Available online at www.sciencedirect.com
ScienceDirect
Journal of Theoretical Biology 243 (2006) 449–454
Journal of Theoretical Biology
www.elsevier.com/locate/yjtbi

Artificial neural networks can learn to estimate extinction rates from molecular phylogenies

Folmer Bokma*,[1]

Groningen Bioinformatics Centre and University Medical Centre, The Netherlands

Received 17 May 2006; accepted 22 June 2006
Available online 11 July 2006

Abstract

Molecular phylogenies typically consist of only extant species, yet they allow inference of past rates of extinction, because recently originated species are less likely to be extinct than ancient species. Despite the simple structure of the assumed underlying speciation–extinction process, parametric functions to estimate extinction rates from phylogenies turned out to be complex and often difficult to derive. Moreover, these parametric functions are specific to a particular process (e.g. complete species level phylogeny with constant birth and death rates) and a particular type of data (e.g. times between bifurcations). Here, it is shown that artificial neural networks can substitute for parametric estimation functions once they have been sufficiently trained on simulated data. This technique can in principle be used for different processes and data types, and because it circumvents the time-consuming and difficult task of deriving parametric estimation functions, it may greatly extend the possibilities to make macro-evolutionary inferences from molecular phylogenies. This novel approach is explained, applied to estimate speciation and extinction rates from a molecular phylogeny of the reef fish genus *Naso* (Acanturidae), and its performance is compared to that of maximum likelihood estimation.

Keywords: Artificial intelligence; Macroevolution; Modelling; Parameter estimation; Speciation

1. Introduction

"If you cannot do it the way you should, you should do it the way you can." Frisian proverb.

Branching processes play an important role at all levels of biology, as they can be used to model for example population dynamics, cell division, as well as the evolutionary diversification of groups of species (Athreya and Ney, 1972; Jagers, 1975). Therefore, biologists often want to estimate the parameters of a branching process. Unfortunately, despite their simple structure, branching processes can be complex to describe mathematically (Kendall, 1948, 1966; Bailey, 1964). Minor modifications to the structure of the process may require derivation of entirely different algorithms for parameter estimation.

An example hereof is the estimation of speciation and extinction rates from molecular phylogenies. Molecular phylogenies present a hypothesis of the historical relationships between species in the form of a binary tree. Though such phylogenies are typically constructed using data from extant species only, they can be used to make inferences about the rate at which species went extinct (Nee et al., 1994a). Because of the central role speciation and extinction play in macroevolution there is a direct interest in estimating those parameters. Even for a narrowly defined branching process the derivation of likelihood functions for parameter estimation is relatively complex, and substantially different estimation algorithms have been developed for the frequent situations in which the phylogeny does not comprise all extant species (Pybus and Harvey, 2000; Bokma, 2003; Paradis, 2003).

Consequently, several distinct estimation methods now exist. The first methods employed changes over time in the times between subsequent bifurcations in the tree. If the logarithmic number of lineages in a phylogeny is plotted against time, the relation is nonlinear, because old lineages

*Tel.: +46 90 786 7121; fax: +46 90 786 6705.
E-mail address: folmer.bokma@emg.umu.se.
[1]Present address: Department of Ecology and Environmental Science, University of Umeå, S-90187, Sweden.

0022-5193/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2006.06.023

have a greater cumulative probability of being extinct. Towards the present the slope of the relationship approaches the speciation rate, as the cumulative probability of being extinct approaches zero towards the present (Harvey et al., 1994; Nee et al., 1994a, b; Kubo and Iwasa, 1995). That is why the so-called lineage-through-time plots contain information about speciation and extinction rates.

Other methods rely on the distribution of species over higher taxa. Fundamental statistical work (Kendall, 1948; Bailey, 1964) indicates that the distribution of species over higher taxa depends on the speciation and extinction rate and on taxon age. Even an incomplete molecular phylogeny may define the ages of many higher taxa (e.g. Bokma, 2003). Therefore, estimation from the distribution of species over higher taxa is especially useful if the phylogeny is not complete at the species level, but at some higher taxonomic level. Times between branching events and the distribution of species over higher taxa can also be combined to maximize estimation power (Paradis, 2003).

