UNIVERSITÉ DU QUEBEC À MONTRÉAL

NETWORK FRAMEWORKS FOR TACKLING ECOLOGICAL AND EVOLUTIONARY PROBLEMS

DISSERTATION
PRESENTED
AS PARTIAL REQUIREMENT
OF THE DOCTORATE OF BIOLOGY

BY

MEHDI LAYEGHIFARD

JANUARY 2014

UNIVERSITÉ DU QUEBEC À MONTRÉAL

DES ANALYSES DE RÉSEAU POUR RÉSOUDRE DES PROBLÈMES EN ÉCOLOGIE ET ÉVOLUTION

THÈSE
PRÉSENTÉE
COMME EXIGENCE PARTIELLE
DU DOCTORAT EN BIOLOGIE

PAR

MEHDI LAYEGHIFARD

JANVIER 2014

ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my advisors Dr. Pedro R. Peres-Neto and Prof. Vladimir Makarenkov for the continuous support of my Ph.D. study and research, for their patience, motivation, enthusiasm, and immense knowledge. Their guidance helped me in all the time of research and writing of this thesis. I cannot imagine a more inspiring and enjoyable environment in which to work. I would also like to thank my lab-mates, whose companionship and support made coming into work fun: Bailey Jacobson, Renato Henriques da Silva, Alix Boc, Frédéric Boivin, Shubha Pandit, Who-Seung Lee, Andrew Smith, Dunarel Badescu, Marie-Hélène Greffard, Marie-Christine Bellemare and Wagner Moreira. Finally, I would like to thank Razi, the love of my life, and family and friends for their support and encouragement.

This thesis was funded 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 an FQRNT PhD grant to M. Layeghifard.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

RESUMÉ

Notre vision de l'écologie et de l'évolution a beaucoup changé au cours des dernières décennies due aux découvertes des mécanismes complexes gouvernant les différents aspects de la vie, des cellules, aux populations, aux espèces et encore, aux communautés et métacommunautés. Par contre, l'analyse de ces phénomènes complexes nécessite le développement de nouveaux concepts de même que de nouveaux outils informatiques rapides et fiables. Un de ces nouveaux concepts, la théorie des graphes, gagne rapidement en popularité dans les domaines de l'écologie et de l'évolution grâce à des avancées théoriques et informatiques. L'objectif principal de ce doctorat est de développer un cadre d'étude basé sur la théorie des graphes afin de résoudre des problèmes possédant des caractéristiques de réseaux en écologie et en évolution (p. ex., évolution réticulée ou connectivité spatiale entre des communautés). Dans cette thèse, quatre problèmes différents sont abordés. Bien que les entités biologiques diffèrent entre les problèmes (variant des espèces aux communautés), ceux-ci peuvent tous être approchés par des approches de réseaux similaires. Ces quatre problèmes (un par chapitre) représentent chacun une contribution originale dans l'application méthodologique des réseaux : 1) construire des réseaux phylogénétiques consensus à partir de données contenant des signaux évolutifs contradictoires ; 2) retracer l'historique de dispersion des espèces ; 3) explorer l'hétérogénéité spatiale des metacommunautés ; et 4) mesurer la connectivité dans des réseaux de métacommunautés. Les résultats obtenus de l'application de ces méthodologies sur des données empiriques et/ou simulées démontrent que la complexité inhérente à plusieurs problèmes en écologie et en évolution peut être explorée et résolue à l'aide d'approches basées sur la théorie des graphes. Ainsi, la théorie des graphes, un outil flexible et robuste pour l'analyse de problèmes complexes, a un grand potentiel pour améliorer notre compréhension des systèmes en écologie et en évolution.

ABSTRACT

Our vision of ecology and evolutionary biology has changed significantly during the past few decades due to the discovery of a plethora of complex mechanisms governing the various aspects of life, from cells to populations to species to even more complex ecological entities (communities and metacommunities). However, the analysis and exploration of such complex problems needs new concepts, as well as reliable as faster computational tools. One of the relatively new and increasingly popular concepts in ecological and evolutionary biology studies is graph theory owing to the recent advances in computer technology. The main objective of this doctoral thesis is to develop frameworks based on graph theory to tackle complex ecological and evolutionary biology problems involving network characteristics (e.g., reticulated evolution, spatial connectivity across ecological communities). In this thesis, I have chosen four different problems involving ecological and evolutionary networks. The biological entities are different (from species to ecological communities) but they can be all tackled by related network approaches. These problems were tackled by four chapters that represent each novel network applications: 1) building consensus phylogenetic networks from datasets containing conflicting evolutionary signals, 2) retracing dispersal history of species, 3) exploring the spatial heterogeneity of metacommunities, and 4) measuring the connectivity of metacommunity networks. The results obtained from the application of these methodologies on real and/or simulated datasets showed that the inherent complexity of many ecological and evolutionary biology problems can be successfully explored, explained and resolved by using graph-theoretical approaches. Network theory has the potential to significantly improve our understanding of ecological systems and evolution because it is a flexible and robust tool to tackle most problems in these fields.

INTRODUCTION

0.1 Background

One of the outstanding characteristics of biological systems (ecological and evolutionary) is that they are complex in both structure and functions due to their dynamic nature, compositional variability and their ability to self-reproduce and self-organize. In the ecological context, one of the main factors contributing to this biological complexity is species interacting with one another and with their surrounding environment. For example, we have just started to understand the relationship between humans and their intestinal bacterial and archaeal flora, which involves many interactions and regulations between the host and symbiont genes (Gill *et al.*, 2006). On the other hand, the recent advances in molecular biology and high-throughput analyses have dramatically changed our vision of evolutionary biology. There are numerous mechanisms contributing to the complexity of molecular biology, such as alternative splicing, post-translational modifications and the presence of micro RNAs and interference RNAs, just to name a few examples. These mechanisms are also likely to play an important role in molecular evolution, thus contributing to its complexity.

The interactions between the components of complex biological systems can be well represented as networks. For example, metabolic networks of biochemical reactions (Karp *et al.*, 2005; Ravasz *et al.*, 2002); protein-protein interaction networks of the physical interactions between proteins (Giot *et al.*, 2003; Li *et al.*, 2004); and the transcriptional (or gene) regulatory networks of the regulatory interactions between various genes (Ihmels *et al.*, 2002; Salgado *et al.*, 2006; Shen-Orr *et al.*, 2002) are among the most well-known biological networks. The above-mentioned biological networks have numerous potential applications within the fields of Biology and

Medicine, such as determining the evolution and functions of the unknown proteins or genes, identifying potential drug targets, unravelling complex biochemical regulatory pathways, and understanding the range and mechanisms of infectious diseases outbreaks (Eubank *et al.*, 2004; Jeong *et al.*, 2003; Samanta and Liang, 2003).

In fact, networks exist at all scales of biological organization, from single cells to large metacommunities and, traditionally, graph theory is the first choice and the most capable tool to investigate such complex networks. Interestingly, many initial efforts to model biological systems involved the use of random graphs (Barabási and Albert, 1999). However, it is too simplistic to think of real networks (i.e., as opposed to artificial or anthropogenic networks such as social networks on the internet) behind such diverse complex systems as random graphs. If these biological networks are not random, then we need to develop tools, measures and frameworks to study and analyze their organization, characteristics and behaviour. Fortunately, the recent technological advances in computer sciences have led to a dramatic growth in the use of graph theory to investigate biological networks.

In this thesis, however, the focus is on two particular types of biological networks: ecological networks and phylogenetic networks. The main goal here is to take advantage of the exceptional potentials of graph theory and computer science as well as available data in order to design and develop novel efficient computational tools and frameworks for tackling some of the complex issues in the fields of ecology and evolutionary biology. Moreover, by using problems from these two fields, one is able to observe how different problems often converge to somewhat similar solutions. Based on these premises, four different ecological and evolutionary biology questions have been chosen to be addressed using graph-theoretical approaches. These problems included 1) resolving gene tree discordancy and detecting unorthodox evolutionary pathways (e.g., horizontal gene transfers, recombination events); 2) retracing species dispersal history; 3) detecting spatial variability in metacommunities; and 4) estimating

the connectivity of biological networks. In common, they share the 'transfer' as a common theme; the transfer of genes among species and the transfer of species among large biogeographic zones and small local ecological communities. The reason for selecting these four seemingly unrelated problems was to showcase the potential and the versatility of network theory in solving complex biological issues across the fields of ecology and evolutionary biology. With the rise of network applications in medicine, social sciences and computer sciences, among others, it seems inevitable for ecologists and evolutionary biologists to take network thinking more seriously (May, 2006). Following the advances in other fields and mainly to keep pace with advances in life sciences and information technologies, we must be on track to design and develop similar tools to tackle the large-scale data problems we face now. In order to understand, organize, model and study large-scale data we need tools far more powerful and complex than classic methods. Moreover, because different ecological and evolutionary problems often require similar computational solutions, my attempt here is also to demonstrate the flexibility of the network based approaches developed in this thesis. This flexibility is particularly useful in the age of data revolution where having access to multi-purpose tools will save us time and money.

Networks are excellent tools to represent many features and processes of ecological and evolutionary systems. Specifically, their incomparable value becomes apparent in cases where the problem involves large datasets in order to reveal patterns behind small and large-scale ecological and evolutionary processes (Proulx *et al.*, 2005). Moreover, the need to move away from a purely reductionist approach in favour of an integrative, systems-oriented approach has been recently promoted by many researchers (see Mason and Verwoerd, 2007 for a review). Since all biological systems are, indeed, sets of interacting components, the application of network theory becomes a natural way to tackle scientific questions within such complex systems. Therefore, this thesis is aimed at contributing to the graph-theoretical toolbox of ecologists and evolutionary biologists and promoting the application of network theory (i.e., network thinking) in

these fields.

In the following sections, brief descriptions of ecological and phylogenetic networks are given. In addition, some of the foundational concepts underlying the four chapters of the thesis are presented. Finally, at the end of the Introduction section an outline of the main four chapters of this thesis is provided.

0.2 Phylogenetic networks

One of the main goals of evolutionary biology is to reconstruct phylogenetic trees which accurately represent the evolutionary history of a group of species. In phylogenetic trees, each leaf represents an existing species, while the internal vertices correspond to hypothetical ancestors, and edges (also called branches) show the relationships between ancestors and their descendants.

Vast progress in the field of molecular biology in the last few decades has profoundly changed the nature of the datasets used in phylogenetic analysis. Initially, the only available data for building evolutionary trees were morphological characters, but nowadays, biological sequence data (nucleotide or amino acid sequences) are mostly used to infer the history of life. These data sets are produced with the aid of efficient DNA and protein sequencing technologies and the comprehensive computer-based analysis of the results. These data are maintained in huge freely available and publicly accessible databases such as GenBank and EMBL among others. Given that the amount of data available in these databases are growing exponentially, it is vital to analyze these data in a fast, efficient, and accurate manner in order to make use of their results to tackle both theoretical and applied questions in evolutionary biology and ecological and societal contemporary problems.

In phylogenetics, this means that algorithms and applications have to be developed with the aim of analyzing and modelling the diverse and complex processes that have occurred during the evolution of any given set of current species. So far, many efforts have been made to develop efficient methods in order to reconstruct phylogenies that best represent the evolutionary history for different sets of taxa. Since evolution just occurred once in the past, there is no direct observational or experimental study that may be used in phylogenetic reconstruction. Moreover, the fossil record is often incomplete and ambiguous. Therefore, evolutionary biologists have to mostly rely upon mathematical and statistical models for analyzing the sequence data of existing species in order to infer phylogenetic trees and understand past events that led to speciation and other evolutionary patterns (Wiens, 2009).

Essentially, there are three types of methods for phylogenetic tree reconstruction: (1) distance-based methods like UPGMA (unweighted pair group method with arithmetic mean) and neighbor-joining, (2) parsimony-based methods like maximum parsimony, and (3) statistical-based methods like maximum likelihood and the closely related Bayesian method. A detailed description of phylogeny reconstruction methods can be found in Felsenstein (2004).

Phylogenetic networks are a generalization of evolutionary trees that make possible the simultaneous visualization of several conflicting or alternating histories of life. In a phylogenetic network, each conflicting or alternative history event is usually represented as an extra branch (or a link between two species or clades involved in the event) added to the phylogenetic tree. Thus, these extra branches or links (also called reticulation events) convert a simple phylogenetic tree, which at best can only represents one dominant hypothesis, to a phylogenetic network which can represent multiple conflicting or alternative historical hypotheses. Indeed, there are several types of events that lead to histories that are not adequately modelled by a single tree (Huson and Bryant, 2006; Legendre and Makarenkov, 2002): (1) horizontal gene transfer in

bacterial evolution; (2) hybridization between species, including allopolyploidy in plants; (3) micro-evolution of local populations within a species, involving genetic differentiation of allopatric populations, gene exchange through migration, or both; (4) homoplasy, the portion of phylogenetic similarity resulting from evolutionary convergence (e.g., parallel evolution and reversals), which can be represented by reticulation branches added to a phylogenetic tree; and non-phylogenetic situations, such as (5) host–parasite relationships involving host transfer and (6) vicariance and dispersal biogeography.

Even if the relationships between species are tree-like, phenomena like sampling error, parallel evolution, or model heterogeneity can also generate difficulties in representing evolution by a single tree (Gascuel, 2005). Generally speaking, there exist two fundamental types of phylogenetic networks, namely: (1) explicit networks that provide a concrete scenario of reticulate evolution and (2) implicit networks that are intended to represent incompatible signals in a data set (see Figure 3 in Huson and Bryant, 2006). An explicit network is generally depicted as a phylogenetic tree with additional edges. The internal nodes in such a network represent ancestral species, and nodes with more than two parents correspond to reticulate events such as hybridization or recombination. Explicit networks model non-tree-like evolution and their purpose is to point out which lineages have undergone reticulation events. Implicit approaches, on the other hand, are often based on split networks which represent all splits contained in a set of gene trees. Each parallelogram of the resulting network corresponds to two incompatible splits. To be able to accommodate incompatible splits, it is often necessary that a split network contains nodes that do not represent ancestral species. Thus, split networks provide only an "implicit" representation of evolutionary history. Phylogenetic networks will be discussed in detail in Chapters I and II.

0.3 Ecological networks

In ecology, the components (i.e., biotic and abiotic objects or entities) that construct a system show varying degrees of interactions. These interactions can be represented as an ecological network in which the components are indicated as nodes (i.e., vertices in graph terminology) and the interactions are depicted as links between the nodes (i.e., edges in graph terminology). These interactions, among other types, can be trophic, competitive, symbiotic, social and geographic connectivity. Ecological networks are very useful models to describe, analyze and compare the structure of ecological systems. For example, they are often used to investigate the effects of network structure (i.e., topology) on the properties of ecological systems such as their stability (Dunne *et al.*, 2002).

Traditionally, ecological networks were first developed and used to model trophic relationships within food webs (Lindeman, 1942; Odum, 1965). Food webs are important components of every ecological system due to the feeding is essential for organisms' survival. In food webs organisms are connected directly through feeding. Networks have been used to model food webs, explore their stability and determine if certain network properties result in more stable networks (MacArthur, 1955). Given that the local extinction of a species within a given ecological system may result in an unstable food web, network analysis have been used to determine how removal of species do influence food webs as a whole (Dunne *et al.*, 2002).

Another type of ecological network is species interaction networks which consist of pairwise interactions between individuals of one or more species. Network analysis of species interactions allows quantifying the associations between individuals and inferring details about the network as a whole. Moreover, the power and flexibility of network approaches allow for the study of various types of interactions (e.g., social, competitive, predatory, cooperative and mutualistic interactions) using the same general approach. As such, ecological networks are useful in analyzing numerous

complex interactions within most ecological systems (Krause *et al.*, 2009; Ryder *et al.*, 2008).

Additional applications of ecological networks include exploring complex interactions at the multi-species levels in terms of both species dispersion and coevolution of pairs of species. In this thesis, network models were developed to study metacommunities, which are particularly complex given their relative large geographic extent, their heterogeneous landscapes and their multi-species composition. Since metacommunities involve large scale problems, there is no direct observational or experimental study that may be used to understand some of the processes (e.g., dispersal history) underlying their structure. These applications are further discussed in Chapters II, III and IV.

0.4 Biogeography

Historical biogeography studies show how ecological processes that happen over long periods of time influence the distributional patterns of living organisms (Cox and Moore, 1993). Conversely, studying the same processes acting in short periods of time is called ecological biogeography. Biogeography as a whole is a multidisciplinary science with a long history. Indeed, the study of plant and animal distributions has a history as long as biology itself.

It is accepted that the scientific theory of biogeography likely grew out of the work of Alfred Russel Wallace (1823-1913) and other early evolutionary scientists. Wallace studied the distribution of flora and fauna of the Malay Archipelago in the 19th century. One of the interesting subjects in historical biogeography has been the study of the effects of Pleistocene glaciations on the distribution of living organisms. However, some authors place this subject between ecological and historical biogeography,

because the processes involved acted for only several thousand years which is not considered a long period of time in geography (Myers and Giller, 1988).

The aim of biogeography is to reveal where species live, why, and at what abundances through the study of the distribution of biodiversity over space and time (Martiny *et al.*, 2006). One of the most impressive features of our planet is the sheer diversity of organisms it contains, and one of the main problems facing scientists is how to explain this diversity, and the reasons for the varying patterns of occurrence of different species over the surface of the planet or in particular large landscapes. Moreover, biogeography is about seeking general rules that can account for distributional patterns and provide a general framework to generate insights that can subsequently be used for predictions about the consequences of upcoming phenomena.

Patterns of species distributions can be usually explained through a combination of historical factors such as speciation, extinction, continental drift, glaciation (and associated variations in sea level, river routes, among other factors), and river capture, in combination with the area and isolation of landmasses (geographic constraints) and available resources. All these factors are the results of the interaction between two great natural phenomena: evolution and plate tectonics. Although, nowadays, biogeography is an independent discipline with a core of accepted knowledge and methodological principles, it is also an adjunct whose status is contingent on other areas of study such as ecology, evolution, taxonomy, molecular systematics, geography, geology, and palaeontology. For instance, phylogenetic networks in which the relations between regions within a landscape are represented by branches could be used to explore the hypothesis that multiple dispersal routes were used by a particular species of interest to migrate from one region to another. The application of network theory on biogeography is the focus of Chapter II.

0.5 Metacommunity

In ecology, a community is a group of populations of two or more different species occupying the same geographical area. Community ecology is primarily concerned with patterns of species distributions, abundance and interactions across different spatial and temporal scales. As an extension, an ecological metacommunity is consisted of a set of local interacting communities that are interconnected through dispersal (Leibold *et al.*, 2004).

Metacommunities have been defined and studied based on four major perspectives: 1) patch dynamics; 2) species sorting; 3) source–sink dynamics (or mass effect); and 4) neutral model. These four theoretical frameworks were developed in order to explore specific processes underlying community patterns. Patch dynamics models are mainly used to describe species composition among multiple habitat patches, such as islands. The focus in patch dynamics is on the possible coexistence due to competition-dispersal, competition-colonization or dispersal-fecundity trade-offs. Conversely, species sorting models try to link the variation in abundance and composition within the metacommunity to similar and differential responses of the species to environmental heterogeneity. Source-sink models, on the other hand, are based on the assumption that dispersal and environmental heterogeneity interact to determine local and regional abundance and composition. Finally, in the neutral framework species are considered essentially equivalent in their competitive and dispersal abilities. Therefore, stochastic demographic processes and dispersal limitation are the primary factors determining the local and regional composition and abundance (Leibold *et al.*, 2004). Spatial heterogeneity and connectivity of metacommunities will be further investigated using networks in Chapter III and Chapter IV, respectively.

0.6 Thesis outline

This thesis is comprised of the following four chapters:

Chapter I	Inferring explicit weighted consensus networks to represent alternative evolutionary histories
Chapter II	Using directed phylogenetic networks to retrace species dispersal history
Chapter III	Spatial networks for inferring dispersal in ecological communities
Chapter IV	A novel connectivity measure for metacommunity networks

Chapter I emphasizes the application of networks in evolutionary biology and phylogenetics. It is comprised of a novel weighted explicit method to construct consensus phylogenetic networks. Moreover, this method is capable of detecting different reticulation events such complete horizontal gene transfers, partial horizontal gene transfers, recombination and hybridizations. This method was also successfully tested and assessed by both empirical and simulated datasets. Chapter II is primarily concerned with the application of networks in biogeography. Specifically, it includes a new network methodology that is developed to retrace species dispersal history. This new method was successfully applied on an empirical dataset in order to reconstruct the historical dispersal events that occurred when fish species left southern refugia to recolonize the northern Ontario province after the last glaciation period. Chapter III focuses on the use of network theory to investigate the spatial heterogeneity within large multi-species ecological systems. In this chapter, a novel graph-theoretical method was developed to capture and explore the spatial variation within metacommunities. This new method was successfully tested on both empirical and simulated datasets. Finally, Chapter IV investigates the application of graph theory in detecting connectivity in metacommunities. In this chapter, a new connectivity

measure was developed to be specially applied on metacommunities. This connectivity measure was successfully tested on simulated datasets.

CHAPTER I

INFERRING EXPLICIT WEIGHTED CONSENSUS NETWORKS TO REPRESENT ALTERNATIVE EVOLUTIONARY HISTORIES

Mehdi Layeghifard, Pedro R. Peres-Neto and Vladimir Makarenkov
Published in BMC Evolutionary Biology.

1.1 Summary

The advent of molecular biology techniques and constant increase in availability of genetic material have triggered the development of many phylogenetic tree inference methods. However, several reticulate evolution processes, such as horizontal gene transfer and hybridization, have been shown to blur the species evolutionary history by causing discordance among phylogenies inferred from different genes. To tackle this problem, we hereby describe a new method for inferring and representing alternative (reticulate) evolutionary histories of species as an explicit weighted consensus network which can be constructed from a collection of gene trees with or without prior knowledge of the species phylogeny. We provide a way of building a weighted phylogenetic network for each of the following reticulation mechanisms: diploid hybridization, intragenic recombination and complete or partial horizontal gene transfer. We successfully tested our method on some synthetic and real datasets to infer the above-mentioned evolutionary events which may have influenced the evolution of many species. Our weighted consensus network inference method allows one to infer, visualize and validate statistically major conflicting signals induced by the mechanisms of reticulate evolution. The results provided by the new method can be used to represent

the inferred conflicting signals by means of explicit and easy-to-interpret phylogenetic networks.

1.2 Introduction

Molecular data have played an instrumental, and usually indispensable, role in many phylogenetic and evolutionary studies in the recent decades. Their increasing availability is due to outstanding advances in the development of fast, efficient and affordable sequencing technologies (Pettersson *et al.*, 2009). Although this growth has triggered the advancements of theoretical informatics aspects of phylogenetics and evolutionary biology via the development of new algorithms, statistical models and software, fast and effective analytical methods have yet to be designed to take advantage of this huge surplus of data. For instance, the field of phylogenetics still faces some key analytical challenges stemming from reticulate evolution. They include: 1) horizontal gene transfer (e.g., in bacterial or viral evolution); 2) hybridization among species (e.g., allopolyploidy in plants); 3) genetic differentiation of allopatric populations and gene exchange through migration; 4) homoplasy (i.e., parallel evolution and reversals); 5) incomplete lineage sorting; and 6) recombination between genes (Huson and Bryant, 2006; Huson *et al.*, 2010; Legendre and Makarenkov, 2002; Posada and Crandall, 2001). All these processes may lead to the incongruity among gene trees (Giribet *et al.*, 2001; Grechko, 2013; Mason-Gamer and Kellogg, 1996; Rokas *et al.*, 2003; Zou and Ge, 2008) inferred from the data affected by reticulate evolutionary mechanisms. Implicit or explicit phylogenetic networks should be used to represent these complex phenomena when the gene tree incongruity is observed (Huson *et al.*, 2010; Makarenkov and Legendre, 2004). Implicit networks are better suited for a general representation of conflicting evolutionary signals present in the data, whereas explicit networks are used for depicting the precise reticulation

events, including their directionality and the species involved. The inference and validation of explicit phylogenetic networks is the main goal of the current study.

Another key factor that contributes to the incompatibility among gene trees is stochastic errors resulting from analytical features such as choice of optimality criterion, taxon sampling and sequence evolution model (Graybeal, 1998; Huelsenbeck, 1995; Yang *et al.*, 1994). These complications not only makes it difficult for researchers to find reliable estimates of the true species phylogenies, but also obstruct such fields as comparative biology and community phylogenetics which rely on phylogenetic trees in their analyses (Harvey and Pagel, 1991; Peres-Neto, 2012; Webb, 2002).

Evidence from many studies conducted on different groups of species, from fruit flies to hominids (Burbrink and Pyron, 2011; Carstens and Knowles, 2007; Ebersberger *et al.*, 2007; Grechko, 2013; Jennings and Edwards, 2005; Pollard *et al.*, 2006; Sánchez-Gracia and Castresana, 2012; Syring *et al.*, 2007; Takahashi *et al.*, 2001), have shown that gene tree discordance is a widespread phenomenon. These studies mostly concluded that rarely a predominate or consistent single-gene-based phylogeny could be perceived or reconstructed for a moderate to large set of species, regardless of the type of phylogenetic data at hand. Among the traditional tree-like techniques developed to solve the gene tree incongruence problem there are two widely used approaches of gene concatenation and consensus tree reconstruction, both of which result in the inference of a single phylogenetic tree as the most probable representation of the evolutionary history of species.

Although, there have been successful cases of using the concatenation approach to elucidate the ancestral relationships among certain groups of species (Baldauf *et al.*, 2000; Chen and Li, 2001; Moreira *et al.*, 2000; Soltis *et al.*, 1999), multi-gene datasets very rarely converge to the same phylogeny, more often providing results which are contradictory or inconsistent with well-known and highly reliable species tree (Giribet

et al., 2001; Hwang *et al.*, 2001; Mossel and Vigoda, 2005; Naylor and Brown, 1998). These statistical inconsistencies in estimating phylogenetic trees using concatenated datasets have been confirmed by simulation studies (Kolaczkowski and Thornton, 2004; Kubatko and Degnan, 2007).

The main idea behind traditional *consensus tree* reconstruction methods is that each of the phylogenetic trees from a given collection of trees should contribute to a consensus tree according to the presence of its clusters. Among the most known and widely used consensus tree reconstruction methods are the majority rule consensus (Margush and McMorris, 1981) and Nelson (often called Nelson-Page) consensus approaches (Nelson, 1979; Page, 1989). The traditional strict majority rule consensus tree includes all the clusters that occur in more than 50% of the considered trees. The major pitfall of this method is that for a set of trees with a poor overall bootstrap support, the 50% cluster occurrence constraint leads to a very weakly resolved phylogeny. On the other hand, in the extended majority rule consensus tree approach, a strict consensus tree is first constructed and then the remaining compatible clusters are added to it following their overall frequency in the considered tree collection. For the collections of trees with a poor overall bootstrap support, the constraint of 50% used when inferring the majority rule and extended majority rule consensus trees can be often inconvenient. Many existing software allow for clusters that are present in less than 50% of the trees. They work downwards in the frequency of the cluster occurrences as long as the new clusters aid to resolve the consensus tree. The extended majority rule consensus method often provides solutions similar to those of the Nelson consensus method, although not necessarily identical to them (Nelson, 1979; Page, 1989). The Nelson consensus method, first described in (Nelson, 1979) and then generalized in (Page, 1989), relies on the graph theory techniques to find maximum cliques of mutually compatible clusters. Its major drawback is that these cliques do not always contain enough compatible clusters to constitute a fully resolved phylogenetic tree (Bryant, 2003).

Moreover, the problem of finding a maximum clique of compatible clusters has been shown to be NP-hard (Abello *et al.*, 1999).

Phylogenetic networks should be used instead of consensus species trees or species trees inferred from concatenated sequences whenever reticulate evolutionary processes are studied (Huson *et al.*, 2010; Legendre and Makarenkov, 2002; Makarenkov and Legendre, 2004). Here, we recall some of the existing phylogenetic network building methods and software based on the cluster support. In an early attempt to build *consensus phylogenetic networks*, Holland *et al.* (2004) developed an implicit consensus network model based on the median network method (Bandelt *et al.*, 1999) to visualize incompatibilities encompassed in the given collection of trees. This method proceeds first by ranking all the splits according to their frequency and then builds a system of compatible splits by adding those splits to the network, one at a time, following their frequency ranking. Holland and colleagues (Holland *et al.*, 2006) further optimized their original greedy consensus network method to incorporate weights from individual trees into the network inference process. Having the length of each split (i.e., branch length of the split branches) in different trees as well as the weights associated with those trees, this method computes an average length for each split and finally selects compatible splits based on their weights to build a consensus network.

In another attempt, Huson (1998) and then Huson and Bryant (2006) have developed a computer program called *SplitsTree* which reconstructs an unrooted splits graph from a collection of phylogenetic trees through selecting all the splits that are present in more than a fixed percentage of all the trees (Holland *et al.*, 2004). However this program provides as result only implicit network structures; the inferred extra links do not usually directly correspond to the tree lineages and the number of nodes and edges of the resulting network can grow exponentially with the number of splits. To address these disadvantages, Huson and Rupp (2008) proposed the *cluster network* approach to

build a phylogenetic network from a collection of gene trees using a modified *tree popping algorithm* which they called *network popping algorithm*. To estimate the support of any reticulation edge, the average support of that edge (computed over all trees) is divided by the average support of the alternative reticulation edges located at the same position and weighted by the average support of all other tree edges (Huson and Rupp, 2008; Huson *et al.*, 2010). The latter authors stated however that no association between clusters and reticulation edges is provided by this method. For instance, the obtained cluster support was not shown in their network representations (Huson and Rupp, 2008). On the other hand, Abby et al. (2010) proposed a horizontal gene transfer inference method called Prunier. Prunier needs a species tree and a gene tree as a reference and does not treat multiple gene trees. Prunier relies on a ranking of branches that are common to the species and gene trees based on the amount of conflicts that is reduced when the branch is removed. This amount of conflicts is a function that depends on the statistical support of the internal branches of the gene tree. For a detailed review of the existing phylogenetic network reconstruction methods the reader is referred to (Huson *et al.*, 2010). Note that the results yielded by most of the existing consensus network building methods are implicit and generally not easy to interpret.

In this study, we present a new algorithm for the inference of explicit weighted consensus networks from a collection of trees (e.g., multiple single-gene phylogenies), with or without prior knowledge of the species phylogeny. Such networks are capable of representing the main historical pattern of the species evolution (i.e., associated with the clusters present in the species tree) as well as the alternative evolutionary routes characterizing the species and genes under consideration. The main advantage of our method is that it allows for visualizing the species evolutionary relationships in a very clear and easy-to-interpret manner. Our algorithm takes advantage of the weights (e.g., least-square scores, posterior probabilities, maximum likelihood scores or *p*-values) assigned to the gene trees as well as the weights associated with the tree clusters (e.g.,

cluster's bootstrap score or posterior probability) to infer the species dominant and alternative evolutionary histories. If a species tree is provided in addition to the collection of gene trees, our algorithm considers it as the dominant evolutionary history (i.e., backbone structure) and uses the collection of gene trees to infer the most significant reticulation events. If only a collection of gene trees is given, the new algorithm first builds a weighted consensus tree as the main evolutionary pattern and then infers the most significant alternative events.

The rest of the article is organized as follows. In the Methods section, a description of the basic concepts of phylogenetic networks and a detailed presentation of our new algorithm are given, followed by the description of the simulation protocol and the three considered real datasets. In the Results section, the results and performances of the new algorithm obtained for both simulated and real data are reported. They are then discussed in detail in the final section of the article.

1.3 Methods

1.3.1 Basic concepts

1.3.1.1 Graph

A graph G (V, E) consists of a collection of vertices (V) which are connected by a collection of edges (E) in a pairwise manner. A path in a graph is a sequence of at least two vertices (v_1, v_2, ..., v_k) such that, for all $i \in \{1, 2, ..., k\text{-}1\}$, there exists an edge $\{v_i, v_{i+1}\}$ in E. A cycle in a graph is a path whose first and last vertices are the same, while all other edges and vertices are pairwise distinct.

1.3.1.2 Phylogenetic tree

A *phylogenetic tree* (T) is an acyclic connected graph whose leaves (i.e., vertices of degree one) are labelled according to the given set of taxa (i.e., species). Phylogenetic trees can be either bifurcating (i.e., all the internal nodes have an indegree of one and an outdegree of two) or multifurcating (i.e., internal nodes can have an outdegree of three and more). Phylogenetic trees can be rooted or unrooted, where the root is a node representing a common ancestor of all the species involved in the analysis.

1.3.1.3 Phylogenetic network

A *phylogenetic network* is a connected graph used either to visualize evolutionary relationships between species or to display conflicting evolutionary signals without such limitations as being acyclic or having a fixed indegree or outdegree of its nodes. Phylogenetic networks can be implicit or explicit: implicit networks such as split graphs are used to represent conflicting and ambiguous signals in a dataset using parallel sets of edges, rather than single branches. These networks often contain nodes that are not representing any ancestral species, hence providing only an *implicit* representation of evolutionary histories (Huson and Bryant, 2006). In explicit networks, in contrast, the internal nodes represent ancestral species and nodes with more than two parents correspond to reticulation events such as hybridization, recombination or horizontal gene transfer. Such networks provide an *explicit* representation of evolutionary history of species (see Huson *et al.*, 2010 for more details). Here, we will first define some basic principles of the weighted consensus tree reconstruction prior to expanding them to phylogenetic networks inferring.

1.3.1.4 Bootstrap-based majority rule consensus tree

The main idea of our approach is that each phylogenetic tree from a given collection of trees should contribute to a consensus tree not simply by the presence, but also by the quality of its clusters (i.e., bipartitions or splits corresponding to the internal tree branches). The quality of a cluster within a given collection of trees can be defined as the sum of bootstrap scores, taken over all the trees in this collection, of the internal branches associated with this cluster. The traditional majority rule consensus tree includes only the clusters that exist in more than 50% of the considered trees (Margush and McMorris, 1981). Note that any other percentage between 50% and 100% can be also specified in most of the existing phylogenetic packages (e.g., in PHYLIP; Felsenstein, 1989). The *bootstrap-based majority rule consensus tree* will include any cluster whose average bootstrap support, i.e., total sum of bootstrap scores, computed over all the trees in the collection, divided by the number of trees in this collection, is greater than 50% (e.g., tree T_{bm} in Figure 1.1 is the bootstrap-based majority rule consensus tree, as well as the strict majority rule consensus tree, of trees T_1, T_2 and T_3). It is easy to prove that all the clusters satisfying such a rule will be pairwise compatible. For this, it will be sufficient to substitute each tree of the original tree collection by the set of its bootstrap replicates (i.e., replicated trees built when carrying out the bootstrap procedure) and then apply the traditional strict majority rule method on this extended set of replicated trees. All the clusters appearing in more than 50% of the replicated trees will be mutually compatible.

1.3.1.5 Bootstrap-based extended majority rule consensus tree

Similar to the traditional extended majority rule method, as implemented in the CONSENS program of the PHYLIP package (Felsenstein, 1989), the *bootstrap-based extended majority rule* method is a two-stage procedure. First, any cluster whose

average bootstrap score is greater than 50% will be included in the consensus tree. Then, the method will consider the remaining clusters following the order of their total sums of bootstrap scores, computed over all the trees in the collection, and gradually add to the consensus tree those that are compatible with the current consensus tree until the tree is fully resolved or no more compatible clusters remains. For instance, tree T_{bem} in Figure 1.1 is the extended bootstrap-based majority rule consensus tree of trees T_1, T_2 and T_3.

1.3.1.6 Bootstrap-based Nelson consensus tree

We also consider the following extension of the traditional Nelson method. To build the *bootstrap-based Nelson consensus tree* each clique will be assigned a score equal to the sum of scores of clusters included in it. The score of each cluster is defined as a sum of bootstrap scores associated with this cluster, computed over the given collection of trees. Unlike the method described by Page (1989), where only the replicated clusters can contribute to the clique scores, our procedure also takes into account the scores of all unreplicated clusters. If a single clique with the highest total bootstrap score is found, the group of compatible clusters included in this clique will define the bootstrap-based Nelson consensus tree. If there exist more than one such clique, then the bootstrap-based Nelson consensus tree will contain only the clusters found in all of the maximal replication cliques. In this case, clusters found in some, but not all, of the maximal-replication cliques can be classified as "ambiguous" (for more details see Felsenstein, 1989; Page, 1989; Swofford, 1991). In some cases, the bootstrap-based extended majority tree and Nelson consensus tree will be identical (e.g., tree T_{bem} in Figure 1.1 is also the Nelson consensus tree of trees T_1, T_2 and T_3), but this equivalence does not hold in general.

