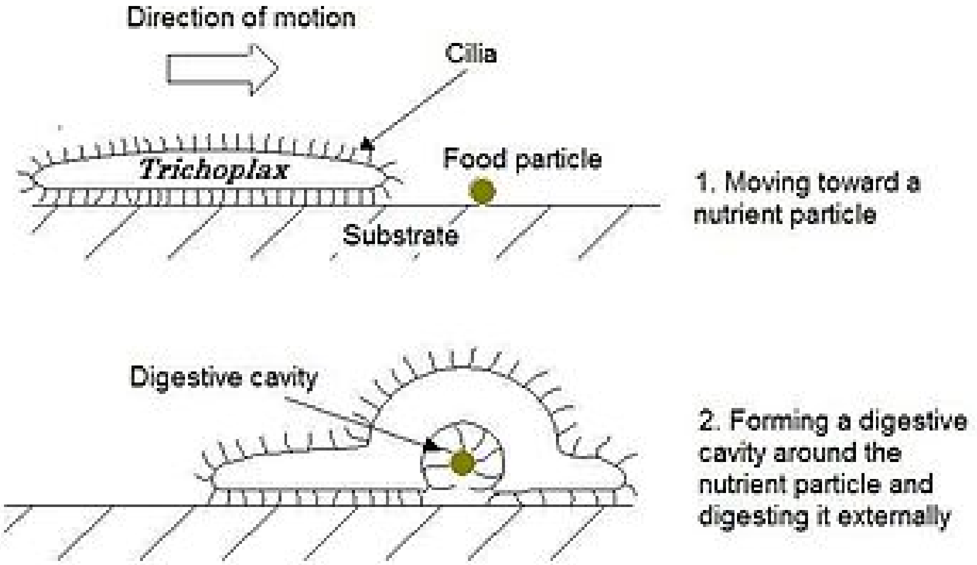


FIG. 1. Six alternative hypotheses for the origin of the Bilateria.




Direction of motion
Cilia
Trichoplax
Food particle
Substrate
1. Moving toward a nutrient particle
Digestive cavity
2. Forming a digestive cavity around the nutrient particle and digesting it externally

