Stochastic Tree Models for Macroevolution - Development, Validation and Application

Stephanie Keller-Schmidt
Dissertation

Summary

Phylogenetic trees capture the relationships between species and can be investigated by morphological and/or molecular data. When focusing on macroevolution, one considers the large-scale history of life with evolutionary changes affecting a single species of the entire clade leading to the enormous diversity of species obtained today. One major problem of biology is the explanation of this biodiversity. Therefore, one may ask which kind of macroevolutionary processes have given rise to observable tree shapes or patterns of species distribution which refers to the appearance of branching orders and time periods. Thus, with an increasing number of known species in the context of phylogenetic studies, testing hypotheses about evolution by analyzing the tree shape of the resulting phylogenetic trees became matter of particular interest. The attention of using those reconstructed phylogenies for studying evolutionary processes increased during the last decades. Many paleontologists (Raup et al., 1973; Gould et al., 1977; Gilinsky and Good, 1989; Nee, 2004) tried to describe such patterns of macroevolution by using models for growing trees. Those models describe stochastic processes to generate phylogenetic trees. Yule (1925) was the first who introduced such a model, the Equal Rate Markov (ERM) model, in the context of biological branching based on a continuous-time, uneven branching process. In the last decades, further dynamical models were proposed (Yule, 1925; Aldous, 1996; Nee, 2006; Rosen, 1978; Ford, 2005; Hernández-García et al., 2010) to address the investigation of tree shapes and hence, capture the rules of macroevolutionary forces. A common model, is the Aldous' Branching (AB) model, which is known for generating trees with a similar structure of "real" trees. To infer those macroevolutionary forces structures, estimated trees are analyzed and compared to simulated trees generated by models. There are a few drawbacks on recent models such as a missing biological motivation or the generated tree shape does not fit well to one observed in empirical trees.

The central aim of this thesis is the development and study of new biologically motivated approaches which might help to better understand or even discover biological forces which lead to the huge diversity of organisms.

The first approach, called age model, can be defined as a stochastic procedure which describes the growth of binary trees by an iterative stochastic attachment of leaves, similar to the ERM model. At difference with the latter, the branching rate at each clade is no longer constant, but decreasing in time, i.e., with the age. Thus, species involved in recent speciation events have a tendency to speciate again. The second introduced model, is a branching process which mimics the evolution of species driven by innovations. The process involves a separation of time scales. Rare innovation events trigger rapid cascades of diversification where a feature combines with previously existing features. The model is called innovation model. Three data sets of estimated phylogenetic trees are used to analyze and compare the produced tree shape of the new growth models. A tree shape statistic considering a variety of imbalance measurements is performed. Results show that simulated trees of both growth models fit well to the tree shape observed in real trees. In a further study, a likelihood analysis is performed in order to rank models with respect to their ability to explain observed tree shapes. Results show that the likelihoods of the age model and the AB model are clearly correlated under the trees in the databases when considering small and medium-sized trees with up to 19 leaves. For a data set, representing of phylogenetic trees of protein families, the age model outperforms the AB model. But for another data set, representing phylogenetic trees of species, the AB model performs slightly better. To support this observation a further analysis using larger trees is necessary. But an exact computation of likelihoods for large trees implies a huge computational effort. Therefore, an efficient method for likelihood estimation is proposed and compared to the estimation using a naive sampling strategy. Nevertheless, both models describe the tree generation

process in a way which is easy to interpret biologically.

Another interesting field of research in biology is the coevolution between species. This is the interaction of species across groups such that the evolution of a species from one group can be triggered by a species from another group. Most prominent examples are systems of host species and their associated parasites. One problem is the reconciliation of the common history of both groups of species and to predict the associations between ancestral hosts and their parasites. To solve this problem some algorithmic methods have been developed in recent years. But only a few host parasite systems have been analyzed in sufficient detail which makes an evaluation of these methods complex. Within the scope of coevolution, the proposed age model is applied to the generation of cophylogenies to evaluate such host parasite reconciliation methods.

The presented age model as well as the innovation model produce tree shapes which are similar to obtained tree structures of estimated trees. Both models describe an evolutionary dynamics and might provide a further opportunity to infer macroevolutionary processes which lead to the biodiversity which can be obtained today. Furthermore with the application of the age model in the context of coevolution by generating a useful benchmark set of cophylogenies is a first step towards systematic studies on evaluating reconciliation methods.