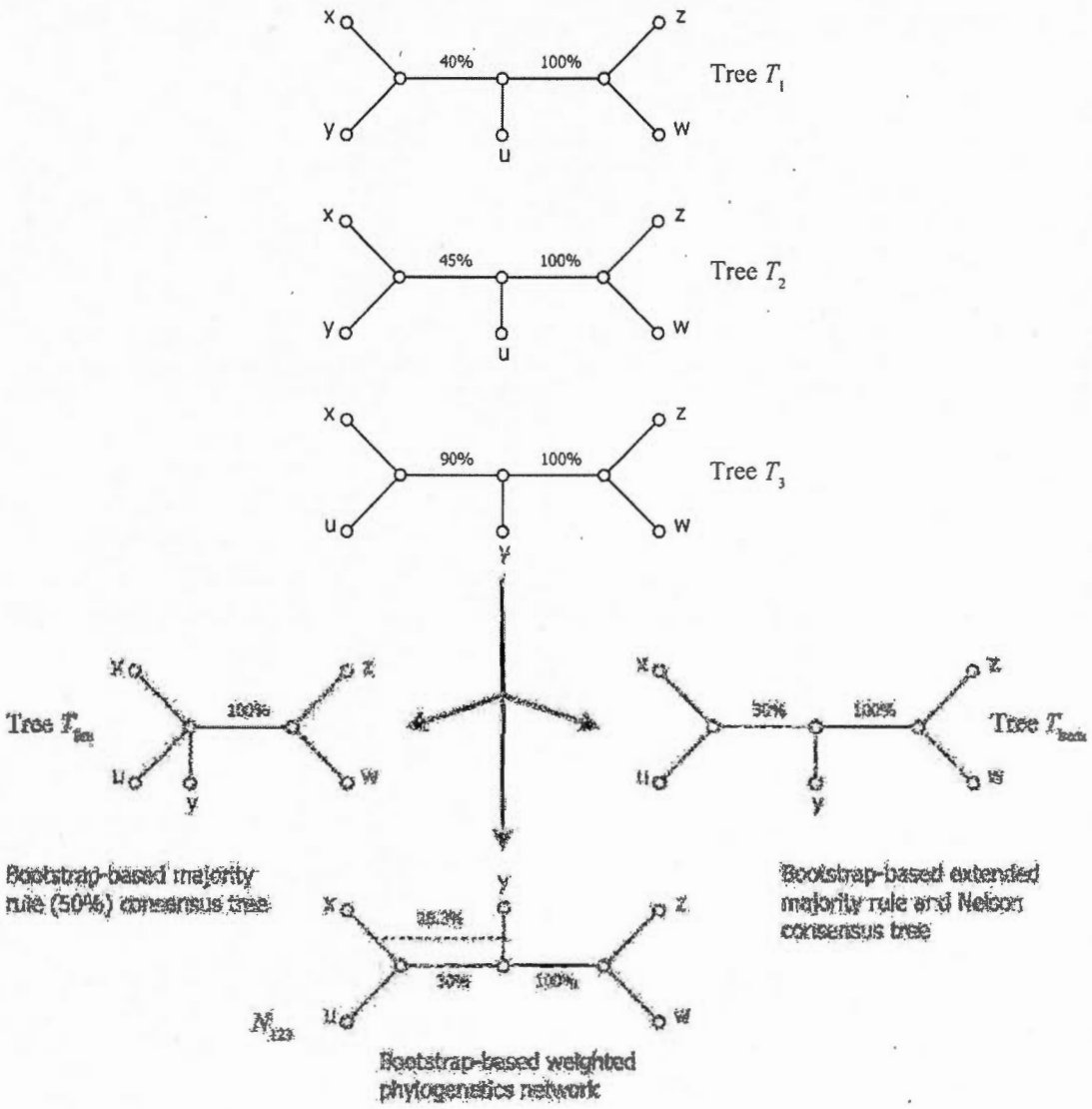


Figure 1.1 Bootstrap-based consensus trees and networks. Bootstrap-based majority rule consensus tree T_{bm}, bootstrap-based extended majority rule consensus tree T_{bem} and weighted implicit phylogenetic network N_{123} for a collection of three binary phylogenetic trees T_1, T_2 and T_3 whose leaves are labelled by the set of 5 taxa (x, y, z, w and u). The bootstrap scores of the internal branches of trees T_1, T_2 and T_3 are indicated. All the trees have the same weight.

In Figure 1.1, a set of three trees is presented (T_1, T_2 and T_3), each of them containing two internal branches with the associated bootstrap scores. The right-hand internal branch (connecting leaves "z" and "w" to the rest of the tree) has bootstrap support of 100% in all three trees. Therefore, it should be included in all consensus trees, or networks, regardless of the reconstruction method used. On the other hand, the left-hand internal branch connecting leaves "x" and "y" to the rest of the tree in T_1 and T_2 has different bootstrap scores in these trees (40 and 45% respectively). In tree T_3, the left-hand internal branch connects leaves "x" and "u" to the rest of the tree. Its bootstrap score, 90%, is higher than the sum of bootstrap scores of the corresponding branch in T_1 and T_2. When using the bootstrap-based majority rule defined above, we obtain a consensus tree (T_{bm} in Figure 1.1) that does not include the left-hand internal branch because neither the sum of scores of T_1 and T_2 nor the bootstrap score of T_3 divided by the number of trees is greater than 50%. The application of the bootstrap-based extended majority rule adds to the consensus tree (tree T_{bem} in Figure 1.1) the left-handed branch of tree T_3, since 90% / 3 = 30% > (40% + 45%) / 3 = 28.3%. Tree T_{bem} is also the bootstrap-based Nelson consensus tree of T_1, T_2 and T_3. Finally, the construction of the bootstrap-based consensus network (N_{123} in Figure 1.1) relies on the same principle as the bootstrap-based extended majority rule, except that it encompasses both left-hand internal branches (that from T_1 and T_2 and that from T_3) characterized by their bootstrap support. Network N_{123} is an implicit consensus network. In this article we will show how such an implicit network can be transformed into explicit one depending on the evolutionary mechanism being studied.

1.3.2 Method description: consensus tree

The method we present and apply here also takes into consideration the weights associated with the given phylogenetic trees in addition to bootstrap scores of the tree clusters (i.e., internal branches). Using one of the three equations presented in the

section "Inferring weights", the method defines a weight of each cluster based on the weights of the trees containing this cluster and on the cluster's bootstrap scores in these trees. Then, after ranking all the clusters based on their weights, it regroups the compatible clusters starting from the top of the list, until a fully resolved consensus tree is built. This method is called here *weight-based extended majority rule consensus tree inference*.

1.3.3 Method description: consensus network

Our *consensus network inference method* accepts two types of input: 1) a species phylogenetic tree and a set of gene phylogenetic trees defined on the same set of species, or 2) only a set of gene trees defined on the same set of species. In phylogenetic studies, gene trees are usually characterized by their weights that reflect the quality of the reconstruction process. Such weights could be an average of bootstrap scores of the tree's internal branches, a maximum parsimony or maximum likelihood score or a Bayesian posterior probability estimate. Thus, we assume that all the phylogenies have bootstrap scores or posterior probabilities (or any other measure of support) for their internal branches. Our algorithm first, breaks down all the gene phylogenies into their relevant clusters and calculates a weight for each cluster based on Equations 1.1, 1.2 or 1.3 presented in the following section. Next, the algorithm ranks all the clusters based on their weights. For this type of input, our algorithm uses the species tree as the backbone of the network and gradually adds to it the highly ranked clusters (i.e., represented by reticulation branches) of the gene phylogenies. For the first type of input, the species tree is accepted as the dominant evolutionary history and the clusters of the gene trees are used to infer the reticulate (alternative) evolutionary events. For the second type of input, our algorithm reconstructs a consensus phylogenetic tree using the weight-based extended majority rule consensus tree method described above and then adds to it the remaining highly ranked incompatible clusters which are

presented as reticulation branches. In the obtained consensus network, the weight-based consensus tree and the reticulation branches can be regarded as the main and alternative evolutionary scenarios, respectively.

Regardless of the input type, the resulting representation is a weighted consensus phylogenetic network with a backbone tree structure and reticulation branches being chosen based on their weights which reflect their contribution to the clustering process. These two algorithmic facets are schematically presented in Figure 1.2, in which the steps depicted by letter *a* correspond to processing the first type of input and those depicted by letter *b* are related to the second type of input. Steps 2 to 4 are common for both types of input.

We present here three network building algorithms (Algorithms I, II and III), each of them being optimized for detecting and representing a specific evolutionary phenomenon. The first algorithm (Algorithm I), which accepts the input of type 2 (a collection of gene trees inferred for various genes), is suitable for inferring either diploid or polyploidy hybridization events occurred among the observed species, or for finding recombination events occurred at the chromosome level. Algorithm I first proceeds by building the weight-based extended majority rule consensus tree followed by finding reticulation events and adding them to the consensus tree with proper direction in order to build the explicit weighted consensus network. The time complexity of Algorithm I is $O(n \times m^2 \times (n+r))$, where n is the number of gene trees in the considered gene tree collection τ, m is the number of leaves in each of these trees and r is the number of reticulation branches (i.e., reticulation events) added to the consensus tree. Note that the cluster inference procedure in Algorithm I (i.e., the first loop *for* in this algorithm) has the time complexity of $O(n \times m^2)$ as we use an optimal algorithm for the tree cluster inference, originally described by Makarenkov and Leclerc (2000), in which each tree cluster is presented as a binary bipartition vector.

The weight computation procedure for the clusters from the gene tree collection τ (i.e., the second loop *for* in Algorithm I) has the time complexity of $O(n^2 \times m^2)$. The time complexity of the second loop *while* in this algorithm, where the reticulation branches are added to the consensus tree, is $O(r \times n \times m^2)$. The function *find_direction* in the same algorithm has the time complexity of $O(n \times m^2)$. A group of clusters (i.e., bipartition vectors) is called *compatible* if altogether these clusters induce a unique phylogenetic tree. A cluster c has the *first degree of incompatibility* with a phylogenetic tree T if there exists an SPR (Subtree Prune and Regraft) move of the branches of T induced by the cluster c that transforms T into another phylogenetic tree. For instance in Figure 1, cluster (xy) has the first degree of incompatibility with tree T_3. In the same way, cluster (xyw) has the second degree of incompatibility with tree T_3, as it requires two SPR moves (i.e., two reticulation branches) to transform T_3 into a tree where cluster (xyw) is present. In the case of a directed phylogenetic network N_h inferred in Algorithm I, cluster c will have the first degree of incompatibility with N_h if it has the first degree of incompatibility with the tree T obtained from N_h after carrying out all SPR moves corresponding to the reticulation branches included in N_h. Mention that in all the three presented algorithms we only need to know whether a given cluster c has the *first degree of incompatibility* with a given network N_h or not.

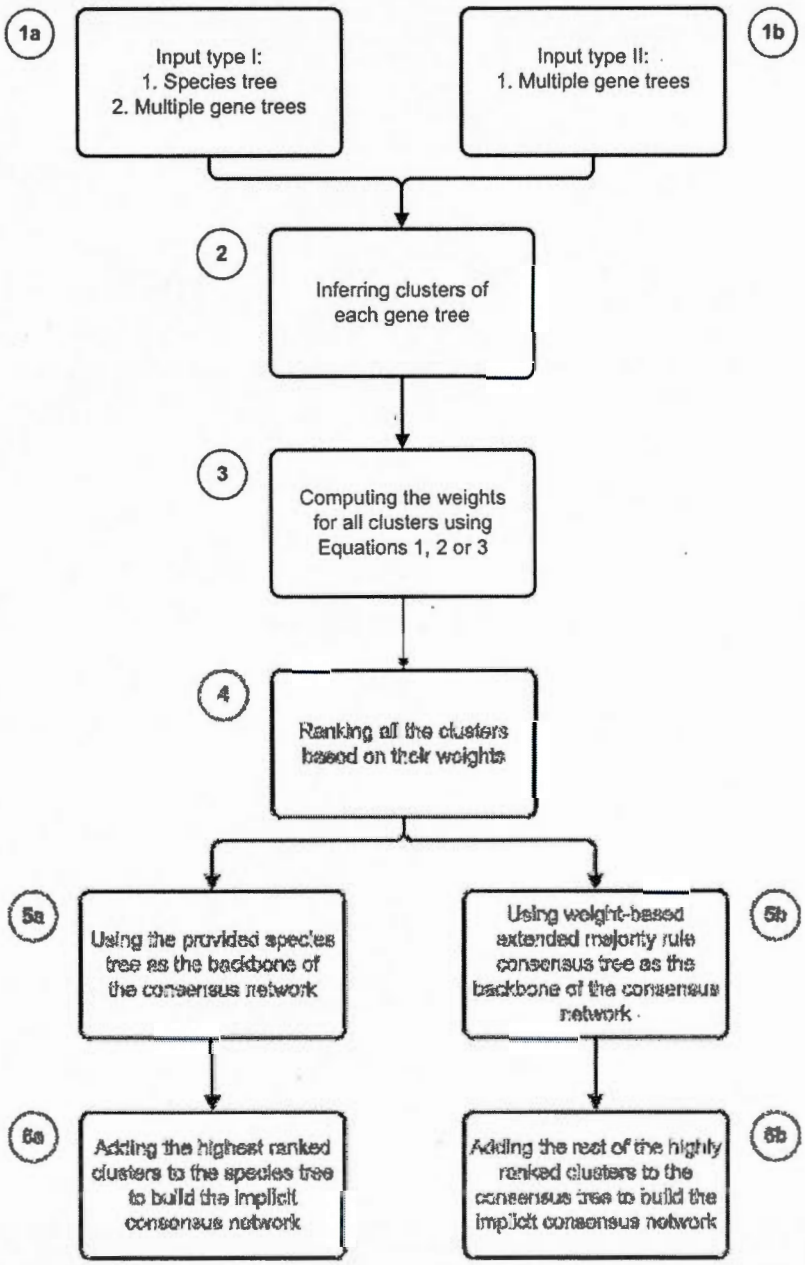


Figure 1.2 Flowchart of the new method for building weighted consensus networks. Facet *a* of the method (indicated by lowercase *a* next to step numbers) uses a species tree as well as a set of gene trees to infer the consensus network. Facet *b* of the method (indicated by lowercase *b* next to step numbers) uses only a set of gene trees to build the consensus network. Step numbers that do not contain any letter are common steps for the two facets.

Algorithm II, on the other hand, is designed to infer intragenic recombination events or partial horizontal gene transfers which lead to the creation of mosaic genes. This algorithm accepts two types of input (a species tree and a multiple sequence alignment, or only a multiple sequence alignment). In cases where a species tree is provided, Algorithm II uses it as a backbone of the network. A sliding window procedure is then carried out for finding the aforementioned reticulation events and adding them to the backbone in order to build an explicit weighted consensus network. Otherwise, if only a multiple sequence alignment is given, a weight-based extended majority rule consensus tree will be built from it and used as the backbone of the network. The time complexity of Algorithm II is $O(|SW| \times (O(PhylInfMeth) + n \times m^2 \times (n+r)))$, where $|SW|$ is the cardinality of the set of *MSA* (multiple sequence alignment) fragments examined by the sliding window procedure and $O(PhylInfMeth)$ is the running time of the phylogeny inference method used to infer the tree T from the *MSA* fragment MSA_f.

Our third algorithm (Algorithm III) is intended for finding complete horizontal gene transfer events. It accepts as input a species tree in addition to one or more gene trees (or multiple sequence alignments). Algorithm III uses the species tree as the backbone for the network and adds to it the most significant clusters (i.e., horizontal gene transfer events) obtained after computing the weights of the gene tree clusters in order to build the weighted consensus horizontal gene transfer network. The time complexity of Algorithm III is $O(PhylInfMeth) + O(n \times m^2 \times (n+r))$.

The resulting phylogenetic network, regardless of the algorithm used, will be an explicit (in the sense that it represents exactly the assumed evolutionary mechanism) weighted and directed consensus network as shown in detail in Figure 1.3. The weight estimates of the obtained backbone and reticulation branches provide statistical support of the inferred speciation and reticulation events.

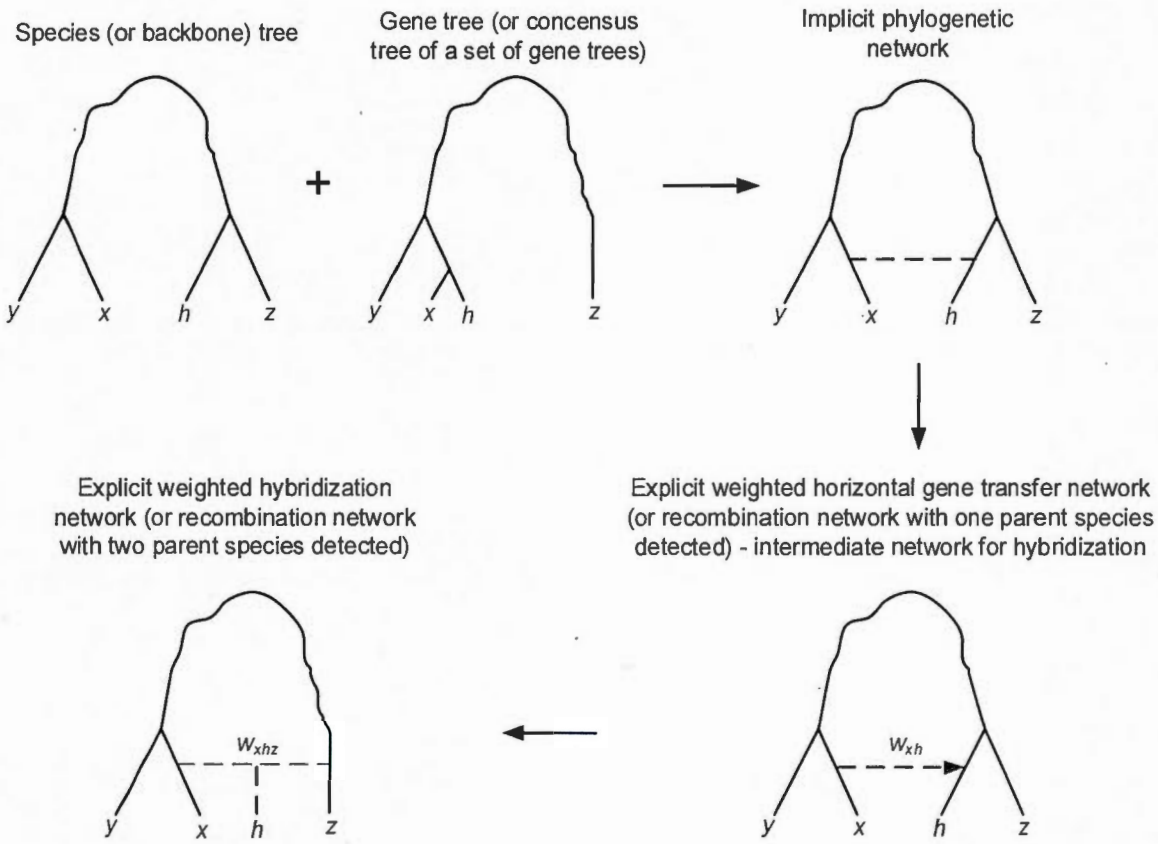


Figure 1.3 Building explicit weighted consensus phylogenetic networks. The explicit network is built from sets of clusters defined by a species (i.e. backbone) phylogenetic tree and a gene tree (or a set of gene trees): An implicit weighted phylogenetic network is first constructed; then, it is transformed into an explicit weighted horizontal gene transfer network, which can be transformed into an explicit hybridization network. Traditional (i.e. complete) horizontal gene transfer, partial horizontal gene transfer and recombination events for which the recombinant organism and only one of its parents can be identified give rise to a horizontal gene transfer network. Diploid and polyploid hybridization along with recombination events for which the recombinant organism and both of its parents can be identified give rise to a hybridization network. Straight lines indicate single tree or network branches, dashed lines - reticulation branches and wavy lines - paths including multiple branches.

1.3.4 Inferring cluster weights

For each cluster from the set of the given gene trees, we have first to compute its overall weight. Every tree cluster can be associated with two types of initial weights, one being its proper bootstrap score or posterior probability in its tree of origin and another characterizing its entire tree of origin. In the case when the input contains only the weights associated with internal tree branches and lacks any measure of support for entire trees, we use the following equation to calculate the overall cluster weights:

$$W_i(C) = (\sum_{j=1}^{n} \sigma_{ij} \times W(C_{ij})) / n, \qquad (1.1)$$

where $W_i(C)$ is the overall weight of cluster i, $W(C_{ij})$ is the weight of cluster i in tree j and n is the total number of trees. If cluster i is absent in tree j, then σ_i equals 0, otherwise it equals 1. Conversely, when the entire tree support is provided for each tree from the given set of trees but the input lacks individual supports for internal branches, we use the following equation to calculate the overall cluster weights:

$$W_i(T) = (\sum_{j=1}^{n} \sigma_{ij} \times W(T_j)) / n, \qquad (1.2)$$

where $W_i(T)$ is the overall weight of cluster i calculated from the tree supports only, $W(T_j)$ is the support of tree j and n is the total number of trees. Finally, when both cluster and tree initial supports are provided in the input, we use the following equation to infer the overall cluster weight, $W_i(C,T)$, for each cluster i:

$$W_i(C,T) = (\sum_{j=1}^{n} \sigma_{ij} \times W(C_{ij}) \times W(T_j)) / n, \qquad (1.3)$$

where $W(C_{ij})$ is the weight of cluster i in tree j, $W(T_j)$ is the support for tree j and n is the total number of trees. These overall cluster weights will be used to build the consensus tree or network as described above.

1.3.5 Assessing the efficiency of the new method

1.3.5.1 Real data

We examined three evolutionary datasets to test the efficiency of our weighted consensus network inference method. The first dataset consisted of 677 bp nucleotide sequences of mitochondrial *cytochrome c oxidase subunit II* of six species of honeybees (subfamily Apinae). The second one comprised eight chloroplast 16S rRNAs (920 nucleotides) from plants, algae and cyanobacterium. These two datasets are well-known and distributed with the SplitsTree program (Huson, 1998) among the data encompassing the events of reticulate evolution. The third considered dataset consisted of amino acid sequences of ribosomal protein *rpL12e* for 14 Archaeal species (Matte-Tailliez, 2002).

We applied four different tree inference methods on both real and simulated (described in the next section) data to produce collections of gene trees. One representative from each of the four main tree reconstruction approaches (i.e., distance-based (Saitou and Nei, 1987), maximum parsimony (Fitch, 1971), maximum likelihood (Felsenstein, 1981) and Bayesian (Rannala and Yang, 1996) approaches) was considered. The exact methods we used were the following: BIONJ (Gascuel, 1997), DNAPARS from the PHYLIP package (Felsenstein, 2005), PhyML (Guindon *et al.*, 2010) and MrBayes (Ronquist *et al.*, 2012).

We applied these tree inference methods on both whole sequences and fragments of sequences (using a sliding window procedure) in order to search for alternative evolutionary events which might have affected either entire gene sequences (e.g., hybridization events) or only small sequence fragments (e.g., partial horizontal gene transfer events). The latter events are usually ignored when analyzing entire genetic sequences during tree or network reconstruction. In the case of horizontal gene transfer events in Archaebacteria, we also computed the directions of complete and partial horizontal gene transfers using a dedicated function based on the Robinson and Foulds topological distance (Robinson and Foulds, 1981); see the function *find_direction* in the end of Algorithm I. Assume that T is the backbone phylogenetic tree and r is the newly found horizontal gene transfer event between clusters C_1 and C_2 (i.e., groups of species related by r). Let T_1 be the tree obtained by an SPR (Subtree Prune and Regraft) move induced by reticulation branch r with direction d_1 (corresponding to the horizontal gene transfer from cluster C_1 to cluster C_2) and T_2 be the tree with r added to represent the gene transfer in the opposite direction (i.e., from C_2 to C_1). Then, the cumulative Robinson and Foulds distance is calculated between T_1 and all the original gene trees containing cluster $C = C_1 \cup C_2$, on one hand, and T_2 and all the original gene trees containing C, on the other hand. Finally, the obtained cumulative Robinson and Foulds distances are weighted by the support of the original gene trees containing C as it is shown in Algorithm I (see the exact formula is in the function *find_direction*) and the resulting inequality indicates the direction of the horizontal gene transfer r.

1.3.5.2 Simulated data

We generated sets of trees encompassing multiple reticulation features to test the efficiency of the proposed consensus network inference method in the context of recombination. First, random binary phylogenetic trees were generated using the procedure originally described by Kuhner and Felsenstein (1994). The branch lengths

of these phylogenies were computed using an exponential distribution. Following the approach of Guindon and Gascuel (2002), we added some noise to the tree branches to create a deviation from the molecular clock hypothesis. All branch lengths were multiplied by $1+ax$, where the variable x was obtained from an exponential distribution ($P(x>k) = \exp(-k)$), and the constant a was a tuning factor accounting for the deviation intensity. The value of a was fixed to 0.8. The random trees generated by this procedure had depth of $O(\log(n))$, where n was the number of species (i.e., number of leaves in a binary phylogenetic tree).

Second, we ran the SeqGen program (Rambaut and Grass, 1997) to generate DNA sequences along the branches of the phylogenies constructed at the first step. SeqGen was used with the HKY model of nucleotide substitution, model of rate heterogeneity assigning different rates to different sites according to a gamma distribution (with the shape parameter equal to 1.0) and (TS/TV) ratio equal to 2.0. These settings were selected in order to render the simulation parameters similar to those used when processing the real datasets. The DNA sequences with 400 nucleotides were generated. Third, using the reticulation events generation procedure described in (Boc and Makarenkov, 2011), we incorporated the blocks of fragments induced by recombination into the generated multiple sequence alignments (MSAs). The sliding window procedure was then employed to recover these recombined blocks of sequences. Forth, for each generated MSA, the BIONJ, DNAPARS, PhyML and MrBayes methods were carried separately to infer phylogenetic trees for the whole MSA and for each MSA fragment corresponding to the fixed position of the sliding window. Finally, we carried the proposed weighted consensus network building method to infer the consensus tree topology (i.e., backbone evolutionary structure representing the most significant speciation events) as well as to recover the most significant (those with the highest weights) recombination events. We repeated this procedure 100 times for each original tree, i.e., 100 different MSAs were generated for the same original tree. The sliding window sizes considered in our simulations were

10, 20, 30, 40 and 50% of the total length of the generated MSAs. The sliding window progress step of 5 nucleotides was adopted. Simulations were carried out with the phylogenies having 16, 24, 32, 48 and 64 leaves and encompassing 1 to 8 recombination events.

1.4 Results

1.4.1 First example: Honeybee data

We applied the BIONJ, DNAPARS, PhyML and MrBayes methods to infer the evolutionary history of the six honeybee species. The inferred trees are shown in Figure 1.4. The BIONJ and PhyML methods provided a single phylogeny (Figures 1.4A and 1.4B, respectively). In contrast, two optimal phylogenies were obtained by each of the DNAPARS and MrBayes methods (Figures 1.4C and 1.4D represent maximum parsimony trees and Figures 1.4E and 1.4F represent Bayesian trees). For the sake of simplicity, we assigned a total weight of 1 to each of the considered methods. Therefore, the BIONJ and PhyML phylogenies received a weight of 1, whereas each of the DNAPARS phylogenies received a weight of 0.5. For the case of Bayesian phylogenies, we also used their specific posterior probabilities whose sum was scaled to 1.

After breaking down the phylogenies into their clusters and calculating the cluster weights using Equation 1.3, we ranked all the clusters according to their weights and put together the compatible clusters to build the backbone of the consensus network based on the clusters ranks. Finally, we added the rest of the highly ranked clusters to the backbone tree to construct a weighted consensus network of the six honeybee species. In this analysis, we found one reticulation branch (alternative event) in addition to the backbone (consensus tree). The explicit weighted consensus network built using

Algorithm I is shown in Figure 1.5A. It depicts one recombination event which might have influenced the evolution of the considered honeybee species.

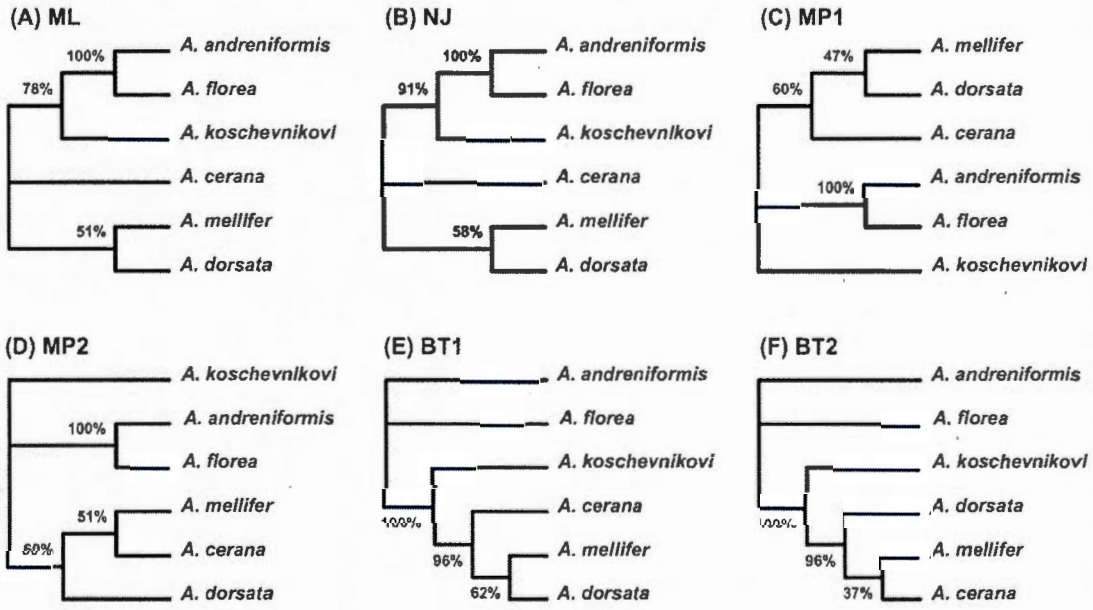


Figure 1.4 The set of six gene trees (A-F) obtained using different tree reconstruction methods for honeybee dataset. ML, NJ, MP and BT abbreviations stand for trees obtained by maximum likelihood, neighbour-joining (here a distance-based approach implemented in BIOINJ), maximum parsimony and Bayesian approaches, respectively.

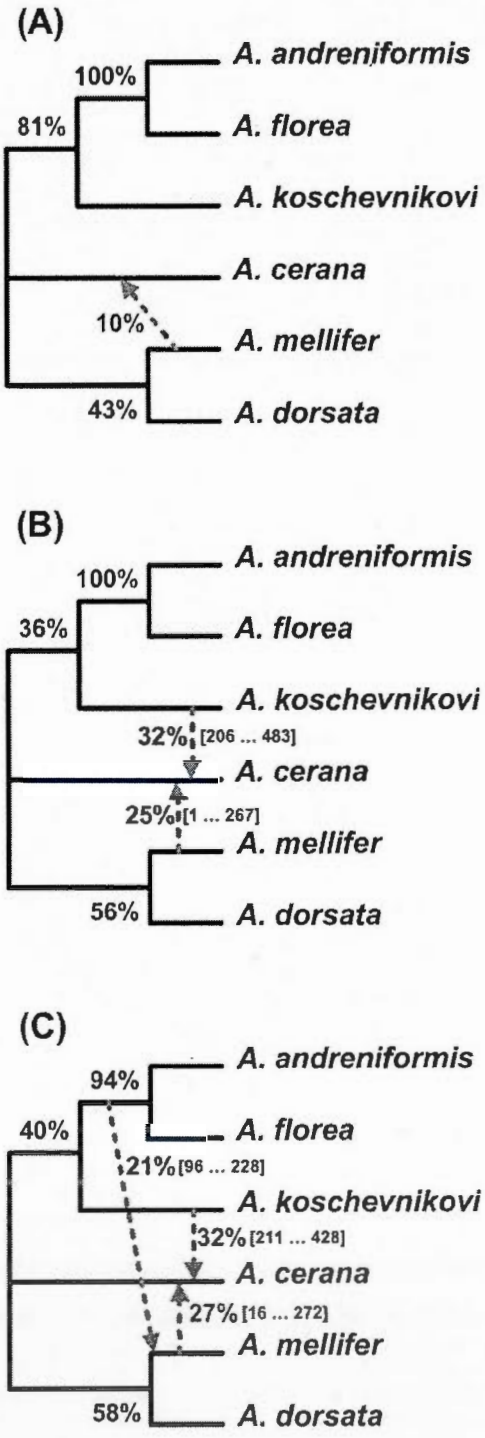


Figure 1.5 Explicit weighted consensus networks inferred for the honeybee dataset. A) network obtained from full-length sequences using all the six trees from Figure 1.4 (which were inferred using the ML, NJ, MP and BT approaches); B) network obtained by the sliding window procedure with a ML method used for tree inference; C) network obtained by the sliding window procedure with a Bayesian method used for tree inference. The bootstrap scores of internal branches of the backbone tree and the weights of reticulation branches are indicated. The sliding window procedure was used to detect smaller-scale reticulation events which are represented by dashed lines in parts B and C of the figure. For each small-scale event, the sequence interval corresponding to this event is given between brackets.

1.4.2 Second example: Chloroplast data

In this example, we used the same four tree inference methods as in the previous section to model evolutionary relationships among the eight plants from the chloroplast dataset. The application of these methods resulted in one maximum likelihood (Figure 1.6A), one distance-based (Figure 1.6B), three maximum parsimony (Figures 1.6C to 1.6E) and two Bayesian phylogenies (Figures 1.6F and 1.6G). Similar to the previous example, we assigned a total weight of 1 to each method. Therefore, the BIONJ and PhyML phylogenies received the weight of 1 while each of the DNAPARS trees received the weight of 0.33. In the case of the MrBayes phylogenies, we also used their corresponding posterior probabilities scaling their sum to 1. We, then, computed the weights of all the clusters presented in at least one of the seven phylogenetic trees using Equation 3. Finally, we built the backbone of the consensus network and added to it the reticulation branches after ranking the clusters as described in Algorithm I.

In this analysis, we found three reticulation branches which represent possible recombination events. The reconstructed weighted consensus network of the plastid 16s rRNAs is shown in Figure 1.7A. Using the cut-off level of 10% and eliminating the two poorly supported reticulation branches (those with the weights of 2% and 3%) would provide us with the weighted consensus network encompassing one probable reticulation event only (that with the weight of 23%).

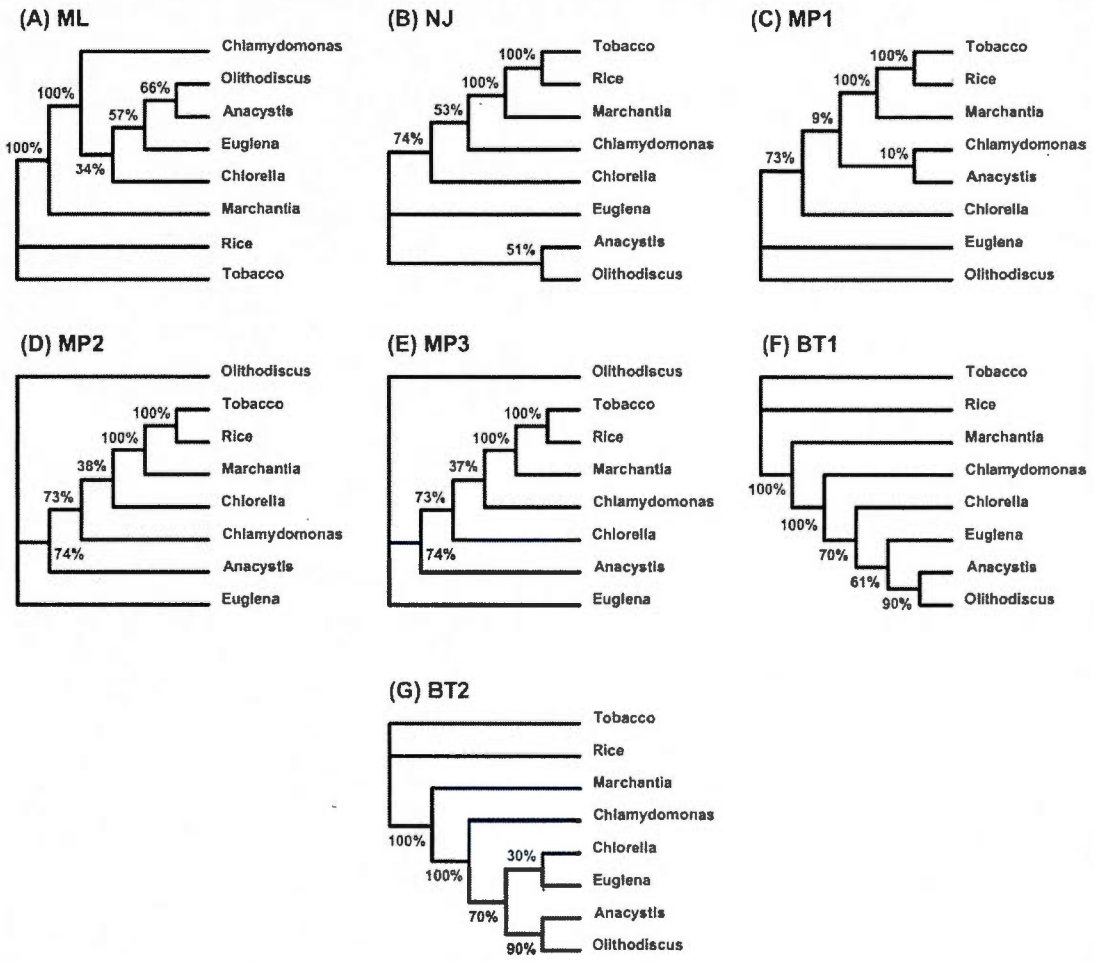


Figure 1.6 The set of seven gene trees (A-G) inferred for the chloroplast dataset. The abbreviations used in Figure 1.4 also apply here.

