Journal of Theoretical Biology 382 (2015) 216–222

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Phylogenetic tree and community structure from a Tangled Nature model

Osman Canko *, Ferhat Taşkın, Kamil Argın

Department of Physics, Erciyes University, Kayseri, Turkey

HIGHLIGHTS

- Phylogenetic trees (pt) estimated from genomes (or morphologies) of extant species cannot be compared with real pt, which is at best imperfectly known from the fossil record.
- One way to assess the accuracy of common estimation methods, such as ML or NJ, would be to apply them to data from in silico evolution models, for which the pt is exactly known.
- The quasi-evolutionary stable strategies' communities are very highly connected and there are no obvious fragmented subgroups among the species in a habitat.

ARTICLE INFO

Article history:
Received 28 January 2015
Received in revised form
25 June 2015
Accepted 7 July 2015
Available online 16 July 2015

Keywords:
Phylogenetic tree
Maximum likelihood method
Neighbor-joining method
Food-web
Modularity

ABSTRACT

In evolutionary biology, the taxonomy and origination of species are widely studied subjects. An estimation of the evolutionary tree can be done via available DNA sequence data. The calculation of the tree is made by well-known and frequently used methods such as maximum likelihood and neighbor-joining. In order to examine the results of these methods, an evolutionary tree is pursued computationally by a mathematical model, called Tangled Nature. A relatively small genome space is investigated due to computational burden and it is found that the actual and predicted trees are in reasonably good agreement in terms of shape. Moreover, the speciation and the resulting community structure of the food-web are investigated by modularity.

1. Introduction

The discovery of inheritance, which is one of the fundamental principles of biology, has caused a revolution in evolutionary systematics (Huxley, 1940; Mayr, 1942; Simpson, 1961; Hennig, 1966). These systematics are based upon a hierarchical layout which is generated from the relationship among groups of living organisms and is very important to interpret evolutionary processes. Evolutionary systematics have been studied at both species and above species levels for the living organism (de Queiroz and Donoghue, 1988). Mayr (1969) and Simpson (1961), who studied at species level, made a significant contribution to the categorization of species. Hennig (1966), who has the tenet of common descent, came to the conclusion that there were higher taxa than species level (de Queiroz and Donoghue, 1990) and changed the concept of evolution in taxonomy (Queiroz and Gauthier, 1992).

In the last few decades, studies about taxonomy have been accelerated by taking advantage of computers. The increase in the capacity of computers has permitted us to study longer and more numerous DNA sequences which are necessary to achieve the real phylogenetic tree. In this perspective, many models have been developed to obtain the best phylogenetic tree. In most models, there are two main groups of approaches to construct the phylogenetic tree (Saitou and Imanishi, 1989). The first group involves searching all of the possible phylogenetic trees and selecting the most correct one according to certain criteria, such as maximizing the probability of evolution. The maximum-parsimony (MP) (Eck and Dayhoff, 1966) and the maximum-likelihood (ML) (Felsenstein, 1981) methods are in this group. The second group involves building the best tree by analyzing the distances between nucleotide sequences. The neighbor-joining (NJ) method (Saitou and Nei, 1987) is a well-known example of this group.

The ML method finds the best possible phylogenetic tree according to the probability of transition (or evolving) which occurred in nucleic acid sequences. The topology and branch

* Corresponding author.
E-mail address: canko@erciyes.edu.tr (O. Canko).

http://dx.doi.org/10.1016/j.jtbi.2015.07.005
0022-5193/

lengths of the tree are of major importance in the ML method. Finding the tree topology and branch lengths is not a suitable approach. This is because, in a direct search, every possible tree topology should be searched and then the optimum value of the branch lengths with the maximum likelihood value should be determined for each topology. However, the number of possible topologies approaches huge numbers when the number of species (tips or nodes) is sufficiently big. Felsenstein (1978) found a procedure to remove this difficulty. He started with two species initially and then added the other species successfully. Hence, the number of possible topologies is systematically reduced. Even though this procedure does not assure the maximum value for the tree being constructed, the results it gives have, in practice, acceptable computational complexity. Since the ML method sets up an algorithm to find the branch lengths rather than using a direct search, the likelihood value of some trees can be equivalent due to the pulley principle. The branch lengths are altered at each step of the algorithm until the highest likelihood value is found. In spite of the fact that the ML method requires too much computational time for large genome sequences, the results it predicts are very appropriate to the phenomenological tree.

The NJ method builds the best tree by using the distance (nucleotide differences) between each species (or nucleotide sequences). The distance matrix of the tree is established from nucleotide sequences which is originally an unresolved tree as a star-like tree. Afterwards, the distance matrix is modified by calculating the differences between the genome sequences and the average divergence of these sequences from all other sequences is taken into account separately. The two sequences which have the smallest value in the modified distance matrix are joined in a single node which is regarded as an ancestor of these two sequences. The single node is replaced by two descendant sequences in the distance matrix and the distance matrix is modified again. The iteration would run $N-3$ times, where N is the number of species (or sequences). The NJ model is fast and gives a unique topology for the best tree because the tree is constructed on the local mathematical relations.