Collins (1998) PNAS

Phylogenetic Analysis using PAUP

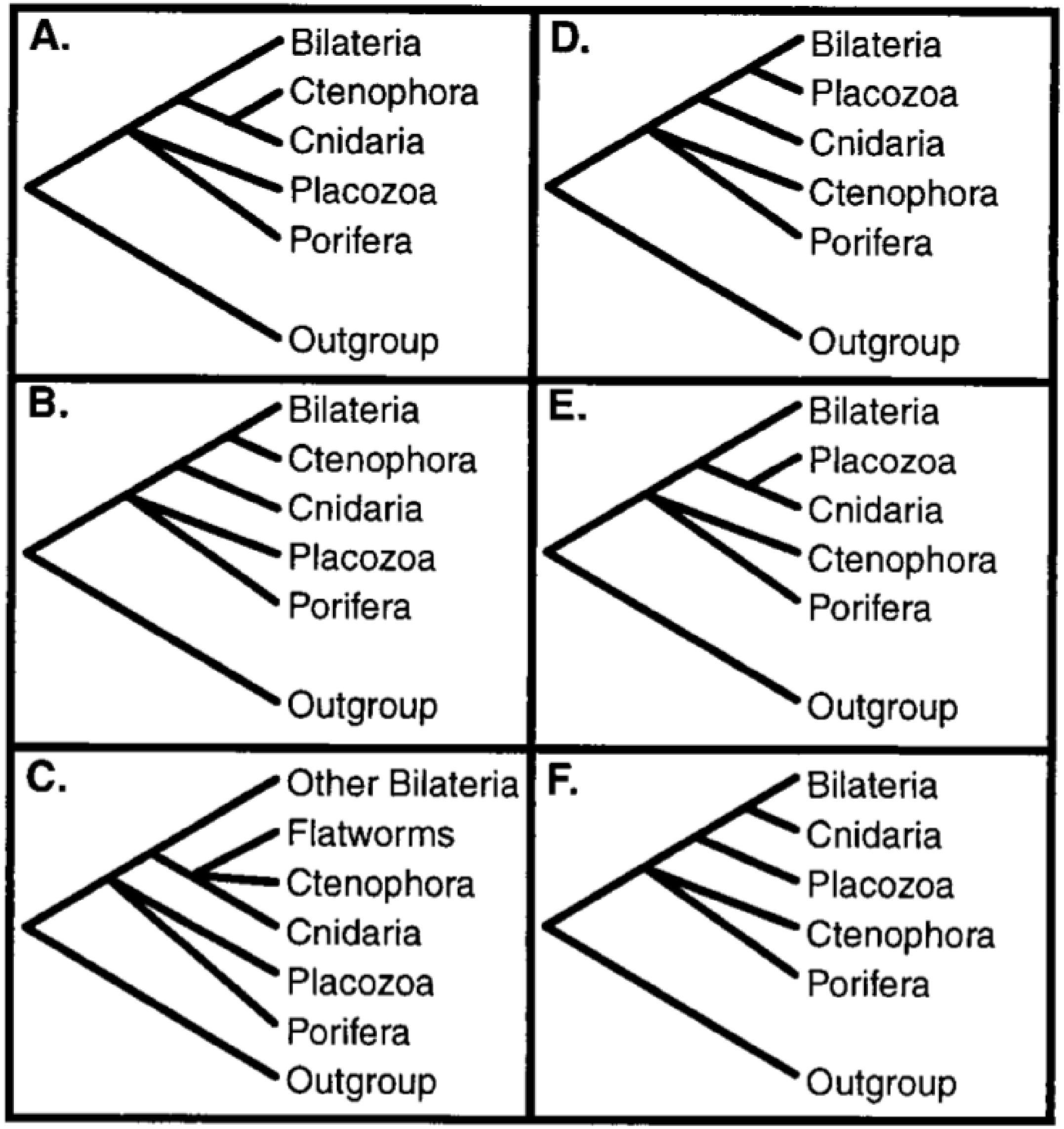


FIG. 1. Six alternative hypotheses for the origin of the Bilateria.

We have a best tree but how confident are we in the results?