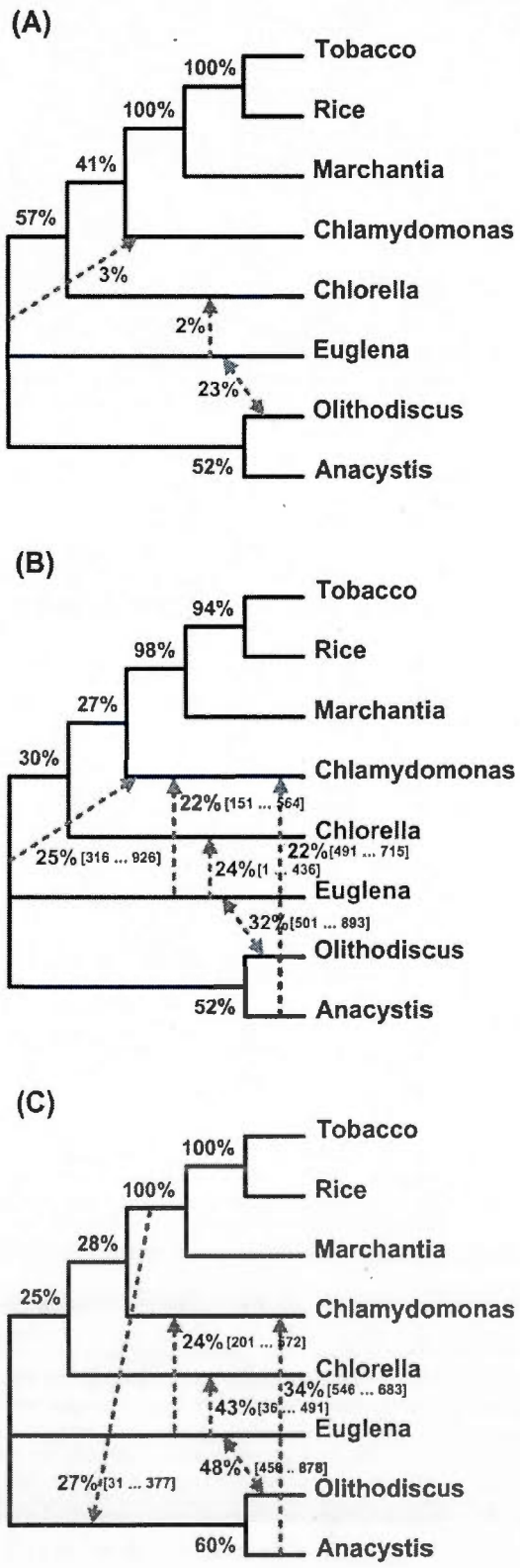


Figure 1.7 Explicit weighted consensus networks obtained for the chloroplast dataset. A) network obtained from full-length sequences using all the seven trees from Figure 1.6 (which were inferred using the ML, NJ, MP and BT approaches); B) network obtained by the sliding window procedure with a ML method used for tree inference; C) network obtained by the sliding window procedure with a Bayesian method used for tree inference. The notations of Figure 1.5 also apply here.

1.4.3 Third example: Archaebacteria data

Similar to the two previous examples we used the four above-mentioned tree inference methods to build multiple phylogenies of the gene *rpl12e* for 14 Archaebacteria species originally analyzed by Matte-Tailliez *et al.* (2002). Thus, one maximum likelihood (Figure 1.8A), one distance-based (Figure 1.8B), five maximum parsimony (Figures 1.8C to 1.8G) and two Bayesian phylogenies (Figures 1.8H and 1.8I) were obtained. Considering the species tree (Figure 1.9A), which was reconstructed using the concatenation approach (Matte-Tailliez *et al.*, 2002), we applied Algorithm III to the obtained phylogenies to infer a horizontal gene transfer network of the gene *rpl12e*. The species tree was used as the backbone topology to which we added the highly ranked incompatible clusters to build the weighted consensus evolutionary network encompassing a scenario of horizontal transfers of *rpl12e*. Using the cut-off level of 30%, we obtained five reticulation branches depicting alternative evolutionary histories (Figure 1.9B). Then, applying the above-discussed strategy for determining horizontal gene transfer direction (see function *find_direction*), we assigned directions to all obtained gene transfer branches. In the case of Transfers 1 and 2 (Figure 1.9B), the transfer direction cannot be retraced without discrepancy because both concurrent transfers are symmetric and lead to the same tree topology.

Note that in Figures 1.5A and 1.7A the supporting weights calculated by our method for the backbone and reticulation branches are given in percentages. For the network presented in Figure 1.9B our method was carried out to calculate the supporting weights of the reticulation branches only, whereas the weights of the internal branches of the backbone (species) phylogeny are the bootstrap scores provided by Matte-Tailliez and colleagues (2002).

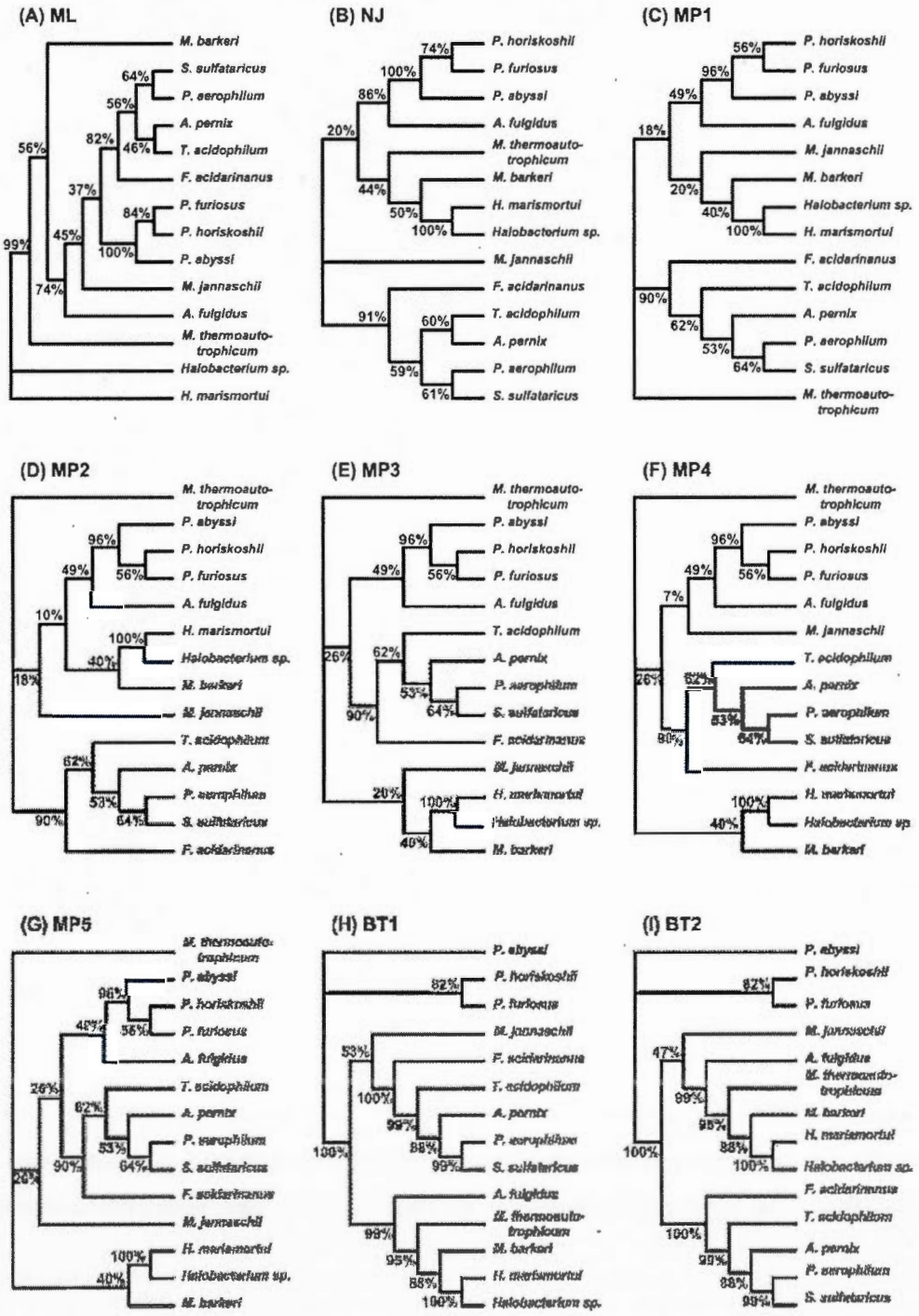


Figure 1.8 The set of nine gene trees (A-I) inferred for the Archaebacteria dataset. The abbreviations used in Figure 1.4 also apply here.

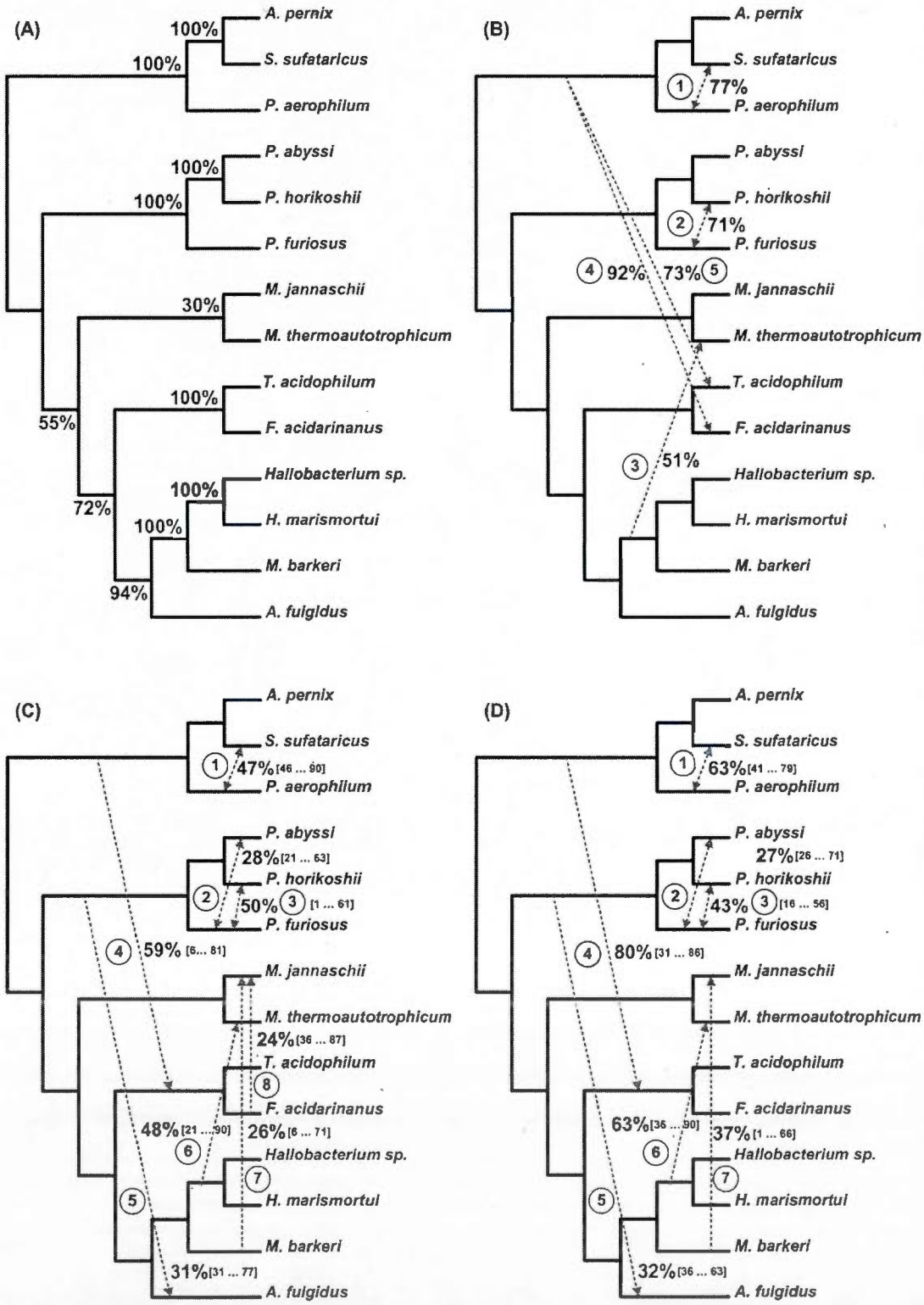

(A)
100%
100%
A. pernix
S. sufataricus
P. aerophilum
100%
100%
P. abyssi
P. horikoshii
P. furiosus
30%
M. jannaschii
M. thermoautotrophicum
100%
T. acidophilum
55%
F. acidarinanus
100%
Hallobacterium sp.
72%
100%
H. marismortui
M. barkeri
94%
A. fulgidus
(B)
A. pernix
S. sufataricus
1
77%
P. aerophilum
P. abyssi
P. horikoshii
2
71%
P. furiosus
4
92%
73%
5
M. jannaschii
M. thermoautotrophicum
T. acidophilum
F. acidarinanus
3
51%
Hallobacterium sp.
H. marismortui
M. barkeri
A. fulgidus
(C)
A. pernix
S. sufataricus
1
47% [46 ... 90]
P. aerophilum
P. abyssi
28% [21 ... 63]
P. horikoshii
2
50%
3
[1 ... 61]
P. furiosus
4
59% [6... 81]
M. jannaschii
M. thermoautotrophicum
24% [36 ... 87]
T. acidophilum
8
F. acidarinanus
48% [21 ... 90]
26% [6 ... 71]
6
Hallobacterium sp.
7
5
H. marismortui
M. barkeri
31% [31 ... 77]
A. fulgidus
(D)
A. pernix
S. sufataricus
1
63% [41 ... 79]
P. aerophilum
P. abyssi
27% [26 ... 71]
P. horikoshii
2
43%
3
[16 ... 56]
P. furiosus
4
80% [31 ... 86]
M. jannaschii
M. thermoautotrophicum
T. acidophilum
F. acidarinanus
63% [36 ... 90]
37% [1 ... 66]
6
Hallobacterium sp.
7
5
H. marismortui
M. barkeri
32% [36 ... 63]
A. fulgidus

Figure 1.9 Explicit weighted consensus horizontal gene transfer networks inferred for the Archaebacteria dataset. A) species tree obtained by Matte-Tailliez *et al.* (2002); B) network obtained from full-length sequences using all the nine gene trees from Figure 1.8 (which were inferred using the ML, NJ, MP and BT approaches) and depicting complete horizontal gene transfer events; C) network obtained by the sliding window procedure with a ML method used for tree inference and depicting complete and partial horizontal gene transfers; D) network obtained by the sliding window procedure with a Bayesian method used for tree inference and depicting complete and partial horizontal gene transfers. The sequence interval corresponding to each partial horizontal gene transfer (see parts C and D of the figure) is given between brackets. The transfer number corresponds to its order of appearance in the gene transfer scenario found by our method. The bootstrap scores of internal branches of the species (backbone) tree and the weights of horizontal gene transfers are also indicated.

1.4.4 Simulation results

The results provided by Algorithm II (inference of recombination events using a sliding window approach) on simulated data are shown in Figures 1.10 and 1.11. For each parameter combination, including the number of taxa, number of reticulation events and sliding window size, 100 datasets were generated and analyzed. The average rates of true and false positives characterizing our weighted consensus network building method are illustrated. Since in our simulations we knew the exact source and target of each reticulation event, we were able to estimate the success and failure rates of the consensus network method in terms of true positives and false positives by measuring the proportion of times when our method was able to identify both the exact source branch and destination branch of the event (i.e., true positive reticulation) and when either the source or destination branch of the detected event, or both of them, were different from the simulated ones (i.e., false positive reticulation). The x-axis depicts either the number of recombination events introduced in the data (Figure 1.10) or the number of taxa (i.e., number of species or tree leaves - Figure 1.11). The results obtained for the sliding windows whose width was equal to 10, 20, 30, 40 and 50% of

the total length of the multiple sequence alignment are illustrated in different panels. The *y*-axis represents the average number of times when our weighted consensus network reconstruction method correctly (true positives - left-hand panels) or incorrectly (false positives - right-hand panels) identified intragenic recombination events.

The obtained results suggest that when the number of recombination events is small, they are more likely to be detected correctly. The best results in terms of both true and false positives were found for longer recombination fragments, i.e., 40 and 50% of the total length of the multiple sequence alignment. Another general trend is that the PhyML and MrBayes methods were much more effective in inferring the correct supporting tree and reticulation events than their BIONJ and DNAPARS counterparts. These results also suggest that it is much easier to detect recombination events in larger (i.e., 32 and 64-species) phylogenies. Furthermore, the probability of finding the correct reticulation events increases as the width of the sliding window becomes closer to the real length of the simulated recombination fragment.

1.4.5 Searching for intragenic recombination and partial horizontal gene transfer events in real data

Considering the results obtained for simulated data, we applied Algorithm II based on the sliding window approach and the two best tree inference methods (PhyML and MrBayes) to reanalyze the honeybee, chloroplast and Archaebacteria data described above. The purpose of this new analysis was to discover alternative evolutionary events of smaller lengths (i.e., intragenic recombination and partial horizontal gene transfer events which trigger the formation of mosaic genes; Boc and Makarenkov, 2011). Those partial evolutionary events, in the sense that they concern only a part of

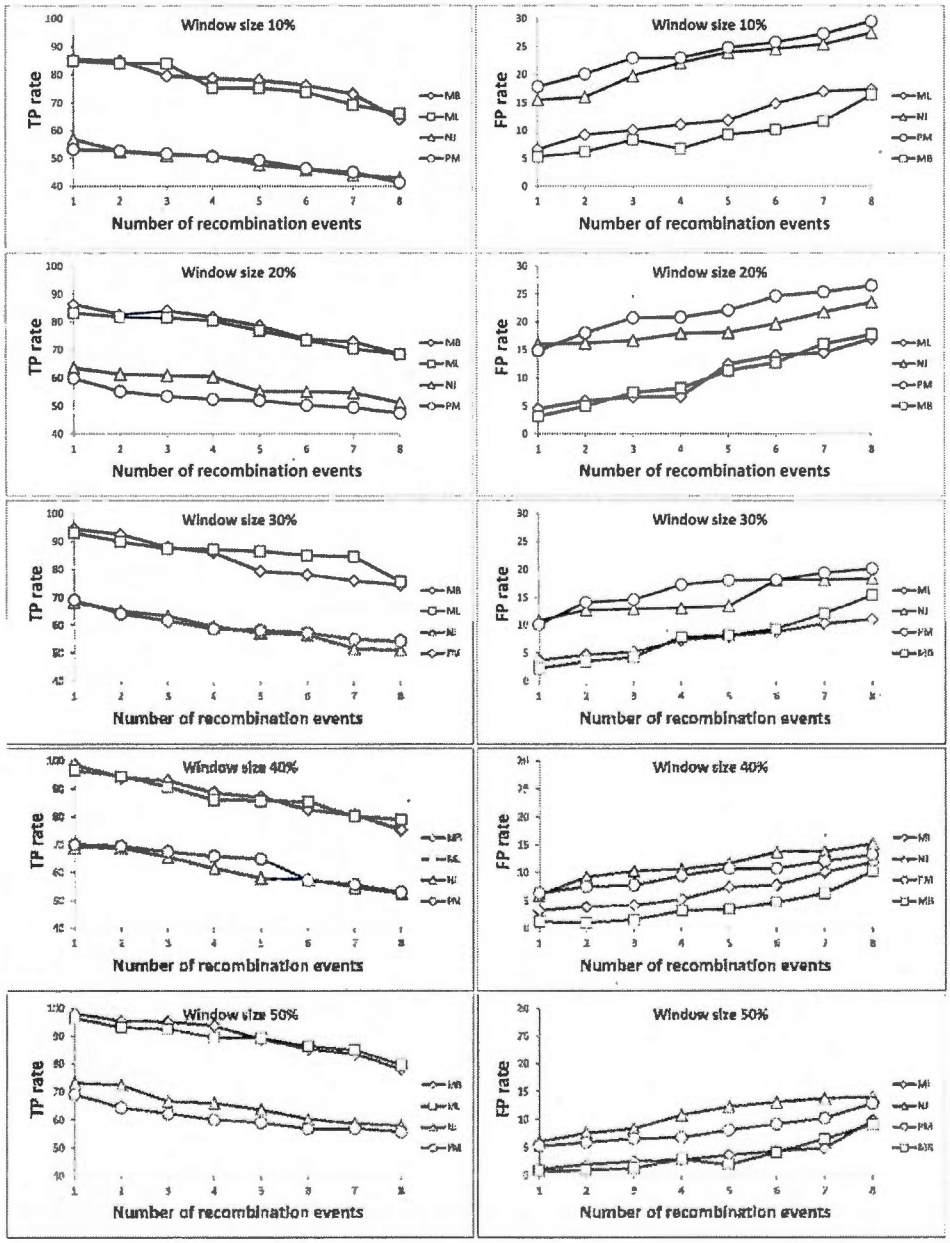


Figure 1.10 Average true-positive (left-hand panel) and false-positive (right-hand panel) rates provided by the weighted consensus network reconstruction method depending on the number of recombination events in the simulated data and the tree inference method used. The presented rates are the averages computed for different sliding window sizes (varying from 10 to 50% of the total MSA length) and different numbers of taxa (ranging from 16 to 64 with the step of 8); 100 datasets were tested for each parameter combination; ML, NJ, MP and BT abbreviations stand for the PhyML, BIOINJ, DNAPARS and MrBayes methods, respectively.

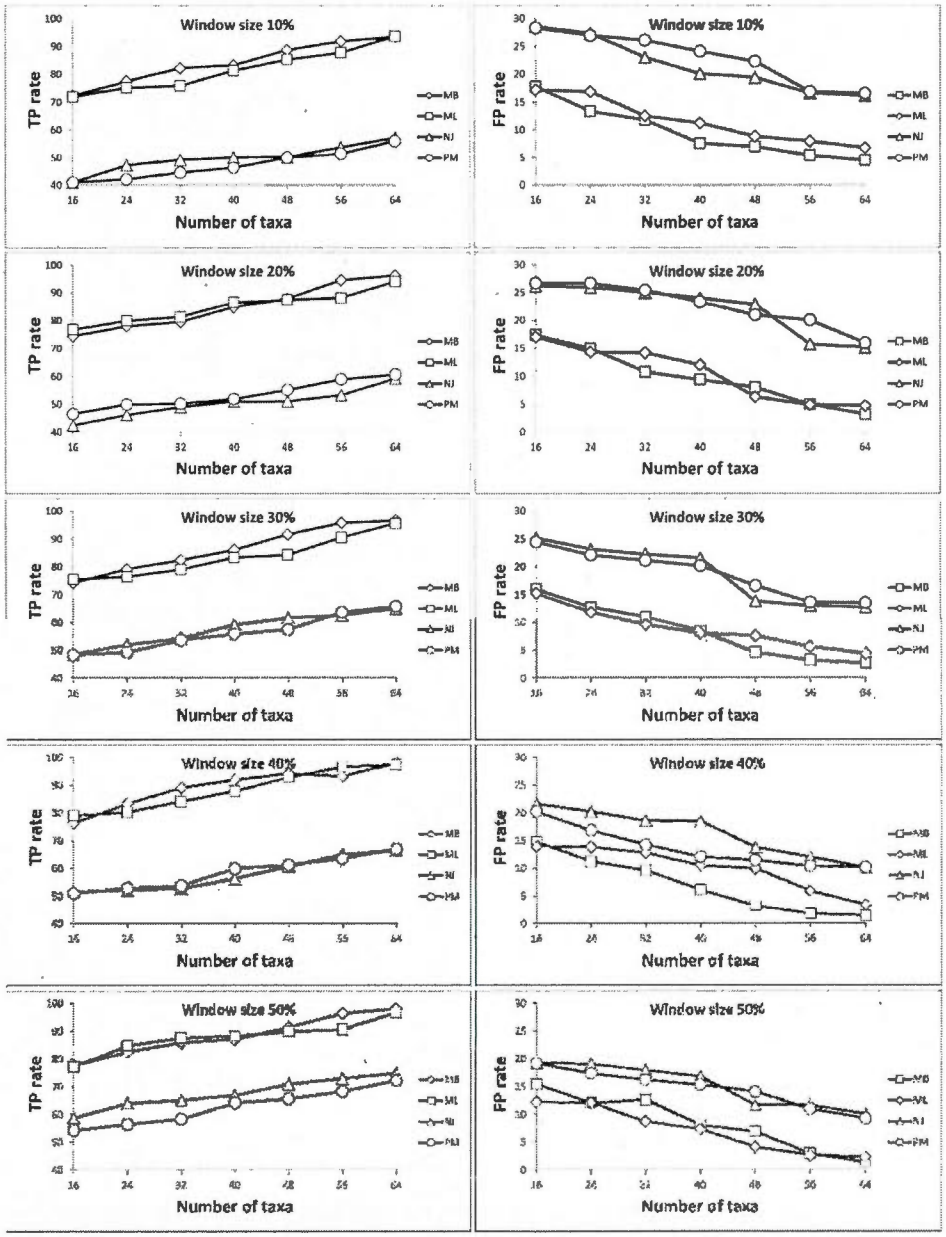


Figure 1.11 Average true-positive (left-hand panel) and false-positive (right-hand panel) rates provided by the weighted consensus network reconstruction method depending on the number of taxa in the simulated data and the tree inference method used. The presented rates are the averages computed for different sliding window sizes (varying from 10 to 50% of the total MSA length) and different numbers of recombination events (ranging from 1 to 8); 100 datasets were tested for each parameter combination; ML, NJ, MP and BT abbreviations stand for the PhyML, BIOINJ, DNAPARS and MrBayes methods, respectively.

the given gene, might have gone unnoticed when analyzing the full-length gene sequences.

For the honeybee example, the PhyML and MrBayes methods allowed us to infer one and two possible recombination events (Figures 1.5B and 1.5C), respectively, in addition to a possible recombination event found in the analysis based on the full-length sequences (i.e., linking the species *A. mellifer* and *A. serana* in Figure 51.A). For the chloroplast data, two additional reticulation events were detected using PhyML (Figure 1.7B) compared to the full-length sequence analysis (Figure 1.7A). Using MrBayes, we inferred four additional recombination events (Figure 1.7C) compared to the full-length sequence analysis three of which were concordant with the results obtained using PhyML.

For the smaller-scale recombination events found using Algorithm II for the honeybee and chloroplast data, the intervals where they were detected are indicated between brackets in addition to their supporting weights (see Figures 1.5B, 1.5C, 1.7B and 1.7C). For the full-sequence analysis events found using Algorithm I (see Figures 1.5A and 1.7A), no intervals are given because the latter events apply to entire genes.

Finally, in the case of the Archaebacteria data, the PhyML and MrBayes methods allowed us to detect eight and seven partial horizontal gene transfers, respectively (Figures 1.9C and 1.9D). Three of the detected partial gene transfers (Transfers 1, 3 and 6 in Figures 1.9C and 1.9D), which were found by both methods, were also reported by Boc *et al.* (2010) (a study dedicated to the detection of complete horizontal gene transfers) and Boc *et al.* (2013) (a study dedicated to the detection of partial horizontal gene transfers). Two other partial gene transfers (Transfers 5 and 8 in Figure 1.9C) detected using PhyML (one of them was also detected using MrBayes; Transfer 5 in Figures 1.9C and 1.9D) were reported only in (Boc *et al.*, 2013), while another gene transfer (Transfer 4 in Figures 1.9C and 1.9D) detected using both PhyML and

MrBayes was a combination of two separate complete gene transfer events (Transfers 3 and 4 in Figure 1.9B) originally detected by Boc *et al.* (2010). Our method also identified two additional partial horizontal gene transfers (Transfers 2 and 7 in Figures 1.9C and 1.9D) that were not indicated in Boc *et al.* (2013).

For comparison purposes, we also inferred splits graphs and cluster networks for the three above-mentioned real datasets using the SplitsTree (Huson, 1998; Huson and Bryant, 2006) and Dendroscope (Huson and Scornavacca, 2012) programs, respectively. Moreover, reticulograms were inferred for the honeybee and chloroplast datasets and a horizontal gene transfer network was constructed for the Archaebacteria dataset, both using the T-Rex web server (Boc *et al.*, 2012). The NeighborNet algorithm (Bryant and Moulton, 2004) from the SplitsTree 4 software was used with the ordinary least-square optimization and convex hull algorithm options. The Dendroscope program (Huson and Scornavacca, 2012) was carried out with the default parameters and the percent threshold equal to 20 to build cluster networks. The reticulogram inference algorithm was carried out using the weighted least-square method MW with global optimization (Makarenkov and Leclerc, 1999) to infer the support tree and the stopping criterion Q_1 (Legendre and Makarenkov, 2002). The HGT-Detection algorithm was performed with the HGT bootstrap option and the species and gene tree roots selected as described in (Boc *et al.*, 2010).

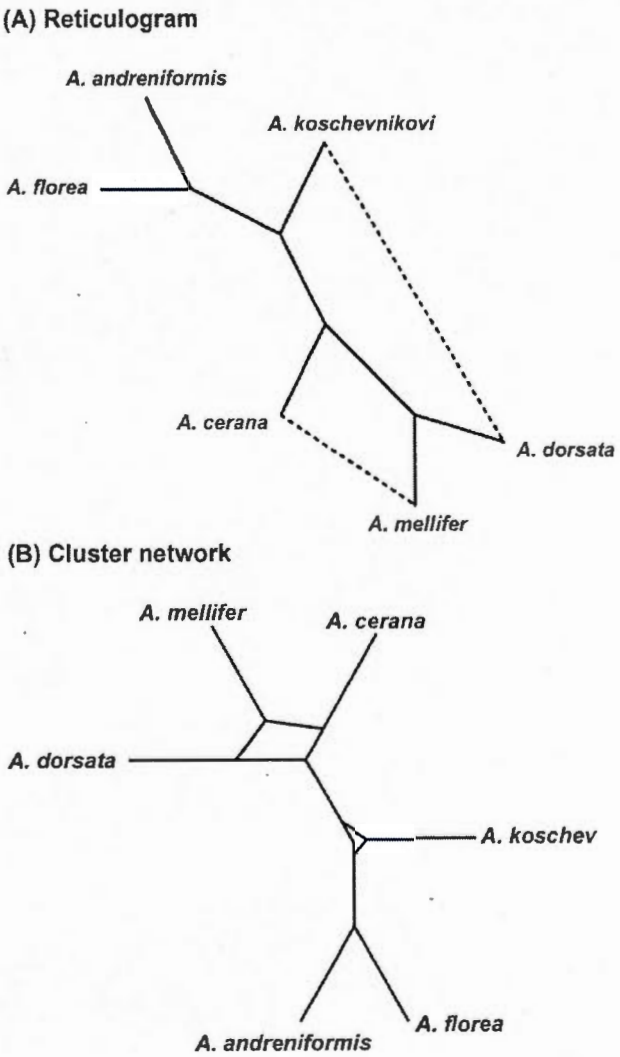
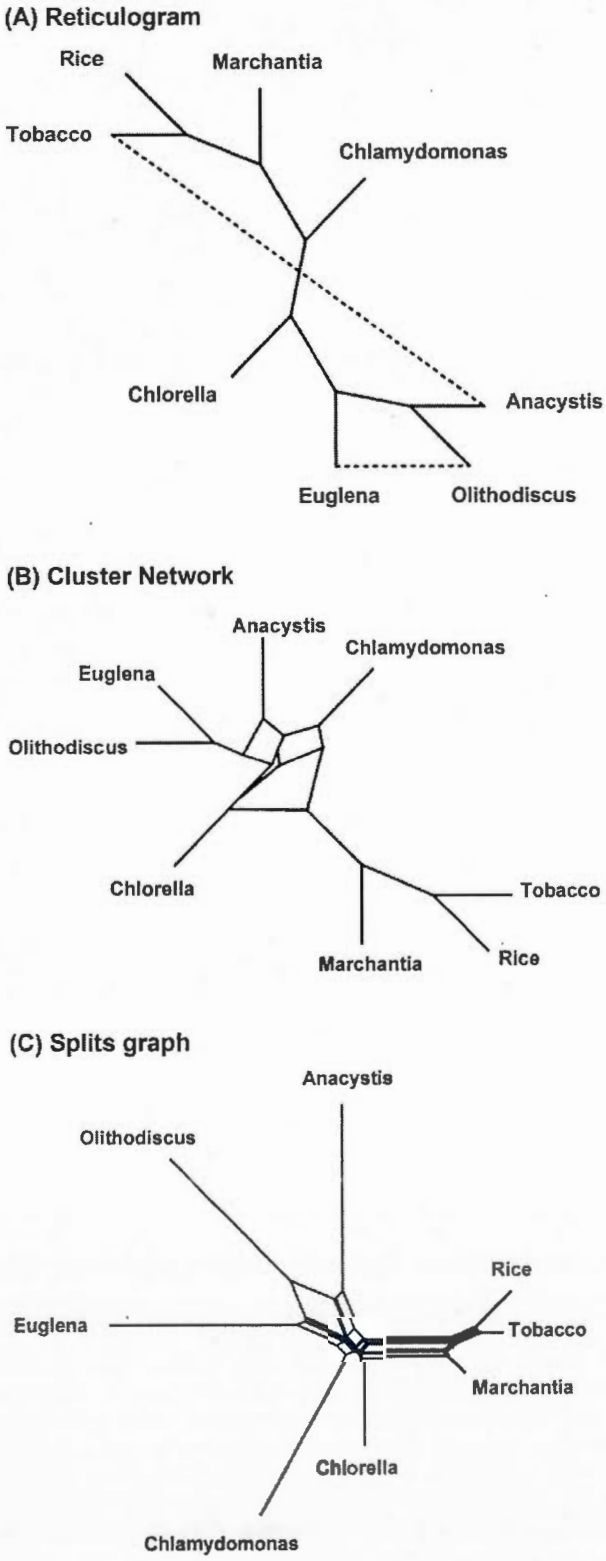


Figure1.12 Alternative network representations of the honeybee dataset. They include: A) reticulogram obtained using the Reticulogram building algorithm from the T-REX web server; B) cluster network obtained by the Cluster network algorithm from the Dendroscope program; C) splits graph obtained by the NeighborNet algorithm from SplitsTree 4.

Figure 1.13 Alternative network representations of the chloroplast dataset. They include: A) reticulogram obtained by the Reticulogram building algorithm from the T-REX web server; B) cluster network obtained by the Cluster network algorithm from the Dendroscope program; C) splits graph obtained by the NeighborNet algorithm from SplitsTree 4.



