phylogenetworks

EXPLORING RETICULATE EVOLUTION AND ITS CONSEQUENCES FOR PHYLOGENETIC RECONSTRUCTION

Bastienne Vriesendorp

Promotor:	Prof. dr. M.S.M. Sosef Hoogleraar Biosystematiek Wageningen Universiteit
Co-promotoren:	Dr. F.T. Bakker Universitair Docent, leerstoelgroep Biosystematiek Wageningen Universiteit Dr. R.G. van den Berg Universitair Hoofddocent, leerstoelgroep Biosystematiek Wageningen Universiteit
Promotiecommissie:	Prof. dr. J.A.M. Leunissen (Wageningen Universiteit) Prof. dr. E.F. Smets (Universiteit Leiden) Prof. dr. P.H. van Tienderen (Universiteit van Amsterdam) Dr. P.H. Hovenkamp (Universiteit Leiden)

Dit onderzoek is uitgevoerd binnen de onderzoekschool Biodiversiteit

phylogenetworks

EXPLORING RETICULATE EVOLUTION AND ITS CONSEQUENCES FOR PHYLOGENETIC RECONSTRUCTION

Bastienne Vriesendorp

Proefschrift

ter verkrijging van de graad van doctor

op gezag van de rector magnificus

van Wageningen Universiteit

Prof. dr. M.J. Kropff

in het openbaar te verdedigen

op woensdag 12 september 2007

des namiddags te vier uur in de Aula

Bastienne Vriesendorp (2007)

Phylogenetworks: Exploring reticulate evolution and its consequences for phylogenetic reconstruction

PhD thesis Wageningen University, The Netherlands
With references – with summaries in English and Dutch

ISBN 978-90-8504-703-2

aan mijn ouders

CONTENTS

chapter 1

GENERAL INTRODUCTION

Reticulate evolution is the collective name for processes such as hybridization, recombination or other events involving gene transfer between individual organisms or lineages. This thesis is about reticulate evolution and its implications for phylogenetic reconstruction.

Reticulate evolution comes from the Latin reticulum, a diminutive for net. So, literally it means net-like evolution, which describes evolutionary events where independent lineages exchange genetic information, causing branches in an evolutionary tree to come together, i.e. forming a "net" (or network). This is in contrast with bifurcating evolution, where branches in an evolutionary tree split into new lineages and evolve independently from another.

Reticulation, hence the exchange of genetic information, can occur at multiple levels: chromosomal level (meiotic recombination), organism level, population level and species level, but the focus in this work is on reticulation at species-level. Evolution at species-level is normally seen as the accumulation of mutational changes within evolutionary lineages (i.e. substitutions, deletions, insertions, etc.) that are passed on to the new generation leading to a bifurcating evolutionary pattern. However, reticulate evolution implies genetic changes that may also arise from *gene transfer* between different lineages.

The most obvious reticulate processes that may occur at species-level are related to hybridization (such as hybrid speciation, introgression, introgressive hybridization) and horizontal gene transfer (the exchange of genetic material via an external vector such as a virus). Other evolutionary processes can mimic species-level reticulation and will produce the same signature, such as incomplete lineage sorting of ancestral alleles and recombination (i.e. reticulation at lower levels).

Whereas a large body of literature has been published to build up insights in the evolutionary and ecological significance of hybrids, the same cannot be said for the general phenomenon of reticulate evolution. Naturally, underlying processes are often discussed in terms of hybridization and throughout this thesis most data sets are based on studies that include putative hybrid species. However, when reticulate phylogenetic patterns are discussed and evaluated in this thesis, this relates to the general case of reticulate evolution and not solely to underlying hybridization events.

While many studies focus on processes, e.g. study the frequency and mechanisms of hybrid speciation, others focus on the resulting character patterns. Also, in phylogenetic reconstruction it can be important to recognize these patterns either before or after an analysis.

When reticulation has occurred, one may expect specific character patterns in the involved species. The "signature" of reticulate evolution may be an "average" of its parents as expected for morphological characters (e.g. pink flowers from a white- and red-flowered parent or sparsely hairy plants from a glabrous and a hairy one), where probably several different genes are involved in the expression of the character. However, characters may also reveal a conflicting signal, when an organism has obtained some characters from one and some from the other parent. Alternatively, additive patterns can be observed, e.g. multiple copies of nrDNA genes may result in the identification of polymorphic sites that represent the base positions of either parent. Also, markers such as AFLPs are expected to reveal an additive pattern, where both unique band positions of the parents might be observed in the hybrid.