Collins (1998) PNAS

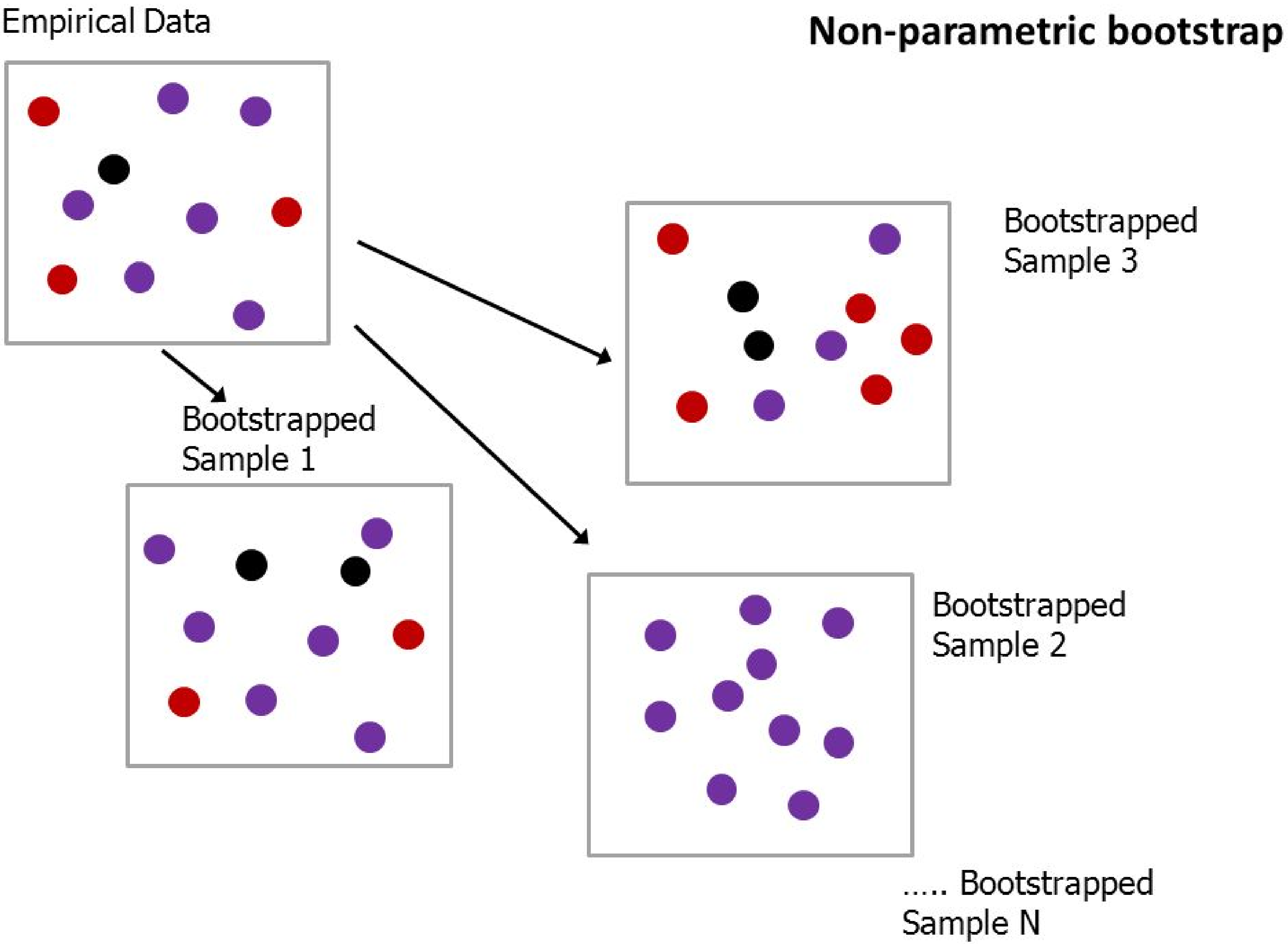
Evaluating a tree: the bootstrap

Data in hand are probably somewhat like the underlying universe of possible data sets

We can simulate other possible data sets by drawing randomly from the data already collected

This seems to get information from nothing, lifting yourself off the ground by your bootstraps...