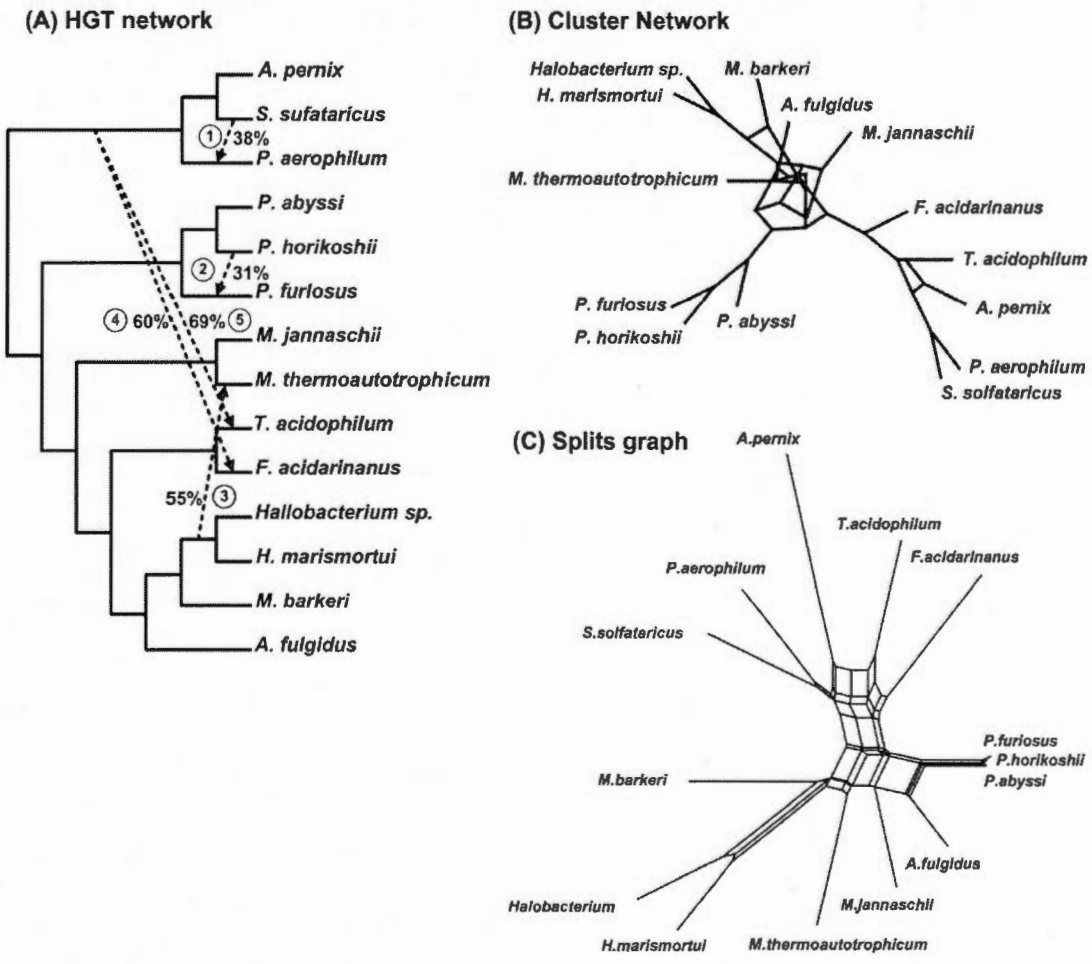


Figure 1.14 Alternative network representations of the Archaebacteria dataset. They include: A) horizontal gene transfer network obtained by the HGT-Detector algorithm from the T-REX web server; B) cluster network obtained by the Cluster network algorithm from the Dendroscope program; C) splits graph obtained by the NeighborNet algorithm from SplitsTree 4.

The obtained network representations are shown in Figures 1.12, 1.13 and 1.14 for the honeybee, chloroplast and Archaebacteria examples, respectively. In Figures 1.12A and 1.13A, one of the reticulation branches (represented by dashed lines) found by reticulogram was also identified by our weighted consensus network building method (i.e., the reticulation branches between (1) *A. mellifer* and *A. cerana* in Figure 1.12A

and between (2) Euglena and Olithodiscus in Figure 1.13A). The similarities between horizontal gene transfer network found by us and by HGT-Detection (Boc *et al.*, 2010; Figure 1.14A) will be discussed in detail in the next section.

1.5 Discussion

Dealing with multiple incompatible phylogenies inferred either through the use of different reconstruction methods or by including multiple genes in the analysis has been always a major issue in phylogenetics. The degree of uncertainty increases in line with the number of various phylogenies inferred for the same set of species (Bryant, 2003). The concatenation approach, which has been widely used as a solution to the single-gene phylogenies discordance problem, has been proven to lead to biased and misleading phylogenies in many practical situations (Hwang *et al.*, 2001; Mossel and Vigoda, 2005; Naylor and Brown, 1998). For instance, Kubatko and Degnan (2007) showed that when the internal branches of a species phylogeny are short (due to adaptive radiation, increased number of taxa from the same group or recent divergences), the concatenation approach usually reduces the accuracy of standard phylogenetic methods. The latter authors also suggested that bootstrap scores obtained from concatenated datasets tend to show moderate to strong support for incorrect trees (Kubatko and Degnan, 2007). In general, the main drawback of the concatenation approach lies in its flawed assumption that all the genes (and in a similar way, the whole genomes) have been subject to the same evolutionary processes at the same evolutionary rate, and consequently, no heterogeneity exists among the genes. Given the broad occurrence of heterogeneity among genes and the high number of phylogenetic mechanisms influencing their evolution, one can argue that in a considerable number of cases the concatenation approach will fail to infer a reliable congruent phylogenetic tree or network. Since incongruence increases with the number of genes included in the analysis, proposing as a final cohesive solution a single

phylogeny reconstructed using either the concatenation or the consensus tree approaches is only an indication of ignoring phylogenetic conflicts, and consequently, ignoring many widespread evolutionary processes such as horizontal gene transfer, recombination, hybridization and deep coalescence, which play major roles in the evolution of many species.

When the heterogeneity among genes is due to reticulate evolution, phylogenetic networks should be used in place of traditional or consensus phylogenetic trees (Huson and Bryant, 2006; Huson *et al.*, 2010; Legendre and Makarenkov, 2002). Phylogenetic networks are generalizations of phylogenetic trees intended to represent both speciation and reticulate evolutionary events characterizing the given group of genes and species (explicit networks) or to display conflicting evolutionary signals present in the data (implicit networks).

To address the gene trees discordancy issue, we described here a new weighted consensus network reconstruction method which is able to infer and validate statistically the dominant evolutionary history of species (consensus tree) as well as the alternative evolutionary scenarios (consensus reticulation events).

Two practical situations are possible: we are either in possession of a reliable species phylogeny or not. In the case when we have a reliable species tree (e.g., when tree topology is confirmed via the Tree of Life project), we can directly define it as the network support structure. Otherwise, averaging the tree clusters present in the given gene trees and using the consensus tree approach as the starting point for building the consensus network is a natural way of computing the support species tree structure in the absence of reliable additional information. The weights are used to take into account the tree cluster support when building an explicit phylogenetic network. The more gene trees we have, even when some of them are affected by different reticulation events, the more reliable the consensus network is. The most difficult practical situation for

our method is when we have only a few gene phylogenies, most of which are affected by the same reticulation event. But there is no any network building method that will infer a correct explicit phylogenetic network is such a situation.

We use both the discrepancy between the gene tree topologies (i.e. between the gene tree clusters) and statistical support of the gene tree branches in order to identify the consensus network branches and reticulation events. Bootstrap scores or posterior probabilities of the gene tree clusters are constantly used to compute weights and thus to validate the selected network braches. The acceptance of some of the clusters and rejection of the other is determined by comparing the cluster weights to a pre-defined threshold. Indeed, like any other phylogenetic method, bootstrapping has its own pitfalls (Morrison, 2013). However, in general, bootstrap scores and posterior probabilities are widely-accepted statistical estimates which have been proven very useful for assessing statistical robustness of phylogenetic trees.

Many studies supported by simulations advocate the use of probabilistic methods over distance- and parsimony-based approaches for inferring phylogenetic trees (Guindon and Gascuel, 2002; Hall, 2005; Huelsenbeck, 1995). Our general conclusion supported by the simulation results is that phylogenetic networks should be preferably reconstructed using maximum likelihood or Bayesian approaches as well. However, in some cases in this study, we used all the four main tree reconstruction approaches since different phylogenetic assumptions, optimality criterions and nucleotide or amino acid substitution models augment the collective probability of finding potential evolutionary conflicts.

In our first example examining the evolution of six honeybee species, we discovered a possible reticulate evolutionary history, suggesting that A. cerana could be a closer relative of A. mellifer, compared to the backbone species phylogeny in which the closest relative of A. mellifer is A. dorsata (Figure 1.5A - network obtained from the

full-length sequences). This finding was consistent with a possible hybridization/recombination hypothesis involving the ancestors of A. cerana and A. mellifer, which was first formulated by Makarenkov and colleagues (Makarenkov et al., 2004). Our weighted hybridization networks constructed using the sliding window procedure (Figure 1.5B and 1.5C) suggest explicitly that A. cerana is a possible hybrid of A. mellifer and A. koschevnikovi (see the arrows stemming from the A. mellifer and A. koschevnikovi branches and entering into the A. cerana branch). The opposite arrows entering into the A. cerana branch concern the intervals that have a very short overlap in both cases (Figure 1.5B and 1.5C) what suggests a possible recombination event. We cannot provide such an easy interpretation for the corresponding reticulogram, cluster network or splits graph (Figure 1.12A to 1.12C, respectively). Note that the backbone phylogeny we built using the bootstrap-based extended majority rule was consistent with the species phylogeny inferred in (Makarenkov et al., 2004).

Similarly, the dominant evolutionary history (i.e., the backbone phylogeny) we inferred when analyzing the chloroplast dataset was consistent with the findings of Makarenkov and Legendre (2004). The most significant reticulation event depicted in the network obtained from the full-length sequences (it is represented by a double-headed arrow in Figure 1.7A showing that each of the involved species might be a parent of the other) suggests a closer relationship between Euglena and Olithodiscus (i.e., stemming from a possible hybridization event involving the ancestors of these species) compared to the dominant scenario in which Olithodiscus is the closest neighbour of Anacystis. The networks inferred using the sliding window procedure (Figure 1.7B and 1.7C) suggest in addition that Chlamydomonas might be a hybrid species whose possible parents include the ancestors of Anacystis and Euglena, and the common ancestor of Tobacco, Rice and Marchantia, and that Euglena might be a parent of Chlorella.

In the horizontal gene transfer example, we considered the maximum likelihood phylogeny of 14 Archaean species inferred by Matte-Tailliez and colleagues (2002) using the gene concatenation approach. This tree played the role of the species tree, representing the dominant evolutionary history, in our analysis (Figure 1.9A). First, using multiple phylogenies of the gene *rpl12e* inferred using the BIONJ, DNAPARS, PhyML and MrBayes methods (Figure 1.8) and Algorithm III, we identified five potential horizontal gene transfer relationships not accounted for by the backbone tree topology (Figure 1.9B). Our findings were consistent with the horizontal gene transfer hypothesis formulated by Boc *et al.* (2011). Four transfer branches we inferred (see Transfers 1, 2, 4 and 5 in Figure 1.9B) were equivalent to those obtained by Boc and colleagues (Figure 1.14A). Furthermore, the fifth horizontal gene transfer we found (Transfer 3 in Figure 1.9B) differs from Transfer 5 in Figure 1.14A only by the presence of *M. bakeri* in the cluster of the donor organisms.

While full-length multiple sequence alignments can be directly used for finding diploid hybridization and complete horizontal gene transfer events, we need to consider the alignment fragments in order to detect smaller-scale evolutionary events, such as intragenic recombination and partial horizontal gene transfer (i.e., in the latter case a horizontal gene transfer is followed by an intragenic recombination leading to the formation of a mosaic gene; Boc *et al.*, 2011). The sliding window approach described above was applied here to search for partial gene transfers. The weighted consensus network of partial horizontal gene transfers built using Algorithm II (Figure 1.9C and 1.9D) allowed us to detect successfully five of the seven partial transfers originally predicted by Boc *et al.* (2013).

In terms of visualization and results interpretation, our explicit network model is easily explicable, while the interpretation of implicit network models (e.g., splits graphs, cluster networks and reticulograms) becomes extremely difficult when dealing with a high number of species or conflicting events (see Figures 1.10B-C, 1.11B-C and 1.12B-

C). Methods and software developed by Huson (1998), Legendre and Makarenkov (2002), Holland and Moulton (2003), Holland *et al.* (2006) and Huson and Rupp (2008) are rather devised to infer and visualize incompatibilities among gene trees without precisely describing the underlying evolutionary events. In contrast, our explicit weighted consensus network inference method is capable of detecting and validating, through the use of the weight function, the following reticulate evolutionary events: diploid or polyploid hybridization (recombination at the chromosome level), intragenic recombination, complete horizontal gene transfer and partial horizontal gene transfer followed by intragenic recombination. In a recent attempt, Guénoche (2013) developed a method to tackle the problem of conflicting evolutionary signals by finding multiple consensus trees instead of a network as a method for separating and representing the evolution of diverging genes. In the future, it would be interesting to verify whether this method could be extended to the inference of multiple consensus phylogenetic networks representing alternative evolutionary hypotheses. The computer program implementing our method is available for download at the following URL: http://www.info2.uqam.ca/~makarenkov_v/ConsensusNetwork.rar.

CHAPTER II

USING DIRECTED PHYLOGENETIC NETWORKS TO RETRACE SPECIES DISPERSAL HISTORY

Mehdi Layeghifard, Pedro R. Peres-Neto and Vladimir Makarenkov
Published in Molecular Phylogenetics and Evolution.

2.1 Summary

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.

2.2 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, 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 de-glaciated 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.

2.3 Materials and Methods

2.3.1 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.

2.3.2 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 (Figure 2.1). These sub-regions were then used in the final dispersal network reconstruction.

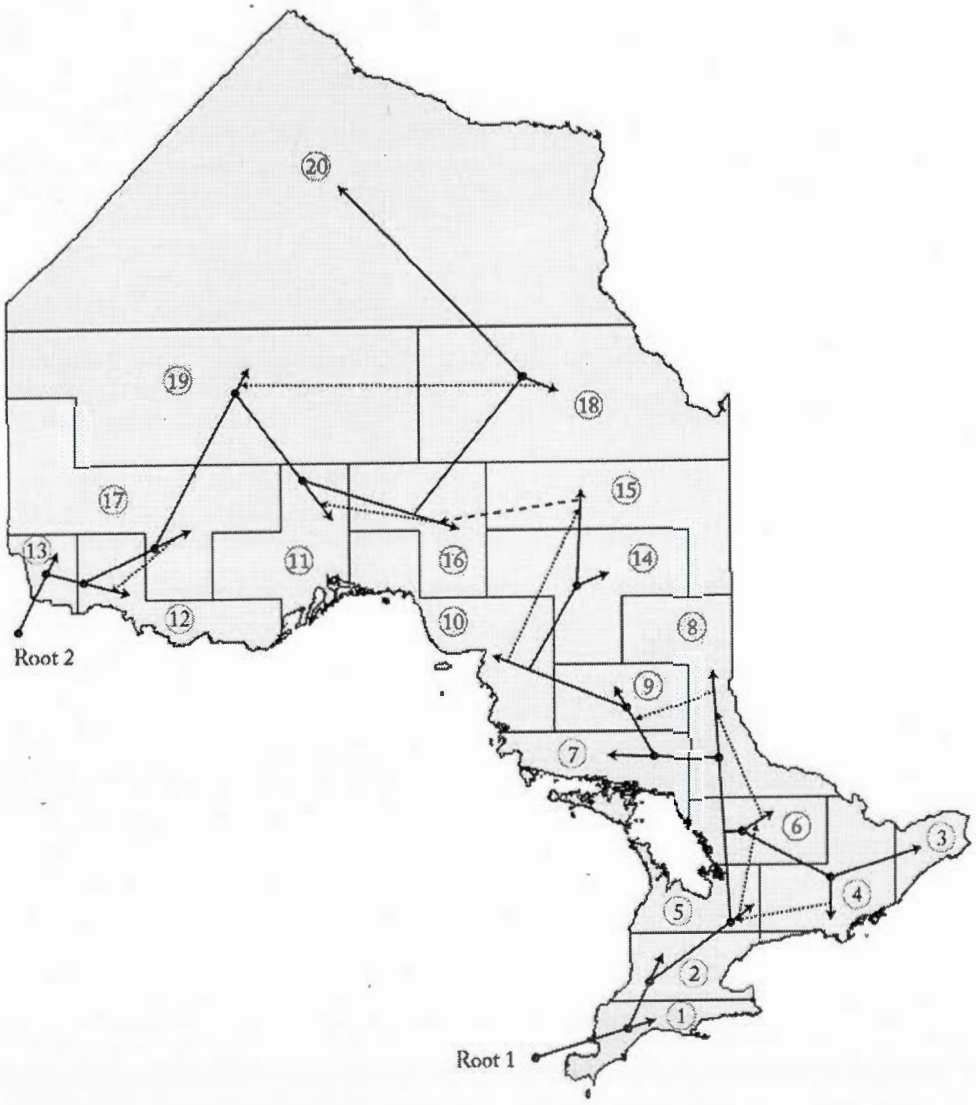


Figure 2.1 Biogeographic history of postglacial dispersal of Ontario fishes represented as a dispersal network. Solid, dotted and dashed lines represent, respectively, the main dispersal routes, alternative routes within both main Ontario regions and an alternative route connecting these two main regions.

2.3.3 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 (Figure 2.2). 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).

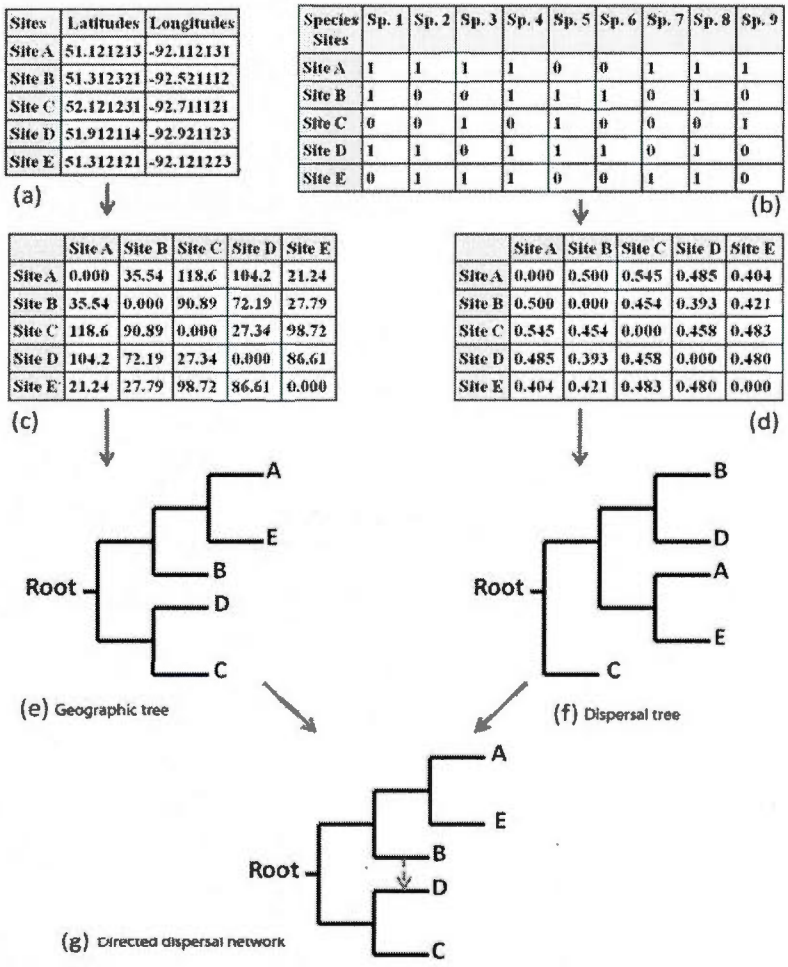


Figure 2.2 Schematic representation of the directional species dispersal network building process based on an artificial data set. (a) Coordinates of geographic sites; (b) presence–absence (or incidence) data set describing the distribution of species across sites; (c) geographic distance (Euclidean) 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.

The second step consists in building a dispersal network (Figure 2.2) 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 horizontal gene transfer (HGT) events. In the remainder of the article, we refer to our method as DSDNs (Directional Species Dispersal Networks; see Table 2.1 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 2.4 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 Figure 2.2).

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 Figure 2.1). 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).

Table 2.1 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

2.3.4 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 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 standardized average distance **Z** and its associated significance value (p-value) using the following formulas:

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

$$\mathbf{P} = (X + 1) / (N + 1),$$

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 2.2.

Table 2.2 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.

Genera	**Z-score**	**p-Value**
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**

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 (Figure 2.3). The species phylogenetic tree was inferred from the DNA sequences of mitochondrial COI genes (Figure 2.3a), whereas the species dispersal pattern tree was inferred from the Sørensen distance matrix calculated from the presence–absence data (Figure 2.3b). 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 (Figure 2.3a) and distributional (Figure 2.3b) 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).

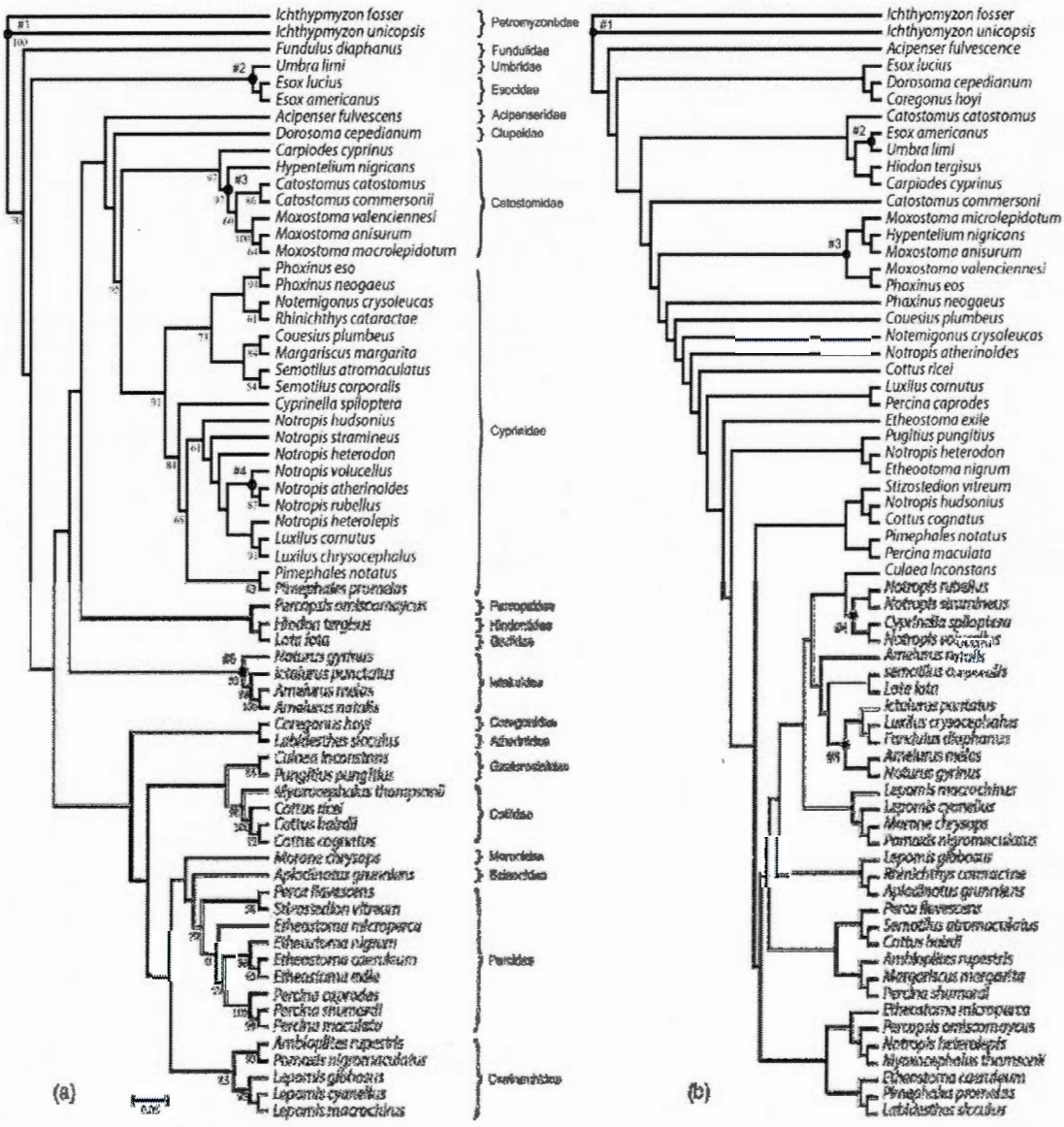


Figure 2.3 Comparison between phylogenetic tree and dispersal pattern tree. (a) Phylogenetic tree for the 66 fish species built using the available mitochondrial COI gene sequences. Family names are included in the species phylogeny. Bootstrap scores greater than 50% are shown on the tree branches and (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.

2.4 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; Figure 2.1). 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 Figure 1). We also found one alternative route between the southern and northern regions (dashed line in Figure 2.1). The dotted and dashed lines in Figure 2.1 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 2.2). 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 Figures 2.3a and 2.3b). 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 Figures 2.3a and 2.3b). 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 similarity (Figures 2.3a and 2.3b) even though the related species were still scattered across the trees.

2.5 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 (Figure 2.1). 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 (Figure 2.1). 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 (Figure 2.3a). 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 (Figure 2.3). 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 2.2). Note that these differences are not related to dispersal

avoidance (Table 2.2), 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.

CHAPTER III

SPATIAL NETWORKS FOR INFERRING DISPERSAL IN ECOLOGICAL COMMUNITIES

Mehdi Layeghifard, Vladimir Makarenkov and Pedro R. Peres-Neto

3.1 Summary

Multiple spatial and non-spatial processes are involved in patterning complex spatial variation in species and their assemblages. This complexity makes modelling and examination of spatial heterogeneity very challenging at the metacommunity level given the logistical limitations in tracking dispersal for multiple species across multiple communities. While metapopulation studies have inferred immigration rates based on landscape connectivity metrics, metacommunity studies instead, have relied on spatial predictors that are built without considering patch connectivity inferred from information on patch occurrence for multiple species at multiple communities. Here, we introduce a novel method to detect and explain spatial variability within metacommunities through the use of a graph-theoretical approach. Our multi-species spatial network (MSSN) method uses both geographic and incidence data as input to infer dispersal within metacommunities. Our simulation results and real data analyses showed that MSSN was more robust in terms of explaining variation in community analysis models than a commonly used method to detect spatial patterns in communities. In addition to its robustness in inferring dispersal within metacommunities, our proposed framework can be also useful in assessing the levels of spatial connectivity for each local community. Finally, our framework is highly

flexible and can incorporate different types of functions to infer spatial and different types of algorithms to infer migration levels and dispersal directionality.

3.2 Introduction

Ecological entities (e.g., individuals, populations, species and communities) show complex patterns of variation in space. This spatiality is often a combination of outcomes generated by endogenous mechanisms, such as dispersal limitation (but also sociality and reproduction) and species interactions, as well as by exogenous factors such as spatially structured environment (e.g., local environment, regional climate) which in turn impose spatial patterns in species distributions via habitat filtering (Dray *et al.*, 2012; Peres-Neto and Legendre, 2010). Therefore, the nature and origin of spatial structures of species and their communities are not always obvious, especially because species distributions are structured by a mix of spatial and non-spatial processes and factors (Gravel *et al.*, 2006; see Leibold *et al.*, 2004 for a review). Moreover, even if only spatial processes were at place, the complex interactions among those may not necessarily leave strong signatures. For example, repulsive interactions between parent trees and their seedlings can generate regular (non-spatial) patterns. Negative spatial autocorrelation (e.g., due to competitive interactions; e.g., Meyer *et al.*, 2008), and positive spatial autocorrelation (e.g., due to dispersal limitation) may actually cancel each other out and generate null spatial patterns (Dray, 2011).

One particular ecological level in which the complex interactions are evident are at the level of metacommunities (Leibold *et al.*, 2004), i.e., spatial networks of local species assemblages connected by dispersal. Metacommunity ecology has become a framework for understanding how dispersal interacts with local community assembly to determine patterns of species distributions among patches. Metacommunity dynamics has been increasing our understanding about complex interactions in

community ecology especially because of local species interactions have long been understood to predict much simpler patterns of community structure at large scales than what we typically observe in natural landscapes (Holyoak *et al.*, 2005; Huston, 1999; Ricklefs, 1987). Nevertheless applying these ideas to natural patterns of community variation is particularly challenging because of the lack of appropriate quantitative frameworks to estimate dispersal and connectivity patterns at the metacommunity level (i.e., multiple species at multiple sites; Jacobson and Peres-Neto, 2010). The main challenge owes to the fact that one cannot possibly estimate dispersal across multiple communities and multiple species directly. Moreover, dispersal dynamics can change through time and current spatial patterns may not necessarily reflect past dispersal history that was important for present-day metacommunity structure. Even in the case of single species distributed across patches by dispersal (i.e., metapopulations), assessing patterns of dispersal (e.g., mark-recapture at several locations) may be technically challenging and they still may not account for the importance of past dispersal (but see Jacobson and Peres-Neto, 2010 for potential genetic methods). Instead, metapopulation ecologists have inferred immigration rates based on connectivity metrics that attempt to estimate the inaccessibility of a patch or site to potential immigrants arriving from other patches and take into consideration the distribution of populations in the landscape (Bender, *et al.*, 2003; Hanski, 1994; Moilanen and Nieminen, 2002). Perhaps the simplest and most common measure of patch connectivity is the distance to the nearest occupied site (e.g., Bender *et al.*, 2003). Metacommunity studies (e.g., Beisner *et al.*, 2006; Cottenie, 2005; Gucht *et al.*, 2007), instead, have relied on spatial predictors (e.g., geographical positioning, geographic polynomials, eigenvector maps; see Legendre *et al.*, 2005 for a review) that are quite robust in detecting spatial patterns in data but are built without considering patch connectivity inferred from information on patch occurrence of multiple species at multiple communities (i.e., a metric homologous to metapopulation connectivity).

In order to address the challenges of assessing connectivity patterns at the metacommunity level, we introduce a novel method to detect and explain spatial

variability within metacommunities through the use of a graph-theoretical approach. A graph or network is a mathematical model of the pairwise relations between members of a given set of objects (here local communities or species assemblies). In ecology, there has been a multitude of studies using graph theory to understand food web structure (Banašek-Richter *et al.*, 2009; Krause *et al.*, 2003; Luczkovich *et al.*, 2003; Pimm, 2002), landscape connectivity (Bodin and Norberg, 2007; Ferrari *et al.*, 2007; Jordán *et al.*, 2007; Urban and Keitt, 2001), conservation biology (Bunn *et al.*, 2000; Naujokaitis-Lewis *et al.*, 2012; Rubio and Saura, 2012; Urban and Keitt, 2001; Yu *et al.*, 2012), and metapopulation ecology (Hanski and Ovaskainen, 2000; Ovaskainen and Hanski, 2001). Urban and Keitt (2001) presented a refined overview of the basic elements of graph theory, focusing especially on meta-population theory in conservation biology.

The aim of this study is to introduce and demonstrate the robustness and utility of a novel framework to investigate spatial patterns of connectivity within metacommunities (i.e., across multiple local communities for multiple species) using a graph-theoretical approach. This graph-theoretical approach, hereafter referred as to multi-species spatial networks (MSSN), uses both geographic data (geographic positions of sites in the form of latitude and longitude values) and incidence data (presence-absence of species across multiple sites) as input to infer dispersal within metacommunities.

3.3 Methods

In graph theory, points or objects (here communities and sites are used interchangeably) in space are referred to as "vertices" or "nodes" and the lines (connections) linking them are called "edges". Therefore, a network is a collection of vertices (points) interconnected by edges (lines). A network is called directed if all the

edges are unidirectional (Figure 3.1b; i.e., amenable to measuring directional dispersal) and undirected if they are bidirectional (Figure 3.1a; i.e., simply connected but no directionality). The basis of our framework is to reconcile the spatial representation of the communities using a geographic tree (i.e., a dendrogram representing the spatial similarity of sites based on their spatial positioning) with the data on their species compositions. If there is a perfect match between the two (closest sites are always more similar in species composition), then there is no need of reconciliation and the spatial tree will perfectly represent the spatial structure in species composition across communities (i.e., metacommunity). Conversely if there are communities that are more similar in species composition than expected by their spatial proximity, then a reconciliation between their spatial differences and species compositions can be performed by adding extra edges (links) connecting the two communities (vertices). Therefore, the final spatial network represents the reconciliation between the spatial distribution of sites and species compositions at those sites (i.e., local communities). A diagrammatic description of the steps involved in our spatial network method is given in Figure 3.2 and is based on two broad steps:

(1) Build a spatial tree using pairwise Euclidean distances between sites, computed from their geographic coordinates, and estimate the root from the incidence data (i.e., species composition across communities). Note that these geographic distances could be also transformed in a way to represent functions that better represent dispersal functions such as the negative exponential ($\exp(-d)$ where d represents the distance between two sites) or other metrics of landscape resistance (see Zeller *et al.*, 2012 for a review).

(2) Use the species distribution data to find extra links (reconciliations) among communities. Community similarity here was measured using the Jaccard similarity coefficient but other indexes can be certainly considered (see Legendre and Legendre, 2012 for a review). Build the spatial network for the

metacommunity by adding extra links to the spatial backbone tree calculated in step (1). The technical details involved in these two steps are explained below.

3.3.1 Step 1: Building the spatial tree

Our spatial network method requires two types of input: an incidence matrix (a matrix of sites-by-species presence-absence) and a geographic positioning matrix (a matrix of sites-by-geographic coordinates). The incidence data matrix is a binary matrix of 1s and 0s indicating the presence or absence information of each species (columns) within each patch (rows).

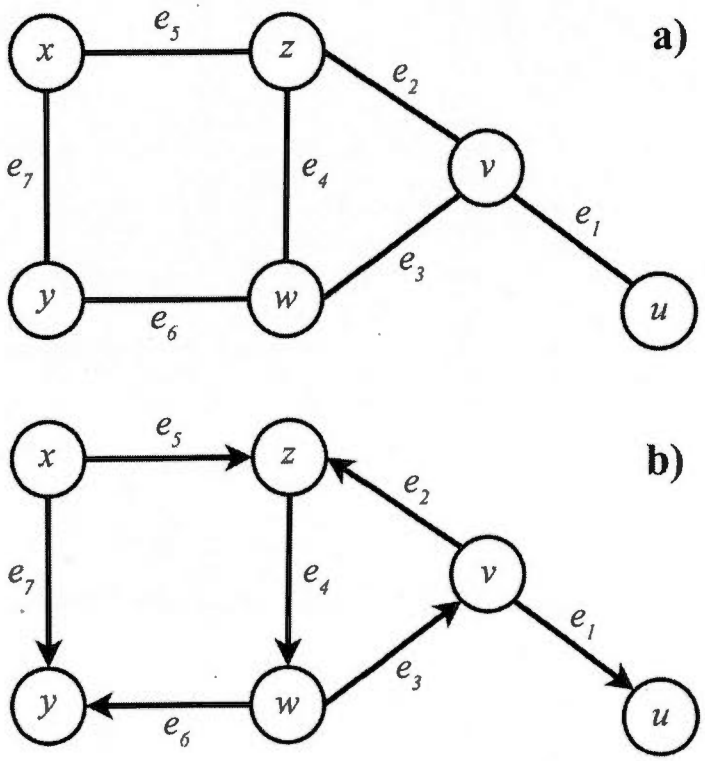


Figure 3.1 A simple representation of two mathematical graphs or networks. a) Shows an undirected graph with six vertices (or nodes) and seven edges. b) A directed graph drawn using the same set of vertices and edges. The only difference between the two graphs is that the edges of graph b have directions.

