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

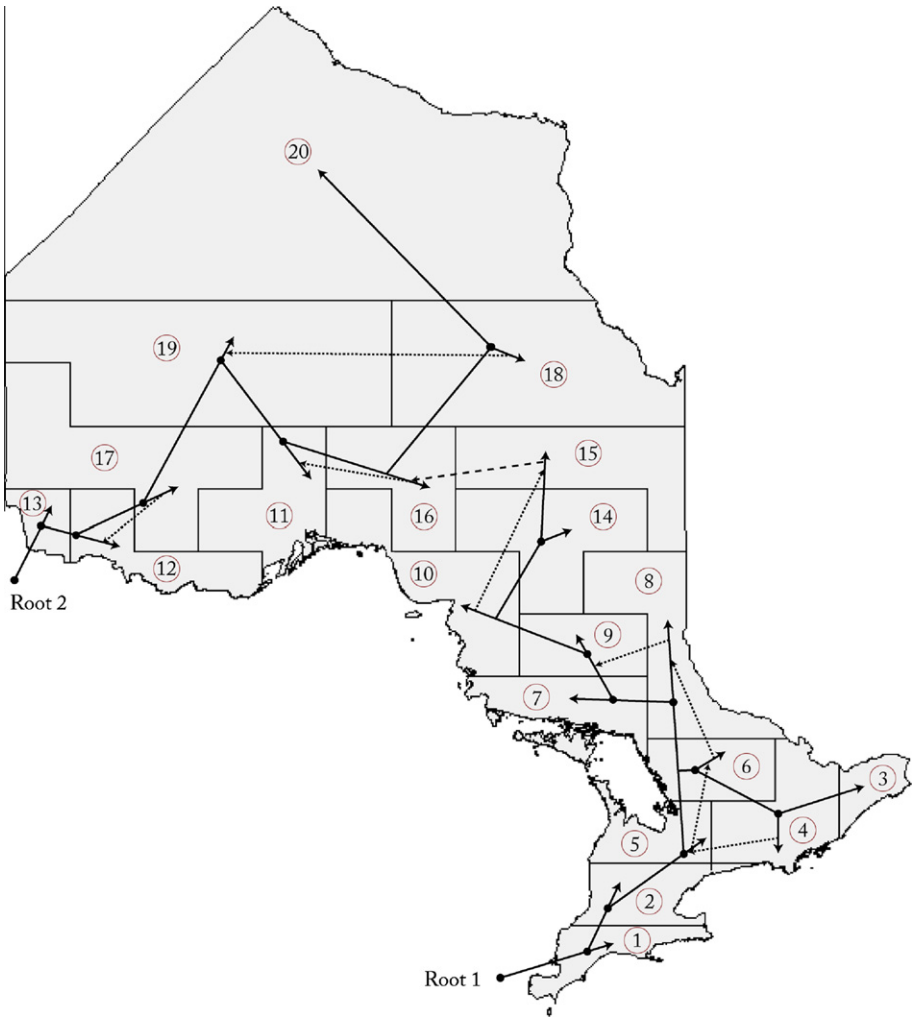

Fig. 1. Biogeographic history of postglacial dispersal of Ontario fishes represented as a dispersal network. Solid, dotted and dashed lines represent, respectively, the main dispersal routes, alternative routes within both main Ontario regions and an alternative route connecting these two main regions.

horizontal gene transfer (HGT) events. In the remainder of the article, we refer to our method as DSDNs (Directional Species Dispersal Networks; see Table 1 for terminological parallels that can be drawn between the HGT and historical biogeography processes). The considered HGT detection method (Boc et al., 2010) uses two trees as input, namely a species tree (representing the non-reticulate history of the species at hand) and a gene tree (representing the evolutionary history of the given gene for the same set of species), and exploits the original discrepancy between their topologies to transform the species tree into the gene tree by an optimal combination of sub-tree moves (i.e., sub-tree prune and regraft operations). It estimates the possibility of an HGT (i.e., reticulation event) between each pair of branches of the species tree and allows for adding new directed branches to the species phylogeny to represent the estimated reticulation events. In contrast, our DSDN method uses geographic (spatial) and Sørensen (distributional) distance matrices in place of the gene and species distance matrices, respectively, considered in the HGT model above. Thus, the DSDN method proceeds by a gradual reconciliation (for more details, see Section 4 and Boc et al., 2010) of the geographic and dispersal (i.e., distributional) trees in order to infer a directed network. The bootstrap scores of the dispersal tree, which is usually obtained from the presence–absence data, can be estimated using the traditional bootstrap procedure (Felsenstein, 1985). Moreover, the bootstrapping of the obtained alternative dispersal routes can be performed by fixing the topology of the geographic tree and by resampling the original presence–absence binary data used to build the dispersal tree. Then, the DSDN method can be performed to calculate the percentage of time that each original alternative dispersal root has been recovered using as input the same geographic tree and, in turn, different dispersal tree phylogenies obtained from the resampled presence–absence matrices. Thus, the DSDN method allows for adding and validating new directed branches to the biogeographic tree to represent these alternative routes (see Fig. 2).