Yet another method calculates a gamma statistic from the relative positions of branching times to measure the curvature of the lineages-through-time data (Pybus and Harvey, 2000). This gamma statistic is calculated so that it follows a standard normal distribution under a pure birth process. This method can be applied to incomplete phylogenies as well.

Here, I introduce a new estimation technique that differs radically from all previous ones because it does not require mathematical derivation of estimation functions. It involves training an artificial neural network to extract the information about speciation and extinction rates contained in simulated branching time data, and makes use of the fact that it is often easier to simulate a branching process than to derive likelihood functions that describe its behaviour.

2. Methods

Originally developed as models of interacting neurons in the brain, artificial neural networks (hereafter ANN) are nowadays often regarded as statistical tools. Comprehensive explanation of the working, properties, and uses of various types of ANN can be found in e.g. Haykin (1999). A typical ANN consists of several layers of nodes, i.e. "neurons". Inputs (in the present example branching times) are connected to the first layer of nodes, and between (but not within) layers nodes are connected. Often the final layer consists of a single node to generate a single output variable. Each node uses a so-called transfer function to convert its input to output. Connections between nodes or between inputs and nodes are associated with weights. When the ANN is being trained, those weights and the parameters of transfer functions are adjusted so that the output of the network approximates a certain target. Thus, an ANN transforms a set of input values (branching times) to a set of output values (speciation or extinction rates), and can be viewed as a universal function approximator.

In the present application an ANN is trained to associate parameter values (speciation or extinction probabilities) with observed data, using simulated data (lineage-through-time plots) that come from the same process as assumed for the observed data. Obviously, the parameter estimates are meaningful only if the process is appropriately modelled. It should be noted, however, that this is also the case with traditional estimation techniques such as likelihood maximization, which similarly yield meaningful estimates only if the estimation functions describe the process that actually generated the data.

2.1. Training the ANN

The estimation consists of a series of steps, as illustrated in Fig. 1. In step one, a model is specified and used with a wide range of parameter values to simulate data. In the present example the model is a constant rates speciation and extinction process, the parameters are the speciation and extinction probabilities (hereafter λ and μ) and the data consist of the branching times in a reconstructed phylogeny (*sensu* Nee et al., 1994b). In step two, the simulated data are partitioned into a training set, a validation set, and a test set, the use of which is explained below. Here, the training set consists of 5000, the validation set of 2000 and the test set of 2750 simulations. In step three, network architecture is chosen. This includes choosing the number of layers, the number of cells per layer, and the types of the transfer functions to be used in each layer. When the data and network are ready, the actual training takes place in step four. During training, the weights and transfer function parameters of the ANN are adjusted so that its output approaches the targets, which are the λ and μ used for simulations in step 1.

Networks with large numbers of neurons can describe the training data so well, that it can not be generalized to other data than the training set. This is called overfitting.

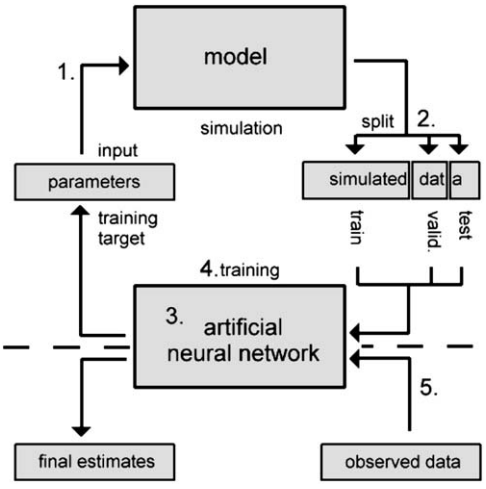


Fig. 1. Schematic illustration of the process of parameter estimation using an artificial neural network. Numbers correspond to the steps of the process detailed in the text.

The validation dataset is used to prevent overfitting, by stopping the training process when the prediction error of the validation set consistently increases. (Typically, the prediction error in both training and validation set initially decrease as the network is being trained. The error in the training set continues to decrease, but when the network starts to overfit the training data, the error in the validation data will increase again.) Finally the performance of the trained network is measured and compared to the performance of other networks using the test data. When the researcher considers the performance to be sufficient, the network is ready for use with observed data. As long as the network still performs poorly, the researcher may want to train it again, using fresh (i.e. random) network parameters, or change its architecture.