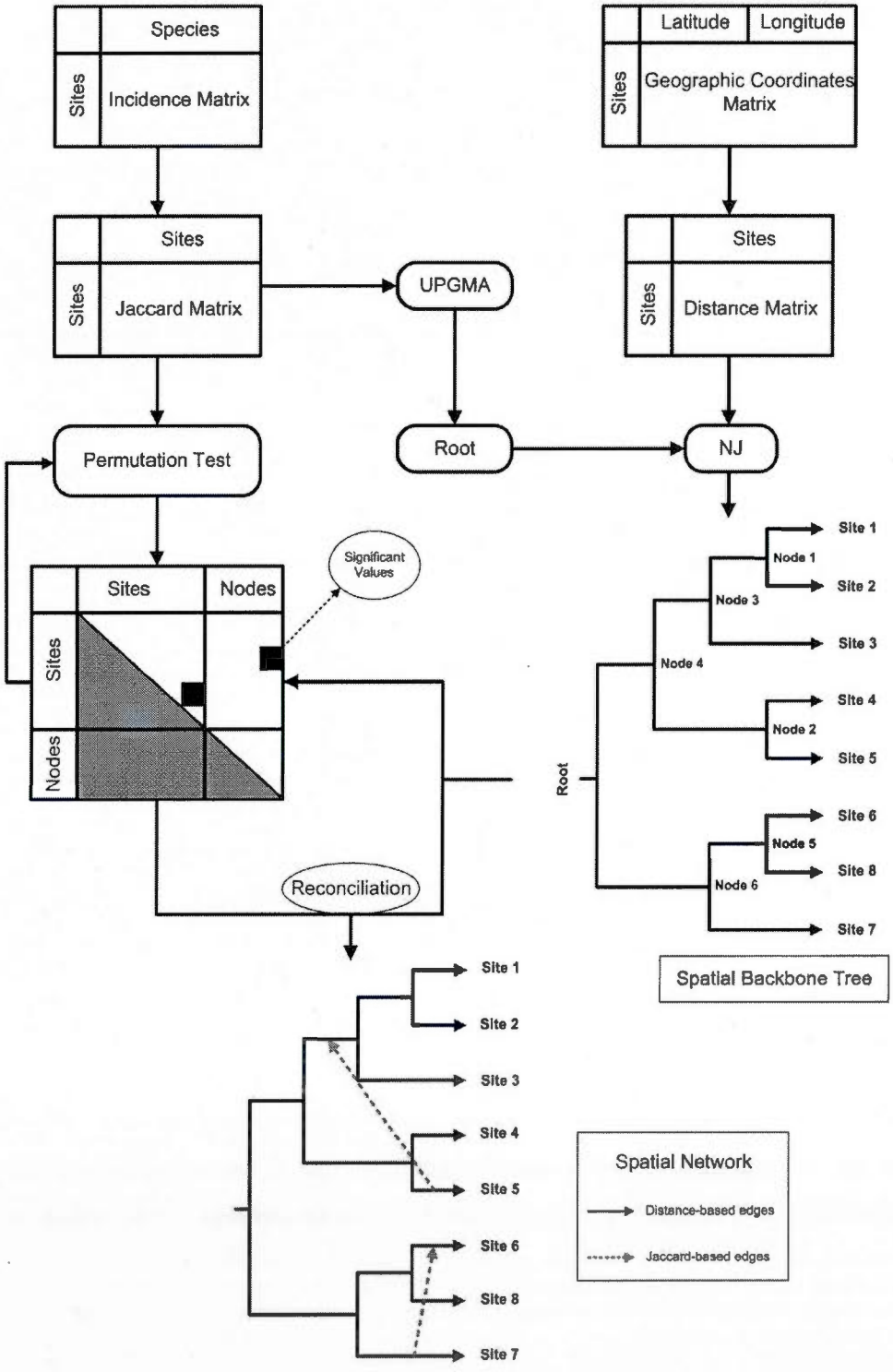


Figure 3.2 Diagrammatic summary of the steps involved in our spatial network method.

An important aspect of inferring dispersal is that any given two communities may not have the same level of dispersal between them in which case dispersal is asymmetric or directional. Therefore, the first step to build the spatial tree is to estimate its root as an unrooted tree has no reference to direction in space. This is akin to unrooted phylogenetic trees in which the direction of time is undetermined. Note, however, that directionality is not an essential component of our method and in cases where directionality is not of interest, it can be simply ignored. In such case, our method could be implemented simply on the basis of an unrooted spatial tree. Note that although the spatial tree was based on geographic distances among sites, its root was determined by the species composition. In this way, the root would represent how the species pool can be best divided across two major clusters of sites that are spatially structured. One way to estimate the root would be via an exhaustive process of determining all possible rooted trees from the unrooted spatial tree (i.e., a root from all possible edges from the unrooted tree) and then assess which rooted tree best fit with species community composition across all communities. Note that the average of all rooted trees is the unrooted one. This exhaustive method, however, is impractical especially given the large number of simulations that we used in this study to assess the performance of our proposed framework. Instead, we used UPGMA (Unweighted Pair Group Method with Arithmetic Mean; Sokal and Michener, 1958), which is a well-known and widely-used hierarchical clustering method.

Once the root has been established (i.e., by an UPGMA on species incidence matrix), we applied the widely-used Neighbor-Joining method for phylogenetic reconstruction (Saitou and Nei, 1987) to build the spatial tree based on a pair-wise Euclidean geographic distance among sites (i.e., local communities). This spatial tree serves, then, as the backbone of the spatial network for the local communities. The reason for using Neighbor-Joining method (instead of the one produced by UPGMA) was that minimum length Steiner tree with 120° between all branches, which is a particular case of a phylogenetic tree, is known to generate the tree connecting all points in the plane and

allows for representing geographic information as a bifurcating minimum length tree (Cavalli-Sforza and Edwards, 1967). While UPGMA is a simple clustering method mainly used in bioinformatics for the creation of phenetic or rooted trees (phenograms and dendrograms, respectively), it is not a well-regarded method for tree inference. Conversely, Neighbor-Joining is well-known for inferring the correct tree as long as the distance matrix is correct and "nearly additive" (Atteson, 1997; Felsenstein, 2004). Although in reality these conditions are rarely satisfied, Neighbor-Joining often constructs the correct tree topology (Mihaescu *et al.*, 2009). Therefore, we combined the strengths of the two methods to construct the rooted spatial tree using the geographic relationships among sites. Note that we could also apply a different combination of methods instead and assess which one performs best for the same data.

3.3.2 Step 2: Building the metacommunity network

To convert the binary spatial tree into a directed (asymmetric) spatial metacommunity network, we needed first to detect strong (significant) connections (similarities) among local communities and then the corresponding direction of these additional connections. Once this was established, we would then add a directional connection to the spatial tree built in step 1. To detect potential connectivity among communities, our framework makes use of the pair-wise similarities between all possible combinations of communities and nodes (i.e., either of two local communities, a node and a community, or two nodes; see Appendix A for computations details) using the Jaccard similarity coefficient (Figure 3.2). We only considered additional links for communities and/or nodes that shared more species than expected by chance alone (an indicative of strong connectivity). In order to estimate this probability, we randomly permuted entire rows (sites) of the incidence matrix in relation to one another and recalculated the Jaccard similarity matrix (i.e., across all possible combinations of communities and nodes) based on the permuted value. We repeated this step 999 times

(Figure 3.2) and computed a p-value for every pair (i.e., either two local communities, a node and a community, or two nodes) as the number of random values greater or equal to their respective observed values plus 1 (i.e., the observed value was also part of the null distribution) divided by 1000. Here we considered that a pair should be connected when they shared a significant number of species (here a confidence limit of 0.05 was used). An alternative (though not considered here) would be to consider greater alpha values (e.g., 0.10, 0.20) and see if they improved model fit in predicting species distributions.

The direction of dispersal is always from the root to the nodes except for the extra branches which do not respect the spatial matrix and need to have their directionality estimated. Once the significance of the connection was established, dispersal directionality of newly added edges was determined by minimizing the topological differences computed by the Robinson and Foulds method (Robinson and Foulds, 1981). For example, in the case of the extra edge found between sites 6 and 7 (dotted arrow line from Site 7 to Site 6), we first attached the newly found edge to Site 6 (representing one direction; Site 6 to Site 7) and calculated the Robinson-Foulds topological distance between the resulting tree and original tree. Then, similarly, we attached the new edge to Site 7 (representing the reverse direction; from Site 7 to Site 6) and computed the Robinson-Foulds topological distance between the resulting tree and original tree. Finally, the smaller distance determined the direction of the newly found edge, which in this case is from Site 7 to Site 6 (see Appendix B for more details).

3.3.3 Building dispersal predictors

In single-species metapopulation studies, the common procedure is to model species distributions (an incidence vector of presences and absences across local populations) against a predictor (or a set of predictors) of site connectivity (Foltête *et al.*, 2012;

Hartel Tibor, 2010; Peres-Neto and Cumming, 2010). Here, in order to build a common set of connectivity predictors for all species across local communities, we first coded the MSSN into a site by edge matrix with rows representing the local communities and columns representing the edges (or branches) of the network (Blanchet *et al.*, 2008; see Appendix A for a complete example with calculations). The site by edge matrix is a site-by-edge binary matrix $H = [h_{ij}]$ in which each entry h_{ij} is set to 1 if edge j was involved in the path connecting the site i to the root and set to 0, otherwise. Next, H was multiplied by a vector (1-by-edge) of edge weights $E = [e_{ij}]$, resulting into a weighted site-by-edge matrix HE. Akin to metapopulation metrics (e.g., distance to nearest occupied site, average distance to all occupied sites), we defined weights as a compromise between the geographic proximity and community composition similarity between two communities (see Appendix A for a complete example with calculations). Finally, a site-by-site Euclidean matrix C was calculated on the basis of HE. In essence, C is a connectivity matrix representing how local communities are spatially connected on the basis of our multi-species spatial network (MSSN). Matrix C was then double-centered as:

$$C_c = (I-11^T/n)C(I-11^T/n))$$

where I is an n-by-n identity matrix, 1 is an n-by-1 vector of ones, T denotes matrix transpose, and n is the number of sites. We then extracted the eigenvectors from C_c, which represents all orthogonal and linearly independent spatial patterns that are possible to produce from C (Griffith and Peres-Neto, 2006). The extracted eigenvectors are then used as dispersal predictors to model species distributions. The extracted eigenvectors are akin and will be referred here as to the asymmetric eigenvector maps (AEMs) approach developed by Blanchet and colleagues (2008) with the difference that the node-by-edge was built on the basis of our MSSN approach.

Our final step, as in metapopulation studies (see Prugh, 2009 for a review), was to model species distributions (i.e., n-by-species incidence matrix) on the basis of our connectivity predictors (i.e., AEMs). Because we have multiple species, we have applied Redundancy Analysis (RDA), a regression modelling technique that can accommodate multiple response variables (species). Model fit was assessed via an adjusted coefficient of determination (adjusted R^2; Peres-Neto *et al.*, 2006) and model performance via a permutation test explained in the next section.

3.3.4 Assessing the performance of MSSN via simulations

Here we compared the performance (type I error, statistical power and model fit – adjusted R^2) of our AEM approach with the most commonly used approach to model the spatial component of multi-species distributions, namely Moran's Eigenvector Maps (MEM; Dray *et al.*, 2006). MEM are the eigenvectors of a non-directional connectivity matrix that simply considers the spatial proximity of sites (see Griffith and Peres-Neto, 2006 for calculation details), thus differing from the AEM approach based on MSSN in which directionality and spatial distributional characteristics of species (as in single species metapopulation models) are used.

In order to estimate the significance of our metacommunity models (RDA with species incidence matrix as response and either AEM and MEM as spatial predictors) in explaining species distributions, we randomly permuted rows (local communities) of the incidence matrix in respect to one another. Because our AEM approach is based on the distributional properties of communities, we re-calculated for each permuted set its multi-species spatial network (MSSN), extracting a new set of 'random' AEMs. The permuted incidence matrix was then modelled (via RDA) against MEMs (which is invariable under permutation) and the AEMs on the basis of the permuted set. For each permuted set and each set of spatial predictors, we calculated their respective adjusted

R^2. We repeated the permutation procedure 999 times and computed a p-value for the each RDA (AEMs or MEMs) as the number of random adjusted R^2 values greater or equal to their respective observed values plus 1 divided by 1000.

In order to test the efficiency of our multi-species spatial network method, it was important to apply the method on simulated data given that we could generate data with known structuring levels (see next section for an assessment based on real data sets). We simulated metacommunities consisting of 2500 local communities (sites) and 50 species were collectively used. Here, local communities were distributed across a squared lattice (50 x 50). The first step was to calculate a pairwise geographic Euclidean distance matrix D = $[d_{ij}]$ among all the 2500 communities in the landscape. Next, in order to generate spatial patterns into the species distributions within the metacommunity (lattice), we created a spatial matrix W as follows:

$$W=[w_{ij}]=\begin{cases}\frac{3d_{ij}}{2a}-0.5\left(\frac{d_{ij}}{a}\right)^3 & \text{if } d_{ij} \le a \\ 0 & \text{if } d_{ij} > a\end{cases}$$

where *a* represents the range parameter. By varying *a* (greater values represent greater autocorrelation, i.e., more spatially structured metacommunities). Next, the Cholesky decomposition was applied to W. By post-multiplying the upper-triangular from the decomposed matrix by a random normally distributed vector N(0,1) with 2500 observations, we created a normally distributed variable X according to a spherical variogram with a specific given range *a*. Because we wanted to simulate species having different levels of similarity in their distributions across local communities, we created a vector b = $[b_i]$ with 50 entries varying in increments of one from -n/2 to +n/2, where n is the number of species. For each species, we created a vector of probabilities P_i

corresponding to the chance that the i^{th} species occupies the j^{th} local community according to the simulated spatial gradient X as follows:

$$\mathbf{P}_i = \frac{1}{1+e^{(-b_o+b_i\mathbf{X}+e)}}$$

where $-b_o$ is a randomly generated number from a uniform distribution that changes for each species i and e is a random normally distributed vector N(0,1) with 2500 observations that introduces further noise to species distributions (i.e., so that species having very similar b values do not end up with extremely similar probabilities). P_i was then converted into a binary vector of presences-absences by drawing for each local community a random value from the binomial distribution according to p_{ji}. By combining all P_i vectors from all n species, we created the incidence matrix (distributional matrix) for any given particular metacommunity. Note that although the approach used here to simulate metacommunities does not simulate dispersal *per se*, our simulation protocol would have led to parallel results if we have actually simulated movement across the landscape instead of constraining species distributions on the basis of a spatialized environment. This is because in our simulation species tracked environmental features that are themselves spatialized. Moreover, simulations based on dispersal are extremely computationally time consuming especially given that our MSSN framework is also time consuming.

In order to contrast the power of our MEM and AEM, X was generated by considering spatial ranges (a) from 1 to 30. For type I error estimates, X was a non-spatial variable N(0,1). For each range and the non-spatial X, we simulated 500 different metacommunities that were then used to infer spatial variability using our MSSN method. Before doing so, however, some non-spatial pattern in local species composition was introduced to each simulated metacommunity through replacing 10 or 20 % of local communities with randomly chosen communities within the

metacommunity. These replacements were intended to emulate unusual non-spatial events and see how the method behaves. Following, from each simulated metacommunity (30 ranges x 500 metacommunities x 3 (0, 10% or 20% replacement = 45 000 metacommunities), one sample containing 50 and 100 local communities, respectively, were randomly selected with replacement, so that the two samples could have communities in common (though unlikely give the large number of local communities, i.e., 2500). Samples were taken because in realistic situations we only estimate patterns of species distributions in a much smaller number of communities in contrast to the metacommunity. The sampled data from the metacommunity represented an incidence matrix of species occurrences across sampled local communities (presences and absences) as well as their matrix of geographic coordinates. These two matrices were then used as input to our MSSN method.

3.3.5 Assessing the performance of MSSN on real datasets

Here we used a dataset on fish communities inhabiting various lakes across Ontario province of Canada. This data set, which was obtained from the Ontario Fish Distribution Database (OFDD) maintained by the Ontario Ministry of Natural Resources (OMNR), contains the presence-absence records of 134 fish species distributed among approximately 9900 lakes as well as the geographic positions of the lakes. We used presence-absence records collected in summers between 1968 and 1985 distributed across 72 independent watersheds (see Henriques-Silva *et al.*, 2012 for complete details). In this study, each watershed was considered as a metacommunity. The number of sites (local communities) and species varied very much across watersheds (between 21 and 280 sites and 17 and 50 species). In order to contrast the two spatial models (MSSN-AEM and MEM), we simply contrasted their adjusted R^2 across all 72 watersheds.

3.4 Results

3.4.1 Simulated data

Figures 3.3 and 3.4 summarize the simulation results comparing our MSSN and MEM approaches. First, in terms of explained variance (adjusted R^2 values), it is clear that our novel approach outperforms MEM (Figure 3.3). Second, both methods are, as expected, sensitive to the level of spatial autocorrelation in which low spatial ranges reduce the ability of both methods in detecting spatial patterns in species community composition (Figures 3.3 and 3.4). Third, both methods are sensitive to the sample size in which large samples increase the performance of both methods. Fourth, both methods were sensitive to the level of random (non-spatial) replacement of communities in which greater levels of non-spatial noise (contrast 20% with 0% replacement; Figures 3 and 4) decreased the performance of both methods. Fifth, despite the fact that our framework generate models with greater variance explained (Figure 3.3), both frameworks (MSSN and MEM) present similar levels of power. Finally, the type I error of our framework is correct (Figure 3.4; a range = 0 provide 5% of significant models as expected under a rejection level of 0.05).

3.4.2 Real ecological data

Figure 3.5 contrasts the adjusted R^2 values for MSSN and MEM frameworks across the 72 watersheds. The results clearly show the advantage of our spatial network method over MEM in detecting spatial patterns in a large set of real data. For the large majority of watersheds (65 watersheds out of 72 or 90.27% of watersheds), the adjusted R^2 values obtained through the use of the MSSN approach were larger than those obtained by MEM. Although our spatial network method failed to surpass the MEM

method in seven cases, it still managed to infer a quite similar amount of variation in these cases.

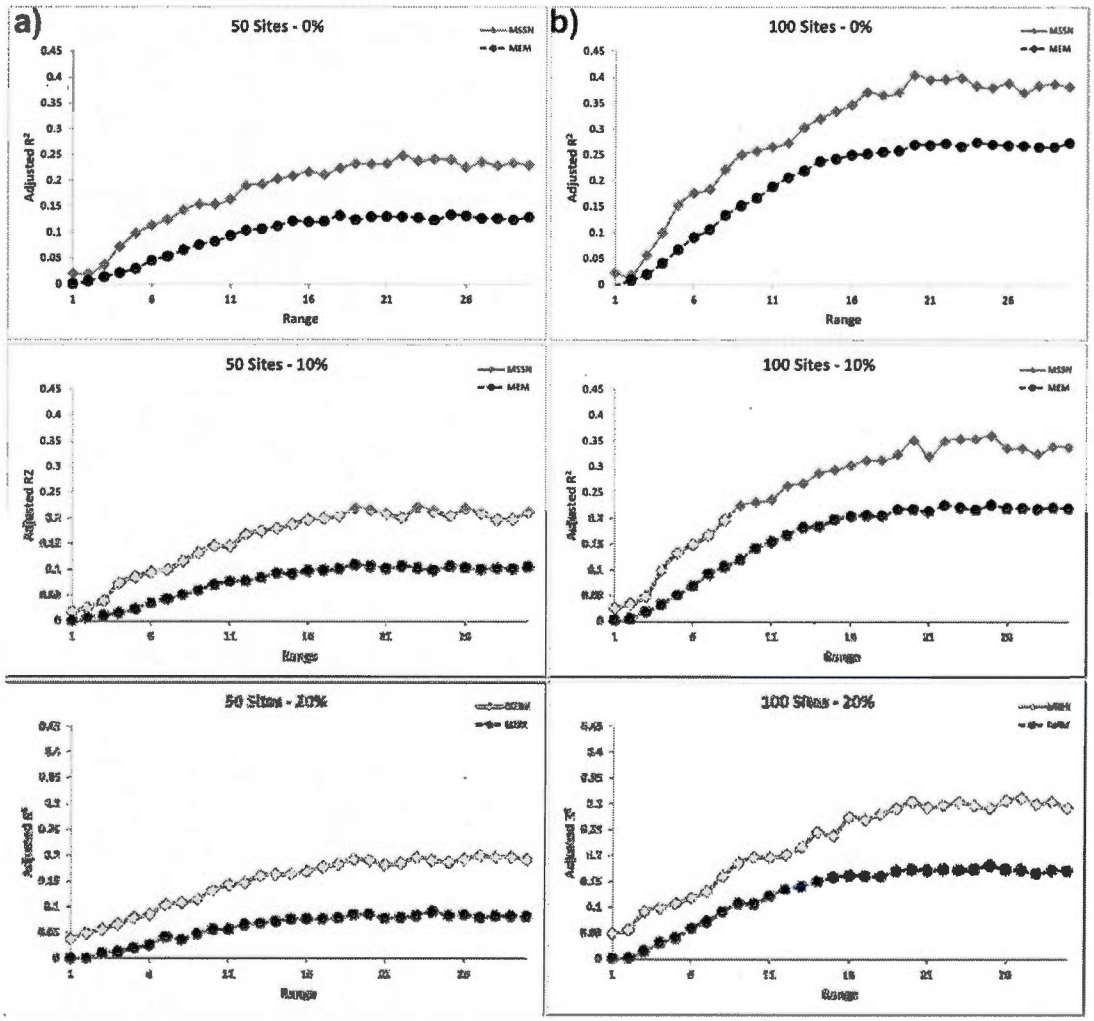


Figure 3.3 Adjusted R^2 values for simulated landscapes with 0, 10 and 20% changes by our spatial network method (SNM) and MEM method. The datasets used in this analysis were consisted of 50- and 100-sites metacommunities.

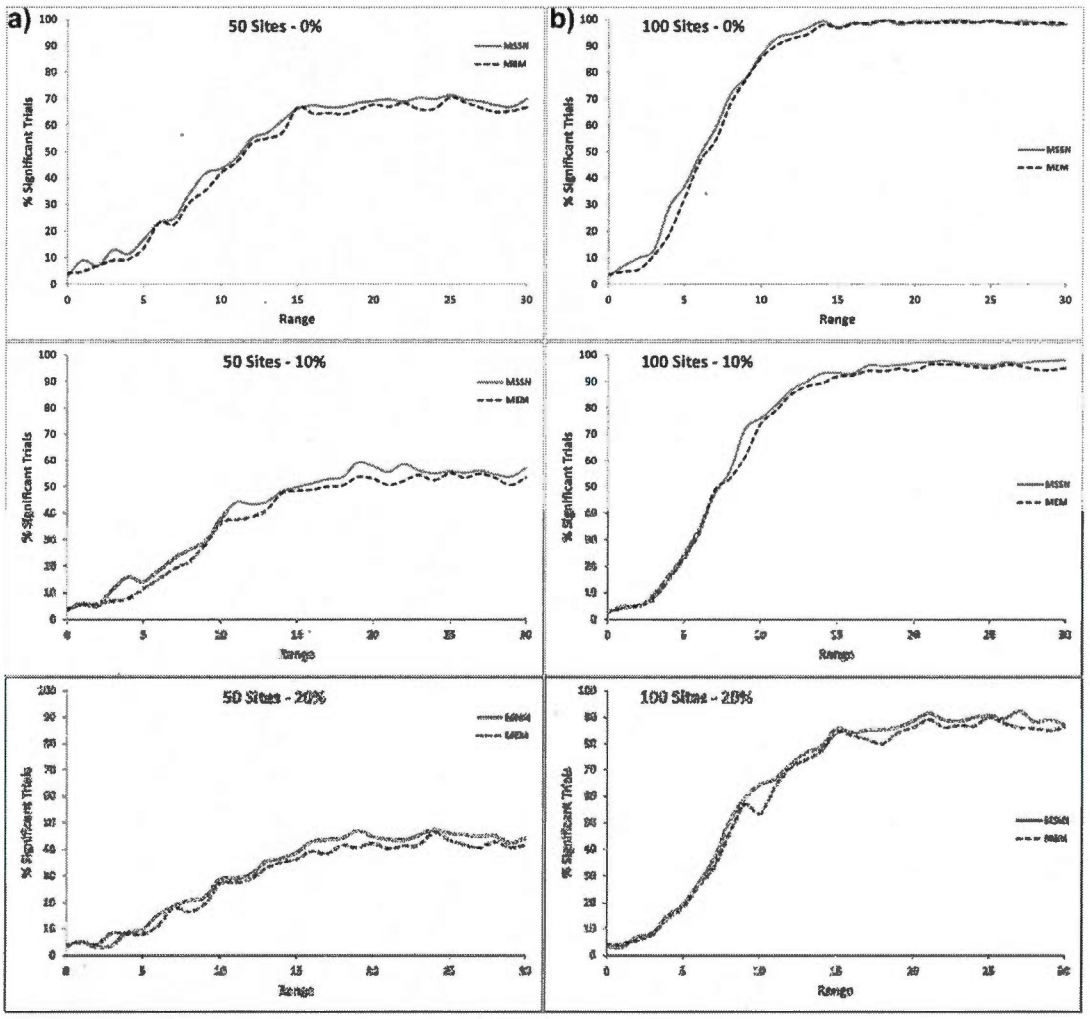


Figure 3.4 Type I error (range = 0) and power (range from 1 to 30) measured as proportion of rejections (alpha=0.05) per 500 tests.

An additional advantage of our framework is that we can make inferences (strength and direction) about the levels of connectivity across local communities. Figure 3.6 contrasts two metacommunities (watersheds) in terms of the spatial patterns of local community connectivity. Each circle in Figure 3.6 represents a site (lake) from the watershed and the size of the circles shows the amount of interaction they have with

the other sites: the size of the circles are proportional to the interaction (connectivity) in the form of the number of links (both inner – immigration events and outer – emigration events) connecting any lake in particular and the rest of the spatial network. The solid black part of the circles represents the number of network links terminated at those sites (immigration from other local communities) and the white part shows the number of links originated from those sites (emigration events toward other local communities).

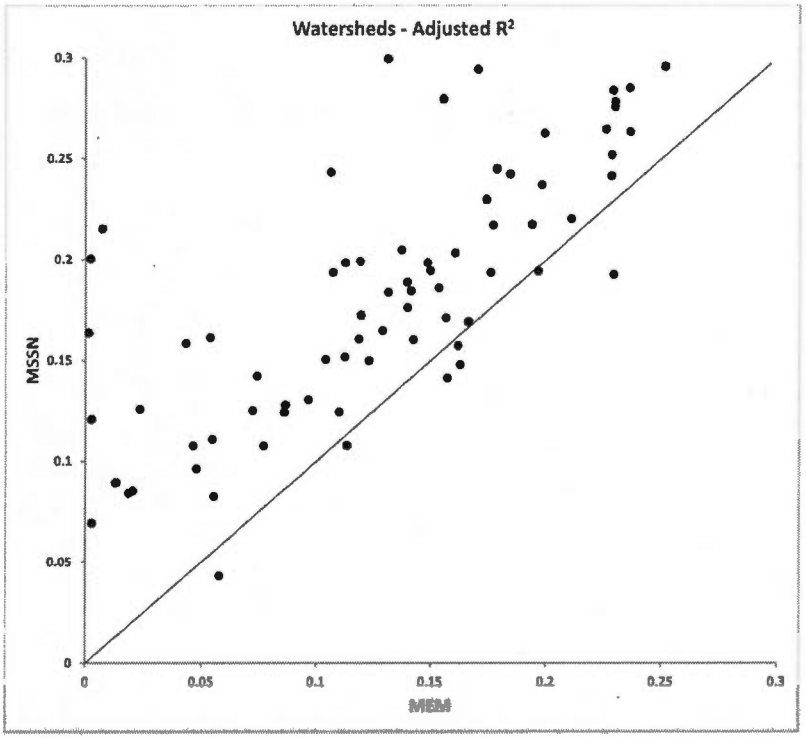


Figure 3.5 Contrast between MSSN and MEM methods on the basis of adjusted R^2 values obtained from real ecological data sets (72 fish metacommunities).

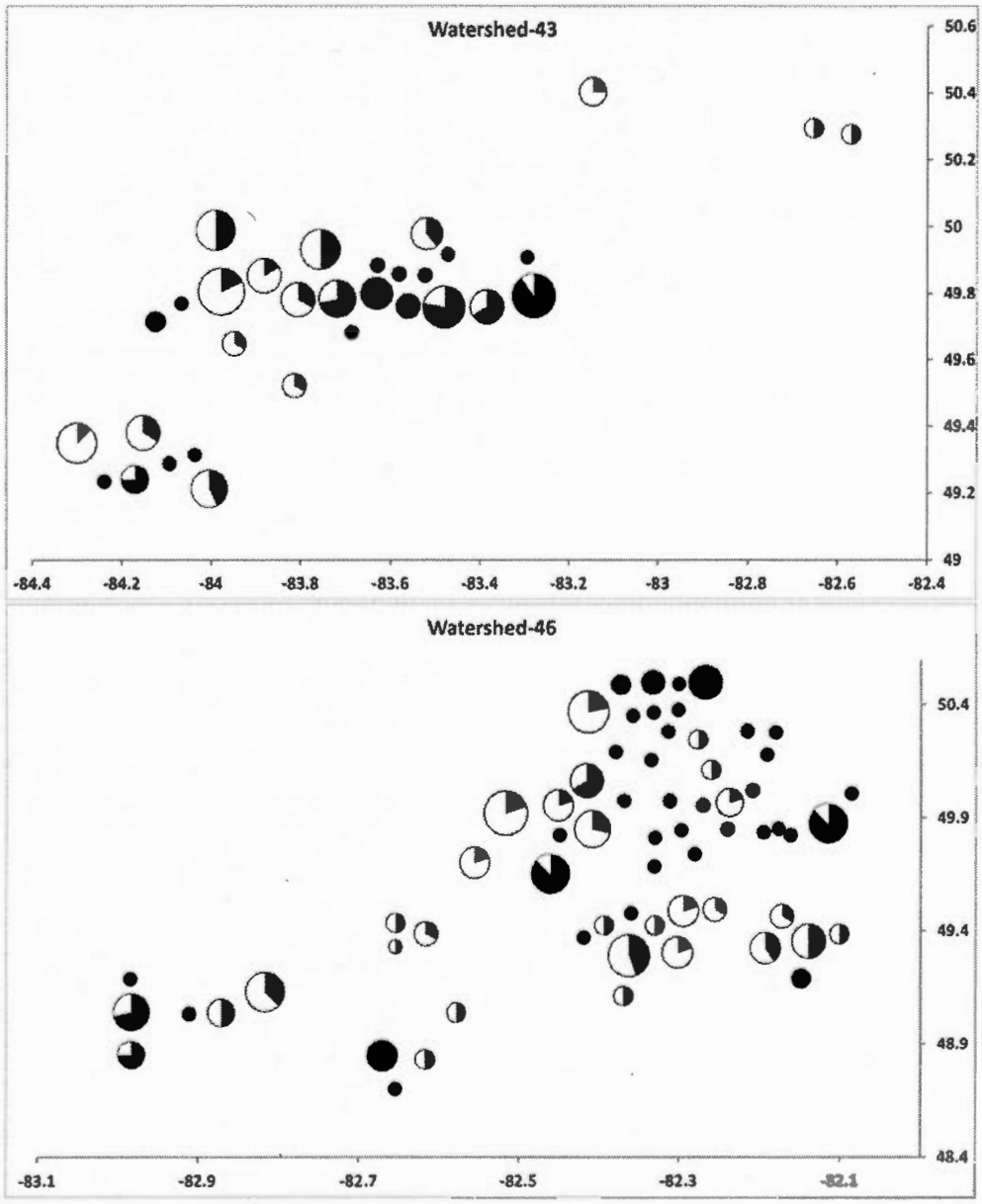


Figure 3.6 Bubble plot maps for lakes of two fish metacommunities (watersheds) representing their levels of connectivity with the other lakes within watersheds. Lakes are plotted according to their geographic positioning. The size of the circles represents the levels of connectivity for any particular lake. The amount of black is proportional to the estimated number of immigration events and the amount of white is proportional to the estimated number of emigration events.

3.5 Discussion

Inferring dispersal in real metacommunities is a daunting task given the logistical limitations in following individuals across a wide range of taxa and geographic locations. Moreover, the processes shaping metacommunities may have been historical and much beyond the temporal scope of the empirical data on species distributions (Layeghifard *et al.*, 2012; Leibold *et al.*, 2010). Our framework is intended to detect the spatial variability in metacommunities and represent them as spatial networks. It is the first such method applied to multi-species communities using a graph-theoretical approach. Graph-theoretical approaches have already been used in landscape ecology to examine the sensitivity of landscape connectivity to changes in landscape configuration (Keitt *et al.*, 1997), to assess overall and individual patch contribution to landscape connectivity (Urban and Keitt, 2001), to quantify levels of compartmentalization in landscapes (Bodin and Norberg, 2007) and to build and analyze spatially implicit models of compartmentalization in trophic structure (Dunne *et al.*, 2002; Pascual and Dunne, 2005). However, it has not been used to detect and explain the spatial variability of metacommunities so far.

Given the challenges of measuring dispersal directly within metacommunities, the proposed framework offers several features. First, it provides a parallel framework used in metapopulation models (Dunham and Rieman, 1999; Hanski, 1994; Hartel Tibor, 2010; Knapp *et al.*, 2003) given that our measure of connectivity is based on functions that represent spatial distributions of occupied versus non-occupied sites for multiple species (homologous to metapopulation metrics; see Bender *et al.*, 2003 and Prugh, 2009 for reviews). Second, it infers dispersal directionality across local communities. This is a major advantage even over metapopulation frameworks based on connectivity metrics (e.g., nearest occupied neighbour site) in which by having distributional information for multiple species, one can infer the likelihood of emigration versus immigration between two local communities. Even metapopulation metrics for

inferring dispersal do not infer directionality (i.e., assume omni-directionality; e.g., Magle *et al.*, 2009) and by using information on multiple species, our framework allows for directionality because we can use similarity in species compositions to estimate the most likely direction regarding species dispersal. The third advantage is that our method infers connectivity matrices that can be then used as spatial predictors into multiple species modelling frameworks. This is akin to single-species modelling that use metapopulation connectivity metrics to estimate site isolation (e.g., Dunham and Rieman, 1999; Prugh, 2009). It also allows estimating how well connected (hot spots) or disconnected (cold spots) local communities are.

The final advantage, in which our MSSM framework is capable of detecting patterns of connectivity that are not necessarily spatialized, deserves some additional attention because it relates to the way that connectivity metrics for metapopulations and ours (metacommunity) make inferences about dispersal. Although we commonly assume that the signatures of dispersal are spatialized, two communities that are spatially close may harbour quite different species and sites that are spatially distant may assemble similar species, thus reducing the ability to infer dispersal solely on the basis of the spatial structure of species distributions. That is the reason why metapopulation uses incidence information across the landscape and ours use community similarity across the metacommunity. However, as in metapopulation metrics, our MSSN framework also weights community similarity in relation to geographic distance by considering a compromise between spatial arrangement and information on community similarity. If there is a high level of similarity between two communities that are quite distant apart, they will not be considered as connected as if the same two communities were nearby. This is an important issue when studying metapopulations and metacommunities inhabiting environmentally heterogeneous landscapes especially those composed of species that are good dispersers but have strong environmental preferences. In this case, species can get anywhere (mass effect perspective; Leibold *et al.*, 2004) but are sorted according to the type of environment (species-sorting perspective). It follows that

metapopulation metrics and our metacommunity framework may infer strong dispersal dynamics across occupied sites (especially in the case in which optimal patches are also spatially structured) whereas in fact the major factor is instead strong environmental filtering instead. One way to separate these two hypotheses is to use a variation partitioning approach (Borcard *et al.*, 1992; Peres-Neto *et al.*, 2006) in which environment and our MSSN-AEM are used and contrasted against each other. In this case, if species have strong environmental affinities in which optimal environments are highly spatially structured, and they are not dispersal limited, then environmental predictors and our MSSN-AEM predictors should covary strongly and serve as an indication that our dispersal predictors are confounded by environment. Note that this is not an issue per se of our method, but an issue of natural landscapes not being able to provide orthogonal designs (i.e., variation in species optima being not spatial structured).

Our simulation and real data applications clearly show the features and advantages of our framework over a widely used method to depict spatial patterns in metacommunities (MEM). The main advantage of our method is that it integrates both geographic and species composition information to infer and explain spatial heterogeneity in species distributions. By integrating both sources of information we can infer about directionality and also non-spatialized dispersal patterns. Although the geographic distance between patches is a fundamental component of any landscape, species dispersal behaviour also plays a key role in shaping the spatial structure of metacommunities. In cases where species distributions closely follow the geographic arrangement of the habitat patches in the landscape, species dispersal can be ignored in practice. However, in reality, species dispersal patterns are much more complex and our network framework aims directly at inferring such patterns. While presence-absence data sets are not always ideal representations of species dispersal patterns in a metacommunity, they are widely available and are commonly used in ecological analyses to understand the processes driving common patterns of species distributions.

Therefore, integrating presence-absence information with geographic distances, we can gain greater insights into the spatial heterogeneity of metacommunities.

In addition to its robustness in inferring dispersal within metacommunities, our proposed framework could also be useful in assessing the impacts of fragmentation or loss of habitats on metacommunity structure (Benton *et al.*, 2003; Fahrig and Merriam, 1994; Meffe *et al.*, 2002). The most immediate effect of habitat fragmentation or loss is the change in the spatial structure of landscapes. This shift in spatial structure typically leads to substantial changes in species dispersal patterns and our framework can be applied using data from temporal surveys to assess changes in patterns of dispersal among local communities considering the entire metacommunity. Because spatial patterns of landscapes are critical to devising habitat conservation plans (Bunn *et al.*, 2000), our multi-species multi-site approach can be applied to infer changes in community connectivity through time and aid in risk analysis and habitat plans as other network based approaches (Keitt *et al.*, 1997).