Once the networks for the southern and northern Ontario regions were built using the new method, we connected them to infer potential alternative routes between their neighbouring sub-regions (i.e., sub-regions 11, 16 and 18 from the northern and sub-regions 10, 14 and 15 from southern region in Fig. 1). All phylogenetic trees in this study were reconstructed using the neighbour-joining method (Saitou and Nei, 1987). The latter method as well as the HGT detection method (Boc et al., 2010) used here are included in the T-Rex package (Makarenkov, 2001; see also the web site: www.trex.uqam.ca).

2.4. Exploring the relationship between dispersal history and species attributes

As pointed out by Wiens and Donoghue (2004), historical biogeography for most parts ignores phylogenetic and ecological characteristics of species and vice versa. Indeed, an important endeavour

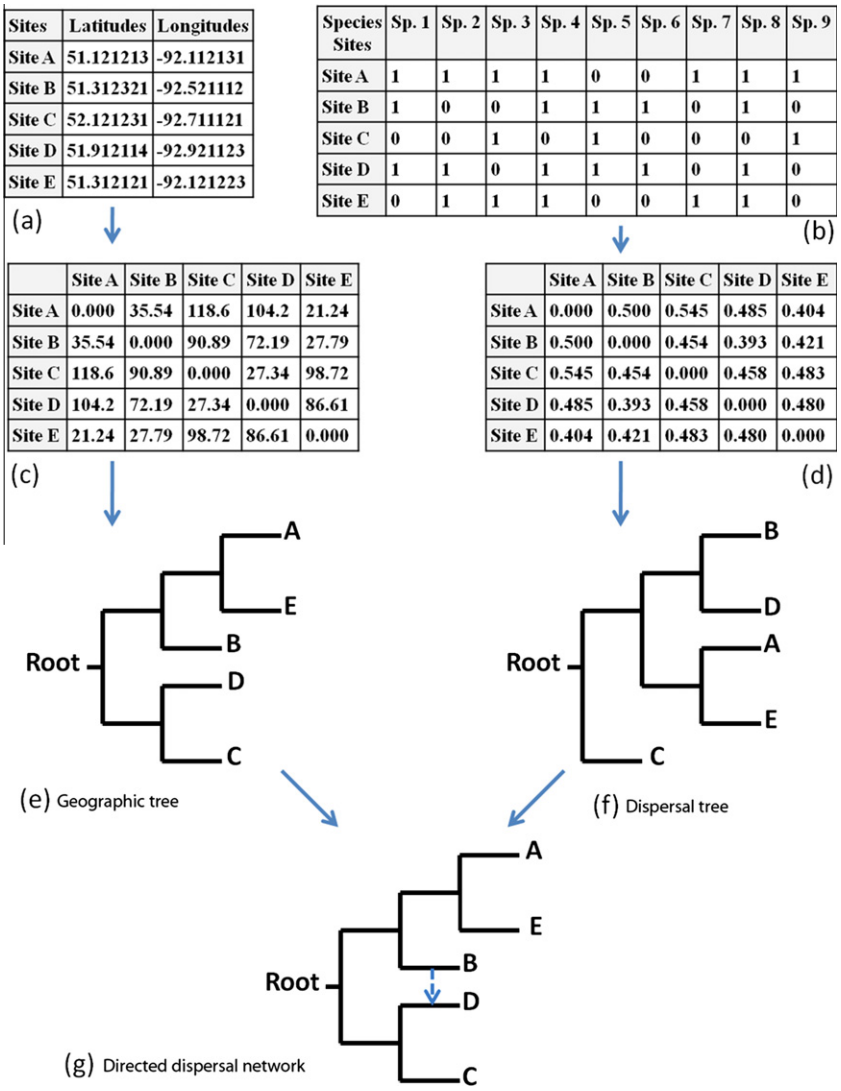

Fig. 2. Schematic representation of the directional species dispersal network building process based on an artificial data set. (a) Coordinates of geographic sites; (b) presence–absence (or incidence) data set describing the distribution of species across sites; (c) geographic distance (Euclidean) matrix calculated from the coordinates of the sites; (d) Sørensen's distance matrix calculated from presence–absence data; (e) geographic tree built from geographic distance matrix; (f) dispersal tree built from the Sørensen distance matrix; (g) the directional dispersal network built from the two above-mentioned (dispersal and geographic) trees. The directed dashed line shows an alternative migratory route (i.e., dispersal or "reticulation" event). The direction of reticulation events can be determined using any of the following optimisation criteria: least-squares, Robinson and Foulds (RF) distance, quartet distance or bipartition dissimilarity.

