Supplemental Material: Annu.Rev.Ecol. Evol. Syst. 2005. 36:445-466
doi: 10.1146/annurev.ecolsys.36.102003.152633
*MODEL SELECTION IN PHYLOGENETICS*
Sullivan and Joyce, 2005

# SUPPLEMENTARY MATERIAL: MODEL SELECTION IN PHYLOGENETICS

Jack Sullivan[1,2,4] & Paul Joyce[2,3]

## A1. THE NEED FOR THOUGHTFUL MODEL SELECTION

There's a huge literature on model selection in other fields, and this highlights the need for thoughtful model choice. To always assume that the most complex model available has the best predictive ability is logically flawed, as illustrated by a simple example from linear regression. In the scatter plot in Figure A1a, eleven points were simulated using the equation $y = 3x + 2 + \varepsilon$, where $\varepsilon$ is a normally distributed random error with mean 0 and variance 2. The input variable $x$ ranges from 0 - 5 and points are equally spaced at one-half unit apart. Using standard linear regression, the best- fit line to the data is $\hat{y} = 2.86x + 1.92$, and prediction of $y$ when $x = 5.5$ using the fitted model would yield y = 17.65, which is well within the predicted margin of error of the expected value of $y$ [given by $E(y) = 3(5.5) + 2 = 18.5$]. A researcher ignorant of the true model may be uncomfortable with such a simple explanation, so instead might fit the data to a quadratic model. The best-fit quadratic model is $\hat{y} = -.01x^2 + 2.91x + 1.88$. While the quadratic model fits slightly better, the quadratic term (-.01 $x^2$) is quite small and the quadratic model provides nearly the same results as the linear model, and one might argue to use the more complex model since it recovers the simple model when the simple model is true. If we take this argument to its logical conclusion then we should consider the most parameter-rich complex model available to describe the relationship between the response variable $y$ and the predictor $x$. For these data, this would be an 11-degree polynomial with equation:

$$\hat{y} = 1.991 - 4.768\,x - 41.974\,x^2 + 214.873\,x^3 - 352.932\,x^4 + 301.407\,x^5 - 151.477\,x^6 + 46.391\,x^7 + 6.391\,x^8 - 8.508\,x^9 + 0.858\,x^{10} - 0.037\,x^{11}.$$

The above model fits the scatter plot data exactly, without error (Fig. A1b). However, since the number of parameters matches the size of the data (both are 11), there are no degrees of freedom available to assess the error and the model has no predictive power. For example, substituting $x = 5.5$ into the above equations gives a predicted value of $y$ = -382.40, a nonsensical result that is nowhere near the correct prediction of 18.5.

Supplemental Material: Annu.Rev.Ecol. Evol. Syst. 2005. 36:445-466
doi: 10.1146/annurev.ecolsys.36.102003.152633
*MODEL SELECTION IN PHYLOGENETICS*
Sullivan and Joyce, 2005