Choosing correct network architecture is the critical step in the estimation procedure. An ANN with the wrong architecture will not learn to associate λ or μ with the branching times, and is effectively useless. There are few rules-of-thumb to design network architecture for a specific task, but there exist computation-intensive methods to search ANN architecture space (Koza and Rice, 1991; Koza, 1993). However, for the present purpose a simple selection from 34 different architectures was found to be sufficient to achieve acceptable results.

2.2. Application

Goal of the present case is to estimate λ and μ from a phylogeny of *Naso*, a genus of reef fish from the family Acanthuridae. Klanten et al. (2004) constructed a molecular phylogeny which contains 19 species and is supposed to be complete at the species level. Klanten et al. (2004) provide estimates and confidence intervals for branching times, which makes it ideal for the present and future studies of macroevolutionary inference.

The most recent common ancestor of the 19 species of *Naso* is estimated to have existed 46.7 mya. (That is the first branching time in the phylogeny.) Thus, for training the ANN, phylogenies were simulated containing 19 species after a first split 46.7 mya. This was achieved by simulating two phylogenies starting with a single species, and combining their branching times only if they together yielded exactly 19 species. The combined phylogenies thus yield 18 branching times, of which the first always occurs at 46.7 mya. Therefore, this branching time does not provide any information to the ANN, and only the remaining 17 branching times were used in ANN analyses.

2.3. Preprocessing of data

The branching times of simulated phylogenies are obviously highly correlated, since the time of the ith branching depends on that of the $(i-1)$th branching. Since it is possible to simulate a large number of branching processes, it would be possible to train a large network that uses all branching times as inputs. Here, however, a principle component analysis (hereafter PCA) was used to obtain orthogonal components that retain 95% of the variance present in the branching times. PCA typically reduced the input from 17 branching times to 4 or 5 principle components. Before PCA all data were $\log_e$-transformed and after PCA data were transformed to a scale of zero to one. Estimates of λ and μ were correspondingly post-processed.

2.4. Comparison with maximum likelihood

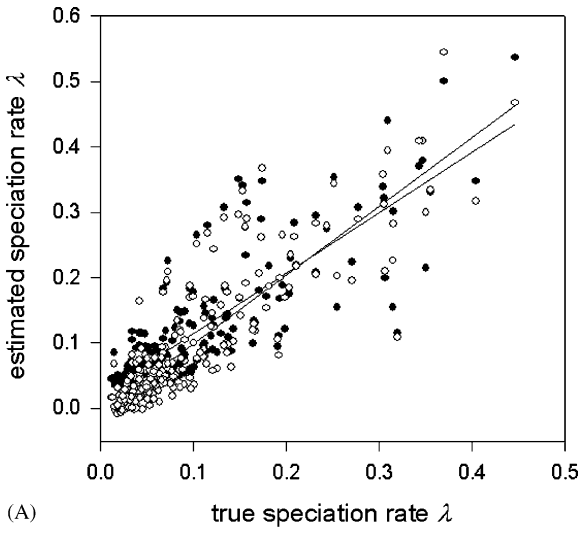
Likelihood function maximization (hereafter ML) is a general technique for parameter estimation with well-understood statistical properties. Nee et al. (1994b) derived likelihood functions to estimate λ and μ from branching times of molecular phylogenies. There exist other techniques that use the same data to the same end (Harvey et al., 1994, Kubo and Iwasa, 1995), but these are rather specific to a particular estimation problem. Because the ANN estimation approach can be applied to more estimation problems than just those involving branching processes, it is most meaningful to compare its performance to that of ML. Moreover, Nee et al.'s (1994b) ML approach is probably the most accurate available for the present purpose.