Table 1
Terminology adopted in this article to draw parallels between the HGT (horizontal gene transfer) detection and DSDN (directional species dispersal network) methods.

HGT terminology		DSDN terminology
Species tree	↔	Geographic tree
Gene tree	↔	Dispersal tree
Phylogenetic network	↔	Dispersal network
Reticulation event	↔	Alternative (dispersal) routes
Clade (cluster)	↔	Biogeographic cluster

in ecology is to understand how ecological species attributes, such as their molecular features or environmental requirements may influence their co-existence (co-occurrence) and dispersal decisions. Two basic processes may be involved in these decisions: (a) species with similar attributes may choose similar dispersal routes on the basis of their common tolerance to the habitats encountered while dispersing (hereafter referred to as dispersal filtering in contrast to environmental filtering in community ecology); and (b) competitive interactions among species, which would limit their co-existence along dispersal routes and perhaps force species to disperse via alternative routes (hereafter referred to as dispersal avoidance). These two processes make contrasting predictions about co-occurrence patterns among species and their phylogenetic relatedness. Under dispersal filtering, closely related species would tend to share similar dispersal histories, whereas under dispersal avoidance, closely related species would tend to have different dispersal histories. Note that species functional traits, when available, can be equally considered, especially in the case where these traits are not phylogenetically conserved.

An interesting extension of our framework is the combination of both biogeographic and phylogenetic information to assess the likelihood of these two processes during dispersal history. In this case, phylogenetic relatedness (e.g., within genera and families) under the assumption of niche conservatism serves as a proxy for the abiotic conditions for which a species can persist given that species sharing common ancestry also tend to share similar ecological attributes. This analysis parallels the work in community phylogenetics by Cavender-Bares et al. (2009) in a biogeographical

setting and may provide additional insights into the mechanisms and factors driving co-existence and dispersal patterns at large spatial scales.

We used the presence–absence incidence matrix to calculate the average phylogenetic distance (APD_{obs}) within each genus or family using Sørensen's similarity index. For each family or genus, we then applied a null model in which we randomly selected a group of species of the same size (e.g., if the genus or family under consideration had x species, then we picked up exactly x species from the entire pool of species, regardless of their taxonomic affiliation). For each randomly chosen group, we calculated its average phylogenetic distance (APD_{rnd}), and finally, the standardised average distance **Z** and its associated significance value (p-value) using the following formulas:

$$\mathbf{Z} = (APD_{obs} - APD_{rnd})/SD_{rnd},$$

$$p = (X + 1)/(N + 1),$$

where X is the number of APD_{rnd} values equal to or greater than APD_{obs} (1 in the formula accounts for the observed value; i.e., the observed value is also considered as one potential outcome of the null model, for more details see Peres-Neto, 2004), N is the number of randomly chosen groups of species (here we used a test based on 999 randomly chosen groups), and SD_{rnd} is the standard deviation of randomly chosen groups. The obtained results are presented in Table 2.

Additionally, we contrasted the species phylogenetic tree against a species dispersal pattern tree in order to identify any potential discrepancy or consistence between the tree clades (Fig. 3). The species phylogenetic tree was inferred from the DNA sequences of mitochondrial COI genes (Fig. 3a), whereas the species dispersal pattern tree was inferred from the Sørensen distance matrix calculated from the presence–absence data (Fig. 3b). The DNA sequences of a 652-bp segment from the 5′ region of the mitochondrial COI (cytochrome C oxidase subunit I) genes of Ontario freshwater fishes were obtained from GenBank using the accession numbers from Hubert et al. (2008). The species phylogenetic tree was built using the neighbour-joining method (Saitou and Nei, 1987). To verify the accuracy of the tree, we also reconstructed the species phylogeny using a maximum likelihood (ML) method, and obtained almost identical results (the ML tree is not presented here). Because the mitochondrial DNA sequences were available for 66 fish species only, we excluded the remaining 11 species from both trees.

We then used the Robinson and Foulds topological distance (Robinson and Foulds, 1981) to compare the phylogenetic (Fig. 3a) and distributional (Fig. 3b) trees and to find possible similarities between the tree topologies. The Robinson and Foulds topological distance is equal to the minimum number of elementary operations, consisting of merging and splitting nodes, necessary to transform one tree into the other. As demonstrated by Robinson and Foulds (1981), it is also the number of bipartitions, or Buneman's splits (1971), that belong to exactly one of the two trees. For two unrooted binary trees whose leaves are labelled according to the same set of n species, the Robinson and Foulds distance between them varies between 0 (when the trees are identical) and $2n - 6$ (when the trees are completely different).

Table 2
Null model results (**Z**-score and probability values) for the Ontario fish genera and families and their associated significance. Probabilities (p-values) smaller than 0.05 were used as indicative of dispersal avoidance, whereas values greater than 0.95 were considered as indicative of dispersal filtering. Significant values are shown in bold.