We certainly hope that ecologists find our approach useful and intuitive. The presented framework is quite flexible and can directly incorporate different types of functions to infer spatial proximity (linear versus non-linear functions), different types of indexes to infer community similarity, and different types of algorithms to infer cluster of sites and dispersal directionality. As such, we expect our MSSN method to become a valuable addition to the spatial ecologists' toolbox and find many interesting applications in metacommunity studies, landscape ecology and conservation biology.

CHAPTER IV

A CONNECTIVITY MEASURE FOR METACOMMUNITY NETWORKS

Mehdi Layeghifard, Vladimir Makarenkov and Pedro R. Peres-Neto

4.1 Summary

Connectivity is an important theme in theoretical, empirical and applied studies of heterogeneous landscapes. Graph theory has recently provided a number of promising methodologies to measure landscape connectivity. Graph-theoretical connectivity measures vary in terms of the assumptions they make as well as the ecological issues they are meant to address. Connectivity within metacommunities (among local communities) is one of the important ecological situations that can be modeled by graph theory. Here, we introduce a novel graph-theoretical approach to define and measure the connectivity of a metacommunity in which local communities are interconnected through species dispersal. Our approach uses species composition similarities among local communities to assess the contribution of each local community to the overall connectivity of the metacommunity. Our results showed that our connectivity measure is quite robust in detecting most significant local communities in terms of their contribution to the overall network connectivity within simulated metacommunities. As such, our connectivity measure for metacommunity networks is a valuable addition to the toolbox of the graph-theoretical connectivity measures and could be used alone or in conjunction with other available measures to investigate metacommunities.

4.2 Introduction

Ever since Merriam (1984) introduced the concept of connectivity to landscape ecology, many studies have been carried out to describe and measure individual patch or overall landscape connectivity (Keitt *et al.*, 1997; Moilanen and Hanski, 2001; Schumaker, 1996; Tischendorf and Fahrig, 2000; Tischendorf and Fahrig, 2001). Nevertheless, the complexity of connectivity has prevented these efforts to converge into a common widely accepted definition. For example, depending on context, connectivity could be defined to be functional or structural (Belisle, 2005). Functional connectivity is behaviour related and corresponds to "the degree to which the landscape facilitates or impedes movement among resource patches" (Taylor *et al.*, 1993). Structural connectivity, on the other hand, ignores organisms' behaviour and only considers the physical connectedness of the landscape elements (With *et al.*, 1997). Moreover, the impacts of connectivity vary across different time scales. For example, it could affect the success of juvenile dispersal, migration or species ability to expand in short, intermediate and large time scales, respectively (Minor and Urban, 2007).

In addition to theoretical and empirical research, investigating connectivity in ecological and conservation studies can also help planning mitigation programs to lessen the outcomes of habitat fragmentation and loss. Habitat fragmentation, caused by anthropological disturbances or natural catastrophes, is one of the most significant causes of populations and species extinction (Hanski, 1998). Habitat fragmentation occurs when discontinuities appear within an otherwise homogeneous landscape, thus reducing species mobility, populations' viability and breeding success (Chetkievicz *et al.*, 2006; Malanson, 2002; Nikolakaki, 2004), among other impacts. Connectivity reduces the harsh effects of habitat fragmentation on populations, species and their communities through facilitating species dispersion and gene flow between suitable patches and reducing the probability of species' extinctions at the landscape level (Haddad *et al.*, 2003; McLaughlin *et al.*, 2002).

Graph theory has recently provided a number of very promising methodologies to measure landscape connectivity. Although graph-theoretical approaches were traditionally considered to be highly sophisticated and computationally prohibitive, recent technological advances have converted them into an increasingly efficient and popular toolbox that can be used to find solutions for many scientific and practical issues (Gross and Yellen, 2006). Graph theory has been widely used to study and schematically represent the connections within natural or anthropogenic entities, including both real (e.g., towns and countries interconnected through roads, railways and airways) and virtual entities (e.g., social networks such as Facebook). In an ecological context, one of the simplest examples is viewing landscapes as a network of habitat patches connected by dispersing individuals (Bunn *et al.*, 2000). Graph-theoretical approaches are quite flexible in such a way that their connectivity measures can vary in terms of both the assumptions they make and the specific ecological questions they were meant to address. As a result, different connectivity measures may be more suitable to tackle different specific tasks. With the increase in the number and popularity of graph-theoretic connectivity measures, efforts have been made to compare these measures and study their performance and properties (Pascual-Hortal and Saura, 2006; Saura and Pascual-Hortal, 2007; refer to Laita *et al.*, 2011 for a recent review on the comparison of various graph-theoretical approaches in terms of their conceptual differences).

Connectivity within metacommunities is one of the important ecological characteristics that can be modeled and explained using graph theory. In metapopulation studies, connectivity is measured based on the links among populations inhabiting landscape patches. As an extension, metacommunities may be seen as a collection of interacting metapopulations. Connectivity among local communities is essential for movement of genes, individuals, populations, and species and therefore critical for their stability, integrity and overall maintenance (Clergeau and Burel, 1997; Collinge, 1998; Raison

et al., 2001; Taylor *et al.*, 1993; With *et al.*, 1997). Graph theory has been shown to be very effective in unravelling the complexities surrounding interactions within metapopulations and metacommunities (Jordan *et al.*, 2003; Urban and Keitt, 2001). It has also provided valuable tools for addressing conservation issues and devising land management and conservation planning policies (e.g., Andersson and Bodin, 2009; Ferrari *et al.*, 2007; Minor and Urban, 2008; Pascual-Hortal and Saura, 2008; Urban and Keitt, 2001).

In this paper, we introduce a novel graph-theoretical approach to measure landscape connectivity. In our method, species composition similarities among local communities are used to assess the overall connectivity of the landscape as well as the contribution of each local community to the overall connectivity of the metacommunity. We start by describing some of fundamental attributes of networks and provide a detailed presentation of our new method, followed by a simulation to demonstrate the performance of the approach.

4.3 Methods

4.3.1 Background

Network. In graph theory, a network is represented as a graph $G = (V, E)$, where V is the set of vertices (nodes; ecological objects such as local communities) and E is the set of edges (links or interactions between nodes). This network of interconnected objects is depicted as a set of dots (or small circles; to represent vertices) which are connected by straight or curved lines (to represent edges).

Node degree. The degree of node i in graph G refers to the number of other nodes in the same graph which are directly connected to node i (i.e., neighbors of node i) in the same graph.

Degree matrix. A degree matrix D of a graph G is a diagonal $n \times n$ matrix where n is the number of nodes and the diagonal entry d_{ii} is the degree of node i, also indicated as $\deg(v_i)$.

Adjacency matrix. An adjacency matrix A of a finite graph G is an $n \times n$ matrix where n is the number of nodes and the entry a_{ij} is the number of edges from node i to node j. The value of the diagonal entry a_{ii} depends on the number of loops connecting node i to itself. In acyclic graphs this value is zero because such graphs are devoid of loops.

Laplacian matrix. The Laplacian matrix L of a graph G is an $n \times n$ symmetric matrix obtained from subtracting the adjacency matrix A from the degree matrix D. Therefore, L represents an undirected, unweighted graph without loops or multiple edges from one node to another. The Laplacian matrix is also called the admittance matrix or Kirchhoff matrix (Babic *et al.*, 2002; Cvetkovic *et al.*, 1998). Moreover, L is a real, non-negative and semi-definite matrix. Therefore, all the eigenvalues of L are real and non-negative.

Second smallest eigenvalue. While the smallest eigenvalue of the Laplacian matrix L is zero (due to normalization, see below), its 2nd smallest eigenvalue is a non-negative value referred to as the algebraic connectivity of graph G in spectral graph theory (Fiedler, 1973; Fiedler, 1975; Chung, 1997). In other words, for any two graphs G_1 and G_2, if G_1 has fewer links than G_2, then the 2nd smallest eigenvalue or the algebraic connectivity of G_1 is smaller than that of G_2. Therefore, 2nd smallest eigenvalue represents a measure of graph connectivity.

4.3.2 Methodology

Assuming that we have a metacommunity (i.e., a group of local communities in a landscape), the graph-theoretical representation would be a network in which the nodes and edges indicate local communities and dispersal routes, respectively. Figure 4.1a shows a simple diagrammatic example of such metacommunity represented as a network, where numbered circles are local communities and the lines connecting the circles are the representations of species dispersal. The main goal of our novel method is to find the significance of each network node to the metacommunity connectedness using a weight-based approach.

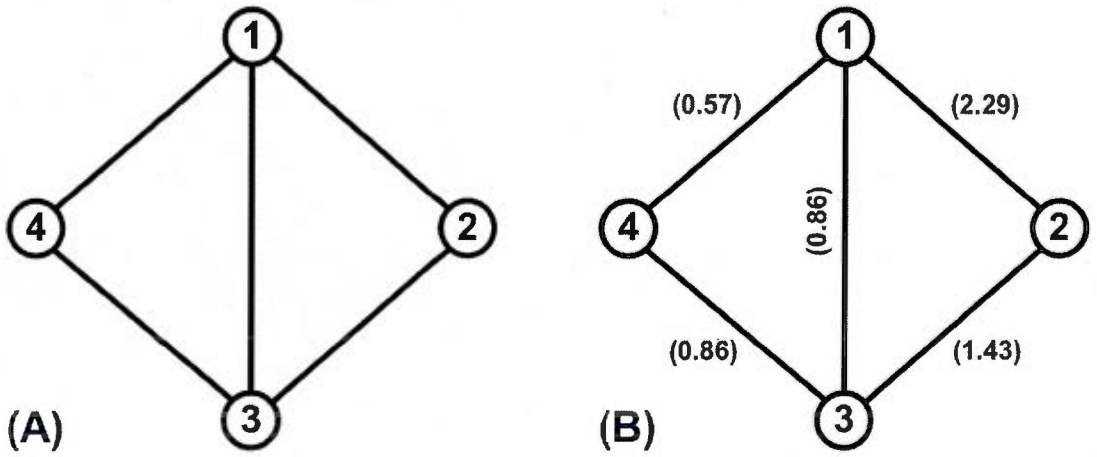


Figure 4.1 A simple example of a metacommunity represented as a network. Numbered circles are local communities and the lines connecting the circles are the representations of species dispersal. (A) A simple metacommunity network. (B) A metacommunity network with species composition similarities shown next to each edge linking every pair of nodes.

Traditionally, the majority of methods developed to investigate network connectivity use unweighted approaches in their computations. In other words, they use a binary system of 1s and 0s to indicate the existence or lack of a connection between every pair of nodes, respectively. Our method, on the other hand, takes advantage of the species composition similarities/dissimilarities to weight the connections between each pair of local communities before calculating the significance of those connections to the network's overall connectivity. If species composition similarity is high, one can assume that there were important dispersal events in the past between these communities. Here, Jaccard's well-known index was used to calculate the dissimilarities in species compositions among local communities, although any other similarity/distance indices could be easily applied.

Our method calculates the pairwise similarities between local communities based on the incidence data (i.e., presence or absence of species within communities) in the first step. These similarities are shown next to the edges connecting the corresponding pairs of nodes in Figure 4.1b. Next, it builds the weighted Laplacian matrix for the metacommunity network and calculates the 2^{nd} smallest eigenvalue of the network. Then, nodes will be removed one at a time the 2^{nd} smallest eigenvalues will be calculated for each of the resulting sub-networks. Finally, the 2^{nd} smallest eigenvalues of sub-networks will be deducted from the 2^{nd} smallest eigenvalue of the initial network and the nodes will be ranked according to the results. Here, in order to make our method easier to understand, we will first describe the unweighted approach of calculating the Laplacian matrix before presenting the weighted approach which is our main objective. A diagrammatic representation of the steps involved in our method is given in Figure 4.2.

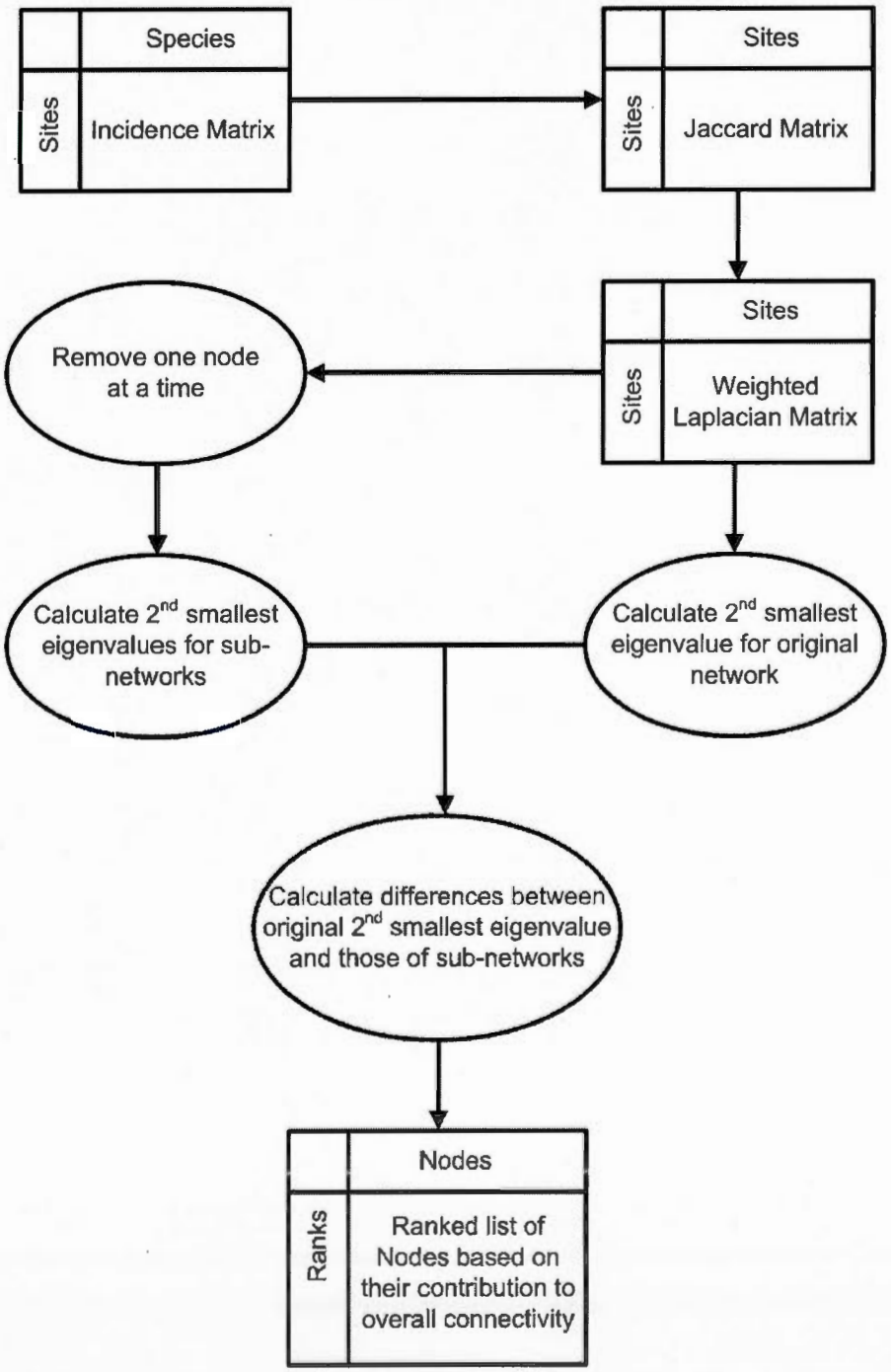


Figure 4.2 A diagrammatic representation of the steps involved in our graph-theoretical connectivity measure methodology. See the text for details.

4.3.2.1 Unweighted approach

For the unweighted network $G = (V, E)$, its Laplacian matrix $L(G) = [l_{i,j}]$ is defined as follows:

$$L(G) = [l_{i,j}] = \begin{cases} \deg(v_i) & if\ i = j \\ -1 & if\ i \neq j\ and\ v_i\ is\ adjacent\ to\ v_j \\ 0 & otherwise \end{cases} \quad (1)$$

where $i,j \in \{1,\ldots,n\}$ are indices of the nodes, v_i and v_j are the ith and jth nodes, respectively and $\deg(v_i)$ is the degree of ith node. Therefore, the resulting Laplacian matrix for the network shown in Figure 1a is the following:

$$L(G) = \begin{bmatrix} 3 & -1 & -1 & -1 \\ -1 & 2 & -1 & 0 \\ -1 & -1 & 3 & -1 \\ -1 & 0 & -1 & 2 \end{bmatrix} \quad (2)$$

The symmetric normalized Laplacian matrix of the same network G, however, is defined as follows:

$$L^{norm}(G) = [l_{i,j}] = \begin{cases} 1 & if\ i = j\ and \deg(v_i) \neq 0 \\ -\frac{1}{\sqrt{\deg(v_i)\deg(v_j)}} & if\ i \neq j\ and\ v_i\ is\ adjacent\ to\ v_j \\ 0 & otherwise \end{cases} \quad (3)$$

Arguably, the most important attribute of the normalized Laplacian matrix is that all its eigenvalues (known as spectra of the normalized Laplacian) are real and non-negative. In fact, if λ is an eigenvalue of L, then $0 = \lambda = 2$. Finally, the normalized Laplacian matrix, $L^{norm}(G)$, calculated for the network shown in Figure 1a is:

$$L^{norm}(G) = \begin{bmatrix} 1 & -0.41 & -0.29 & -0.41 \\ -0.41 & 1 & -0.41 & 0 \\ -0.33 & -0.41 & 1 & -0.41 \\ -0.41 & 0 & -0.41 & 1 \end{bmatrix} \quad (4)$$

4.3.2.2 Weighted approach

For any weighted network $G_w = (V, E)$, its Laplacian matrix $L(G_w) = [l_{i,j}]$ is generally defined as follows:

$$L^{weighted}(G) = [l_{i,j}] = \begin{cases} deg(v_i) & if\ i = j\ and\ \deg(v_i) \neq 0 \\ -W(i,j) & if\ i \neq j\ and\ v_i\ is\ adjacent\ to\ v_j \\ 0 & otherwise \end{cases} \quad (5)$$

where W is the weight of the edge connecting the two nodes v_i and v_j. In our approach, however, instead of adding up the number of edges which coincide on any particular node to calculate the degree of that node, we add up the weight of those coinciding edges to calculate the weight for that node. This weighting approach is very similar to the approaches used in designing or analyzing networks with non-uniform traffic, such as computer networks (Liu *et al.*, 2009). Since the species dispersal is also non-uniform (i.e., the rate of dispersal varies across landscape), this weighting approach is more meaningful than simply counting the number of links. Therefore, our weighted Laplacian matrix is defined as follows:

$$L^{weighted}(G) = [l_{i,j}] = \begin{cases} W(v_i) & if\ i = j\ and\ \deg(v_i) \neq 0 \\ -W(i,j) & if\ i \neq j\ and\ v_i\ is\ adjacent\ to\ v_j \\ 0 & otherwise \end{cases} \quad (6)$$

where $W(v_i)$ is the weight of the node v_i. Then, the weighted Laplacian matrix calculated for the network shown in Figure 1b is:

$$L^{weighted}(G) = \begin{bmatrix} 3.72 & -0.27 & -0.29 & -0.43 \\ -0.27 & 3.72 & -0.29 & 0 \\ -0.33 & -0.29 & 3.15 & -0.47 \\ -0.43 & 0 & -0.47 & 1.43 \end{bmatrix} \quad (7)$$

4.3.2.3 Assessing node contribution to the network connectivity

As described above, since our weighted Laplacian matrix is symmetric and normalized, all the eigenvalues are real and non-negative. It is also well-known that the 2nd smallest eigenvalue (the smallest eigenvalue is 0) is referred to as the algebraic connectivity of the graph. To assess the significance of each node for the overall connectedness of the network, we first calculate the 2nd smallest eigenvalue for the network as a whole. Next, we remove the nodes one at a time and recalculate the 2nd smallest eigenvalue for each of the resulting sub-networks. Finally, by subtracting the 2nd smallest eigenvalues obtained for each sub-network from the 2nd smallest eigenvalue of the full network, we are able to calculate the level of the contribution of each node to the overall connectivity of the network.

4.3.3 Assessing the performance of the metric

We used four different types of networks (i.e., random, regular, exponential and scale-free networks) to simulate the metacommunities required to assess the performance of our metric. Random network model which was first developed by Erdős and Rényi (thus the so-called " Erdős-Rényi model"; Erdős and Rényi, 1959) consists of N_V

vertices, connected by N_E (undirected) edges that are chosen randomly from the set of $N_V(N_V - 1)/2$ possible edges (excluding multiple connections and loops). A regular network is a graph where each vertex has the same number of neighbors (i.e., same degree of connectivity). For example, in a regular network of degree 3 each node has three neighbors. An exponential network is a graph whose degree distribution follows an exponential function. Finally, a scale-free network is a graph whose degree distribution follows a power law function. In other words, when building the scale-free network the new edges are preferentially added to vertices with higher degrees. Therefore, a fraction of vertices in a scale-free network will have very large degrees compared to other vertices.

Using each of the above-mentioned network types we first constructed networks with 100 nodes and then used them as blueprints along with a colonization-extinction model to simulate metacommunities consisting of 100 local communities each (collectively containing 50 species). For the simulation process we used a colonization probability of 0.1 and extinction rates of 0.1, 0.2, 0.3, 0.4, 0.5 and 0.6. For each possible combination of network types, colonization probability and extinction rates we simulated 1000 metacommunities or 24,000 metacommunities in total.

After applying our weighted metric method on every one of the simulated metacommunities, we calculated the correlation coefficients (Pearson coefficients) between our estimates (levels of contribution to overall connectivity calculated for each node using dissimilarities between species composition of local communities) and closeness centrality estimates (based solely on the topology of constructed networks without considering metacommunity data). Since metacommunities were simulated based on the topology of the constructed networks of different types, the correlation between our metric estimates for network nodes and a node centrality measure such closeness centrality would be a reasonable indicator of our methods efficiency in finding nodes contributions to the overall connectivity of a metacommunity. In

connected graphs, closeness of a node is defined as the inverse of the sum of its distances (shortest paths) to all other nodes. Therefore, closeness centrality is used in network analysis as a measure of the relative importance of a node within a graph.

4.4 Results

The correlation coefficients between estimates obtained from our weighted connectivity metric and closeness centrality measures for different types of networks are presented in Figures 4.3 to 4.6. Histograms in Figure 4.3 show the correlation coefficients between our connectivity estimates for all metacommunity nodes and the closeness centrality measures of the underlying regular networks. Each histogram summarizes 1000 metacommunities simulated for a specific extinction rate. Similarly, Figure 4.4 to 4.6 present same measurements for random, exponential and scale-free network types, respectively. As mentioned above, all the nodes in regular networks have the same degree of connectivity. Therefore, all the nodes have almost same amount of contribution to the network overall connectivity. This homogeneity in degree distribution among the nodes is efficiently detected by our method as shown in Figure 4.3, where correlation coefficients are closely centered around zero in all histograms. In random networks, on the other hand, there exists some level of variation in degree distribution, but this random variation is not enough to structuralise networks and metacommunities. This is clearly attested by our method as shown in Figure 4.4, where correlation coefficients centered around -1 or -2 in the histograms.

In Figure 4.5 the correlation results for exponential network are presented. In this type of network the degree distribution follows an exponential function. Therefore, networks and the metacommunities simulated based on them are structured. In other words, some nodes in the network have larger contribution to the overall connectivity than others and, in this case, our weighted connectivity metric method was highly

efficient in detecting these nodes as shown in Figure 4.5. Similar to exponential networks, the metacommunities simulated based on scale-free networks are also

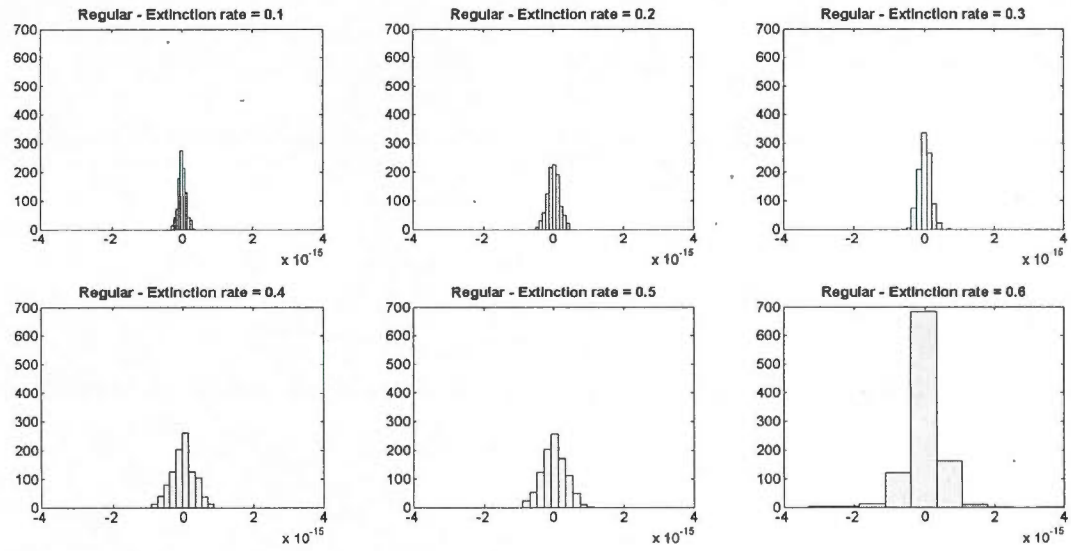


Figure 4.3 Correlation results between our algorithm estimates and closeness centrality measures of nodes relative importance within metacommunities simulated based on regular networks. 1000 simulated metacommunities were analyzed per each extinction rate. Correlation coefficients shown here are between -4×10^{-15} and 4×10^{-15}.

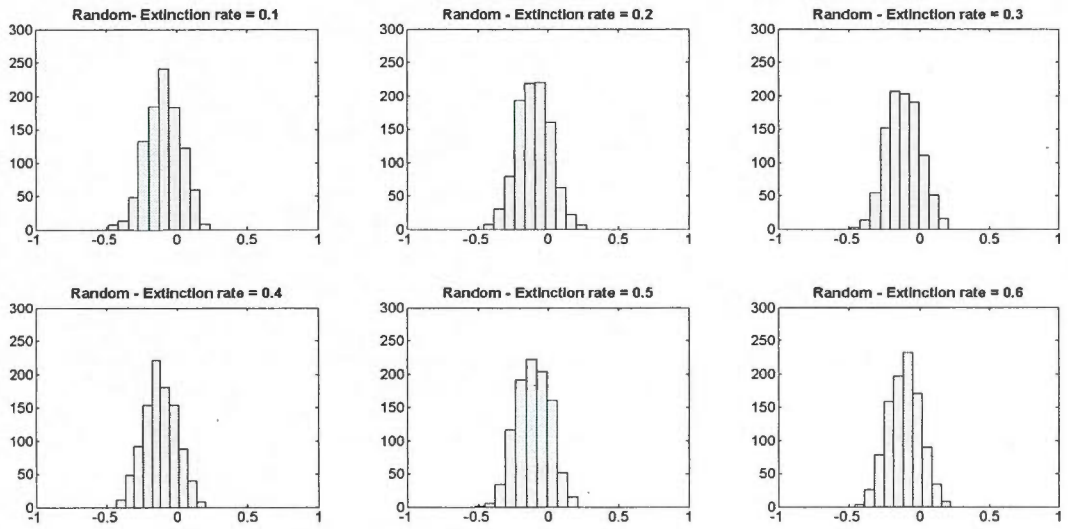


Figure 4.4 Correlation results between our algorithm estimates and closeness centrality measures of nodes relative importance within metacommunities simulated based on random networks. 1000 simulated metacommunities were analyzed per each extinction rate.

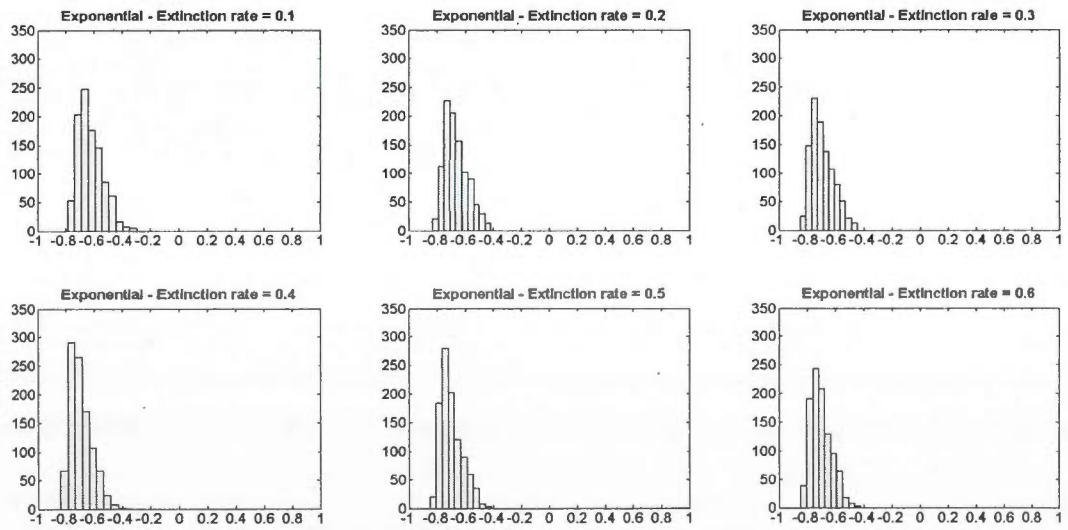


Figure 4.5 Correlation results between our algorithm estimates and closeness centrality measures of nodes relative importance within metacommunities simulated based on exponential networks. 1000 simulated metacommunities were analyzed per each extinction rate.

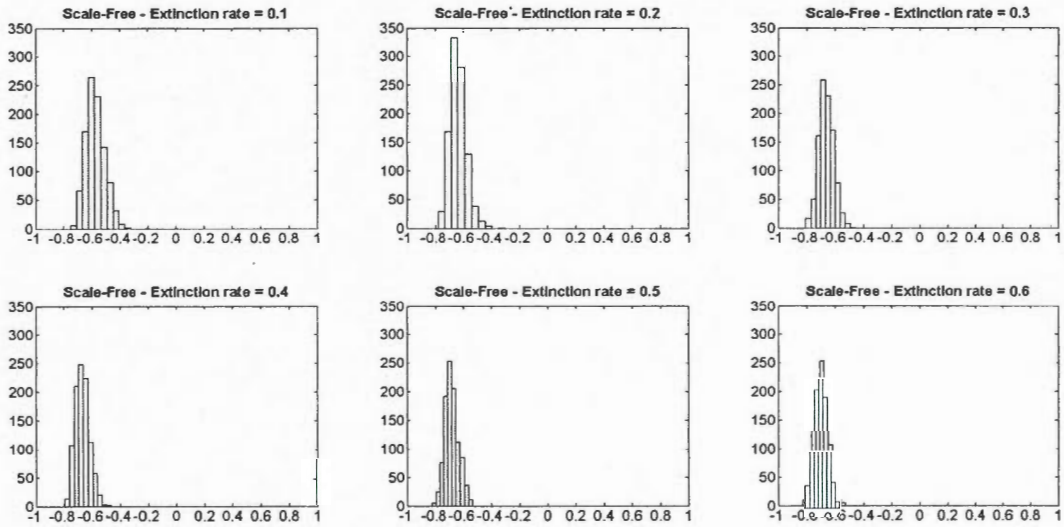


Figure 4.6 Correlation results between our algorithm estimates and closeness centrality measures of nodes relative importance within metacommunities simulated based on scale-free networks. 1000 simulated metacommunities were analyzed per each extinction rate.

structured, although the degree distribution follows a different function (i.e., a power law function). The correlation results for scale-free metacommunities are presented in Figure 4.6. Again, our weighted connectivity metric method was highly successful in finding the nodes with high levels of contribution to the overall network connectivity. As shown in Figures 4.5 and 4.6, in both exponential and scale-free cases, the vast majority of correlation coefficients are between -7 and -8. This shows that our metric is highly efficient in detecting the nodes that are most vital for the overall connectivity of metacommunity networks. It must be noted that the correlation sign is negative because we used Jaccard's distance index. As we mentioned above, any similarity/dissimilarity measure could be used instead of the Jaccard's distance. Using a similarity measure would therefore result in positive signs.

Furthermore, we also tested our metric finding using Spearman and weighted Pearson correlation methods which yielded in similar results (data not shown). In summary, the results from regular and random network simulations show the robustness of our method the results obtained from exponential and scale-free networks demonstrate the high efficiency of our method in detecting the nodes (local communities) highly important for the overall connectivity of metacommunities.

4.5 Discussion

Graph theory has been widely used to represent landscapes as networks where habitat patches are depicted as nodes and connections between patches are depicted as links connecting the corresponding nodes (Bunn *et al.*, 2000; Fall *et al.*, 2007; Urban and Keitt, 2001; Urban *et al.*, 2009). Using graphs to represent landscapes has resulted in the development of numerous connectivity measures as well as the adoption of many connectivity measures developed for purposes other than landscape ecology (Laita *et al.*, 2011).

Graph-theoretical connectivity measures have been successfully used in landscape ecology to design reserve networks (Fuller *et al.*, 2006), to conserve endangered species (Fall *et al.*, 2007) and to find most valuable habitat patches and their roles in landscapes (Jordan *et al.*, 2003; Minor and Urban, 2007; Opdam *et al.*, 2006; Pascual-Hortal and Saura, 2006; Pascual-Hortal and Saura, 2008; Rothley and Rae, 2005). Similarly, these measures have the potential to explain the complex processes occurring within metacommunities including dynamics of species dispersal and the significance of the individual local communities and their corresponding connections to the maintenance of the metacommunity.

It must be noted, however, that connectivity measures developed based on graph theory do not behave in a similar manner. So, they must not be treated equally or expected to result in similar outcomes. Different connectivity measures were developed with different underlying assumptions in order to address different issues. In addition, various connectivity measures take advantage of different attributes of graphs which might be more suitable for a variety of specific purposes. For example, Expected Cluster Size (O'Brien *et al.*, 2006) represents an area-weighted mean habitat size which carries information on the amount of habitat within a landscape component, but its value increases with the loss of small isolated patches or components, although the total habitat area in the landscape reduces. On the other hand, Landscape Coincidence Probability (Pascual-Hortal and Saura, 2006) which is the probability that two points located randomly within a landscape reside in the same habitat component shows a decrease in its value with increasing fragmentation. Contrary to the two graph-theoretical measures described above, graph diameter (Bunn *et al.*, 2000; Ferrari *et al.*, 2007) is purely a topological measure which is sensitive to the number of patches. Therefore, the value of graph diameter increases due to fragmentation, because usually habitat fragmentation leads to a higher number of patches. Graph diameter is in fact the longest path between any two nodes in the graph, where the path length between those nodes is itself the shortest possible length.

One of the most appealing characteristics of graph theory to ecologists is that it can be used efficiently on small data sets and further refinements can be made if more data is available in the future (Urban and Keitt, 2001; Urban *et al.*, 2009). In other words, our method allows us to start with a small (e.g., incomplete) dataset and fine-tune the outcomes while we gather more data. Particularly in the case of large landscapes, graph-theoretical approaches are suggested to show the greatest benefit-to-effort ratio for conservation purposes, because they are quite capable of providing detailed results even from modest datasets (Calabrese and Fagan, 2004).

As is often the case with novel methods, our connectivity measure for metacommunity networks should be regarded as a valuable addition to the toolbox of the graph-theoretical connectivity measures. Our method could be used alone or in conjunction with other available metrics (or even alternative approaches to graph theory) in order to investigate highly complex ecological systems, such as metacommunities, and expand our knowledge of processes that shape and govern them.

One of the main advantages of using graph theory to develop connectivity measures is that it can consider both structural and functional aspects of connectivity at the same time (i.e., the topology of network represent the structure and the weighted links represent the functions; Urban and Keitt, 2001). Here, we adopted and modified the application of graph theory in landscape ecology in order to investigate the functional (i.e., species similarities) connectivity of metacommunities. As a result, instead of habitat patches, here we considered a set of local communities interconnected by dispersal. In real situations, when two local communities show similar species compositions, the main conclusion one can draw is that these two local communities are (were) connected to each other through important dispersal routes or corridors.

Here we only used the functional aspect of connectivity in which we used dissimilarities among local communities to find the local communities contributing the most to the overall connectivity of a metacommunity. Our results clearly showed the efficiency of our metric in detecting highly connected local communities in structured metacommunities (e.g., exponential and scale-free networks) as well as its robustness in not erroneously finding such local communities in unstructured metacommunities (e.g., regular and random networks). However, our measure can be expanded in different ways. We could easily add structural factors such as geographic distribution of local communities representing different dispersal functions and other functional aspects such as species' dispersal range and behaviour-related attributes to movement, among others.