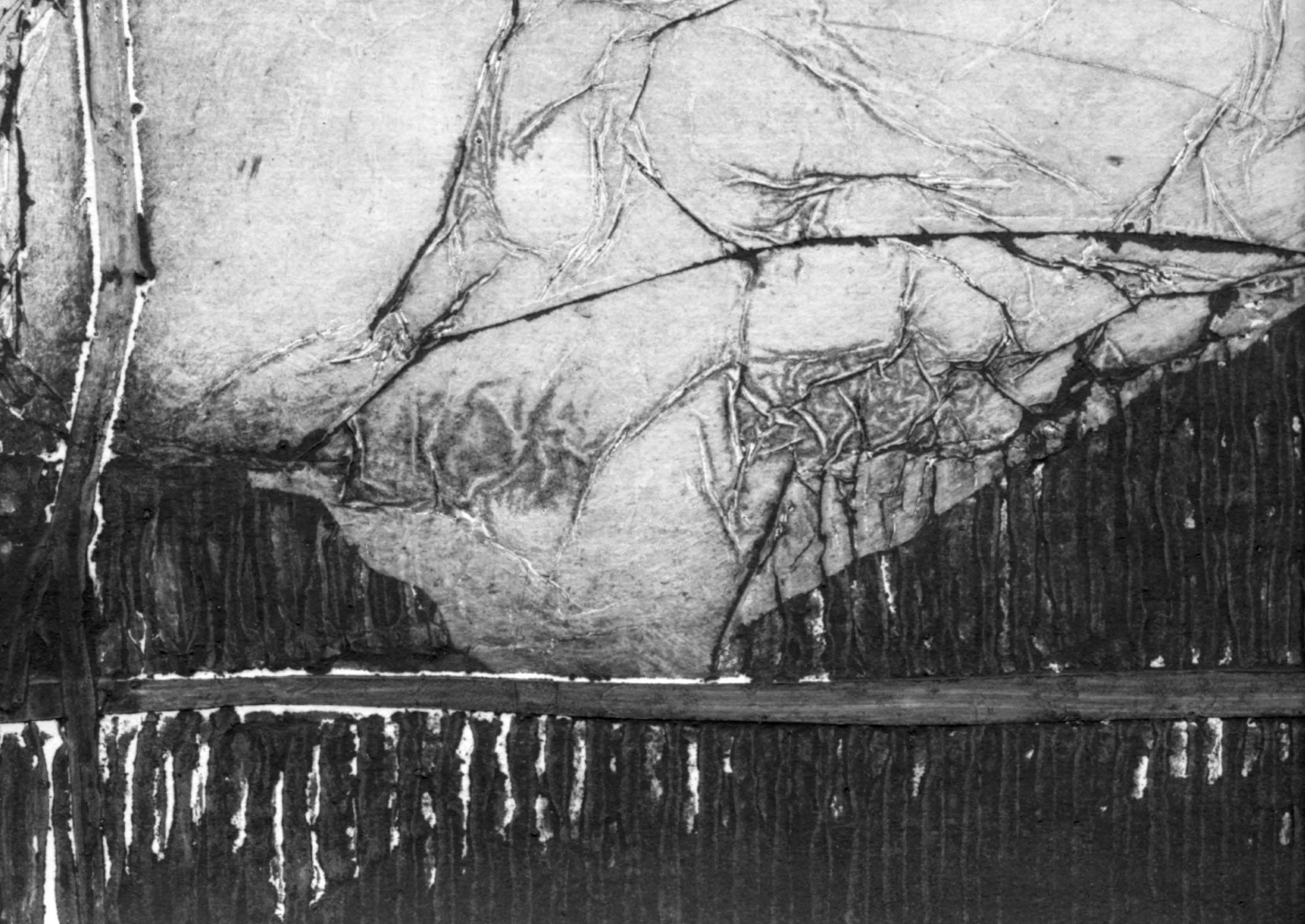
Therefore, it seems good to concentrate on what patterns may evolve from reticulate evolution and how to properly interpret them given the data. The main focus in this thesis is on data incongruence as representation for an underlying reticulate pattern, using different parts of DNA sequence alignments that are incompatible with each other.

A reticulate pattern does of course not necessarily implicate underlying reticulate historical processes. Different sources of data error or random noise may cause the same pattern of incongruence, such as taxonomic and character sampling artefacts, random error in site readings, compositional bias, etc.