	Z-score	p-Value
Genera		
Ameiurus sp.	−0.8489	0.8610
Catostomus sp.	0.5809	0.1574
Coregonus sp.	1.7250	0.0875
Cottus sp.	−0.8473	**0.9900**
Esox sp.	0.9791	0.0995
Etheostoma sp.	−0.9989	**0.9900**
Hiodon sp.	0.8136	0.1194
Ichthyomyzon sp.	6.9705	0.0018
Lepomis sp.	−1.0255	**0.9630**
Luxilus sp.	−0.5806	0.6311
Moxostoma sp.	0.2023	0.2523
Notropis sp.	−1.1097	0.9120
Percina sp.	−0.8471	**0.9950**
Phoxinus sp.	−0.0102	0.3723
Pimephales sp.	−0.6234	0.6471
Rhinichthys sp.	−0.6517	0.7501
Semotilus sp.	−0.7025	**0.9950**
Stizostedion sp.	−0.7091	0.9310
Families		
Catostomidae	0.8101	0.1974
Centrarchidae	−1.1631	**0.9940**
Cottidae	−0.9624	**0.9990**
Cyprinidae	−1.7853	**0.9900**
Gasterosteidae	−0.6760	0.8081
Ictaluridae	−1.0371	0.9470
Percidae	−1.5722	**1.0000**

3. Results

The k-means method suggested separating the Ontario map into 20 sub-regions which can be divided into two regions (i.e., southern and northern; Fig. 1). However, it should be noted that in two cases, k-means grouped together two geographically distant cells instead of neighbouring cells. Given the large total number of cells (i.e., 96), these two inconsistencies (errors) were considered negligible and were corrected manually. The data analysis showed that 56 species, out of a total of 77 fish species, inhabit both the northern and southern regions of the Ontario province, while 18 and three fish species are unique to the southern and northern regions, respectively, confirming the fact that the southern region presents a greater species diversity.

In searching for alternative routes, our directional species dispersal network method identified five and three such routes in the southern and northern regions of Ontario, respectively (dotted lines in Fig. 1). We also found one alternative route between the southern and northern regions (dashed line in Fig. 1). The dotted and dashed lines in Fig. 1 show the potential different routes taken by Ontario fish species during the postglacial dispersal.

The null model analysis performed for all fish genera and families showed a significant correlation between the dispersal patterns and phylogenetic relationships for only six genera and four families (Table 2). Among them, all but one genus (i.e., *Ichthyomyzon*) was consistent with dispersal filtering rather than dispersal avoidance, as species in these genera and families tended to have similar distributions.

By comparing the two 66-species trees (dispersal pattern and molecular phylogeny) using the Robinson and Foulds topological distance, we found five similar species clusters (numbered from #1 to #5 in Fig. 3a and b). The Cyprinidae family appeared to be the largest (23 species) and the most vastly distributed group of fishes in Ontario, though four members of this family were grouped together in the distributional tree, suggesting a similar pattern of dispersal for these species (see cluster #4 in Fig. 3a and b). Conversely, members of the Percidae family (nine species) were scattered across the distributional tree showing different dispersal patterns. A similar trend was found for the Cottidae family (four species). In the remaining families having at least three members, the distributional patterns across species showed a higher