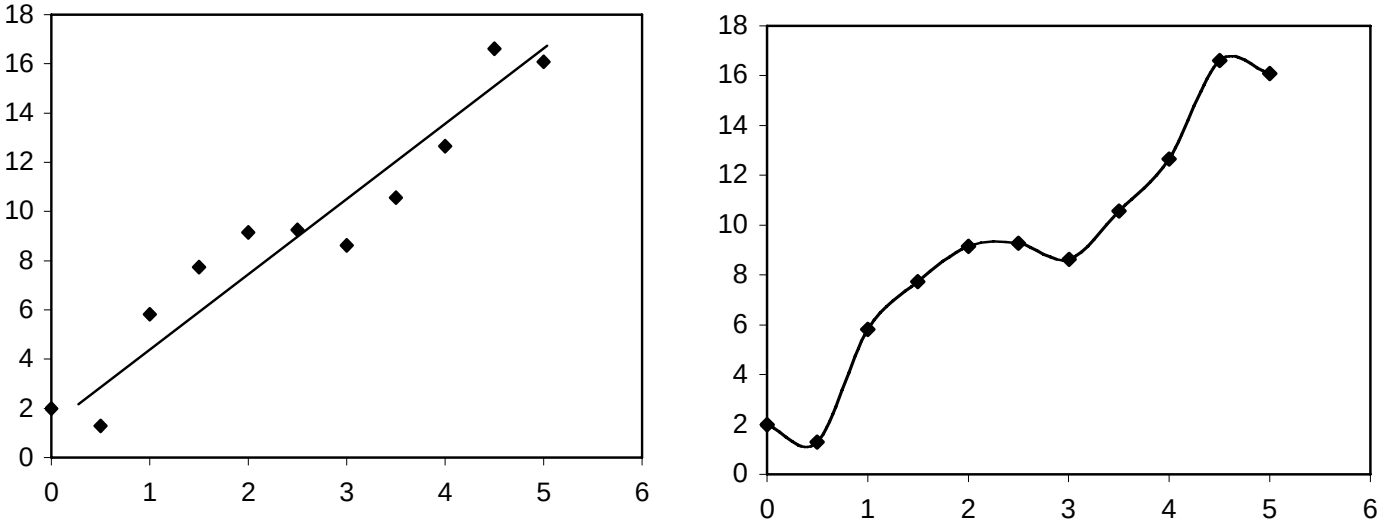


Figure A1. An example of overfitting. The 11 data points in the scatter plots were generated under a linear model (a). However, an 11$^{th}$ order polymonial fits the data perfectly (b), although this model has no predictive ability; there are as many parameters as data points.

As computer power increase, more complex parameter-rich models can be considered in phylogenetics and this might be expected to necessarily lead to better inferences. We disagree; one must, at the very least, consider the trade-off between degrees of freedom and model fit. Given the ability to partition data extremely finely and apply complex models to each partition independently, the number of parameters in the global model may begin to approach the number of independent data points. It may therefore be appropriate to choose the simpler model because we may reach a point where the improvement in fit of the complex model no longer compensates for the loss of degrees of freedom. Model-selection approach should address this trade-off.

## A2. DECISION THEORY UNIFYING MODEL SELECTION

Although decision theory (DT) is often viewed in the context of Bayesian statistics, it also provides a unified framework for statistical inference that can accommodate both the Bayesian and frequentist statistical philosophies. Furthermore, the well-established model-selection criteria described in the text (AIC, BIC, LRT) naturally arise in a decision theoretic context. We can demonstrate DT by the following; suppose one is playing a game, the object of which is to choose an evolutionary model. At the end of the game the true state of nature will be revealed

and penalties will be assessed according to how far off one's guess is from the truth. Fortunately, the game is not played in complete ignorance because the data provide information that may lead to a reasonable choice that receives a low penalty. This penalty is referred to as the *loss function.* More precisely, if there are $M_1, M_2, \cdots, M_K$ evolutionary scenarios (or models) under review, we can denote the true state of nature (i.e., the true model) by $M_T$. Then if one chooses model $i$, the loss or penalty assessed at the end of the game will be denoted by $l(M_i, M_T)$. Now we can develop a data based decision rule for model choice, $d(D)$, for a data set denoted by $D$. If we observe $D$, which leads to model choice $d(D)$, then our penalty will be $l(d(D), M_T)$. For some data sets, $d(D)$ may choose models with a small penalty, but for other data sets the penalty may be larger. In order to assess the overall effectiveness of a proposed decision rule, we calculate the *risk*, which is defined to be the expected loss under the true model, denoted by $R_d(M_T) = E\left(l(d(D), M_T) \middle| M_T\right)$. Note that $R_d(M_T)$ can never be directly calculated without knowing the true state of nature and the true model is never actually revealed. The best we can do is use the data to estimate the risk function, (or in some cases, a portion of the risk function), and then choose the model with low estimated risk. Therefore, there are two relevant aspects of decision theory that will determine the model-selection strategy.

1. *The form of the loss function.* Since we argue that hidden in the background of any model-selection criterion is a loss function, one can argue for or against a particular method based on the relative merits of the different loss functions.
2. *The method for estimating risk.* Bayesians and frequentists will have different approaches to estimating risk, and it's here that the differences in the two philosophies toward statistical inference will be elucidate most clearly.