Evaluating a tree: the bootstrap



Non-parametric bootstrap

Empirical Data

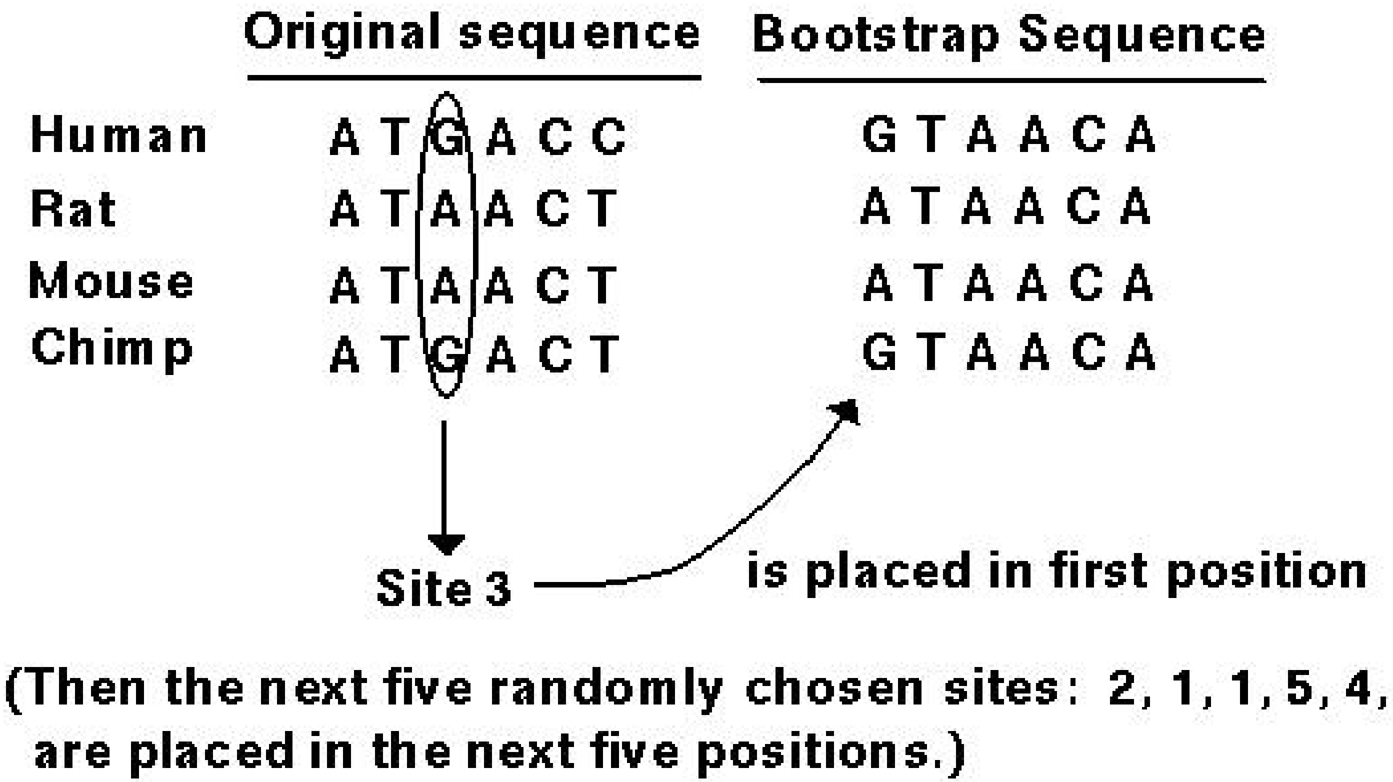
Bootstrapped Sample 1

Bootstrapped Sample 3

Bootstrapped Sample 2

….. Bootstrapped Sample N

Evaluating a tree: the bootstrap

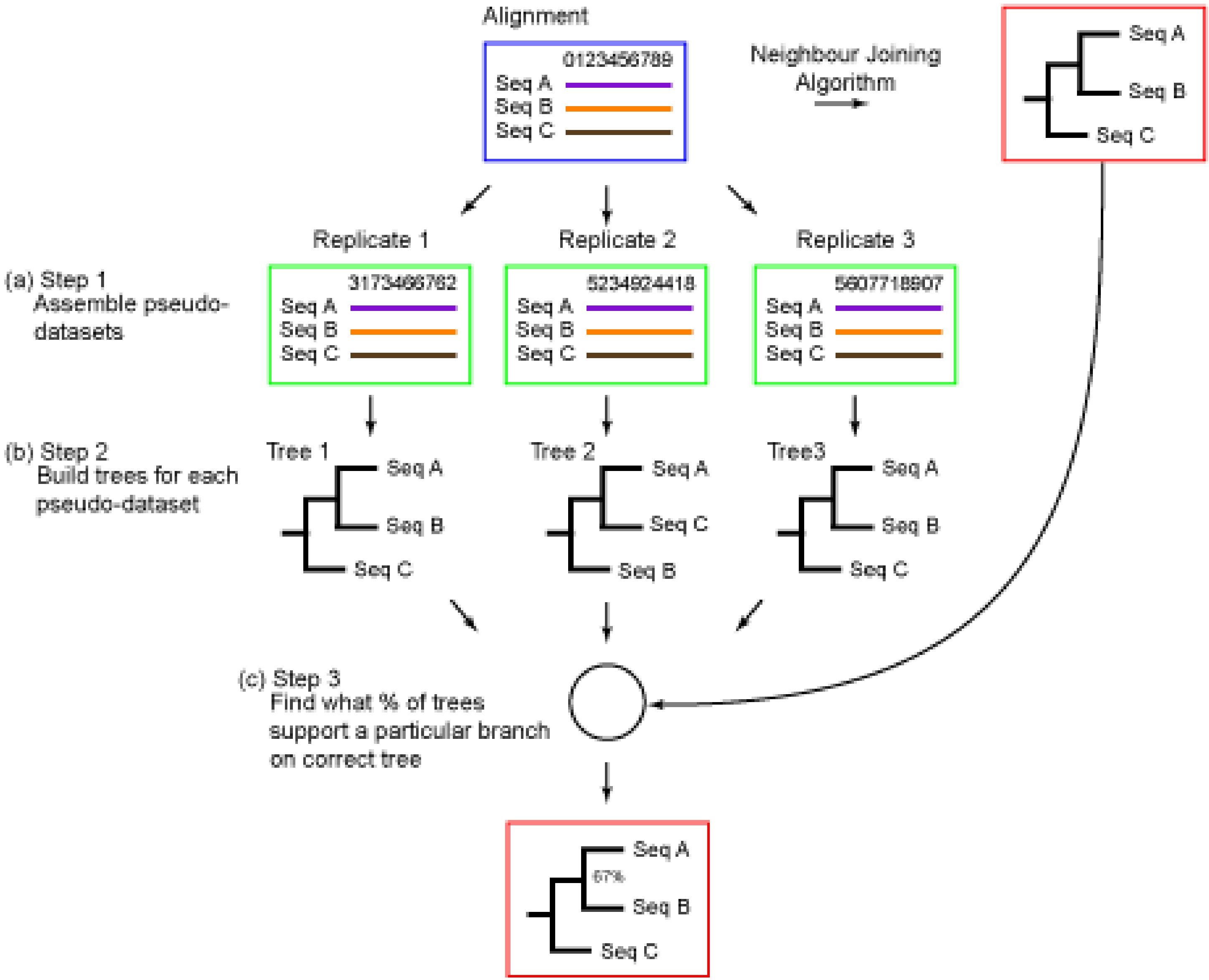


	Original sequence	Bootstrap Sequence
Human	A T G A C C	G T A A C A
Rat	A T A A C T	A T A A C A
Mouse	A T A A C T	A T A A C A
Chimp	A T G A C T	G T A A C A

Site 3 is placed in first position

(Then the next five randomly chosen sites: 2, 1, 1, 5, 4, are placed in the next five positions.)