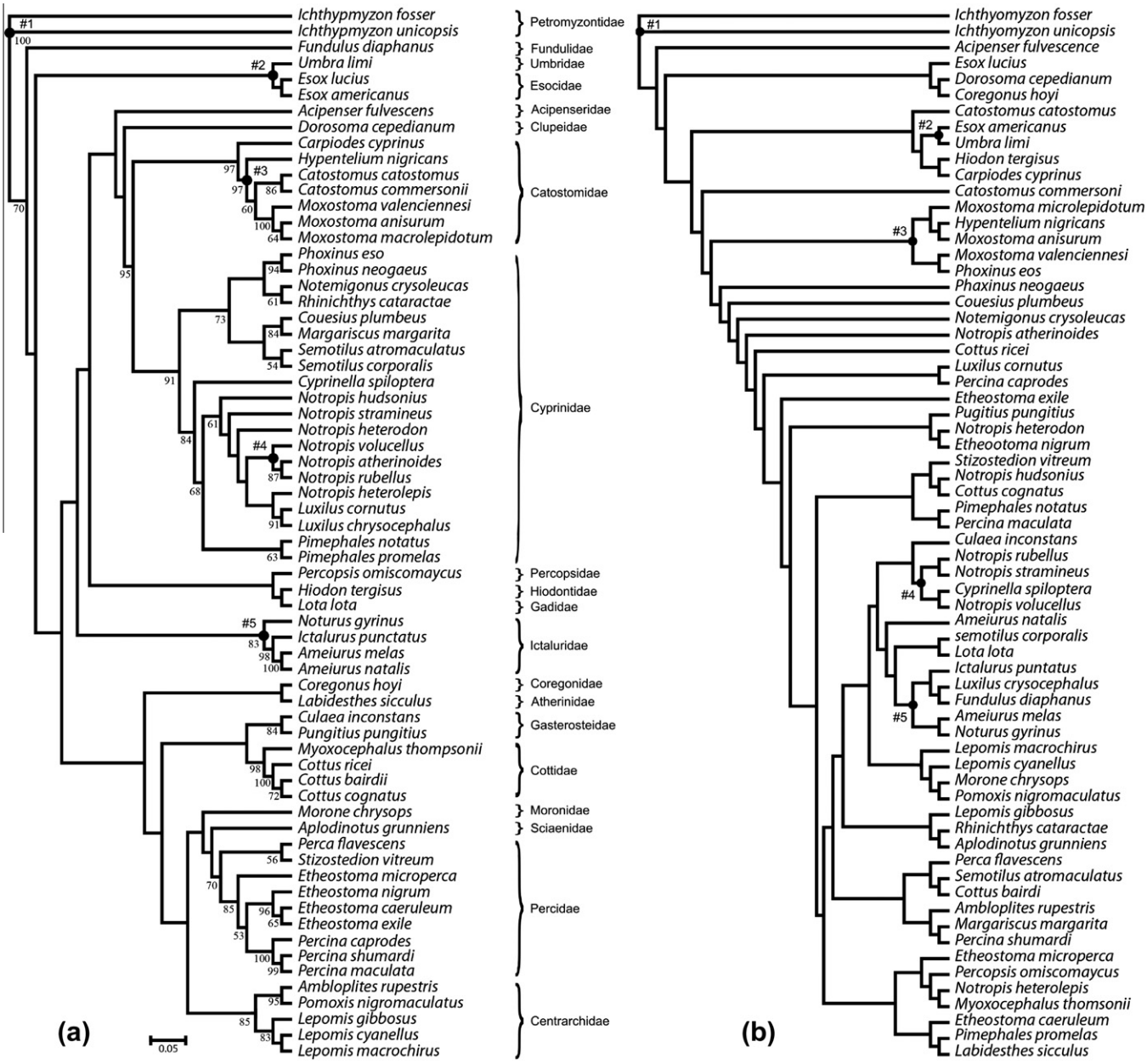

Fig. 3. (a) Phylogenetic tree for the 66 fish species built using the available mitochondrial COI gene sequences. Family names are included in the species phylogeny. Bootstrap scores greater than 50% are shown on the tree branches and (b) dispersal pattern tree for the same set of species inferred from presence–absence data. Convergent biogeographic clusters between the two trees are indicated by the numbers #1–#5.

similarity (Fig. 3a and b) even though the related species were still scattered across the trees.

4. Discussion

In this article we drew parallels between the processes of horizontal gene transfer, which can be represented by directed phylogenetic networks, and historical species dispersal, which can be represented by biogeographic dispersal networks. We introduced a framework that allows directional network analysis in historical biogeographic reconstruction, and as an illustration, we applied the new method to explore the historical patterns of biogeography of Ontario fishes as well as the possible relationships of those patterns with the species phylogeny. Although trees have been proven to be useful in reconstructing biogeographic history (Legendre and Legendre, 1984), they provide a much simplified view of what most likely took place. In order to estimate the possibility of other major dispersal events and the related routes used by species during these events, a more comprehensive method is needed. To the best of our knowledge, our method is the first one to allow the construction of a directional network to estimate such alternative dispersal events.

In our DSDN framework, a phylogenetic tree built from the geographic distances between regional centres is used as the backbone for the method, because fish, as a number of other organisms, are likely to migrate from a region to its bordering regions, and then to the next bordering regions and so on, as in a stepping stone process (Olden et al., 2001). Thus, the inferred backbone tree represented the shortest possible way for fish to disperse throughout Ontario. However, as mentioned above, fishes could have also used alternative dispersal routes, which would have been neglected by