Based on the variety of reticulate patterns and their causes, a range of different methods exist to detect and/or represent reticulation. In phylogenetic reconstruction at species-level, hybrids are often left out of the analysis, because they are considered to disrupt the reconstruction of (bifurcating) phylogenetic trees that do not allow for reticulate patterns. However, specific network methods have been developed that may detect reticulation between lineages and may represent the (reticulate) evolutionary relationships in a graphical way. These methods allow conflicting phylogenetic signals to be displayed in a network and are mainly based on the incongruence between different molecular markers. While most of these network methods were originally applied at population-level to simplify the non-hierarchical relationships, they can also be applied to species-level to represent the reticulate phylogenetic patterns.

This thesis is aimed at describing the different aspects of reticulate evolution, the patterns in the data and the resulting phylogeny that may emerge from it, and its consequences for phylogeny reconstruction in general.

In plant systematics, reticulate evolution usually involves hybridization. To first clarify the possible confusion in terminology related to hybrids and hybridization processes, the first part of the introductory Chapter 2 presents an overview of the terminology. The second part provides a brief history on studies related to hybrids and describes the different views on the evolutionary significance of plant hybridization. The rest of the thesis is outlined as follows: the third chapter treats the implications of plant hybrids for systematics and phylogeny reconstruction, describing the treatment of hybrids in recent phylogenetic studies, possible problems, and providing a conceptual framework for hybrid terminology. Testing of the actual influence of hybrid terminals (represented by mosaic terminals) on phylogenetic analysis is performed in Chapter 4, while Chapter 5 involves testing of the performance of a selection of network methods. In the following Chapter 6 the behaviour and suitability of AFLPs in hybrid studies is explored using an ancient hybrid species of *Solanum* plus its re-synthesized F1 hybrid. Chapter 7 explores the possibilities of methods that reconcile gene trees into a species tree to contribute to the representation of organisms instead of character-level networks. Finally, Chapter 8 provides a summary of the most important conclusions and describes the consequences of reticulate evolution for taxonomy and classification.



chapter 2

HYBRIDIZATION: HISTORY, TERMINOLOGY AND EVOLUTIONARY SIGNIFICANCE

INTRODUCTION

Hybrids are being studied in various contexts, such as speciation processes, population genetics, molecular evolution of characters in hybrids, occurrence of hybrids in phylogenies, or within crossing experiments, with applications across many different fields of expertise (i.e. genetics, ecology and systematics). Accordingly, it is not surprising that the concepts and terminology relating to hybrids is not uniform across these fields and therefore not always straightforward. Here I will first give an overview of the wealth of definitions and concepts that are used to describe hybrids or implicitly implied in studies on hybrids in section 2.1. Subsequently, in the second part (section 2.2) I will describe the evolutionary significance of plant hybridization together with a general historical overview of the occurrence of hybrids in (scientific) studies as well as views on the evolutionary significance of hybrids in a historical context.

Terms and examples are mostly related to hybrids in plant species (with the emphasis on angiosperms). However, some descriptions or examples are from studies in other groups. Especially in literature on evolutionary consequences of hybridization, often no distinction is made between plants, animals, fungi or other groups. If applicable, I will mention the specific implications for plant hybridization and mention this explicitly in the text.

2.1. PLANT HYBRIDS: HISTORICAL OVERVIEW AND DEFINITIONS

Historical overview

The term hybrid is probably deduced from the greek word "hubris" meaning arrogant pride or insolence against the gods, indicating the general feeling that the production of new forms are seen as a criticism to the work of the Gods (Zirkle, 1935). Originally, the word hybrid was used to assign mongrel animals, with many examples in Greek mythology, but it was also applied to human beings, i.e. the offspring of a roman father and an Asiatic or African mother was considered a hybrid (Zirkle, 1935).

While spontaneous hybridization must always have been frequent in the various cultivated crops and in the different breeds of domestic animals, for 2500 years these hybrids were not recognized as such because they could not be distinguished easily from true species (Zirkle, 1935). Real progress only came at the beginning of the 18th century with the discovery of sex in plants and the first crossing experiments in plants. After the publication of Camerarius in 1694, being the first scientific investigation into the question of the existence of sex in plants, many botanists were seeking to prove or disprove the new theory. Numerous botanists recorded the production of hybrids, not

so much because of the plant hybrid themselves as for the proof of the fact that sexual reproduction occurred in the vegetable kingdom (Roberts, 1929).