## A2A. LRT & MINIMUM RISK

Under the LRT approach, where the simple model is denoted by $M_s$ and the more complex model by $M_c$, we can consider a simple binary loss function, where we lose one point for choosing $M_s$ when $M_c$ is true and one point for choosing $M_c$ when $M_s$ is true. There is no loss for correct choice. Note that for any decision rule $d$ under this loss function, the risk associated with $d$ when $M_s$ is assumed to be the true model is $R_d(M_s) = pr(d(D) = M_c \mid M_s)$ which is commonly referred to as the probability of a type I error. Similarly, the risk associated with the $d$ when the complex model is the true model is the probability of a type II error. We must first

acknowledge that there is no uniformly optimal decision strategy, even in this simple case. The reason for this is simple; the naïve strategy of always choosing the simple model (regardless of the data) will be the best strategy when the simple model is correct, so it is impossible to devise a data based strategy that performs best in all circumstances. Neyman & Pearson (cited in Bickel and Doksum 2001), who invented the LRT, recognized this dilemma and decided to restrict their decision rules to a subclass and use the decision rule that is optimal among that class. The Neyman-Pearson rule is to fix the risk when the simple model is assumed to be true (i.e., fix the Type I error) and use the decision rule within this class that minimizes the risk when the complex model is true (i.e., minimizes the Type II error); they proved that the LRT is the optimal decision rule for this restrictive class. Note that by requiring the risk associated with the simple model to be low, the procedure is biased toward simple models; we therefore fail to reject the simple model unless there is substantial evidence to the contrary. However, the LRT will assess that evidence in the most efficient manner possible and should therefore be best at detecting departures from the null hypothesis (or evidence against the simple model) among all decision rules in this class.

Unfortunately, extending this optimality criterion beyond the two-model case is not obvious. However, by viewing likelihood ratio in a decision theoretic context we gain a better appreciation for the rationale behind the use of LRTs. Its optimality property derives from assuming that one of the models under review is true and the chance of incorrectly picking the simple model is set in advance and is small. The LRT detects departures from the null optimally among this set of conservative decision rules, and thus fits well into the decision theoretic framework.

### A2B. THE AIC APPROACH TO MINIMIZING RISK

Rather than compare several models where we assume that the true model is among the models being considered, the AIC assumes that the data were generated via a stochastic mechanism under a complex model denoted by $M_T$, where $M_T$ is outside the candidate set. The AIC is based on the Kullback-Leibler (K-L) Information function. We view this as a loss function and define the loss of choosing $M_i$ when $M_T$ is correct as

$$l(M_i, M_T) = E\left( \ln \frac{pr(D \mid M_T)}{pr(D \mid M_i)} \middle| M_T \right).$$

This loss function comes from information theory and is viewed as the amount of information lost when $M_i$ is used to approximate $M_T$ (Burnham and Anderson 2002). We can rewrite the K-L loss function by

$$l(M_i, M_T) = E\left(\ln pr(D \mid M_T)\middle| M_T\right) - E\left(\ln pr(D \mid M_i)\middle| M_T\right)$$

and note that, in order to minimize loss associated with the decision rule, we need only consider minimizing $-E\left(\ln pr(D \mid M_i)\middle| M_T\right)$. To recognize explicitly the fact that each model under review contains a set of unknown parameters $\boldsymbol{\theta}_i = (\theta_{1i}, \theta_{2i}, \cdots, \theta_{k_i i})$, we will now write $pr(D \mid M_i)$ as $pr(D \mid M_i, \boldsymbol{\theta}_i)$. For a sufficiently large sample and under the assumption that all models under review are sufficiently 'close' to the true model $M_T$, we can further approximate $-E\left(\ln pr(D \mid M_i, \boldsymbol{\theta}_i)\middle| M_T\right)$ by $-E\left(\ln pr(D \mid M_i, \boldsymbol{\theta}_i)\middle| M_i, \boldsymbol{\theta}_i\right)$, which can be further approximated by $-\ln(pr(D \mid M_i, \hat{\boldsymbol{\theta}}_i) + k_i$, where $k_i$ is the number of parameters in the model and $\hat{\boldsymbol{\theta}}_i$ is the maximum likelihood estimate of $\boldsymbol{\theta}_i$. For historical reasons we multiply by 2 to get the AIC score as

$$AIC_i = -2\ln(pr(D \mid M_i, \hat{\boldsymbol{\theta}}_i) + 2k_i .$$

Thus, the AIC conforms quite precisely to decision theory, with the risk function assessing the loss in KL information.