traditional methods based on traditional phylogenetic trees. Using a dispersal tree, built from the distance matrix calculated from the presence–absence data, the DSDN method searches for discrepancies between the trees and transforms them into estimates for alternative dispersal routes. As stressed earlier, a great advantage of our method over the reticulograms introduced by Legendre and Makarenkov (2002) is that it also shows the directions of reticulation events. Moreover, in the process of reticulogram reconstruction, a phylogenetic tree is first built from a single distance matrix (e.g., using the neighbour-joining method), and supplementary branches (reticulation events) are then added to that tree, once at a time, in order to minimise a least-squares or weighted least-squares loss function (based either on the same distance matrix used to reconstruct the original phylogenetic tree or on an alternative one), whereas our DSDN algorithm proceeds by a progressive reconciliation of two phylogenetic trees (one for each distance matrix). The described method uses the "bipartition dissimilarity" between two trees for inferring and validating horizontal gene transfer (HGT) events (Boc et al., 2010). This measure of proximity can be considered as a refinement of the Robinson and Foulds distance (Robinson and Foulds, 1981), which takes into account only identical bipartitions in the compared phylogenies. Boc et al. (2010) showed that the use of the bipartition dissimilarity as an optimisation criterion offers important improvements over the well-known least squares (used when building reticulograms as in Legendre and Makarenkov, 2002), Robinson and Foulds distance, and quartet distance measures. They also showed that this algorithm outperforms other well-known horizontal gene transfer detection methods such as LatTrans (Hallett and Lagergren, 2001) and RIATA-HGT (Nakhleh et al., 1992) in many aspects. Moreover, it includes a bootstrap validation procedure allowing one to assess the reliability of obtained HGT events (i.e., alternative dispersal routes in the biogeographic context). As horizontal gene transfers can be inferred directly from sequence data (Boc and Makarenkov, 2011), alternative dispersal routes could be also inferred from an available matrix of presence–absence data without transforming these data into a dispersal tree. However, the geographic tree, which is the backbone structure of the new method, must be always inferred or provided.

At present, only two matrices are used as input in our method, though it would be plausible to consider multiple sources of information, such as combining species composition, geographic distances and species' ecological characteristics (e.g., environmental affinities, dispersal capability, body size) to provide a more complete analysis of the processes that drove and constrained past dispersal events and current faunal distribution (see, Wiens and Donoghue (2004) for a discussion). Moreover, the integration of faunal composition (our approach) with species phylogenetic evidence is certainly interesting in the sense of thinking about the diversity of historical processes that may have taken place (Esselstyn et al., 2010) and the association of geological and speciation patterns and events. Note, however, that in our case study, there has been no speciation in the area after the last glaciation event. Finally, our method could be certainly applied to small-scale dispersal events. While dispersal dynamics for multiple species at small scales are certainly interesting, recent ecological events across large areas may produce a large noise to signal ratio in presence–absence matrices (i.e., many absences within a given species geographic range) that may obscure historical dispersal. As a result, we used cluster analyses prior to applying our method to cluster sampling units (lakes) and ensure that well-delimited faunal units were used in the method.

Our case study well illustrated the utility and robustness of the proposed method, indicating that the most important events were a south-to-north dispersal pattern, as one would expect, with secondary faunal interchange among sub-regions. Moreover, in the southern region of Ontario, most of the alternative routes (four out of five routes) were found between neighbouring sub-regions (Fig. 1). This scenario is indeed extremely plausible because these sub-regions have both the greatest concentration of water bodies and the highest fish biodiversity. The only alternative route that did not link two bordering sub-regions was the one between sub-regions 10 and 15 (Fig. 1). This exception suggests that some fishes migrated from sub-region 10 to sub-region 15, most likely through sub-region 14, and that, subsequently, fishes in the latter sub-region went extinct. The only alternative route detected between the two Ontario regions was the one from sub-region 15 to sub-region 16. This event also seems quite plausible because migration occurred from the southern region, with higher diversity, to the northern region, with less diversity. The frequency of the alternative routes found in both this study (directed networks) and that conducted by Legendre and Makarenkov (2002; undirected networks) shows that the detection of alternative dispersal pathways uncovers much more detailed information on biogeographic history and provides a better estimate of the major dispersal events that led to the main biogeographic patterns observed in present times. The large-scale patterns found in this study are particularly strong and most likely due to the fact that small-scale environmental conditions may have played a reduced role in structuring the fish faunal distribution in Ontario province. Jackson and Harvey (1989), using a much reduced data set based on only six sub-regions in Ontario (286 lakes in total), showed that the local environmental characteristics of lakes cannot explain present-day fish distribution and that postglacial dispersal likely played the most important role in structuring their fish assemblages.

Several refugia and dispersal corridors have been suggested to explain the re-colonisation and dispersal patterns of fishes into Ontario after the last glaciation (Mandrak and Crossman, 1992). However, our results indicated only two major detectable dispersal events. One of them took place in the southern and eastern sub-regions of Ontario, when the other in the northern and western sub-regions. In both regions (southern and northern), the number of species decreased moving from south to north. This is most likely due to the fact that moving northward, the weather becomes increasingly colder, and only a few species would have been able to survive in harsh environments. The southern sub-regions of the southern region of Ontario have the greatest diversity among all of the sub-regions in Ontario along with those of British Columbia (Chu et al., 2003).