Evaluating a tree: the bootstrap



(a) Step 1
Assemble pseudo-datasets

(b) Step 2
Build trees for each pseudo-dataset

(c) Step 3
Find what % of trees support a particular branch on correct tree

How about maximum likelihood?

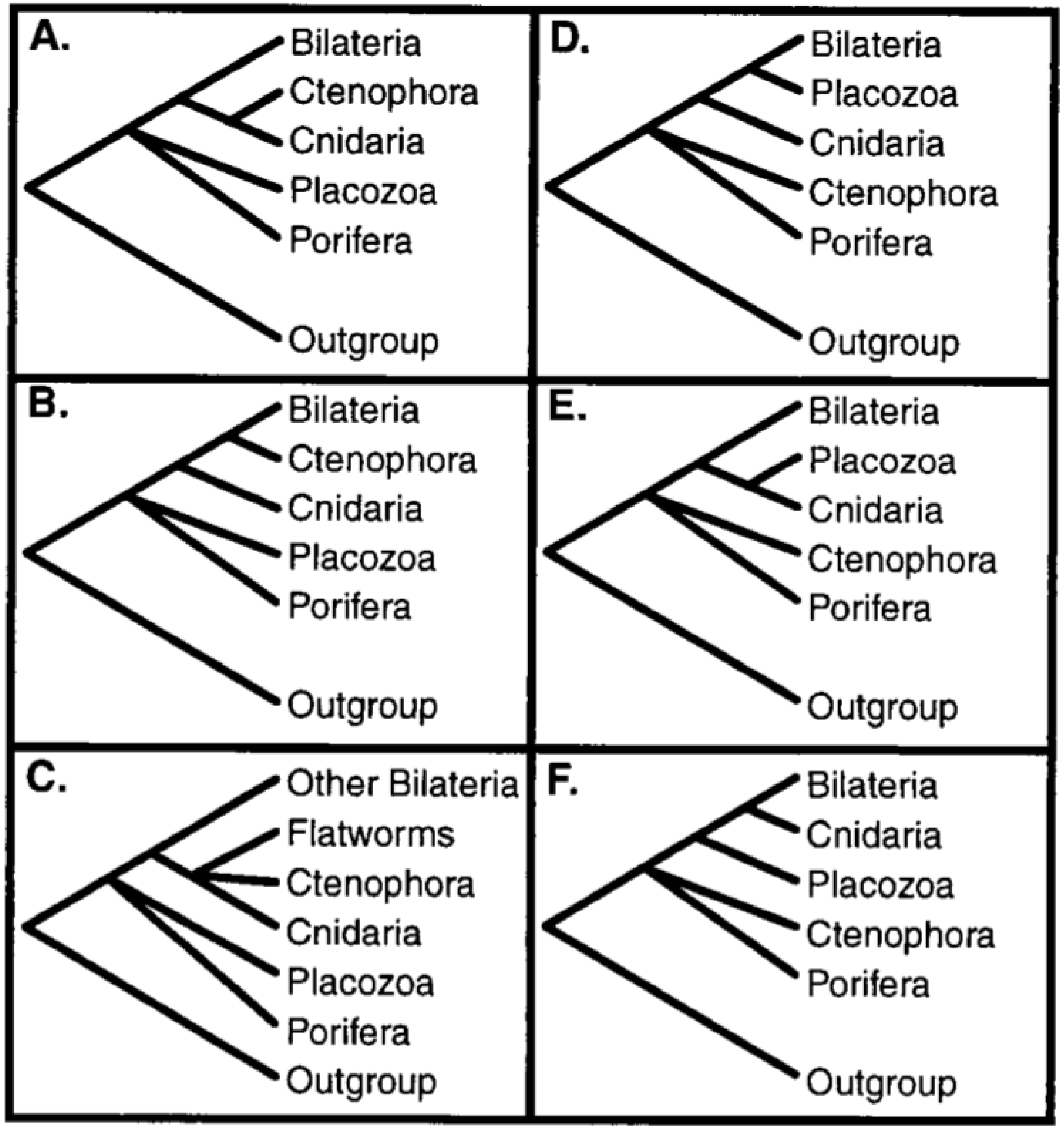


FIG. 1. Six alternative hypotheses for the origin of the Bilateria.