A trained ANN obviously performs well on the data used to train or evaluate it. Therefore, after networks had been trained and the best network had been selected, a new set of 250 simulations was created to compare the performance of ANN and ML. Training of ANN by minimizing the mean squared prediction error can be regarded as regressing true values of λ on their estimates (see "confidence intervals" below). For comparison with ML it is more intuitive to view estimates of λ as a function of their true values. Therefore, the linear regression of the true value of λ on its estimates $\lambda_{true} = a\lambda_{est} + b$ was used to transform the estimates as $\lambda_{trsf} = \lambda_{est}/a - b/a$. This linear transformation can result in some estimates becoming negative, but that is of no concern for the comparative purposes.

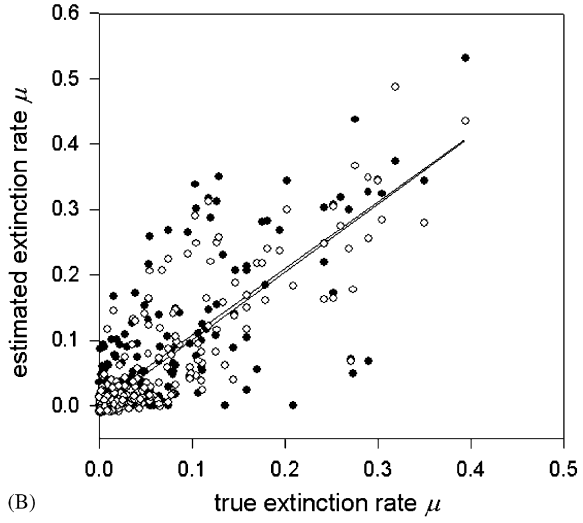
3. Results

3.1. Comparison with likelihood

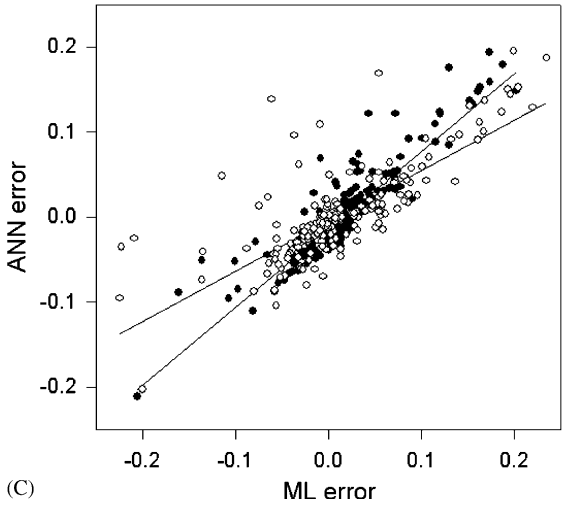
For 250 simulations of phylogenies with 19 species 47.6 years after the first bifurcation, estimates of λ and μ are shown as a function of their true values used for simulation in Fig. 2. The performances of the ML and ANN procedures were investigated by linear regression of estimated on true probabilities. Both procedures appear to yield unbiased estimates of both speciation and extinction rates, since linear regression coefficients are close to one and intercepts close to zero (Fig. 2 and Supplementary Table 1). The correlation of estimated and true probabilities is an index of estimation accuracy. The correlation coefficients of ANN and ML approaches are



(A)



(B)



(C)

Fig. 2. (A) Relation between true and estimated values of λ obtained by ML (closed dots) and ANN (open dots). (B) *Idem dito* for μ. (C) Relation between estimation errors from ML and ANN for λ (closed dots) and μ (open dots). All lines by linear regression. Regression statistics are in Supplementary Table 1.

very similar (Supplementary Table 1), which indicates that the ANN can estimate speciation and extinction probabilities as well as ML can.

Since both ANN and ML approaches yield approximately unbiased estimates of the true probabilities of speciation and extinction, the estimates from both procedures will be correlated. Interestingly however, also the estimation errors are highly correlated (Fig. 2c and Supplementary Table 1). This suggests that the ANN has learned to distinguish the same features in the lineage-through-time data as employed by ML to estimate λ and μ.