The phylogenetic tree built from the COI gene sequences appears robust given that, without exception, members of each genus and family were grouped together (Fig. 3a). The main purpose of reconstructing the species phylogenetic tree along with the species dispersal pattern tree was to reveal possible relationships between the phylogenetic patterns and biogeographic distribution of Ontario fishes. There are two main processes involved in determining distributional patterns of closely related species within a biota: the positive co-occurrence of closely related species due to similar physiological limitations and niche conservatism (Weiher and Keddy, 1995; Weiher et al., 1998) and repulsion (negative co-occurrence) of species due to competitive interactions or differential environmental affinities (Chesson, 1991; Elton, 1946; Leibold, 1998; MacArthur and Levins, 1964). These two processes are referred to as phylogenetic attraction and phylogenetic repulsion, respectively (Cavender-Bares et al., 2009). A secondary aim of this study was to incorporate this ecological framework within the context of historical biogeography, in which these processes are referred to as dispersal filtering and dispersal avoidance, respectively.

Comparing the species dispersal tree with the phylogenetic trees built for 66 species, we found five similar biogeographic

clusters in the two trees. However, most of the clusters in these two trees were quite different (Fig. 3). The Robinson and Foulds distance between the two trees, which should be between 0 (if the trees are identical) and 126 (if the trees are completely different), was 109, thus suggesting that these trees are not topologically equivalent. Indeed, our phylogenetic null models showed a strong relationship between phylogeny and dispersion for only five genera and four families of the Ontario fishes (Table 2). Note that these differences are not related to dispersal avoidance (Table 2), but rather to random patterning regarding phylogenetic relationships. Perhaps, these species share similar dispersal histories that are related to environmental affinities, which, in turn, are not phylogenetically conserved. Indeed, there is evidence that environmental preferences are not necessarily phylogenetically conserved (Diniz-Filho et al., 2010), including those of fish (Peres-Neto, 2004; see also Helmus et al. (2007) for more complex analyses). Moreover, if these phylogenetic patterns are driven by complex interactions between environmental filtering, competitive interactions and biogeographic events, regions composed by a species that underwent a mix of these processes may appear as being non-structured (Leibold et al., 2010). Finally, it is arguable that a lack of strong correspondence between distributional and phylogenetic patterns may provide data that are more suitable for biogeographic reconstruction.

In conclusion, we attempted to show that, as has been found in evolutionary studies where phylogenetic networks have been proven advantageous over phylogenetic trees, the use of network-like structures, such as our DSDN framework, instead of tree-like structures, do provide much greater and detailed information about the biogeographic history of dispersals. This study should serve as a starting point for adopting or developing more versatile network reconstruction methods that could take into account other factors affecting biogeographic dispersal, such as geographic barriers, environmental conditions, climate, and species characteristics.

Acknowledgments

We would like to thank Associate Editor Naruya Saitou and two anonymous reviewers for their helpful comments and suggestions. We would like to thank the Ontario Ministry of Natural Resources for making available the fish distributional data analysed in this paper. This study was supported by the FQRNT (Fonds de Recherche sur la Nature et les Technologies du Québec) team research grant to V. Makarenkov and P. Peres-Neto and the FQRNT PhD grant to M. Layeghifard.