Collins (1998) PNAS

Testing Model Fit

Pick a more complex model only when the gain in likelihood is higher than expected if the simpler model were true.

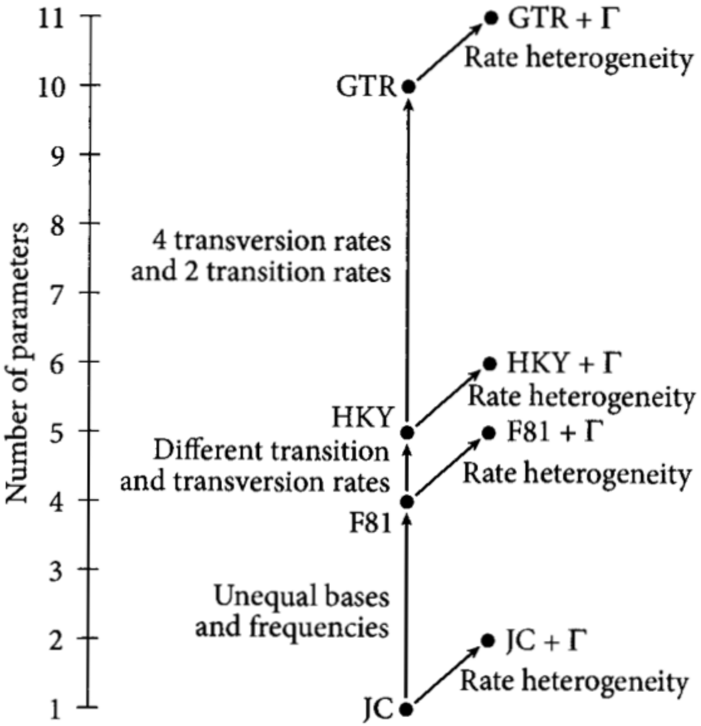
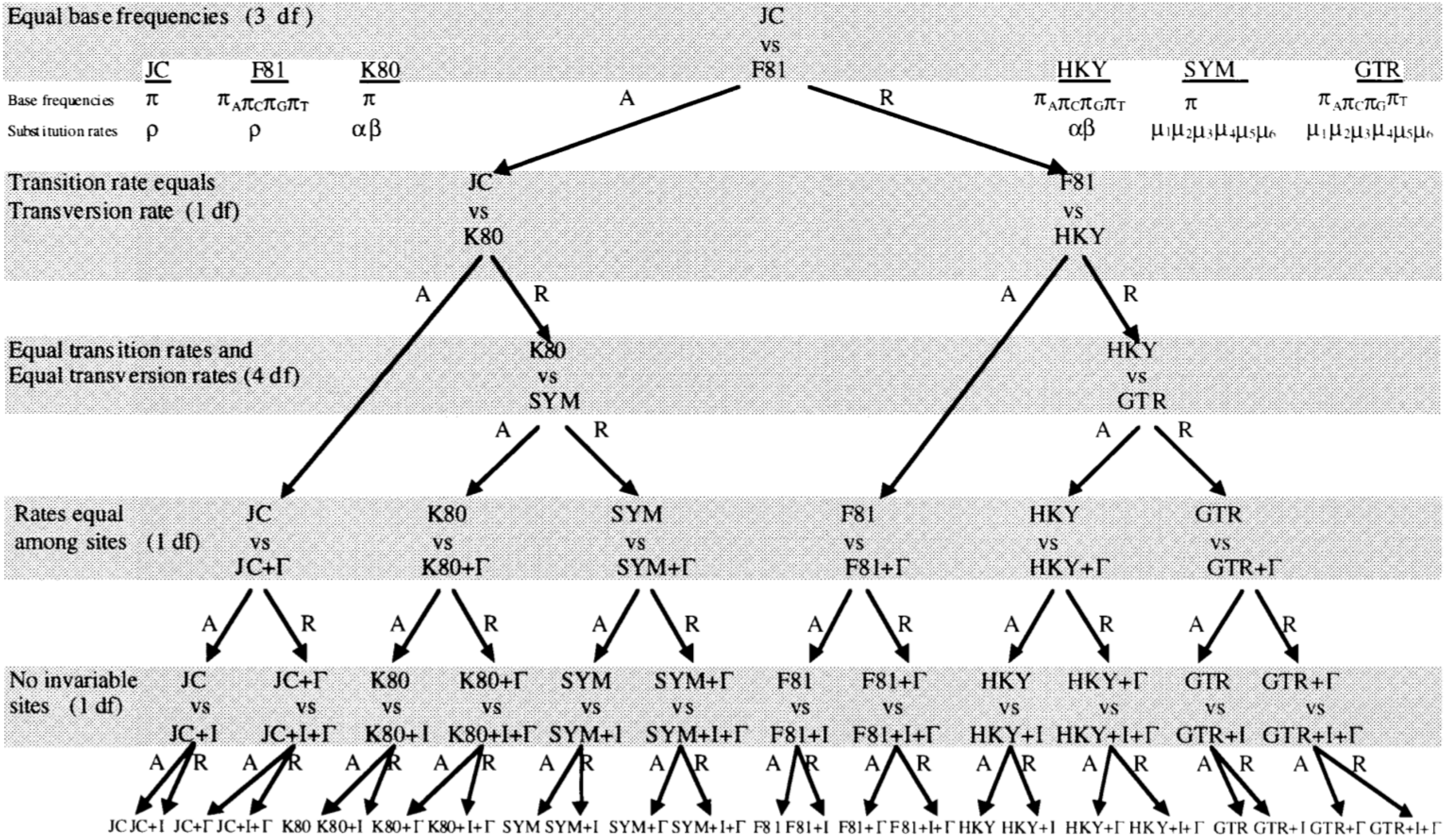