3.2. Confidence intervals

ML does not only yield estimates of λ and μ but also likelihood-ratio-based confidence regions. Thus for the *Naso* example, likelihood estimation yields an (approximate) 95% confidence region of λ and μ as shown in supplementary Fig. 1. For estimation using ANN there is no such mathematical framework to delimit confidence regions. However, the simulation approach does provide some information about the uncertainty in the estimates of λ and μ. This is shown in Fig. 3a where true values of μ are plotted as a function of their estimates. Lines can be drawn (by eye) that delimit the scatter of the estimates above and below the ideal relation $x = y$. The positions of these two lines at the best estimate of the extinction rate ($\log_e \mu = -5.6$) delimit a confidence interval for the estimate. For the speciation rate this interval is approximately symmetric about the best estimate (not shown), but for the extinction rate it is clearly asymmetric (Fig. 3).

The confidence intervals constructed for ANN estimates of λ and μ agree well with likelihood ratio based confidence intervals. The magnitude of confidence intervals is approximately equal, and the confidence interval of the extinction rate is similarly asymmetric (Fig. 3b). Thus, estimation using ANN not only provides similar parameter estimates as ML, it also produces similar confidence intervals. Confidence regions like in supplementary Fig. 1 are still lacking, however.

4. Discussion

Branching processes are called critical if $\lambda = \mu$, supercritical if $\lambda > \mu$, and subcritical if $\lambda < \mu$. Some biologists seem to believe that the process of speciation and extinction should be modelled as a critical or supercritical process. However, a subcritical process can yield species rich taxa if the birth rate happens to exceed the death rate (λ and μ are *probabilities* of speciation and extinction, so *rates* of speciation and extinction show stochastic fluctuations). Existing procedures to estimate speciation and extinction rates are sometimes limited to supercritical cases, albeit for mathematical rather than theoretical reasons. The ANN approach allows us to investigate subcritical parameter space as well. One can very well simulate phylogenies with 19 species after 47.6 million

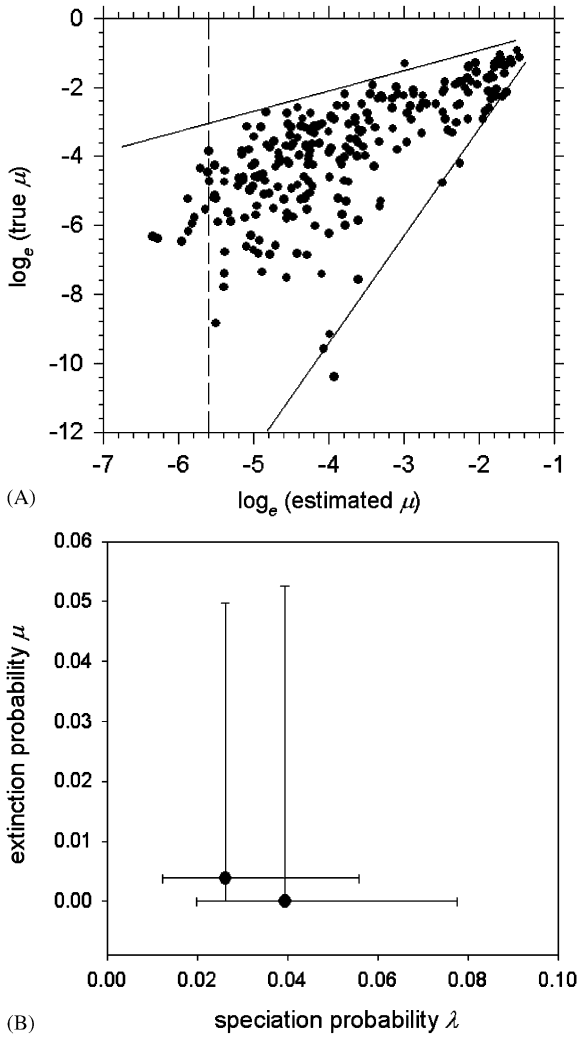


Fig. 3. (A) Relation between the true value of μ and its ANN estimate, showing the uncertainty in the estimates. Solid lines were drawn by eye delimiting the scatter of the data. The intersections of these lines with the best estimate of mu (dashed line) are confidence bounds for the estimate of μ. (B) Comparison of confidence regions for λ and μ obtained like in panel A (upper left) with likelihood ration-based confidence intervals (lower right).

years with $\lambda<\mu$. However, in order to be able to compare estimation using ANN to other estimation techniques, simulations were limited here to supercritical parameter space.