CONCLUSIONS

The primary goal of this thesis was to explore the power and flexibility of graph (network) theory in finding solutions for complex biological problems, particularly in the fields of ecology and evolutionary biology. In doing so, each chapter focused on one particular ecological or evolutionary biological problem and the methods developed in this thesis were assessed for efficiency and robustness using empirical and/or simulated datasets. The main logic behind choosing four different problems to be addressed by graph theory was to demonstrate the inherent capacities of networks in dealing with complex biological issues regardless of their scope and field.

The primary result of Chapter I was a novel method to reconstruct weighted explicit consensus network from a collection of species and gene trees. Given the broad occurrence of heterogeneity among genes and the high number of phylogenetic mechanisms influencing their evolution, having a method to resolve the incongruence among gene trees was the main objective of this chapter. In addition, this powerful and flexible network reconstruction method allows one to infer, visualize and statistically validate major conflicting signals induced by various mechanisms of reticulate evolution. Moreover, the inferred conflicting signals could be presented by means of explicit and easy-to-interpret phylogenetic networks.

The main conclusion of Chapter II was that graph-theoretical approaches (i.e., network-like structures) were shown to be more advantageous than tree-like structures in investigating and reconstructing dispersal histories due to their capabilities in providing much greater and more detailed information about the biogeographic history of dispersals. This was consistent with results obtained from the application of networks in evolutionary biology, where networks easily outperform phylogenetic trees in providing detailed information about various

evolutionary mechanisms. This study could 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.

In Chapter III a new multi-species spatial network method for modelling the spatial heterogeneity of metacommunities was developed. Results from both simulated and real data analyses showed that this method was more robust in terms of explaining variation in community analysis models than the predominant model being used today. Moreover, this newly developed framework is useful in assessing the levels of spatial connectivity for each local community within a metacommunity. Finally, this spatial network framework is highly flexible and can incorporate different types of functions to infer spatial variation and different types of algorithms to infer migration levels and dispersal directionality.

The final chapter, Chapter IV, resulted in the development of a new graph-theoretical connectivity measure for metacommunities that can be easily generalized to any other network system. This new connectivity measure was capable of successfully detecting the most important local communities (in terms of connectivity) within a metacommunity using species composition similarity/dissimilarity. These local communities are essential for the survival of species through dispersal and subsequent colonization of habitat patches across heterogeneous landscapes. Moreover, the extra information gained through the application of this connectivity measure could play an important role in designing conservation plans for metacommunities.

In conclusion, this thesis showed that network-based approaches can provide a way to describe complex biological systems such as metacommunities. They can also improve our understanding of many biological systems as diverse as conflicting

evolutionary histories, biogeography and community dynamics. The hope is that this thesis and similar works will pave the way for further advances in biological networks so that every scientist can have access to an efficient, fast and easy-to-use toolbox of network-based methods.

ALGORITHM I

Inference of hybridization events (Diploid or Polyploid hybridization) – recombination at the chromosome level

Input: *set of unrooted gene trees τ defined on the same set of taxa X*
Output: *explicit weighted consensus network N_h on X representing diploid or polyploid hybridization events*

begin

define p – cut-off level
define C - set of all clusters of τ and C_w - weighted set of clusters of τ
define C_b - set of clusters (splits or bipartitions) of backbone tree T_b

for each T of τ
infer all clusters of T
add clusters to the set C

for each cluster c of C
compute weight $W(c)$ of c using Equations 1, 2 or 3
add c to C_w

sort C_w according to the weight magnitude

while (there exist clusters in C_w compatible with all clusters in C_b) **do**
consider cluster c from C_w with the highest weight $W(c)$
if (($W(c) \geq p$)) && (c is compatible with all clusters in C_b)) **then**
add c to C_b
eliminate c from C_w

$N_h = T_b$ // network is first defined as backbone tree with the cluster set C_b
among remaining clusters in C_w, identify clusters with the 1st degree of incompatibility with N_h

while (there exists a cluster c from C_w such that:
(($W(c) \geq p$)) && (c has the 1st degree of incompatibility with N_h)) **do**
find cluster c from C_w with the highest weight $W(c)$ such that:
(($W(c) \geq p$)) && (c has the 1st degree of incompatibility with N_h))
d = *find_direction* (reticulation branch r_c, N_h, τ)
add r_c, representing cluster c, to N_h with direction d and weight $W(c)$
eliminate c from C_w
among remaining clusters in C_w, identify clusters with the 1st degree of incompatibility with N_h
transform N_h into an *explicit weighted hybridization network* (see Figure 3)

end

Function *find_direction* (reticulation branch r_c, N_h, τ)

begin

define T_{d1} - tree obtained from N_h and induced by reticulation branch r_c with direction d_1 (a directed reticulation branch corresponds to an SPR move)
define T_{d2} - tree obtained from N_h and induced by reticulation branch r_c with direction d_2 (opposite to d_1)

if (N_h contains some other directed reticulation branches, apart from r_c) **then**
obtain T_{d1} and T_{d2} by carrying out SPR moves corresponding to these reticulation branches

if $(\sum(W(T_i) / RF(T_i,T_{d1})) < \sum(W(T_i) / RF(T_i,T_{d2})))$ **then**
return d_2
else
return d_1

// here RF denotes the Robinson and Foulds distance and $W(T_i)$ is the weight of tree T_i
// the sums are taken over all trees in τ that include cluster c

end *find_direction*

ALGORITHM II

Inference of intragenic recombination or partial horizontal gene transfer events followed by intragenic recombination – two or more genes recombine to create a mosaic gene

Input: *unrooted species phylogenetic tree T_s and multiple sequence alignment MSA (or only multiple sequence alignment MSA) defined on the same set of taxa X*
Output: *explicit weighted consensus network N_r on X representing recombination or partial HGT events*

begin
define p - cut-off level
define $C(T)$ - set of clusters of tree T and $C_w(T)$ - weighted set of clusters of T
define SW - set of MSA fragments examined by sliding window procedure

if T_s is not given **then**
infer weight-based consensus T_s from *MSA* (e.g., using PhyML, RaxML or BIONJ)

for each MSA fragment, MSA_f, from SW
infer a phylogenetic tree T from MSA_f
compute bootstrap scores of internal branches of T
infer $C(T)$, set of all clusters of T

for each cluster c of $C(T)$
compute weight $W(c)$ of c using Equation 1 (based on bootstrap scores)
add c to $C_w(T)$

sort $C_w(T)$ according to the weight magnitude
$N_r(SW) = T_s$ *// network is first defined as species tree*
among remaining clusters in $C_w(T)$, identify clusters with the 1st degree of incompatibility with $N_r(SW)$

while (there exists a cluster c from $C_w(T)$ such that:
$((W(c) \geq p))$ && (c has the 1st degree of incompatibility with $N_r(SW)$)) **do**
find cluster c from $C_w(T)$ with the highest weight $W(c)$ such that:
$((W(c) \geq p))$ && (c has the 1st degree of incompatibility with $N_r(SW)$))
d = *find_direction* (reticulation branch r_c, $N_r(SW)$, T)
add r_c, representing cluster c, to $N_r(SW)$ with direction d and weight $W(c)$
eliminate c from $C_w(T)$
among remaining clusters in $C_w(T)$, identify clusters with the 1st degree of incompatibility with $N_r(SW)$

if (recombination network $N_r(SW)$ obtained for the MSA fragment MSA_f is identical to that obtained for the previous interval MSA_{f-1}) **then**
merge MSA_{f-1} and MSA_f as intervals providing the identical solutions
if (recombination is studied and both parents of recombinant species are identified) **then**
transform N_h into an *explicit weighted hybridization network* (see Figure 3)

end

ALGORITHM III

Inference of horizontal gene transfer events (the case of a complete gene transfers when the whole gene is transferred from donor to host; Input data: Species tree + gene tree or Species tree + MSA)

Input: *unrooted species phylogenetic tree T_s and unrooted gene phylogenetic tree T_g (or multiple gene sequence alignment MSA) defined on the same set of taxa X*
Output: *explicit weighted consensus horizontal gene transfer network N_{hgt} on X*

begin

define p - cut-off level
define $C(T_s)$ - set of clusters of tree T_s and $C_w(T_g)$ - weighted set of clusters of tree T_g

if (T_g is not given) **then**
infer weight-based consensus T_g from *MSA* (e.g., using PhyML, RaxML or BIONJ)
compute bootstrap scores of internal branches (i.e. clusters) of T_g

for each cluster c of T_g
compute weight $W(c)$ of c using Equation 1 (based on bootstrap scores)
add c to $C_w(T_g)$

sort $C_w(T_g)$ according to the weight magnitude
$N_{hgt} = T_s$ *// network is first defined as backbone tree*
among remaining clusters in $C_w(T_g)$, identify clusters with the 1st degree of incompatibility with N_{hgt}

while (there exists a cluster c from $C_w(T_g)$ such that:
$((W(c) \geq p))$ && (c has the 1st degree of incompatibility with N_{hgt})) **do**
find cluster c from $C_w(T_g)$ with the highest weight $W(c)$ such that:
$((W(c) \geq p))$ && (c has the 1st degree of incompatibility with N_{hgt}))
$d = find\_direction$ (reticulation branch r_c, N_{hgt}, T_g)
add r_c, representing cluster c, to N_{hgt} with direction d and weight $W(c)$
eliminate c from $C_w(T_g)$
among remaining clusters in $C_w(T_g)$, identify clusters with the 1st degree of incompatibility with N_{hgt}

end

APPENDIX A

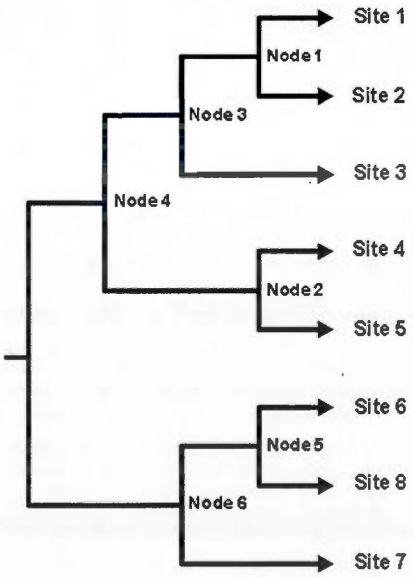
CALCULATION OF WEIGHTS FOR THE EDGES OF SPATIAL NETWORK

In order to calculate the weights for all edges of the spatial network, we need A) the pair-wise Euclidean distances between local communities (sites), B) the spatial tree topology, and C) the spatial network:

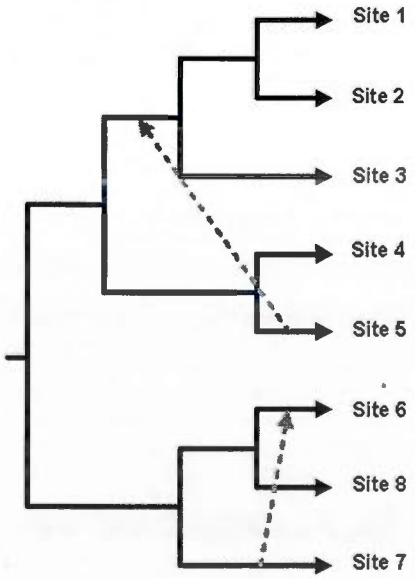
A) Euclidean geographic matrix

	Site 1	Site 2	Site 3	Site 4	Site 5	Site 6	Site 7	Site 8
Site 1	0	0.12	0.15	0.56	0.61	0.89	0.99	1.00
Site 2	0.12	0	0.14	0.47	0.52	0.77	0.88	0.88
Site 3	0.15	0.14	0	0.42	0.47	0.85	0.96	0.95
Site 4	0.56	0.47	0.42	0	0.07	0.67	0.80	0.73
Site 5	0.61	0.52	0.47	0.07	0	0.63	0.75	0.67
Site 6	0.89	0.77	0.85	0.67	0.63	0	0.12	0.12
Site 7	0.99	0.88	0.96	0.80	0.75	0.12	0	0.13
Site 8	1.00	0.88	0.95	0.73	0.67	0.12	0.13	0

B) The backbone spatial tree.



C) The final spatial network



In order to calculate the edge weights, we need to know the distances and similarities between all the sites and nodes, because edges are the connections between those sites and nodes. Here, we only show an example of the Euclidean distance matrix obtained from geographic coordinates. The same approach (detailed calculations now shown here) is used for the community similarity matrix (Jaccard).

Calculation of the distance between Node 1 and the other communities:

d[Site 3, Node 1] = d[Site 3, (Site 1, Site 2)]

= 1/2 * [d(Site 3, Site 1) + d(Site 3, Site 2) - d(Site 1, Site 2)]

= 1/2 * [0.15 + 0.14 – 0.12]

= 0.09

Distances between Node 1 and the other nodes (sites) are calculated in the same way.

	Site 1	Site 2	Site 3	Site 4	Site 5	Site 6	Site 7	Site 8	Node 1
Site 1	0	0.12	0.15	0.56	0.61	0.89	0.99	1.00	0
Site 2	0.12	0	0.14	0.47	0.52	0.77	0.88	0.88	0
Site 3	0.15	0.14	0	0.42	0.47	0.85	0.96	0.95	0.09
Site 4	0.56	0.47	0.42	0	0.07	0.67	0.80	0.73	0.45
Site 5	0.61	0.52	0.47	0.07	0	0.63	0.75	0.67	0.50
Site 6	0.89	0.77	0.85	0.67	0.63	0	0.12	0.12	0.77
Site 7	0.99	0.88	0.96	0.80	0.75	0.12	0	0.13	0.88
Site 8	1.00	0.88	0.95	0.73	0.67	0.12	0.13	0	0.88
Node 1	0	0	0.09	0.45	0.50	0.77	0.88	0.88	0

Calculation of the distance between Node 2 and the other communities:

d[Site 3, Node 2] = d[Site 2, (Site 4, Site 5)]

= 1/2 * [d(Site 3, Site 4) + d(Site 3, Site 5) - d(Site 4, Site 5)]

= 1/2 * [0.42 + 0.47 - 0.07]

= 0.41

Distance calculations between Node 2 and the other nodes (sites) follow the same rule. For the edge between Node 1 and Node 2, we proceed as follows:

d[Node 1, Node 2] = d[Node 1, (Site 4, Site 5)] or d[(Site 1, Site2), Node 2]

= 1/2 * [d(Node 1, Site 4) + d(Node 1, Site 5) - d(Site 4, Site 5)]

= 1/2 * [0.45 + 0.5 - 0.07]

= 0.44

	Site 1	Site 2	Site 3	Site 4	Site 5	Site 6	Site 7	Ste 8	Node 1	Node 2
Site 1	0	0.12	0.15	0.56	0.61	0.89	0.99	1.00	0	0.55
Site 2	0.12	0	0.14	0.47	0.52	0.77	0.88	0.88	0	0.46
Site 3	0.15	0.14	0	0.42	0.47	0.85	0.96	0.95	0.09	0.41
Site 4	0.56	0.47	0.42	0	0.07	0.67	0.80	0.73	0.45	0
Site 5	0.61	0.52	0.47	0.07	0	0.63	0.75	0.67	0.50	0
Site 6	0.89	0.77	0.85	0.67	0.63	0	0.12	0.12	0.77	0.62
Site 7	0.99	0.88	0.96	0.80	0.75	0.12	0	0.13	0.88	0.74
Site 8	1.00	0.88	0.95	0.73	0.67	0.12	0.13	0	0.88	0.67
Node 1	0	0	0.09	0.45	0.50	0.77	0.88	0.88	0	0.44
Node 2	0.55	0.46	0.41	0	0	0.62	0.74	0.67	0.44	0

Calculation of the distance between Node 3 and other communities:

d[Site 4, Node 3] = d[Site 4, (Site 3, Node 1)]

= 1/2 * [d(Site 4, Site 3) + d(Site 4, Node 1) - d(Site 3, Node 1)]

= 1/2 * [0.42 + 0.45 – 0.09]

= 0.39

Distance calculations between Node 3 and the other nodes (sites) follow the same rule. For the possible edge between Node 3 and Node 2, we proceed as follows:

d[Node 3, Node 2] = d[Node 3, (Site 4, Site 5)] or d[(Node 1, Site2), Node 2]

= 1/2 * [d(Node 3, Site 4) + d(Node 3, Site 5) - d(Site 4, Site 5)]

= 1/2 * [0.39 + 0.45 – 0.07]

= 0.38

	Site 1	Site 2	Site 3	Site 4	Site 5	Site 6	Site 7	Site 8	Node 1	Node 2	Node 3
Site 1	0	0.12	0.15	0.56	0.61	0.89	0.99	1.00	0	0.55	0
Site 2	0.12	0	0.14	0.47	0.52	0.77	0.88	0.88	0	0.46	0
Site 3	0.15	0.14	0	0.42	0.47	0.85	0.96	0.95	0.09	0.41	0
Site 4	0.56	0.47	0.42	0	0.07	0.67	0.80	0.73	0.45	0	0.39
Site 5	0.61	0.52	0.47	0.07	0	0.63	0.75	0.67	0.50	0	0.45
Site 6	0.89	0.77	0.85	0.67	0.63	0	0.12	0.12	0.77	0.62	0.77
Site 7	0.99	0.88	0.96	0.80	0.75	0.12	0	0.13	0.88	0.74	0.88
Site 8	1.00	0.88	0.95	0.73	0.67	0.12	0.13	0	0.88	0.67	0.87
Node 1	0	0	0.09	0.45	0.50	0.77	0.88	0.88	0	0.44	0
Node 2	0.55	0.46	0.41	0	0	0.62	0.74	0.67	0.44	0	0.38
Node 3	0	0	0	0.39	0.45	0.77	0.88	0.87	0	0.38	0

Calculation of the distance between Node 4 and other communities:

d[Site 6, Node 4] = d[Site 6, (Node 2, Node 3)]

= 1/2 * [d(Site 6, Node 2) + d(Site 6, Node 3) - d(Node 2, Node 3)]

= 1/2 * [0.62 + 0.77 – 0.38]

= 0.50

Distance calculations between Node 4 and the other nodes (sites) follow the same rule.

	Site 1	Site 2	Site 3	Site 4	Site 5	Site 6	Site 7	Site 8	Node 1	Node 2	Node 3	Node 4
Site 1	0	0.12	0.15	0.56	0.61	0.89	0.99	1.00	0	0.55	0	0
Site 2	0.12	0	0.14	0.47	0.52	0.77	0.88	0.88	0	0.46	0	0
Site 3	0.15	0.14	0	0.42	0.47	0.85	0.96	0.95	0.09	0.41	0	0
Site 4	0.56	0.47	0.42	0	0.07	0.67	0.80	0.73	0.45	0	0.39	0
Site 5	0.61	0.52	0.47	0.07	0	0.63	0.75	0.67	0.50	0	0.45	0
Site 6	0.89	0.77	0.85	0.67	0.63	0	0.12	0.12	0.77	0.62	0.77	0.50
Site 7	0.99	0.88	0.96	0.80	0.75	0.12	0	0.13	0.88	0.74	0.88	0.62
Site 8	1.00	0.88	0.95	0.73	0.67	0.12	0.13	0	0.88	0.67	0.87	0.58
Node 1	0	0	0.09	0.45	0.50	0.77	0.88	0.88	0	0.44	0	0
Node 2	0.55	0.46	0.41	0	0	0.62	0.74	0.67	0.44	0	0.38	0
Node 3	0	0	0	0.39	0.45	0.77	0.88	0.87	0	0.38	0	0
Node 4	0	0	0	0	0	0.50	0.62	0.58	0	0	0	0

Calculation of the distance between Node 5 and other communities:

d[Site 7, Node 5] = d[Site 7, (Site 6, Site 8)]

= 1/2 * [d(Site 7, Site 6) + d(Site 7, Site 8) - d(Site 6, Site 8)]

= 1/2 * [0.12 + 0.13 – 0.12]

= 0.07

Distance calculations between Node 5 and the other nodes (sites) follow the same rule. For the possible edges between Node 5 and any of Nodes 1, 2, 3 and 4, we proceed as follows:

d[Node 1, Node 5] => d[Node 1, (Site 6, Site 8)] or d[(Site 1, Site2), Node 5]

= 1/2 * [d(Node 1, Site 6) + d(Node 1, Site8) - d(Site 6, Site 8)]

= 1/2 * [0.77 + 0.88 – 0.12]

= 0.77

d[Node 2, Node 5] => d[Node 2, (Site 6, Site 8)] or d[(Site 4, Site 5), Node 5]

= 1/2 * [d(Node 2, Site 6) + d(Node 2, Site8) - d(Site 6, Site 8)]

= 1/2 * [0.62 + 0.67 – 0.12]

= 0.58

d[Node 3, Node 5] => d[Node 3, (Site 6, Site 8)] or d[(Site 3, Node 1), Node 5]

= 1/2 * [d(Node 3, Site 6) + d(Node 3, Site8) - d(Site 6, Site 8)]

= 1/2 * [0.77 + 0.87 – 0.12]

= 0.76

d[Node 4, Node 5] => d[Node 4, (Site 6, Site 8)] or d[(Node 2, Node 3), Node 5]

= 1/2 * [d(Node 4, Site 6) + d(Node 4, Site8) - d(Site 6, Site 8)]

= 1/2 * [0.5 + 0.58 – 0.12]

= 0.48

	Site 1	Site 2	Site 3	Site 4	Site 5	Site 6	Site 7	Site 8	Node 1	Node 2	Node 3	Node 4	Node 5
Site 1	0	0.12	0.15	0.56	0.61	0.89	0.99	1.00	0	0.55	0	0	0.89
Site 2	0.12	0	0.14	0.47	0.52	0.77	0.88	0.88	0	0.46	0	0	0.77
Site 3	0.15	0.14	0	0.42	0.47	0.85	0.96	0.95	0.09	0.41	0	0	0.84
Site 4	0.56	0.47	0.42	0	0.07	0.67	0.80	0.73	0.45	0	0.39	0	0.64
Site 5	0.61	0.52	0.47	0.07	0	0.63	0.75	0.67	0.50	0	0.45	0	0.59
Site 6	0.89	0.77	0.85	0.67	0.63	0	0.12	0.12	0.77	0.62	0.77	0.50	0
Site 7	0.99	0.88	0.96	0.80	0.75	0.12	0	0.13	0.88	0.74	0.88	0.62	0.07
Site 8	1.00	0.88	0.95	0.73	0.67	0.12	0.13	0	0.88	0.67	0.87	0.58	0
Node 1	0	0	0.09	0.45	0.50	0.77	0.88	0.88	0	0.44	0	0	0.77
Node 2	0.55	0.46	0.41	0	0	0.62	0.74	0.67	0.44	0	0.38	0	0.58
Node 3	0	0	0	0.39	0.45	0.77	0.88	0.87	0	0.38	0	0	0.76
Node 4	0	0	0	0	0	0.50	0.62	0.58	0	0	0	0	0.48
Node 5	0.89	0.77	0.84	0.64	0.59	0	0.07	0	0.77	0.58	0.76	0.48	0

Calculation of the distance between Node 6 and other communities:

d[Site 1, Node 6] = d[Site 1, (Site 7, Node 5)]

= 1/2 * [d(Site 1, Site 7) + d(Site 1, Node 5) - d(Site 7, Node 5)]

= 1/2 * [0.99 + 0.89 – 0.07]

= 0.91

Distance calculations between Node 6 and the other nodes (sites) follow the same rule. For the possible edges between Node 6 and any of Nodes 1, 2, 3 and 4, we follow the same procedure presented for Node 5.

	Site 1	Site 2	Site 3	Site 4	Site 5	Site 6	Site 7	Site 8	Node 1	Node 2	Node 3	Node 4	Node 5	Node 6
Site 1	0	0.12	0.15	0.56	0.61	0.89	0.99	1.00	0	0.55	0	0	0.89	0.91
Site 2	0.12	0	0.14	0.47	0.52	0.77	0.88	0.88	0	0.46	0	0	0.77	0.79
Site 3	0.15	0.14	0	0.42	0.47	0.85	0.96	0.95	0.09	0.41	0	0	0.84	0.87
Site 4	0.56	0.47	0.42	0	0.07	0.67	0.80	0.73	0.45	0	0.39	0	0.64	0.68
Site 5	0.61	0.52	0.47	0.07	0	0.63	0.75	0.67	0.50	0	0.45	0	0.59	0.64
Site 6	0.89	0.77	0.85	0.67	0.63	0	0.12	0.12	0.77	0.62	0.77	0.50	0	0
Site 7	0.99	0.88	0.96	0.80	0.75	0.12	0	0.13	0.88	0.74	0.88	0.62	0.07	0
Site 8	1.00	0.88	0.95	0.73	0.67	0.12	0.13	0	0.88	0.67	0.87	0.58	0	0
Node 1	0	0	0.09	0.45	0.50	0.77	0.88	0.88	0	0.44	0	0	0.77	0.79
Node 2	0.55	0.46	0.41	0	0	0.62	0.74	0.67	0.44	0	0.38	0	0.58	0.63
Node 3	0	0	0	0.39	0.45	0.77	0.88	0.87	0	0.38	0	0	0.76	0.79
Node 4	0	0	0	0	0	0.50	0.62	0.58	0	0	0	0	0.48	0.51
Node 5	0.89	0.77	0.84	0.64	0.59	0	0.07	0	0.77	0.58	0.76	0.48	0	0
Node 6	0.91	0.79	0.87	0.68	0.64	0	0	0	0.79	0.63	0.79	0.51	0	0

Finally, in order to obtain the weight for edges between an internal node (e.g., Node 1) and its successor nodes (e.g., Site 1 and Site 2) we divide the distance between the successors by 2 and assign the result to each of the edges. For example:

d(Site 1,Site 2) = 0.12 => W(Node 1-Site 1) = W(Node 1-Site 2) = 0.06

	Site 1	Site 2	Site 3	Site 4	Site 5	Site 6	Site 7	Site 8	Node 1	Node 2	Node 3	Node 4	Node 5	Node 6
Site 1	0	0.12	0.15	0.56	0.61	0.89	0.99	1.00	0.06	0.55	0.10	0.30	0.89	0.91
Site 2	0.12	0	0.14	0.47	0.52	0.77	0.88	0.88	0.06	0.46	0.10	0.30	0.77	0.79
Site 3	0.15	0.14	0	0.42	0.47	0.85	0.96	0.95	0.09	0.41	0.04	0.24	0.84	0.87
Site 4	0.56	0.47	0.42	0	0.07	0.67	0.80	0.73	0.45	0.03	0.39	0.23	0.64	0.68
Site 5	0.61	0.52	0.47	0.07	0	0.63	0.75	0.67	0.50	0.03	0.45	0.23	0.59	0.64
Site 6	0.89	0.77	0.85	0.67	0.63	0	0.12	0.12	0.77	0.62	0.77	0.50	0.06	0.09
Site 7	0.99	0.88	0.96	0.80	0.75	0.12	0	0.13	0.88	0.74	0.88	0.62	0.07	0.03
Site 8	1.00	0.88	0.95	0.73	0.67	0.12	0.13	0	0.88	0.67	0.87	0.58	0.06	0.09
Node 1	0.06	0.06	0.09	0.45	0.50	0.77	0.88	0.88	0	0.44	0.04	0.24	0.77	0.79
Node 2	0.55	0.46	0.41	0.03	0.03	0.62	0.74	0.67	0.44	0	0.38	0.19	0.58	0.63
Node 3	0.10	0.10	0.04	0.39	0.45	0.77	0.88	0.87	0.04	0.38	0	0.19	0.76	0.79
Node 4	0.30	0.30	0.24	0.23	0.23	0.50	0.62	0.58	0.24	0.19	0.19	0	0.48	0.51
Node 5	0.89	0.77	0.84	0.64	0.59	0.06	0.07	0.06	0.77	0.58	0.76	0.48	0	0.03
Node 6	0.91	0.79	0.87	0.68	0.64	0.09	0.03	0.09	0.79	0.63	0.79	0.51	0.03	0

Site by Edge matrix indicating the edges between each local community and the Root:

	N1-S1	N1-S2	N2-S4	N2-S5	N3-N1	N3-S3	N4-N2	N4-N3	N5-S6	N5-S8	N6-S7	N6-N5	R-N4	R-N6
Site 1	1	0	0	0	1	0	0	1	0	0	0	0	1	0
Site 2	0	1	0	0	1	0	0	1	0	0	0	0	1	0
Site 3	0	0	0	0	0	1	0	1	0	0	0	0	1	0
Site 4	0	0	1	0	0	0	1	0	0	0	0	0	1	0
Site 5	0	0	0	1	0	0	1	0	0	0	0	0	1	0
Site 6	0	0	0	0	0	0	0	0	1	0	0	1	0	1
Site 7	0	0	0	0	0	0	0	0	0	1	0	1	0	1
Site 8	0	0	0	0	0	0	0	0	0	0	1	0	0	1

Site by Edge matrix including the extra branches:

	N1-S1	N1-S2	N2-S4	N2-S5	N3-N1	N3-S3	N4-N2	N4-N3	N5-S6	N5-S8	N6-S7	N6-N5	R-N4	R-N6	S5-N3	S7-S6
Site 1	1	0	0	1	1	0	1	1	0	0	0	0	1	0	1	0
Site 2	0	1	0	1	1	0	1	1	0	1	0	1	1	1	1	0
Site 3	0	0	0	1	0	1	1	1	0	0	0	0	1	0	1	0
Site 4	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0
Site 5	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0
Site 6	0	0	0	0	0	0	0	0	1	0	1	1	0	1	0	1
Site 7	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0
Site 8	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0

Vector of distance-based weights for all edges for our spatial network:

The values are obtained from the final site (node) by site (node) distance matrix. For instance, the value for column N1-S1, which is the distance-based weight for the edge between Node 1 and Site 1, is the distance between Node 1 and Site 1 in the final distance matrix.

	N1-S1	N1-S2	N2-S4	N2-S5	N3-N1	N3-S3	N4-N2	N4-N3	N5-S6	N5-S8	N6-S7	N6-N5	R-N4	R-N6	S5-N3	S7-S6
WD	0.06	0.06	0.03	0.03	0.04	0.04	0.19	0.19	0.06	0.06	0.03	0.03	0.26	0.26	0.47	0.12

Vector of similarity-based weights for all edges for our spatial network:

Same approach described above in detail could be used to calculate a weight vector based on the Jaccard similarity instead of Euclidean distance.

	N1-S1	N1-S2	N2-S4	N2-S5	N3-N1	N3-S3	N4-N2	N4-N3	N5-S6	N5-S8	N6-S7	N6-N5	R-N4	R-N6	S5-N3	S7-S6
WD	0.93	0.93	0.92	0.92	0.88	0.88	0.94	0.94	0.75	0.75	0.93	0.93	0.87	0.87	0.81	0.72

Finally, the overall weights will be computed using the following equation applied on the two weight vectors above:

$$W_i = (1 - W(D)_i) * W(J)_i$$

Where W_i is the final weight of edge i, $W(D)_j$ is the distance-based weight of edge i and $W(J)_i$ is the Jaccard-based weight of edge i. Therefore, the final weigh vector is:

	N1-S1	N1-S2	N2-S4	N2-S5	N3-N1	N3-S3	N4-N2	N4-N3	N5-S6	N5-S8	N6-S7	N6-N5	R-N4	R-N6	S5-N3	S7-S6
WD	0.87	0.87	0.88	0.88	0.84	0.84	0.76	0.76	0.7	0.7	0.9	0.9	0.65	0.65	0.42	0.63

Weighted Site by Edge matrix which is the product of the Site by Edge matrix and the final weight vector.

	N1-S1	N1-S2	N2-S4	N2-S5	N3-N1	N3-S3	N4-N2	N4-N3	N5-S6	N5-S8	N6-S7	N6-N5	R-N4	R-N6	S5-N3	S7-S6
Site 1	0.87	0	0	0.88	0.84	0	0.76	0.76	0	0	0	0	0.65	0	0.42	0
Site 2	0	0.87	0	0.88	0.84	0	0.76	0.76	0	0.7	0	0.9	0.65	0.65	0.42	0
Site 3	0	0	0	0.88	0	0.84	0.76	0.76	0	0	0	0	0.65	0	0.42	0
Site 4	0	0	0.88	0	0	0	0.76	0	0	0	0	0	0.65	0	0	0
Site 5	0	0	0	0.88	0	0	0.76	0	0	0	0	0	0.65	0	0	0
Site 6	0	0	0	0	0	0	0	0	0.7	0	0.9	0.9	0	0.65	0	0.63
Site 7	0	0	0	0	0	0	0	0	0	0.7	0	0.9	0	0.65	0	0
Site 8	0	0	0	0	0	0	0	0	0	0	0.9	0	0	0.65	0	0

The **Euclidean distance matrix**, calculated for the weighted Site by Edge matrix, to be used in RDA analysis.

	Site 1	Site 2	Site 3	Site 4	Site 5	Site 6	Site 7	Site 8
Site 1	0	1.80	1.47	1.94	1.49	2.63	2.39	2.29
Site 2	1.80	0	1.97	2.34	1.98	2.49	2.00	2.47
Site 3	1.47	1.97	0	1.73	1.21	2.48	2.23	2.11
Site 4	1.94	2.34	1.73	0	1.24	2.17	1.87	1.73
Site 5	1.49	1.98	1.21	1.24	0	2.17	1.87	1.73
Site 6	2.63	2.49	2.48	2.17	2.17	0	1.48	1.30
Site 7	2.39	2.00	2.23	1.87	1.87	1.48	0	1.45
Site 8	2.29	2.47	2.11	1.73	1.73	1.30	1.45	0

APPENDIX B

F INDING THE DISPERSAL DIRECTION OF NEWLY ADDED EDGES.

The dispersal directionality of newly added edges was determined by minimizing the topological differences computed by the Robinson and Foulds method. For example, assuming that T_0 in Figure B.1 is the backbone spatial tree and e is the newly found edge between Sites 3 and 4, then, T_1 will be the backbone tree with the new edge, e, added to Site 3 to represent direction from Site 3 to Site 4. Conversely, T_2 will be the backbone tree with e added to Site 4 to represent the reverse direction (From Site 4 to Site 3). Then, the Robinson-Foulds topological distance between T_0 and each of T_1 and T_2 (RF_1 and RF_2, respectively) will be computed, separately. Finally, the smaller distance will determine the direction of the newly found link. For instance, if RF_2 is smaller than RF_1, the direction will be from Site 4 to Site 3 in the final network. In other words, minimizing the topological distance proves that the new link is more connected to Site 4 than to Site 3 showing that this migratory route is more probably originated from Site 4. Eventually, as shown in Figure B.1, after adding the newly-found significant link to the backbone spatial tree, the final multi-species spatial network (MSSN) is built.

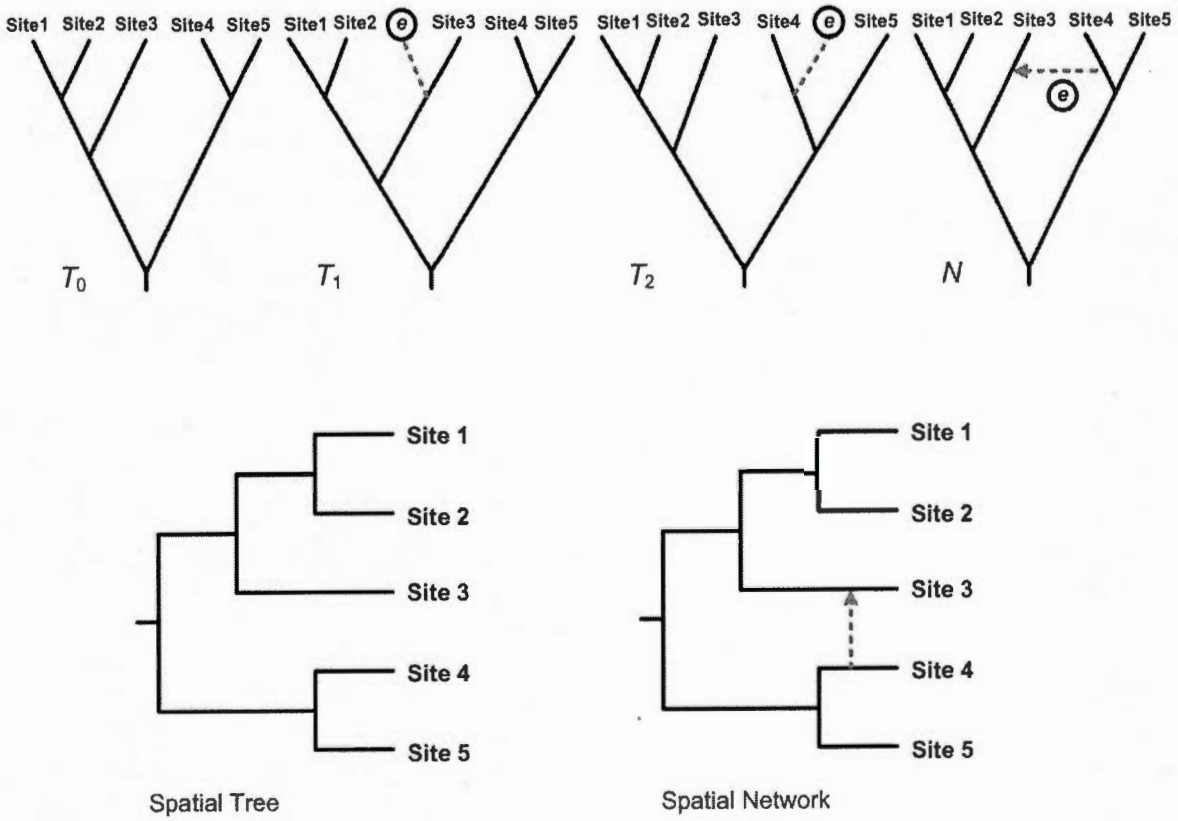


Figure B.1 Finding the dispersal direction of newly added edges

REFERENCES

Abello, J., Pardalos, P. M., et Resende, M. G. C. (1999). On Maximum Clique Problems In Very Large Graphs. In *In External Memory Algorithms* (pp. 119–130). American Mathematical Society.