FIGURE 8.10 Depiction of the relationship between some commonly used models of evolution. Any two models that are connected by arrows that proceed in the same direction are nested: the simpler model (closer to the bottom of the chart) contains a subset of the free parameters of the more complex model. The axis on the left shows the number of free rate parameters in the model. The figure assumes that site-to-site rate heterogeneity is modeled using a discrete approximation to a gamma (Γ) distribution, which adds one free parameter to the model.

Testing Model Fit



	JC	F81	K80	HKY	SYM	GTR
Base frequencies	π	$\pi_A\pi_C\pi_G\pi_T$	π	$\pi_A\pi_C\pi_G\pi_T$	π	$\pi_A\pi_C\pi_G\pi_T$
Substitution rates	ρ	ρ	$\alpha\beta$	$\alpha\beta$	$\mu_1\mu_2\mu_3\mu_4\mu_5\mu_6$	$\mu_1\mu_2\mu_3\mu_4\mu_5\mu_6$

Equal base frequencies (3 df)

JC vs F81

A R

Transition rate equals Transversion rate (1 df)

JC vs K80 | F81 vs HKY

A R A R

Equal transition rates and Equal transversion rates (4 df)

K80 vs SYM | HKY vs GTR

A R A R

Rates equal among sites (1 df)

JC vs JC+Γ | K80 vs K80+Γ | SYM vs SYM+Γ | F81 vs F81+Γ | HKY vs HKY+Γ | GTR vs GTR+Γ

A R A R A R A R A R A R

No invariable sites (1 df)

JC vs JC+I | JC+Γ vs JC+I+Γ | K80 vs K80+I | K80+Γ vs K80+I+Γ | SYM vs SYM+I | SYM+Γ vs SYM+I+Γ | F81 vs F81+I | F81+Γ vs F81+I+Γ | HKY vs HKY+I | HKY+Γ vs HKY+I+Γ | GTR vs GTR+I | GTR+Γ vs GTR+I+Γ

A R A R A R A R A R A R A R A R A R A R A R A R

JC JC+I JC+Γ JC+I+Γ K80 K80+I K80+Γ K80+I+Γ SYM SYM+I SYM+Γ SYM+I+Γ F81 F81+I F81+Γ F81+I+Γ HKY HKY+I HKY+Γ HKY+I+Γ GTR GTR+I GTR+Γ GTR+I+Γ

Posada & Crandall (1998)

Testing Model Fit

Likelihood Ratio:

Chi-square can be used as approximation for likelihood ratio test

$$\Lambda = \frac{\max\ [L_0\ (Null\ \ Model \mid Data)]}{\max\ [L_1\ (Alternative\ \ Model \mid Data)]}$$

Testing Model Fit

Akaike Information Criterion (AIC)

Suppose data generated by some unknown process **f**. Two candidate models represent **f**: **g1** and **g2**.

If we knew **f**, then we could find the
information lost from using **g1**; similarly, the information lost from using **g2** to represent **f** could be found. We would then choose the candidate model that minimized the information loss.