Among these botanists was Linnaeus, contributing to an essay-contest initiated by the Academy of Sciences at St. Petersburg for the best determination of the problem of sex in the plant kingdom. He wrote the winning article named "Disquisitio de Sexu Plantarum" (1760) where he describes some experimentally produced hybrids and also reflects upon the constancy of hybrid species (Roberts, 1929).

Notwithstanding the numerous hybridization experiments, the work of Kölreuter, published in 1761, was the first real scientific application of Camerarius discoveries of sex in plants. In this work the results of 136 distinct experiments in the crossing of plants were reported, including the protocols of his experiments and a discussion on their importance. Kölreuter found that most hybrid forms changed back towards the parents and considered hybridization as a process mainly occurring between congeneric species, contradicting Linnaeus' theory on constant species. His work was continued by the physician Gärtner (1827) who performed 25 years of extensive experimental work in hybridization. Gärtner carried out nearly ten thousand crossing experiments among seven hundred species, obtaining about three hundred and fifty different hybrids (Roberts, 1929).

At the beginning of the nineteenth century new investigations on hybridization were started, most of them closely associated with breeding experiments. The horticulturalist Thomas Andrew Knight contributed to the knowledge of hybrids by performing experiments in raising new varieties of fruits and vegetables. He was also the first to apply the science of plant hybridization to plant improvement (e.g. in 1806 and 1809, see Roberts, 1929 for a list of all his work). A contemporary of Knight, William Herbert, was also a practical plant hybridizer and he conducted experiments to improve florists' flowers and some agricultural plants. Contrary to the prevalent view of that time among botanists (such as Knight) that hybrids between different species were sterile, Herbert concludes from his experiments that "the fertility of the hybrid depends more upon the constitutional than the botanical affinities of the parents and there did not exist any decided line of absolute sterility in hybrid vegetables" (Herbert, 1837, cited in Roberts, 1929). He also emphasized that species and varieties were but arbitrary and artificial distinctions in the plant kingdom, and that "the question of whether a wild plant is a new species or a variety of a known species is a waste of intellect". Even today, systematists would agree with this view, see for example the discussion in Hamilton & Reichard (1992).

Other examples of scientists active in hybrid research at that time are Goss, Seton, Laxton, Sageret and Wiegmann (see Roberts 1929 for descriptions of their

work), followed by four more hybridizers in the middle of the century (Godron, Vilmorim, Regel and Lecoq) and later followed by Darwin (1859).

Darwin dedicated one chapter of his Origin of Species to hybridism and concluded that most hybrids were sterile, discussing hybrids mostly in the context of the distinctiveness of species. He summarized results from different studies (from e.g. Kölreuter, Gärtner and Herbert) to conclude that "the capacity in any two species to cross is often completely independent of their systematic affinity" (p.274) and stated that this supports "the view, that there is no fundamental distinction between species and varieties" (p.290), but he does not contemplate the evolutionary role of the generated hybrids themselves.

About a century after the publications of Kölreuter there was a revival in hybridization studies, e.g. with the publications of Mendel (1865), Wichura (1865), Nägeli (1865), Naudin (1861), Godron (1863), Laxton (1866) and Shireff (1873) (for an overview see Roberts, 1929). In 1881 Focke reviewed and summarized all these explorations of plant hybridization in his comprehensive work "Die Planzen-Mischlinge" (1881). This increasing interest in hybridization at the end of the 19th century facilitated the rediscovery of the Mendelian laws in 1900 (Zirkle, 1935), which provided a tremendous stimulus to the study of hybrids, although early genetics chiefly focussed on crossing experiments between closely related organisms (Stebbins, 1959). The theoretical work of Lotsy (1916) also intensified the interest in the role of natural hybridization in evolution. He attempted to explain all evolution in terms of crossing and although most of his thoughts were not shared by his contemporaries his work did serve as a stimulus for further studies (Heiser, 1949).

In the same period, the influential theoretical paper of Winge (1917) was published in which he postulates that hybrid sterility can be surmounted by chromosome doubling. This was relevant for further experiments on polyploidy and studies on the influence of hybridization on the course of evolution. In 1928, Ostenfeld summarized the available knowledge on hybridization, mainly focussing on hybridization experiments. Examples are the tetraploids in *Primula* and *Nicotiana* (Clausen & Goodspeed, 1925), the famous crossing between *Raphanus* and *Brassica* of Karpechenko (1927) and some examples of homoploid hybrids. In addition to describing examples, Ostenfeld also reflected on the possible evolutionary importance of natural hybrids.

