Molecular Phylogenetics and Evolution 64 (2012) 190–197

Contents lists available at SciVerse ScienceDirect

Molecular Phylogenetics and Evolution

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

Using directed phylogenetic networks to retrace species dispersal history

Mehdi Layeghifard [a,b], Pedro R. Peres-Neto [a], Vladimir Makarenkov [b,*]

[a] *Département des Sciences biologiques, Université du Québec à Montréal (UQÀM), CP 8888, Succ. Centre Ville, Montréal, QC, Canada H3C 3P8*
[b] *Département d'Informatique, Université du Québec à Montréal (UQÀM), CP 8888, Succ. Centre Ville, Montréal, QC, Canada H3C 3P8*

ARTICLE INFO

Article history:
Received 29 October 2011
Revised 8 March 2012
Accepted 26 March 2012
Available online 2 April 2012

Keywords:
Biogeographic reconstruction
Dispersal network
Historical biogeography
Horizontal gene transfer (HGT)
Phylogenetic network
Phylogenetic tree

ABSTRACT

Methods designed for inferring phylogenetic trees have been widely applied to reconstruct biogeographic history. Because traditional phylogenetic methods used in biogeographic reconstruction are based on trees rather than networks, they follow the strict assumption in which dispersal among geographical units have occurred on the basis of single dispersal routes across regions and are, therefore, incapable of modelling multiple alternative dispersal scenarios. The goal of this study is to describe a new method that allows for retracing species dispersal by means of directed phylogenetic networks obtained using a horizontal gene transfer (HGT) detection method as well as to draw parallels between the processes of HGT and biogeographic reconstruction. In our case study, we reconstructed the biogeographic history of the postglacial dispersal of freshwater fishes in the Ontario province of Canada. This case study demonstrated the utility and robustness of the new method, indicating that the most important events were south-to-north dispersal patterns, as one would expect, with secondary faunal interchange among regions. Finally, we showed how our method can be used to explore additional questions regarding the commonalities in dispersal history patterns and phylogenetic similarities among species.

1. Introduction

The minimum length Steiner tree with 120° between all branches, which is a particular case of a phylogenetic tree, is known to give the tree connecting all points in the plane. It allows for representing geographic information as a bifurcating minimum length tree (Cavalli-Sforza and Edwards, 1967). Methods designed for inferring phylogenetic trees have been widely used to reconstruct biogeographic history (e.g., Anderson, 2002; Brooks, 1990; Graham et al., 2004; Legendre and Legendre, 1984; Legendre and Makarenkov, 2002). In many biogeographic applications, the goal is to apply methods used for characterising the evolutionary relationships among species (or genes) in the context of inferring dispersal scenarios among geographical units (i.e., terminal species or genes become regions). However, biogeographic reconstruction has not kept pace with new developments in phylogenetics. Current phylogenetic methods used in biogeographic reconstruction are based on trees rather than networks, thus following the strict assumption that different branches of a dispersal tree have evolved independently from one another. In the same way that we know that the independent evolution of different branches of a phylogeny is considered to be an unrealistic assumption for reconstructing the phylogenetic history of many taxa (e.g., bacteria, hybrids), dispersal among geographical units has, most likely, not occurred on the basis of independent single dispersal routes. While species might have taken multiple dispersal routes to migrate from one region to another, most of the current phylogenetic methods used in biogeographic reconstruction assume a lack of trade-offs between territorial units (geographic regions) during dispersal periods; i.e., current methods assume that one single dispersal route is always optimal for all species between any two given regions. Indeed, simple tree-like structures only show one dispersal scenario (one dispersal route) out of several that might have been occurred during dispersal events (akin to hybridization in reticulated evolution). While phylogenetic networks have been widely employed in the analysis of reticulate evolution, their use should be encouraged as well when constructing biogeographic dispersal hypotheses to represent multiple alternative dispersal patterns that explain present day species distribution.

Phylogenetic networks are a generalisation of phylogenetic trees allowing for simultaneous representation of several conflicting or alternative forces shaping evolutionary histories (Huson and Bryant, 2006), such as horizontal gene transfer (HGT) in bacterial evolution, evolution through allopolyploidy in plants, hybridisation events between related species, and homoplasy (i.e., evolutionary convergence). Phylogenetic networks inference methods can be also used to address non-phylogenetic questions, such as host–parasite relationships, vicariance and dispersal biogeography. Legendre and Makarenkov (2002) were the first to use *reticulograms* in historical biogeography while studying the postglacial dispersal of freshwater fishes in the Quebec peninsula. However,

---

* Corresponding author. Fax: +1 514 987 8477.
E-mail address: makarenkov.vladimir@uqam.ca (V. Makarenkov).

1055-7903/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ympev.2012.03.016

reticulograms are undirected graphs (reticulation branches show no direction), not allowing one to infer the direction of dispersal and migration events among regions. The goal of this study is to introduce a new method for inferring *directed phylogenetic networks* that can be used to model multiple dispersal events among regions in biogeographic reconstruction. As a case study, we reconstruct the biogeographic history of the postglacial dispersal of freshwater fishes in the Ontario province. We chose Ontario as the case study because of the availability of a large and detailed dataset on fish distribution for this province. Ontario is the second largest Canadian province after Quebec in both total and water-covered area, and it is also second to Manitoba in the percentage of total area covered by water. Finally, Ontario contains the greatest biodiversity of freshwater fishes in Canada along with British Columbia (Chu et al., 2003).