We do not know **f**. We can estimate how much more (or less) information is lost by g1 than by g2; ie, which model fits the data better (Akaike 1974)

$$\mathrm{AIC} = 2k - 2\ln(\hat{L})$$

k is the number of estimated parameters; ln(L) the log likelihood

Smaller AIC indicates better fit of the model to the data

Testing Model Fit

MrModeltest:

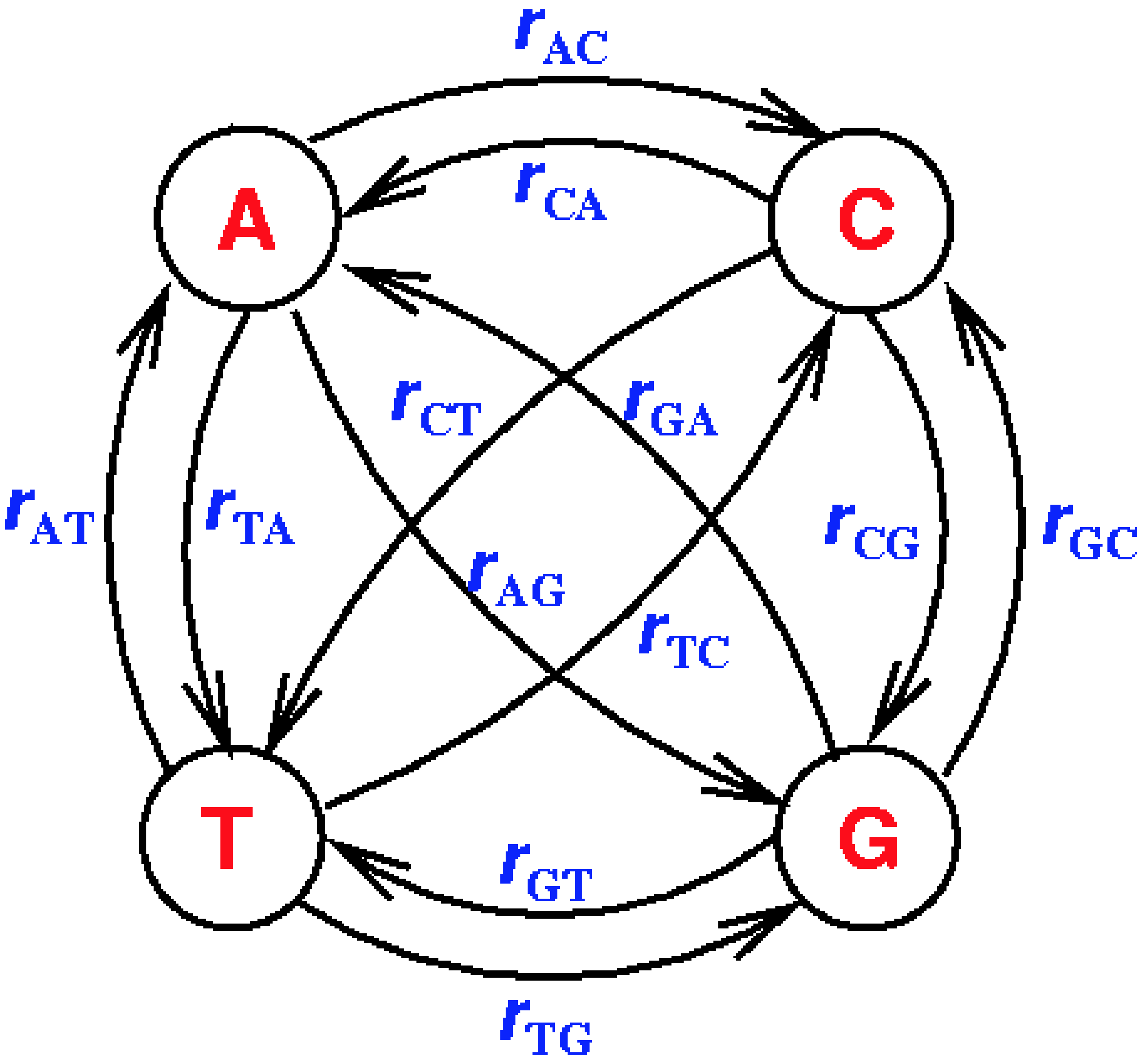
Execute your data file in PAUP*.

Execute the file MrModelblock in PAUP*. A file called mrmodel.scores will appear in the current directory.

This file is the input for mrmodeltest2.

```
mrmodeltest2 < mrmodel.scores > out
```

Testing Model Fit



r_{AC}
r_{CA}
A
C
r_{CT}
r_{GA}
r_{AT}
r_{TA}
r_{AG}
r_{CG}
r_{GC}
r_{TC}
T
G
r_{GT}
r_{TG}