A little later, Anderson started a series of studies on natural hybrid populations, with the main emphasis on the role of introgressive hybridization (e.g. Anderson, 1949; Anderson & Hubricht, 1938, and a review by Heiser, 1949). This provided a stimulus for the later work of the influential evolutionary botanist Stebbins, who

devoted several chapters of his book on variation and evolution in plants (1950) to hybridization and polyploidy. Additionally, Stebbins performed many experimental studies (e.g. in *Ophrys* (1956) and *Elymus* (1957)) and wrote several publications on the evolutionary significance of hybridization, describing the consequences of hybridization in general (e.g. Anderson & Stebbins, 1954, Stebbins, 1959) or specifically referring to plant hybrids (e.g. Stebbins, 1969).

Another prominent contributor to the theory on the impact of hybridity in plant speciation is Grant who treated this subject extensively in his work of 1981. His findings served as an important springboard for further research (Arnold et al., 2004).

In the last decades many more studies on hybrids have been performed, both experimental and theoretical, investigating the different processes concerning hybrid speciation, the genetic aspects, ecological implications and fitness or frequency of hybrids.

Comprehensive reviews have been published on homoploid hybrid speciation (Rieseberg, 1997), on the frequency of spontaneous hybrids (Ellstrand et al., 1996), and many papers on polyploidy processes (e.g. Thompson and Lumaret, 1992; Leitch and Bennett, 1997; Ramsey and Schemske, 1998, 2002; Soltis and Soltis, 1999, 2000; Otto and Whitton, 2000; Soltis et al., 2004). Arnold published several studies with emphasis on fitness of hybrids (e.g. Arnold and Hodges, 1995; Arnold et al., 2001; Burke and Arnold, 2001) and he performed many experiments e.g. on hybrid formation, speciation and introgression in *Iris* (Iridaceae) (e.g. Arnold et al., 1991, 1992; Arnold, 1993). Furthermore several publications or reviews on evolutionary aspects of hybrids in general are published by e.g. Rieseberg & Carney (1998), Arnold (1992, 1997, 2004) and Mallet (2007), while an overview on the use of current molecular technologies in relation to hybrid studies is presented by Hegarty & Hiscock (2005).

Most of the studies mentioned above examine the processes occurring during or after hybridization. The very important aspects related to the phylogenetic consequences of hybridization often remain underexposed, although there are several such studies, for instance with the focus on (molecular) characters and their behaviour in hybrids using the expected character patterns to reflect upon possible ways to detect hybrids. These aspects are discussed and reviewed in e.g. Rieseberg et al. (1996), McDade (1995), Rieseberg & Ellstrand (1993), Rieseberg & Soltis (1991) and Vriesendorp & Bakker (2005, Chapter 3).

Hybrid definitions

In the different crossing experiments performed by the early plant breeders of the 18th and 19th century, hybrids are mentioned without explaining what they really mean or what definition of a hybrid is (unconsciously) adopted. They use this term merely as another word for the result of a crossing between their taxa under study (either species, subspecies, varieties or races). However, the general view of 19th century evolutionary biologists was that a hybrid refers to the offspring between different *species* (see Harrison, 1993). In the statements on general sterility of hybrids many workers also implicitly imply that a hybrid is a crossing between species (see above, for example Godron (1844) and Knight (1809) described in Roberts 1929). Exemplarily, Darwin (1859) mainly treats crosses between different species in his chapter on hybridism (although he does not discuss the species definition problem) and explicitly mentions in his glossary that a hybrid is "the offspring of the union of two distinct species".

Similarly, in many general texts on systematics and evolution (e.g. Judd et al., 2002; Mineli, 1993; Ridley, 2004) a hybrid is referred to as the offspring between different species. This also seems the most general usage in phylogenetic studies, which generally concern interspecific hybrids. Some researchers add the adjective "interspecific", but this does not seem convenient because of its dependence on the choice of species concepts. However, hybridization in its broadest sense can also be seen as the "crossing of any two genetically unlike individuals" (Stebbins, 1950). This view was also held by Lotsy (1916), who designated hybridization as "any cross-mating of genetically different individuals". Since this definition can be applied to nearly all individuals of a sexually reproducing species, this is not a very useful definition either. Other definitions were formulated to overcome the latter problem, see e.g. Stebbins (1959), Mayr (1963) and Bigelow (1965) in the overview of definitions given in Table 2.1. These last three definitions do not rely on a particular species concept, but are still not very practical since they consist of the rather vague terms "unlike populations" and "different adaptive norms". Woodruff (1973) formulated another definition that was further adjusted by Harrison (1990,1993) and this slightly modified version of Harrison is often cited by later reviews: "Hybridization is the interbreeding of individuals from two populations, or groups of populations, which are distinguishable on the basis of one or more heritable characters". In this definition the parental entities only need to differ in one trait, but they need to originate from genetically distinct populations. This definition is more realistic than Lotsy's description and more practical than the other definitions, since it is easier to identify heritable different traits than different adaptive norms (Stebbins) or to determine whether the taxa under study are from two "unlike natural populations" (Mayr).