### A2C. BAYESIAN MODEL SELECTION AND RISK

The Bayesian perspective provides much more flexibility for developing model-selection criteria. Because the risk associated with any decision rule depends on true state of nature, it's impossible to develop an optimality criterion without either restricting the class of decision rules allowed (as in LRTs) or making certain approximations about the closeness of the models under review to the true model (as in AIC). In the Bayesian framework, we place priors on each of the models under review $pr(M_1), pr(M_2), \cdots pr(M_k)$, and then calculate the average risk associated with a decision rule $d$. This leads to the Bayes risk $B(d)$ defined by

$$\begin{aligned} B(d) &= R_d(M_1)\, pr(M_1) + \cdots + R_d(M_K)\, pr(M_k) \\ &= E_M\left(E_{D|M}(l(d(D), M) \mid M)\right) \\ &= E_D\left(E_{M|D}(l(d(D), M) \mid D)\right) \end{aligned}$$

Supplemental Material: Annu.Rev.Ecol. Evol. Syst. 2005. 36:445-466
doi: 10.1146/annurev.ecolsys.36.102003.152633
*MODEL SELECTION IN PHYLOGENETICS*
Sullivan and Joyce, 2005

The Bayes decision rule is the rule $d$ that makes $B(d)$ minimum. For each data set, $D$, we see from the above equations that if we choose $d$ so as to minimize $E_{M|D}(l(d(D), M_i) \mid D)$, then this will be the rule that minimizes the Bayes risk. We refer to $E_{M|D}(l(d(D), M) \mid D)$ as the posterior loss. For example, if the loss function is binary, that is $l(M_i, M_j) = 1$ if model *j* is correct and model *i* is chosen and $l(M_i, M_i) = 0$, then the posterior loss of the decision to choose model $i$ ($d(D) = M_i$) is given by $1 - pr(M_i \mid D)$. Thus, minimizing the posterior loss is the same as maximizing the posterior probability; Bayesian model selection is a DT approach, with a binary loss function.

Under such a binary loss function, the Bayes rule is defined by calculating the posterior probability of the model (given the data) and choosing the model with the maximum posterior probability. Each model is given a weight in light of the data and the model is chosen based on which is most likely in light of the data. However, each model has a number of parameters. If we assume that q($\boldsymbol{\theta}_i$) is the prior distribution on the parameters associated with model $M_i$ then, following Bayes Theorem

$$pr(M_i \mid D) = \frac{\int pr(D \mid M_i, \boldsymbol{\theta}_i) q(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i \, pr(M_i)}{\sum_i \int pr(D \mid M_i, \boldsymbol{\theta}_i) q(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i \, pr(M_i)}.$$

Note that the denominator is a fixed constant and the Bayes decision rule requires only that the numerator be maximized. If we assume a uniform prior on models, that is $pr(M_i) = 1/k_i$, then the above reduces to maximizing $\int pr(D \mid M_i, \boldsymbol{\theta}_i) q(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i$. The approach suggested by Huelsenbeck et.al. (2004) estimates this posterior probability by using an MCMC sampler, and this approach can be viewed as a Bayesian model selection rule under a 0-1 loss function. Using a different approximation technique, Laplace's method (Raftery 1995), one can approximate

$$\ln \int pr(D \mid M_i, \boldsymbol{\theta}_i) q(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i \approx \ln pr(D \mid M_i, \hat{\boldsymbol{\theta}}_i) - (k_i / 2) \ln n,$$

where $\hat{\boldsymbol{\theta}}_i$ is the maximum likelihood under $M_i$ and $n$ is the sample size. If the above approximation holds, then the Bayes decision rule under a binary loss function can be approximated by minimizing the BIC score where the BIC score is defined as above:

$$BIC \approx -2\ln \int pr(D \mid M_i, \boldsymbol{\theta}_i) q(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i = -2\ln pr(D \mid M_i, \hat{\boldsymbol{\theta}}_i) + 2(k_i / 2)\ln n .$$

As long as the Laplace approximation is sufficient, Baysian model selection using MCMC and using the BIC should be equivalent, and both represent the special case of the DT approach with a binary loss function.