It is possible to create a neural network that estimates two (or more) parameters simultaneously (i.e. a network with two or more output neurons). Thus, we could construct an ANN with two output neurons that estimates both λ and μ simultaneously, just like Nee et al.'s (1994b) likelihood functions yield simultaneous estimates of the speciation and extinction rates. ANN with more than one output neuron however often suffer from "cross-talk" in the hidden layers. That is, parametrization of transfer functions of hidden neurons beneficial for estimating λ may adversely affect estimation of μ. Cross-talk can be eliminated by constructing larger networks and the use of larger training data sets, but it is more convenient to use separate networks for both parameters to be estimated.

The results show that we do not need exact parametric estimation functions like ML to estimate the parameters of a process. This applies to estimation of λ and μ from lineage-through-time data as shown here, but also to various other processes. For example, ANN trained to associate sets of 10 Gaussian random numbers with the variance used to simulate them, can be used to estimate the variance of 10 Gaussian random numbers. The only requirements are (i) that we can specify the model, (ii) that we can simulate data using the model, and (iii) that we can present the data to a network in such a way that the network can be trained. The third requirement can be challenging, as it may require careful inspection of the data to find out what aspect of it conveys information about parameter values (though for the speciation, extinction, as well as variance estimation examples above it is sufficient to sort the random numbers).

It is emphasized that when alternative estimation methods such as those of moments or likelihood are available, there is little need to involve the computational complexity of ANN. However, for many, sometimes even relatively simple processes parametric estimation functions are lacking, and it is often a time-consuming and difficult exercise to derive them. In such cases, estimation with ANN provides a fast alternative: evaluation of 34 networks including PCA on a set of nearly ten thousand simulations took the author's computer about an hour. Modern software packages render ANN analysis so simple that one needs little or no understanding of ANN to do all analyses shown in this paper, except the simulation of data.

Appendix A. Supporting Materials

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jtbi.2006.06.023

References

Athreya, K.B., Ney, P., 1972. Branching Processes. Springer, Berlin.
Bailey, N.T.J., 1964. The Elements of Stochastic Processes with Applications to the Natural Sciences, first ed. Wiley, New York.
Bokma, F., 2003. Testing for equal rates of cladogenesis in diverse taxa. Evolution 57, 2469–2474.
Harvey, P.H., May, R.M., Nee, S., 1994. Phylogenies without fossils. Evolution 48, 523–529.
Haykin, S., 1999. Neural Networks: A Comprehensive Foundation. Prentice-Hall, Upper Saddle River, NJ.
Jagers, P., 1975. Branching Processes with Biological Applications. Wiley, London.
Kendall, D.G., 1966. Branching processes since 1873. J London Math. Soc. 41, 385–406.
Kendall, D.G., 1948. On the generalized "birth-and-death" process. Ann. Math. Stat. 19, 1–15.

Klanten, S.O., Van Herwerden, L., Choat, J.H., Blair, D., 2004. Patterns of lineage diversification in the genus *Naso* (Acanthuridae). Mol. Phyl. Evol. 32, 221–235.

Koza, J.R., 1993. Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press, Cambridge.

Koza, J.R., Rice, J.P., 1991. Genetic generation of both the weights and architecture for a neural network. In: Proceedings of International Joint Conference on Neural Networks, II, pp. 397–404.

Kubo, T., Iwasa, Y., 1995. Inferring the rates of branching and extinction from molecular phylogenies. Evolution 49, 694–704.

Nee, S., Holmes, E.C., May, R.M., Harvey, P.H., 1994a. Extinction rates can be estimated from molecular phylogenies. Philos. Trans. R. Soc. London B 344, 77–82.

Nee, S., May, R.M., Harvey, P.H., 1994b. The reconstructed evolutionary process. Philos. Trans. R. Soc. London B 344, 305–311.

Paradis, E., 2003. Analysis of diversification: combining phylogenetic and taxonomic data. Proc. R. Soc. London B 270, 2499–2505.

Pybus, O.G., Harvey, P.H., 2000. Testing macro-evolutionary models using incomplete molecular phylogenies. Proc. R. Soc. London B 267, 2267–2272.