References

Anderson, F.E., 2002. Phylogeny and historical biogeography of the loliginid squids (Mollusca: Cephalopoda) based on mitochondrial DNA sequence data. Mol. Phylogenet. Evol. 15, 191–214.
Boc, A., Philippe, H., Makarenkov, V., 2010. Inferring and validating horizontal gene transfer events using bipartition dissimilarity. Syst. Biol. 59, 195–211.
Boc, A., Makarenkov, V., 2011. Towards an accurate identification of mosaic genes and partial horizontal gene transfers. Nucleic Acids Res. 39, e144.
Brooks, D.R., 1990. Parsimony analysis in historical biogeography and coevolution: Methodological and theoretical update. Syst. Zool. 39, 14–30.
Buneman, P., 1971. The recovery of trees from measures of dissimilarity. In: Hodson, F.R., Kendall, D.G. (Eds.), Mathematics in archaeological and historical sciences. Edinburgh Uni. Press, Edinburgh, pp. 387–395.
Cavalli-Sforza, L.L., Edwards, A.W.F., 1967. Phylogenetic analysis: Models and estimation procedures. Am. J. Hum. Genet. 19, 233–257.
Cavender-Bares, J., Kozak, K.H., Fine, P.V.A., Kembel, S.W., 2009. The merging of community ecology and phylogenetic biology. Ecol. Lett. 12, 693–715.
Chesson, P., 1991. Stochastic population models. In: Kolasa, J., Pickett, S.T.A. (Eds.), Ecological heterogeneity. Springer-Verlag, pp. 123–143.
Chu, C., Minns, C.K., Mandrak, N.E., 2003. Comparative regional assessment of factors impacting freshwater fish biodiversity in Canada. Can. J. Fish. Aquat. Sci. 60, 624–634.
Diniz-Filho, J.A.F., Terribile, L.C., Cruz, M.J.R., Vieira, L.C.G., 2010. Hidden patterns of phylogenetic non-stationarity overwhelm comparative analyses of niche conservatism and divergence. Global Ecol. Biogeogr. 19, 916–926.
Elton, C.S., 1946. Competition and the structure of ecological communities. J. Anim. Ecol. 15, 54–68.
Esselstyn, J.A., Oliveros, C.H., Moyle, R.G., Peterson, A.T., McGuire, J.A., Brown, R.M., 2010. Integrating phylogenetic and taxonomic data illuminates complex biogeographic patterns along Huxley's modification of Wallace's line. J. Biogeogr. 37, 2054–2066.
Felsenstein, J., 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39, 738–791.
Graham, C.H., Ron, S.R., Santos, J.C., Schneider, C.J., Moritz, C., 2004. Integrating phylogenetics and environmental niche models to explore speciation mechanisms in dendrobatid frogs. Evolution 58, 1781–1793.
Hallett, M., Lagergren, J., 2001. Efficient algorithms for lateral gene transfer problems. In: Proceedings of the Fifth Annual International Conference on Computational Biology, ACM, New York, pp. 149–156.
Helmus, M.R., Savage, K., Diebel, M.W., Maxted, J.T., Ives, A.R., 2007. Separating the determinants of phylogenetic community structure. Ecol. Lett. 10, 917–925.
Hubert, N., Hanner, R., Holm, E., Mandrak, N.E., Taylor, E., Burridge, M., Watkinson, D., Dumont, P., Curry, A., Bentzen, P., Zhang, J., April, J., Bernatchez, L., 2008. Identifying Canadian freshwater fishes through DNA barcodes. PLoS ONE 3, e2490.
Huson, D.H., Bryant, D., 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23, 254–267.
Jackson, D.A., Harvey, H.H., 1989. Biogeographic associations in fish assemblages: Local versus regional processes. Ecology 70, 1472–1484.
Legendre, P., 1987. Constrained clustering. In: Legendre, P., Legendre, L. (Eds.), Developments in Numerical Ecology, NATO ASI Series, pp. 289–307.
Legendre, P., Legendre, V., 1984. The postglacial dispersal of freshwater fishes in the Quebec peninsula. Can. J. Fish. Aquat. Sci. 41, 1781–1802.
Legendre, P., Makarenkov, V., 2002. Reconstruction of biogeographic and evolutionary networks using reticulograms. Syst. Biol. 51, 199–216.
Leibold, M.A., 1998. Similarity and local co-existence of species in regional biotas. Evol. Ecol. 12, 95–110.
Leibold, M.A., Economo, E.P., Peres-Neto, P., 2010. Metacommunity phylogenetics: Separating the roles of environmental filters and historical biogeography. Ecol. Lett. 13, 1290–1299.
MacArthur, R.H., Levins, R., 1964. Competition, habitat selection and character displacement in a patchy environment. Proc. Natl. Acad. Sci. USA 51, 1207–1210.
Makarenkov, V., 2001. T-REX: Reconstructing and visualizing phylogenetic trees and reticulation networks. Bioinformatics 17, 664–668.
Mandrak, N.E., Crossman, E.J., 1992. Postglacial dispersal of freshwater fishes into Ontario. Can. J. Zool. 70, 2247–2259.
MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. In: Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, Uni. California Press, pp. 281–297.
Nakhleh, L., Ruths, D., Wang, L.S., 2005. RIATA-HGT: a fast and accurate heuristic for reconstructing horizontal gene transfer. In: Wang, L. (Ed.), Proceedings of the Eleventh International Computing and Combinatorics Conference (COCOON 05). LNCS #3595, pp. 84–93.
Olden, J.D., Jackson, D.A., Peres-Neto, P.R., 2001. Spatial isolation and fish communities in drainage lakes. Oecologia 127, 572–585.
Peres-Neto, P.R., 2004. Patterns in the co-occurrence of fish species in streams: The role of site suitability, morphology and phylogeny versus species interactions. Oecologia 140, 352–360.
Robinson, D.R., Foulds, L.R., 1981. Comparison of phylogenetic trees. Math. Biosci. 53, 131–147.
Saitou, N., Nei, M., 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406–425.
Sørensen, T., 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. Biol. Skr. Dan. Vid. Sel. 5, 1–34.
Strauss, R.E., 2001. Cluster analysis and the identification of aggregations. Anim. Behav. 61, 481–488.
Weiher, E., Clarke, G.D.P., Keddy, P.A., 1998. Community assembly rules, morphological dispersion, and the coexistence of plant species. Oikos 81, 309–321.
Weiher, E., Keddy, P.A., 1995. Assembly rules, null models, and trait dispersion: New questions from old patterns. Oikos 74, 159–164.
Wiens, J.J., Donoghue, M.J., 2004. Historical biogeography, ecology, and species richness. Trends Ecol. Evol. 19, 639–644.