Note that the BIC approximation does not depend on the prior probability distribution. Schwarz (1978) showed for a large class of models that if the true model is among those under review, the BIC will choose the true model in the limit as the size of the data increases with probability approaching 1 (i.e., BIC is consistent if the true model is in the candidate pool). It is somewhat ironic that the BIC, derived from Bayesian principles, is the only method that has been proven to be consistent, a distinctly frequentist property. Because Schwarz (1978) illustrated this frequentist property of the BIC, it is sometimes referred to as the Schwarz Information Criteria (SIC). This property does not mean, however, that BIC assumes that the true model be in the candidate set. In real applications, the BIC will select the quasi-true model: the model in the candidate pool that best approximates the true model.

Both BIC and AIC score models according to how well the model fits the data with a penalty for over fitting. Both methods were derived as approximately optimal under different decision theoretic criteria, and the approximations require the use of asymptotic theory. As was noted in the text, asymptotic theory may often fail to produce appropriate approximations in phylogenetic contexts; therefore the critical tests for how well various model-selection criteria work necessarily involve simulations under more complex models than any being evaluated.

### A2D. DECISION THEORY INCORPORATING PERFORMANCE

For a phylogeny described by an unrooted binary tree with $k$ terminal nodes, there are $2k$-3 branches. The branch lengths can be denoted by vector $\mathbf{B} = (B_1, B_2, \ldots, B_{2k-3})$, and $\hat{\mathbf{B}}_i$ is the *estimated* branch lengths under the assumptions of model $M_i$. That is, $\hat{\mathbf{B}}_i$ is a function of the data $D$, the model $M_i$, and the maximum-likelihood estimates of the parameters $\hat{\boldsymbol{\theta}}_i$ under model $M_i$. Instead of the 0 or 1 loss function described above, a DT approach may penalize models according to their performance with regard to branch-length estimation.

Consider the estimated vector of branch lengths under model $M_i$ and $M_j$. The squared Euclidean distance between the branch length estimates is given by

Supplemental Material: Annu.Rev.Ecol. Evol. Syst. 2005. 36:445-466
doi: 10.1146/annurev.ecolsys.36.102003.152633
*MODEL SELECTION IN PHYLOGENETICS*
Sullivan and Joyce, 2005

$$\left\|\hat{\mathbf{B}}_i - \hat{\mathbf{B}}_j\right\|^2 = \sum_{l=1}^{2n-3}\left(\hat{B}_{il} - \hat{B}_{jl}\right)^2,$$

and the posterior loss of choosing model $M_i$ is given by

$$R_i = \sum_{j=1}^{K}\left\|\hat{\mathbf{B}}_i - \hat{\mathbf{B}}_j\right\|^2 pr(M_j \mid D)$$

$$\propto \sum_{j=1}^{K}\left\|\hat{\mathbf{B}}_i - \hat{\mathbf{B}}_j\right\|^2 \frac{e^{-BIC_j/2}}{\sum_{l=1}^{K} e^{-BIC_l/2}} \quad .$$

This is the risk factor that incorporates the non-binary loss function developed by Minin et al. (2003), and $R_i$ can be calculated for each model. That model with the minimum posterior risk, $R_i$, is chosen under this model-selection criterion. Thus, use of the BIC described above is precisely analogous to a special case of the DT method developed by Minin et al. (2003), one with the loss function in that case assumed to be binary.

The following theoretical comparisons point out the advantages to this approach. Each model is weighted according to the posterior probability of the model conditional on the data. Since any model of evolution is only a crude approximation to reality, rather than focus attention on trying to find the 'correct model', we have a measure of how plausible a model is given the data. In addition, the decision theoretic framework allows for much flexibility. One can decide based on biologically relevant criteria, what makes a model useful and use this criterion to assess a higher penalty to models that do not meet the criterion than to those that do.

References:

Bickel PJ, Doksum KA. 2001. *Mathematical Statistics*, 2nd ed., New Jersey, Prentice Hall 2001.

Minin V, Abdo Z, Joyce P, Sullivan J. 2003. Performance-based selection of likelihood models for phylogeny estimation. *Syst. Biol.* 52: 674 - 683.

Raftery AE. 1995. Bayesian model selection in social research (with discussion by A. Gelman, D. B. Rubin, and R. M. Hauser). In *Sociological Methodology* ed. PV Marsden, pp. 111 - 196. Oxford, U.K.: Blackwell.

Schwarz G. 1978. Estimating the dimensions of a model. *Ann. Stat*. 6:461-4.