The current distributional patterns of freshwater fishes in Canada are the result of active processes following the Wisconsinian glacial period, which occurred 8000–10,000 years ago (Mandrak and Crossman, 1992). During the maximum extent of the Wisconsinian ice sheet, there were no known freshwater habitats in Canada. During the period in which Canada was being gradually covered by ice, fishes either died out or migrated to refugia in warmer southern water bodies. The present-day fishes living in water bodies across Canada reinvaded the country as lakes and rivers were created by the melt-water of the receding glacial ice sheet. Because these water bodies were first developed along the southern margin of the glacial ice sheet, they were easily linked to the southern refugia and provided water routes acting as dispersal corridors into increasingly deglaciated areas for fish and other aquatic organisms. Given that present-day fish distributions are entirely due to relatively new dispersal events in the region, the biogeographic reconstruction of this area should be relatively simpler and thus regarded as a relevant case test for our framework.

2. Materials and methods

2.1. Biogeographic data and study area

The fish distributional dataset used in this study came from the Ontario Ministry of Natural Resources (OMNRs) and comprises presence–absence records and geographic positioning for more than 9000 lakes. Ontario province is located in east-central Canada and is bordered by the provinces of Manitoba to the west, Quebec to the east, and the US states (from west to east) of Minnesota, Michigan, Ohio and Pennsylvania (both across Lake Erie), and New York to the south and east. Ontario ranges roughly from 74° to 95° longitudinally and from 42° to 57° latitudinally. The presence–absence data for 77 species (excluding introduced and hybrid species) in 9372 lakes of Ontario were analysed in this study.

2.2. Defining geographical units

Because of the very large number of lakes included in this analysis, we grouped adjacent lakes together to make the analysis more computationally effective. Moreover, the interest in biogeography is often to estimate the faunal exchange among large geographic units rather than dispersal events at smaller scales. Given that we did not have any *a priori* expectation regarding important geographic units or regions that would represent major patterns of biogeographic differentiation among them, we decided to distribute the lakes into regions using somewhat artificial biogeographic boundaries. The new method we will present can be applied in either situation (i.e., natural – by the recognition of natural geographic boundaries or biogeographic events, or artificial – by geographical proximity as in this study). We first converted the map of Ontario into a 15-by-15-cell grid map, and then assigned each lake to one of these cells based on its geographical coordinates. From the total of 225 cells, only 96 cells contained one or more lakes for which we had data. Note that other methods could be certainly used to arrange lakes into large geographic units based on objective criteria such as the identification of groups using permutation procedures (Strauss, 2001) or space-constrained algorithms (Legendre, 1987). Then, a *k*-means least-squares partitioning method (the software we used is available at http://www.bio.umontreal.ca/Casgrain/en/labo/k-means.html; one can also use the function '*k*-means' from the R package) was carried out to partition the 96 Ontario cells according to their levels of species' similarities. *K*-means is a method of cluster analysis that aims at partitioning *n* observations (here the 96 geographic cells) into *k* clusters based on attributes (here faunal composition) (MacQueen, 1967). The clustering is performed by minimising the sum-of-squares of the distances between the cells in each cluster and the corresponding cluster centroid. This analysis indicated that the geographic cells should be divided into two large groups, indicating that the species composition of the southern and northern Ontario regions were significantly different. We then conducted an additional *k*-means analysis for each region separately that allowed us to further amalgamate the geographic cells into 12 and 8 geographic sub-regions within the southern and northern regions, respectively (Fig. 1). These sub-regions were then used in the final dispersal network reconstruction.

2.3. Directional species dispersal networks

The method discussed here to reconstruct a dispersal network (which comprises, for example, all possible migration routes taken by fish species to reoccupy the newly de-glaciated areas) includes two main steps (Fig. 2). The first step consists in reconstructing two different phylogenetic trees (see algorithm below) for each of the two regions in Ontario identified earlier – one spatial, based on the geographic distances (Euclidean) between the sub-regions, and another distributional, based on the presence–absence of fishes in the sub-regions within each region (i.e., southern and northern regions). As a starting point, we needed to know the approximate locations of the refugia (i.e., network roots) and the first regions through which the fish entered Ontario to root the trees. Mandrak and Crossman (1992) proposed several possible dispersal corridors into Ontario from three different refugia. Here we adopted the two refugia that coincided with the southern and northern regions defined earlier as roots. For instance, the third major possible refugium suggested by Mandrak and Crossman (1992) has multiple corridors spreading all over the Great Lakes and entering into various geographic units of Ontario. Considering the wide geographic range of this multi-corridor refugium, we decided not to include it in our analysis. Moreover, a finer scale of the two refugia that we considered contributes to the accuracy of our analysis compared to a broader scale of the third refugium which is more suitable for analyses involving a much larger geographic region.

We calculated a pairwise geographic distance matrix among the sub-regions (8 northern and 12 southern sub-regions determined by *k*-means) using the geographic coordinates of the centre of each sub-region. The resulting matrix was then used to build the geographic distance tree. The distributional tree was built using a matrix of Sørensen distances (Sørensen, 1948) between the sub-regions based on the distributional data (i.e., presence–absence data).

The second step consists in building a dispersal network (Fig. 2) for each of the southern and northern regions of Ontario separately. In order to build these dispersal networks, we adapted a recent method developed by Boc et al. (2010) for detection of