The phylogenetic tree of related species covers implicit information on how species evolved and adapted to the nature of their environment throughout different time periods. However, the accuracy of the assessments of evaluated trees is seldom investigated. The central question is whether or not the predicted phylogenetic tree is correct and reliable, because the methods for obtaining the phylogenetic tree only use the DNA sequence of species whose life forms are observed today. The main contribution of this study is to compare and test a simulated tree with the estimated evolutionary tree obtained from the above-mentioned statistical methods. For this purpose, the actual phylogenetic or evolutionary tree is produced from an individual based model. The simulation model considered here is called the Tangled-Nature (TaNa) model (Christensen et al., 2002; Hall et al., 2002) which emphasizes the co-evolution of individuals. The TaNa model has proven to be successful for use in the evolutionary phenomena seen in nature such as punctuated equilibrium, gradually decreasing extinction rate, and increasing diversity and power-law lifetimes (Christensen et al., 2002; Hall et al., 2002; Rikvold and Zia, 2003; Rikvold and Sevim, 2007).

The remaining part of the paper is organized as follows: The TaNa model is briefly explained in Section 2. In Section 3, the phylogenetic tree of the model is created from the simulation result and transitional forms are depicted in it. Then, the trees of the ML and the NJ methods are constructed using the Molecular Evolutionary Genetics Analysis (MEGA) program (Tamura et al., 2011). In Section 4, the interaction network among the species seen in the phylogenetic tree is investigated and the last section is devoted to the summary and conclusion.

2. The model

The TaNa model is an individual-based stochastic model of evolutionary ecology. As in the case of DNA sequence, the species are represented by binary strings whose elements are purine and pyrimidine. Because of computational burden, genome length is small (only 30 bits) in comparison to the real genome and when a mutation occurs (a change in the genome sequence), a new species appears in the system. In other words, genetic variety, namely phenotype, is ignored. The success of offspring probability, i.e., the reproduction ability or fitness of an individual i, is given as

$$P_i(t) = \frac{\exp[H(n_i,t)]}{1+\exp[H(n_i,t)]} \in [0,1], \tag{1}$$

where the weight function, H, is given by

$$H(n_i,t) = \frac{1}{cN(t)} \sum_{j=0}^{2^L-1} J_{ij} n_j(t) - \mu N(t). \tag{2}$$

Here c specifies the constant interaction strength, $N(t)$ is the total population at time t, the pair interaction term between species, J_{ij}, has a non-zero coupling with 0.25 probability. The non-zero elements of the fixed interaction matrix, J_{ij}, are taken as random distribution whose range is $[-1,+1]$ at the beginning of simulation. Self-interaction, namely cannibalism ($J_{ii}=0$), is ignored. If the individual i interacts individuals at position j, as either a prey or predator, the occupancy, $n_j(t)$, makes a contribution to the weight function. Since $n_j(t)$ is the total population of a species j, the normalization, $n_j(t)/N(t)$, corresponds to the population density. μ determines physical environment and the average sustainable total population size of habitat. μ corresponds to the inverse of Verhulst carrying capacity.

A time step of the model consists of the following dynamics: first a randomly chosen individual is killed with a constant probability, p_{kill}. At a reproduction step following this annihilation event, a randomly selected individual reproduces asexually with an offspring probability, Eq. (1). The successful individual gives two offspring before it dies and the genes of each offsprings are exposed to a low mutation rate, p_{mut}, as well. The occurrence of a mutation does depend on the current state of genome, as is in the memoryless Markov process. One generation contains the $N(t)/p_{kill}$ time steps. The model evolving the above steps finds quasi-evolutionary stable strategies (qESS) and these long periods are interrupted by short evolutionary active, *hectic*, periods. This feature is seen in Fig. 1.

Simulation starts with a population on a randomly assigned position in the genome space and a rapid diversification occurs by mutations to the neighboring sites. A relatively stable ecosystem is

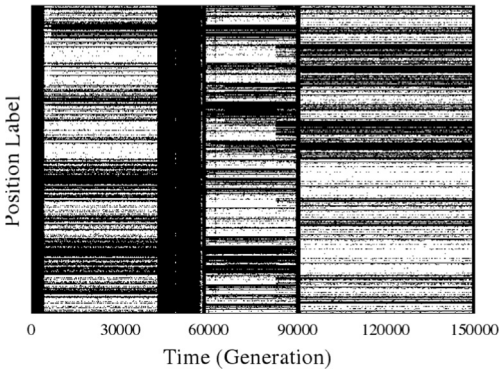

Fig. 1. Time series of occupation of genome space. A dot is placed for each of the occupied positions in the genome space. The genotypes are enumerated in an arbitrary way along the y-axis. Parameters are $c=0.5$, $L=30$, $p_{mut}=0.002$, $p_{kill}=0.2$ and $\mu=0.0002$.