Anderson, F. E. (2000). Phylogeny and historical biogeography of the loliginid squids (Mollusca: cephalopoda) based on mitochondrial DNA sequence data. *Molecular phylogenetics and evolution*, *15*(2), 191–214.

Andersson, E., et Bodin, Ö. (2009). Practical tool for landscape planning? An empirical investigation of network based models of habitat fragmentation. *Ecography*, *32*(1), 123–132.

Babić, D., Klein, D. J., Lukovits, I., Nikolić, S., et Trinajstić, N. (2002). Resistance-distance matrix: a computational algorithm and its application. *International Journal of Quantum Chemistry*, *90*(1), 166–176.

Baldauf, S. L., Roger, A. J., Wenk-Siefert, I., et Doolittle, W. F. (2000). A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science*, *290*(5493), 972–977.

Banašek-Richter, C., Bersier, L.-F., Cattin, M.-F., Baltensperger, R., Gabriel, J.-P., Merz, Y., … Naisbit, R. E. (2009). Complexity in quantitative food webs. *Ecology*, *90*(6), 1470–1477.

Bandelt, H. J., Forster, P., et Röhl, A. (1999). Median-joining networks for inferring intraspecific phylogenies. *Molecular biology and evolution*, *16*(1), 37–48.

Barabási, A.-L., et Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509–512.

Beisner, B. E., Peres-Neto, P. R., Lindström, E. S., Barnett, A., et Longhi, M. L. (2006). The role of environmental and spatial processes in structuring lake communities from bacteria to fish. *Ecology*, *87*(12), 2985–2991.

Bélisle, M. (2005). Measuring landscape connectivity: the challenge of behavioral landscape ecology. *Ecology*, *86*(8), 1988–1995.

Bender, D. J., Tischendorf, L., et Fahrig, L. (2003). Using patch isolation metrics to predict animal movement in binary landscapes. *Landscape Ecology*, *18*(1), 17–39.

Benton, T. G., Vickery, J. A., et Wilson, J. D. (2003). Farmland biodiversity: is habitat heterogeneity the key? *Trends in Ecology et Evolution*, *18*(4), 182–188.

Blanchet, F. G., Legendre, P., et Borcard, D. (2008). Modelling directional spatial processes in ecological data. *Ecological Modelling*, *215*(4), 325–336.

Boc, A., Diallo, A. B., et Makarenkov, V. (2012). T-REX: a web server for inferring, validating and visualizing phylogenetic trees and networks. *Nucleic acids research*, *40*(Web Server issue), W573–579.

Boc, A., Legendre, P., et Makarenkov, V. (2013). An efficient algorithm for the detection and classification of horizontal gene transfer events and identification of mosaic genes. In B. Lausen, D. V. den Poel, et A. Ultsch (Eds.), *Algorithms from and for Nature and Life* (pp. 253–260). Springer International Publishing.

Boc, A., et Makarenkov, V. (2011). Towards an accurate identification of mosaic genes and partial horizontal gene transfers. *Nucleic acids research*, *39*(21), e144.

Boc, A., Philippe, H., et Makarenkov, V. (2010). Inferring and validating horizontal gene transfer events using bipartition dissimilarity. *Systematic biology*, *59*(2), 195–211.

Bodin, O., et Norberg, J. (2007). A network approach for analyzing spatially structured populations in a fragmented landscape. *Landscape Ecology*, *22*(1), 31–44.

Borcard, D., Legendre, P., et Drapeau, P. (1992). Partialling out the spatial component of ecological variation. *Ecology*, *73*(3), 1045–1055.

Brooks, D. R. (1990). Parsimony analysis in historical biogeography and coevolution: methodological and theoretical update. *Systematic Zoology*, *39*(1), 14–30.

Bryant, D. (2003). A classification of consensus methods for phylogenies. Bioconsensus: DIMACS Working Group Meetings on Bioconsensus.

Buneman, P. (1971). The recovery of trees from measures of dissimilarity. In D. Kendall et P. Tautù (Eds.), *Mathematics the the Archeological and Historical Sciences* (pp. 387–395). Edinburgh University Press.

Bunn, A. G., Urban, D. L., et Keitt, T. H. (2000). Landscape connectivity: a conservation application of graph theory. *Journal of Environmental Management*, *59*(4), 265–278.

Burbrink, F. T., et Pyron, R. A. (2011). The impact of gene-tree/species-tree discordance on diversification-rate estimation. *Evolution; international journal of organic evolution*, *65*(7), 1851–1861.

Calabrese, J. M., et Fagan, W. F. (2004). A comparison-shopper's guide to connectivity metrics. *Frontiers in Ecology and the Environment*, *2*(10), 529–536.

Carstens, B. C., et Knowles, L. L. (2007). Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: an example from Melanoplus grasshoppers. *Systematic biology*, *56*(3), 400–411.

Castillo-Ramírez, S., et González, V. (2008). Factors affecting the concordance between orthologous gene trees and species tree in bacteria. *BMC Evolutionary Biology*, *8*(1), 1–12.

Cavalli-Sforza, L. L., et Edwards, A. W. F. (1967). Phylogenetic analysis: models and estimation procedures. *Evolution*, *21*(3), 550–570.

Cavender-Bares, J., Kozak, K. H., Fine, P. V. A., et Kembel, S. W. (2009). The merging of community ecology and phylogenetic biology. *Ecology Letters*, *12*(7), 693–715.

Chen, F. C., et Li, W. H. (2001). Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *American journal of human genetics*, *68*(2), 444–456.

Chesson, P. (1991). Stochastic population models. In J. Kolasa et S. T. A. Pickett (Eds.), *Ecological Heterogeneity* (pp. 123–143). Springer New York.

Chetkiewicz, C.-L. B., St. Clair, C. C., et Boyce, M. S. (2006). Corridors for conservation: integrating pattern and process. *Annual Review of Ecology, Evolution, and Systematics*, *37*(1), 317–342.

Chu, C., Minns, C. K., et Mandrak, N. E. (2003). Comparative regional assessment of factors impacting freshwater fish biodiversity in Canada. *Canadian Journal of Fisheries and Aquatic Sciences*, *60*(5), 624–634.

Chung, F. (1997). *Spectral graph theory*. Amer Mathematical Society.

Clergeau, P., et Burel, F. (1997). The role of spatio-temporal patch connectivity at the landscape level: an example in a bird distribution. *Landscape and Urban Planning*, *38*(1–2), 37–43.

Collinge, S. K. (1998). Spatial arrangement of habitat patches and corridors: clues from ecological field experiments. *Landscape and Urban Planning*, *42*(2–4), 157–168.

Cvetkovic, D. M., Doob, M., et Sachs, H. (1999). *Spectra of graphs: theory and applications*. Wiley.

Diniz-Filho, J. A. F., Terribile, L. C., da Cruz, M. J. R., et Vieira, L. C. G. (2010). Hidden patterns of phylogenetic non-stationarity overwhelm comparative analyses of niche conservatism and divergence. *Global Ecology and Biogeography*, *19*(6), 916–926.

Doerr, D., Gronau, I., Moran, S., et Yavneh, I. (2012). Stochastic errors vs. modeling errors in distance based phylogenetic reconstructions. *Algorithms for Molecular Biology : AMB*, *7*, 22.

Dray, S., Pélissier, R., Couteron, P., Fortin, M.-J., Legendre, P., Peres-Neto, P. R., … Wagner, H. H. (2012). Community ecology in the age of multivariate multiscale spatial analysis. *Ecological Monographs*, *82*(3), 257–275.

Dray, Stéphane. (2011). A new perspective about moran's coefficient: spatial autocorrelation as a linear regression problem. *Geographical Analysis*, *43*(2), 127–141.

Dray, Stephane, Legendre, P., et Peres-Neto, P. (2006). Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbor matrices (PCNM). *Ecological Modelling*, *196*(3-4), 483–493.

Dunham, J. B., et Rieman, B. E. (1999). Metapopulation structure of bull trout: influences of physical, biotic, and geometrical landscape characteristics. *Ecological Applications*, *9*(2), 642–655.

Dunne, J. A., Williams, R. J., et Martinez, N. D. (2002). Network structure and biodiversity loss in food webs: robustness increases with connectance. *Ecology Letters*, *5*(4), 558–567.

Ebersberger, I., Galgoczy, P., Taudien, S., Taenzer, S., Platzer, M., et von Haeseler, A. (2007). Mapping human genetic ancestry. *Molecular biology and evolution*, *24*(10), 2266–2276.

Elton, C. (1946). Competition and the structure of ecological communities. *Journal of Animal Ecology*, *15*(1), 54–68.

Erdos, P. et Rényi, A. (1959). On random graphs I. *Publicationes Mathematicae*, *6*, 290–297.

Esselstyn, J. A., Oliveros, C. H., Moyle, R. G., Peterson, A. T., McGuire, J. A., et Brown, R. M. (2010). Integrating phylogenetic and taxonomic evidence illuminates complex biogeographic patterns along Huxley's modification of Wallace's Line. *Journal of Biogeography*, *37*(11), 2054–2066.

Eubank, S., Guclu, H., Anil Kumar, V. S., Marathe, M. V., Srinivasan, A., Toroczkai, Z., et Wang, N. (2004). Modelling disease outbreaks in realistic urban social networks. *Nature*, *429*(6988), 180–184.

Fahrig, L., et Merriam, G. (1994). Conservation of fragmented populations. *Conservation Biology*, *8*(1), 50–59.

Fall, A., Fortin, M.-J., Manseau, M., et O'Brien, D. (2007). Spatial Graphs: Principles and Applications for Habitat Connectivity. *Ecosystems*, *10*(3), 448–461.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of molecular evolution*, *17*(6), 368–376.

Felsenstein, J. (1989). PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics*, *5*, 164–166.

Felsenstein, J. (2005). *PHYLIP (Phylogeny Inference Package) version 3.6*. Department of Genome Sciences, University of Washington: Disributed by Author.

Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, *27*(4), 401–410.

Felsenstein, Joseph. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, *39*, 783–791.

Felsenstein, Joseph. (2004). *Inferring phylogenies*. Sinauer Associates, Incorporated.

Ferrari, J. R., Lookingbill, T. R., et Neel, M. C. (2007). Two measures of landscape-graph connectivity: assessment across gradients in area and configuration. *Landscape Ecology*, *22*(9), 1315–1323.

Fiedler, M. (1973). Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, *23*, 298–305.

Fiedler, M. (1975). A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory.

Fitch, W. M. (1971). Toward defining the course of evolution: minimum change for a species tree topology. *Systematic Biology*, *20*(4), 406–416.

Foltête, J.-C., Clauzel, C., Vuidel, G., et Tournant, P. (2012). Integrating graph-based connectivity metrics into species distribution models. *Landscape Ecology*, *27*(4), 557–569.

Gascuel, O. (1997). BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular biology and evolution*, *14*(7), 685–695.

Gilarranz, L. J., Bascompte, J. (2012). Spatial network structure and metapopulation persistence. *Journal of Theoretical Biology*, *297*, 11–16.

Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., … Rothberg, J. M. (2003). A Protein Interaction Map of Drosophila melanogaster. *Science*, *302*(5651), 1727–1736.

Giribet, G., Edgecombe, G. D., et Wheeler, W. C. (2001). Arthropod phylogeny based on eight molecular loci and morphology. *Nature*, *413*(6852), 157–161.

Graham, C. H., Ron, S. R., Santos, J. C., Schneider, C. J., et Moritz, C. (2004). Integrating phylogenetics and environmental niche models to explore speciation mechanisms in dendrobatid frogs. *Evolution; international journal of organic evolution*, *58*(8), 1781–1793.

Graybeal, A. (1998). Is it better to add taxa or characters to a difficult phylogenetic problem? *Systematic biology*, *47*(1), 9–17.

Grechko, V. V. (2013). The problems of molecular phylogenetics with the example of squamate reptiles: Mitochondrial DNA markers. *Molecular Biology*, *47*(1), 55–74.

Griffith, D. A., et Peres-Neto, P. R. (2006). Spatial modeling in ecology: The flexibility of eigenfunction spatial analyses. *Ecology*, *87*(10), 2603–2613.

Gross, J. L., et Yellen, J. (2006). *Graph theory and its applications*. Chapman et Hall/CRC.

Gucht, K. V. der, Cottenie, K., Muylaert, K., Vloemans, N., Cousin, S., Declerck, S., … Meester, L. D. (2007). The power of species sorting: Local factors drive bacterial community composition over a wide range of spatial scales. *Proceedings of the National Academy of Sciences*, *104*(51), 20404–20409.

Guénoche, A. (2013). Multiple consensus trees: a method to separate divergent genes. *BMC bioinformatics*, *14*, 46.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., et Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic biology*, *59*(3), 307–321.

Guindon, S., et Gascuel, O. (2002). Efficient biased estimation of evolutionary distances when substitution rates vary across sites. *Molecular biology and evolution*, *19*(4), 534–543.

Haddad, N. M., Bowne, D. R., Cunningham, A., Danielson, B. J., Levey, D. J., Sargent, S., et Spira, T. (2003). Corridor use by diverse taxa. *Ecology*, *84*(3), 609–615.

Hall, B. G. (2005). Comparison of the accuracies of several phylogenetic methods using protein and DNA sequences. *Molecular biology and evolution*, *22*(3), 792–802.

Hallett, M. T., et Lagergren, J. (2001). Efficient algorithms for lateral gene transfer problems. In *Proceedings of the fifth annual international conference on Computational biology* (pp. 149–156). New York, NY, USA: ACM.

Hanski, I. (1994). A practical model of metapopulation dynamics. *Journal of Animal Ecology*, *63*(1), 151–162.

Hanski, I. (1998). Metapopulation dynamics. *Nature*, *396*(6706), 41–49.

Hanski, I., et Ovaskainen, O. (2000). The metapopulation capacity of a fragmented landscape. *Nature*, *404*(6779), 755.

Hartel, T., Nemes, S., Öllerer, K., Cogalniceanu, D., Moga, C., et Arntzen, J. W. (2010). Using connectivity metrics and niche modeling to explore the occurrence of the Northern Crested Newt (Amphibia, Caudata) in a traditionally managed landscape. *Environmental Conservation*, *37*, 195–200.

Harvey, P. H., et Pagel, M. D. (1991). *The comparative method in evolutionary biology*. Oxford; New York: Oxford University Press.

Helmus, M. R., Savage, K., Diebel, M. W., Maxted, J. T., et Ives, A. R. (2007). Separating the determinants of phylogenetic community structure. *Ecology letters*, *10*(10), 917–925.

Henriques-Silva, R., Lindo, Z., et Peres-Neto, P. R. (2012). A community of metacommunities: exploring patterns in species distributions across large geographical areas. *Ecology*, *94*(3), 627–639.

Holland, B., et Moulton, V. (2003). Consensus networks: a method for visualising incompatibilities in collections of trees. In G. Benson et R. D. M. Page (Eds.), *Algorithms in Bioinformatics* (pp. 165–176).

Holland, B. R., Huber, K. T., Moulton, V., et Lockhart, P. J. (2004). Using consensus networks to visualize contradictory evidence for species phylogeny. *Molecular biology and evolution*, *21*(7), 1459–1461.

Holland, B. R., Jermiin, L. S., Moulton, V., et SMBE Tri-National Young Investigators. (2006). Proceedings of the SMBE Tri-National Young

Investigators' Workshop 2005. Improved consensus network techniques for genome-scale phylogeny. *Molecular biology and evolution*, *23*(5), 848–855.

Holyoak, M., Leibold, M. A., et Holt, R. D. (2005). *Metacommunities: spatial dynamics and ecological communities*. Chicago; London: University of Chicago.

Hubert, N., Hanner, R., Holm, E., Mandrak, N. E., Taylor, E., Burridge, M., ... Bernatchez, L. (2008). Identifying Canadian freshwater fishes through DNA barcodes. *PloS one*, *3*(6), e2490.

Huelsenbeck, J. P. (1995). Performance of phylogenetic methods in simulation. *Systematic Biology*, *44*(1), 17–48.

Huson, D H. (1998). SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics (Oxford, England)*, *14*(1), 68–73.

Huson, Daniel H, Rupp, R., et Scornavacca, C. (2010). *Phylogenetic networks: concepts, algorithms and applications*. Cambridge, UK; New York: Cambridge University Press.

Huson, Daniel H, et Scornavacca, C. (2012). Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Systematic biology*, *61*(6), 1061–1067.

Huson, Daniel H., et Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, *23*(2), 254–267.

Huson, Daniel H., et Rupp, R. (2008). Summarizing multiple gene trees using cluster networks. In K. A. Crandall et J. Lagergren (Eds.), *Algorithms in Bioinformatics* (pp. 296–305). Springer Berlin Heidelberg.

Huston, M. A. (1999). Local processes and regional patterns: appropriate scales for understanding variation in the diversity of plants and animals. *Oikos*, *86*(3), 393–401.

Hwang, U. W., Friedrich, M., Tautz, D., Park, C. J., et Kim, W. (2001). Mitochondrial protein phylogeny joins myriapods with chelicerates. *Nature*, *413*(6852), 154–157.

Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y., et Barkai, N. (2002). Revealing modular organization in the yeast transcriptional network. *Nature genetics*, *31*(4), 370–377.

Jackson, D. A., et Harvey, H. H. (1989). Biogeographic associations in fish assemblages: local vs. regional processes. *Ecology*, *70*(5), 1472–1484.

Jacobson, B., et Peres-Neto, P. R. (2010). Quantifying and disentangling dispersal in metacommunities: how close have we come? How far is there to go? *Landscape Ecology*, *25*(4), 495–507.

Jennings, W. B., et Edwards, S. V. (2005). Speciational history of Australian grass finches (Poephila) inferred from thirty gene trees. *Evolution; international journal of organic evolution*, *59*(9), 2033–2047.

Jeong, H., Oltvai, Z. N., et Barabetaacute;si, A.-L. (2003). Prediction of protein essentiality based on genomic data. *Complexus*, *1*(1), 19–28.

Jordán, F., Báldi, A., Orci, K.-M., Rácz, I., et Varga, Z. (2003). Characterizing the importance of habitat patches and corridors in maintaining the landscape

connectivity of a Pholidoptera transsylvanica (Orthoptera) metapopulation. *Landscape Ecology*, *18*(1), 83–92.

Jordán, Ferenc, Magura, T., Tóthmérész, B., Vasas, V., et Ködöböcz, V. (2007). Carabids (Coleoptera: Carabidae) in a forest patchwork: a connectivity analysis of the Bereg Plain landscape graph. *Landscape Ecology*, *22*(10), 1527–1539.

Karp, P. D., Ouzounis, C. A., Moore-Kochlacs, C., Goldovsky, L., Kaipa, P., Ahrén, D., ... López-Bigas, N. (2005). Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Research*, *33*(19), 6083–6089.

Keitt, T., Urban, D. L., et Milne, B. T. (1997). Detecting critical scales in fragmented landscapes. *Conservation Ecology*, *1*(1), 4.

Kolaczkowski, B., et Thornton, J. W. (2004). Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature*, *431*(7011), 980–984.

Krause, A. E., Frank, K. A., Mason, D. M., Ulanowicz, R. E., et Taylor, W. W. (2003). Compartments revealed in food-web structure. *Nature*, *426*(6964), 282.

Krause, J., Lusseau, D., et James, R. (2009). Animal social networks: an introduction. *Behavioral Ecology and Sociobiology*, *63*(7), 967–973.

Kubatko, L. S., et Degnan, J. H. (2007). Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic biology*, *56*(1), 17–24.

Kuhner, M. K., et Felsenstein, J. (1994). A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular biology and evolution*, *11*(3), 459–468.

Laita, A., Kotiaho, J. S., et Mönkkönen, M. (2011). Graph-theoretic connectivity measures: what do they tell us about connectivity? *Landscape Ecology*, *26*(7), 951–967.

Layeghifard, M., Peres-Neto, P. R., et Makarenkov, V. (2012). Using directed phylogenetic networks to retrace species dispersal history. *Molecular phylogenetics and evolution*, *64*(1), 190–197.

Legendre, P, et Legendre, V. (1984). Postglacial dispersal of freshwater fishes in the Québec peninsula. *Canadian Journal of Fisheries and Aquatic Sciences*, *41*, 1781–1802.

Legendre, P, et Makarenkov, V. (2002). Reconstruction of biogeographic and evolutionary networks using reticulograms. *Systematic biology*, *51*(2), 199–216.

Legendre, P., et Legendre, L. (1987). Developments in numerical ecology.

Legendre, Pierre, Borcard, D., et Peres-Neto, P. R. (2005). Analyzing beta diversity: partitioning the spatial variation of community composition data. *Ecological Monographs*, *75*(4), 435–450.

Legendre, Pierre, et Legendre, L. (2012). *Numerical ecology*. Elsevier.

Leibold, M. A., Holyoak, M., Mouquet, N., Amarasekare, P., Chase, J. M., Hoopes, M. F., ... Gonzalez, A. (2004). The metacommunity concept: a framework for multi-scale community ecology. *Ecology Letters*, *7*(7), 601–613.

Leibold, Mathew A. (1998). Similarity and local co-existence of species in regional biotas. *Evolutionary Ecology*, *12*(1), 95–110.

Leibold, Mathew A., Economo, E. P., et Peres-Neto, P. (2010). Metacommunity phylogenetics: separating the roles of environmental filters and historical biogeography. *Ecology Letters*, *13*(10), 1290–1299.

Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., ... Vidal, M. (2004). A map of the interactome network of the metazoan C. elegans. *Science (New York, N.Y.)*, *303*(5657), 540–543.

Lindeman, R. L. (1942). The trophic-dynamic aspect of ecology. *Ecology*, *23*(4), 399–417.

Liu, W., Sirisena, H., et Pawlikowski, K. (2009). Weighted algebraic connectivity metric for non-uniform traffic in reliable network design. In *International Conference on Ultra Modern Telecommunications Workshops, 2009. ICUMT '09* (pp. 1–6). Presented at the International Conference on Ultra Modern Telecommunications Workshops, 2009. ICUMT '09.

Luczkovich, J. J., Borgatti, S. P., Johnson, J. C., et Everett, M. G. (2003). Defining and measuring trophic role similarity in food webs using regular equivalence. *Journal of Theoretical Biology*, *220*(3), 303–321.

MacArthur, R. (1955). Fluctuations of animal populations and a measure of community stability. *Ecology*, *36*(3), 533–536.

MacArthur, R., et Levins, R. (1964). Competition, habitat selection, and character displacement in a patchy environment. *Proceedings of the National Academy of Sciences of the United States of America*, *51*(6), 1207–1210.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations (Vol. 1, pp. 281–297). Presented at the Proc. Fifth Berkeley Symp. on Math. Statist. and Prob., Univ. of Calif. Press.

Magle, S. B., Theobald, D. M., et Crooks, K. R. (2009). A comparison of metrics predicting landscape connectivity for a highly interactive species along an urban gradient in Colorado, USA. *Landscape Ecology*, *24*(2), 267–280.

Makarenkov, V. (2001). T-REX: reconstructing and visualizing phylogenetic trees and reticulation networks. *Bioinformatics*, *17*(7), 664–668.

Makarenkov, V., et Legendre, P. (2004). From a phylogenetic tree to a reticulated network. *Journal of computational biology: a journal of computational molecular cell biology*, *11*(1), 195–212.

Makarenkov, V., Legendre, P., et Desdevises, Y. (2004). Modelling phylogenetic relationships using reticulated networks. *Zoologica Scripta*, *33*(1), 89–96.

Malanson, G. P. (2002). Extinction-debt trajectories and spatial patterns of habitat destruction. *Annals of the Association of American Geographers*, *92*(2), 177–188.

Mandrak, N. E., et Crossman, E. J. (1992). Postglacial dispersal of freshwater fishes into Ontario. *Canadian Journal of Zoology*, *70*(11), 2247–2259.

Margush, T., et McMorris, F. R. (1981). Consensus n-trees. *Bulletin of Mathematical Biology*, *43*(2), 239–244.

Mason-Gamer, R. J., et Kellogg, E. A. (1996). Testing for phylogenetic conflict among molecular data sets in the tribe Triticeae (Gramineae). *Systematic Biology*, *45*(4), 524–545.

Matte-Tailliez, O., Brochier, C., Forterre, P., et Philippe, H. (2002). Archaeal phylogeny based on ribosomal proteins. *Molecular biology and evolution*, *19*(5), 631–639.

McLaughlin, J. F., Hellmann, J. J., Boggs, C. L., et Ehrlich, P. R. (2002). Climate change hastens population extinctions. *Proceedings of the National Academy of Sciences*, *99*(9), 6070–6074.

Meyer, K. M., Ward, D., Wiegand, K., et Moustakas, A. (2008). Multi-proxy evidence for competition between savanna woody species. *Perspectives in Plant Ecology, Evolution and Systematics*, *10*(1), 63–72.

Mihaescu, R., Levy, D., et Pachter, L. (2009). Why neighbor-joining works. *Algorithmica*, *54*(1), 1–24.

Minor, E. S., et Urban, D. L. (2007). Graph theory as a proxy for spatially explicit population models in conservation planning. *Ecological applications: a publication of the Ecological Society of America*, *17*(6), 1771–1782.

Minor, E. S., et Urban, D. L. (2008). A graph-theory framework for evaluating landscape connectivity and conservation planning. *Conservation biology: the journal of the Society for Conservation Biology*, *22*(2), 297–307.

Moilanen, A., et Hanski, I. (2001). On the use of connectivity measures in spatial ecology. *Oikos*, *95*(1), 147–151.

Moreira, D., Le Guyader, H., et Philippe, H. (2000). The origin of red algae and the evolution of chloroplasts. *Nature*, *405*(6782), 69–72.

Mossel, E., et Vigoda, E. (2005). Phylogenetic MCMC algorithms are misleading on mixtures of trees. *Science (New York, N.Y.)*, *309*(5744), 2207–2209.

Nakhleh, L., Ruths, D., et Wang, L.-S. (2005). RIATA-HGT: a fast and accurate heuristic for reconstructing horizontal gene transfer. In L. Wang (Ed.), *Computing and Combinatorics* (pp. 84–93).

Naylor, G. J., et Brown, W. M. (1998). Amphioxus mitochondrial DNA, chordate phylogeny, and the limits of inference based on comparisons of sequences. *Systematic biology*, *47*(1), 61–76.

Nelson, G. (1979). Cladistic analysis and synthesis: principles and definitions, with a historical note on Adanson's Familles des plantes (1763-1764). *Systematic Zoology*, *28*(1), 1–21.

Nikolakaki, P. (2004). A GIS site-selection process for habitat creation: estimating connectivity of habitat patches. *Landscape and Urban Planning*, *68*(1), 77–94.

Odum, H. T. (1956). Primary production in flowing waters. *Limnology and Oceanography*, *1*(2), 102–117.

Olden, J., Jackson, P., et Peres-Neto, P. (2001). Spatial isolation and fish communities in drainage lakes. *Oecologia*, *127*(4), 572–585.

Opdam, P., Steingröver, E., et Rooij, S. van. (2006). Ecological networks: a spatial concept for multi-actor planning of sustainable landscapes. *Landscape and Urban Planning*, *75*(3–4), 322–332.

Ovaskainen, O., et Hanski, I. (2001). Spatially structured metapopulation models: global and local assessment of metapopulation capacity. *Theoretical Population Biology*, *60*(4), 281–302.

Page, R. D. M. (1989). Comments on component-compatibility in historical biogeography. *Cladistics*, *5*(2), 167–182.

Pascual, M., et Dunne, J. A. (2005). *Ecological networks: linking structure to dynamics in food webs: linking structure to dynamics in food webs*. Oxford University Press.

Pascual-Hortal, L., et Saura, S. (2006). Comparison and development of new graph-based landscape connectivity indices: towards the priorization of habitat patches and corridors for conservation. *Landscape Ecology*, *21*(7), 959–967.

Pascual-Hortal, L., et Saura, S. (2008). Integrating landscape connectivity in broad-scale forest planning through a new graph-based habitat availability methodology: application to capercaillie (Tetrao urogallus) in Catalonia (NE Spain). *European Journal of Forest Research*, *127*(1), 23–31.

Peres-Neto, P., et Cumming, G. (2010). A multi-scale framework for the analysis of fish metacommunities (Vol. 73, pp. 000–000). Presented at the American Fisheries Society Symposium.

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*(2), 352–360.

Peres-Neto, P. R., et Legendre, P. (2010). Estimating and controlling for spatial structure in the study of ecological communities. *Global Ecology and Biogeography*, *19*(2), 174–184.

Peres-Neto, P. R., Legendre, P., Dray, S., et Borcard, D. (2006). Variation partitioning of species data matrices: estimation and comparison of fractions. *Ecology*, *87*(10), 2614–2625.

Peres-Neto, P. R., Leibold, M. A., et Dray, S. (2012). Assessing the effects of spatial contingency and environmental filtering on metacommunity phylogenetics. *Ecology*, *93*(sp8), S14–S30.

Pettersson, E., Lundeberg, J., et Ahmadian, A. (2009). Generations of sequencing technologies. *Genomics*, *93*(2), 105–111.

Pierre Legendre, V. M. (2002). Reconstruction of biogeographic and evolutionary networks using reticulograms. *Systematic biology*, *51*(2), 199–216.

Pollard, D. A., Iyer, V. N., Moses, A. M., et Eisen, M. B. (2006). Widespread discordance of gene trees with species tree in Drosophila: evidence for incomplete lineage sorting. *PLoS genetics*, *2*(10), e173.

Posada, et Crandall. (2001). Intraspecific gene genealogies: trees grafting into networks. *Trends in ecology et evolution*, *16*(1), 37–45.

Prugh, L. R. (2009). An evaluation of patch connectivity measures. *Ecological Applications*, *19*(5), 1300–1310.

Raison, R. J., Brown, A. G., et Flinn, D. W. (2001). *Criteria and indicators for sustainable forest management*. CABI.

Rambaut, A., et Grassly, N. C. (1997). Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer applications in the biosciences: CABIOS*, *13*(3), 235–238.

Rannala, B., et Yang, Z. (1996). Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *Journal of molecular evolution*, *43*(3), 304–311.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., et Barabási, A.-L. (2002). Hierarchical organization of modularity in metabolic networks. *Science*, *297*(5586), 1551–1555.

Robinson, D. F., et Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, *53*(1–2), 131–147.

Rokas, A., King, N., Finnerty, J., et Carroll, S. B. (2003). Conflicting phylogenetic signals at the base of the metazoan tree. *Evolution et development*, *5*(4), 346–359.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., ... Huelsenbeck, J. P. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic biology*, *61*(3), 539–542.

Rothley, K. D., et Rae, C. (2005). Working backwards to move forwards: graph-based connectivity metrics for reserve network selection. *Environmental Modeling et Assessment*, *10*(2), 107–113.

Rubio, L., et Saura, S. (2012). Assessing the importance of individual habitat patches as irreplaceable connecting elements: An analysis of simulated and real landscape data. *Ecological Complexity*, *11*(0), 28–37.

Ryder, T. B., McDonald, D. B., Blake, J. G., Parker, P. G., et Loiselle, B. A. (2008). Social networks in the lek-mating wire-tailed manakin (Pipra filicauda). *Proceedings of the Royal Society B: Biological Sciences*, *275*(1641), 1367–1374.

Saitou, N., et Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, *4*(4), 406–425.

Salgado, H., Santos-Zavaleta, A., Gama-Castro, S., Peralta-Gil, M., Peñaloza-Spínola, M. I., Martínez-Antonio, A., ... Collado-Vides, J. (2006). The comprehensive updated regulatory network of Escherichia coli K-12. *BMC bioinformatics*, *7*, 5.

Samanta, M. P., et Liang, S. (2003). Predicting protein functions from redundancies in large-scale protein interaction networks. *Proceedings of the National Academy of Sciences*, *100*(22), 12579–12583.

Sánchez-Gracia, A., et Castresana, J. (2012). Impact of deep coalescence on the reliability of species tree inference from different types of DNA markers in mammals. *PloS one*, *7*(1), e30239.

Schumaker, N. H. (1996). Using landscape indices to predict habitat connectivity. *Ecology*, *77*(4), 1210–1225.

Shen-Orr, S. S., Milo, R., Mangan, S., et Alon, U. (2002). Network motifs in the transcriptional regulation network of Escherichia coli. *Nature genetics*, *31*(1), 64–68.

Sokal, R. R., et Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, *38*, 1409–1438.

Soltis, P. S., Soltis, D. E., et Chase, M. W. (1999). Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. *Nature*, *402*(6760), 402–404.

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. *Biologiske Skrifter*, *5*, 1–34.

Strauss, R. E. (2001). Cluster analysis and the identification of aggregations. *Animal Behaviour*, *61*(2), 481–488.

Swofford, D. (1991). *PAUP: Phylogenetic Analysis Using Parsimony, Macintosh Version 3.0r*. Champaign, Illinois: Illinois Natural History Survey.

Syring, J., Farrell, K., Businský, R., Cronn, R., et Liston, A. (2007). Widespread genealogical nonmonophyly in species of Pinus subgenus Strobus. *Systematic biology*, *56*(2), 163–181.

Takahashi, K., Terai, Y., Nishida, M., et Okada, N. (2001). Phylogenetic relationships and ancient incomplete lineage sorting among cichlid fishes in Lake Tanganyika as revealed by analysis of the insertion of retroposons. *Molecular biology and evolution*, *18*(11), 2057–2066.

Taylor, P. D., Fahrig, L., Henein, K., et Merriam, G. (1993). Connectivity is a vital element of landscape structure. *Oikos*, *68*(3), 571–573.

Tischendorf, L., et Fahrig, L. (2000). How should we measure landscape connectivity? *Landscape Ecology*, *15*(7), 633–641.

Tischendorf, L., et Fahrig, L. (2001). On the use of connectivity measures in spatial ecology. A reply. *Oikos*, *95*(1), 152–155.

Urban, D., et Keitt, T. (2001). Landscape connectivity: a graph-theoretic perspective. *Ecology*, *82*(5), 1205–1218.

Urban, D. L., Minor, E. S., Treml, E. A., et Schick, R. S. (2009). Graph models of habitat mosaics. *Ecology Letters*, *12*(3), 260–273.

Webb, C. O., Ackerly, D. D., McPeek, M. A., et Donoghue, M. J. (2002). Phylogenies and community ecology. *Annual Review of Ecology and Systematics*, *33*(1), 475–505.

Weiher, E., Clarke, G. D. P., et Keddy, P. A. (1998). Community assembly rules, morphological dispersion, and the coexistence of plant species. *Oikos*, *81*(2), 309–322.

Weiher, E., et Keddy, P. A. (1995). Assembly rules, null models, and trait dispersion: new questions from old patterns. *Oikos*, *74*(1), 159–164.

Wiens, J. J., et Donoghue, M. J. (2004). Historical biogeography, ecology and species richness. *Trends in ecology et evolution*, *19*(12), 639–644.

With, K. A., Gardner, R. H., et Turner, M. G. (1997). Landscape connectivity and population distributions in heterogeneous environments. *Oikos*, *78*(1), 151–169.

Yang, Z., Goldman, N., et Friday, A. (1994). Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Molecular Biology and Evolution*, *11*(2), 316–324.

Zeller, K. A., McGarigal, K., et Whiteley, A. R. (2012). Estimating landscape resistance to movement: a review. *Landscape Ecology*, *27*(6), 777–797.

Zou, X.-H., et Ge, S. (2008). Conflicting gene trees and phylogenomics. *Journal of Systematics and Evolution*, *46*(6), 795–807.