Table 2.1. Summary of definitions of hybridization in publications where the definition is explicitly mentioned. Most definitions apply to both animal and plant species.

Publication	Definition
Lotsy, 1916	any cross-mating of genetically different individuals
Darlington, 1937	a zygote produced by the union of dissimilar gametes
Miller, 1949	crossing or interbreeding organisms that are different, whether of varieties, races, species, or genera
Stebbins, 1959	crossing between individuals belonging to separate populations which have different adaptive norms
Mayr, 1963	crossing of individuals belonging to two unlike natural populations that have secondarily come into contact
Short, 1969	interbreeding of individuals of morphologically and presumably genetically distinct populations
Bigelow, 1965	crossing between natural populations that are sufficiently divergent to render the effects of genetic incompatibility recognizable as such
Woodruff, 1973	interbreeding of individuals from two populations, or groups of populations, which are distinguishable on the basis of one or more characters[a]
Grant, 1981	spontaneous interbreeding between populations which have undergone a previous history of divergence to the level of disjunct races, semispecies, or species, and which are separated by partial ecological or reproductive isolation or both
Harrison, 1990, 1993	interbreeding of individuals from two populations, or groups of populations, which are distinguishable on the basis of one or more heritable characters
Arnold, 1997	successful matings in nature between individuals from two populations, or groups of populations, which are distinguishable on the basis of one or more heritable characters
Rieseberg, 1997	crosses between different species[b]
Rieseberg & Carney, 1998	crosses between individuals of different species[c]

[a] Characters include genetically controlled morphological, ecological, physiological and ethological features
[b] species are defined as reproductively isolated entities, referring to the BSC (Mayr, 1963)
[c] (Biological) species are referred to as groups of interbreeding populations that are genetically isolated rather than reproductively isolated: modified version of Mayr's species definition (1963)

Only a few reviews on hybridization provide an explicit hybrid definition. For example Arnold (1997), Rieseberg (1997) and Rieseberg & Carney (1998) use Harrison's definition (1993). However, the last two reviews focus on hybrids between different species, with different species considered as, respectively, reproductively isolated, or genetically isolated groups of interbreeding populations, i.e. referring to Mayr's biological species concept of 1963. Dowling & Secor (1997) review the

significance of hybridization in animals and use the definition of Harrison (1993). However, in their study they focus on hybrid *taxa*, which they define as "an independently evolving, historically stable population or group of populations possessing a unique combination of heritable characteristics derived from interbreeding of representatives from two or more discrete units (e.g. races, subspecies, species, etc.)".

In recent phylogenetic studies that include putative hybrids in angiosperm phylogenies the term hybrid or hybridization is usually not explicitly explained or defined. In most cases the word hybrid is just used to indicate the result of a crossing between any entity (either species, populations or any other category).

Introgression

While the definitions mentioned above (Table 2.1) emphasize the differences of the parental taxa involved, there are also other terms that need to be considered when discussing hybrid concepts. Considering the process of hybrid formation itself, it is inevitable have a closer look at a highly related process, introgression. In Table 2.2 several definitions of introgression are summarized, mostly in the context of plant species.

Table 2.2. Definitions of introgression (or introgressive hybridization).

Publication	Definition
Anderson & Hubricht, 1938	Infiltration of germ plasm of one species into another through repeated backcrossing of hybrids to the parental species.
Anderson, 1949	Repeated backcrossing of hybrids to one or both parents (where the hybrid nature becomes less apparent with each backcross and the end result of hybridization is mainly an increased variability in the participating species)
Stace, 1989	Repeated backcrossing of a hybrid to one or the other parent, the hybrid products coming to resemble that parent quite closely after even a few generations but differing from it by some characteristics of the other species
Rieseberg & Wendel, 1993	The permanent incorporation of genes from one set of differentiated populations into another, (i.e. the incorporation of alien alleles into a new, reproductively integrated population system).
Rieseberg & Carney, 1998	The movement of genes between species mediated by backcrossing or more 1998 broadly the transfer of genes between genetically distinguishable populations.

The terms hybridization and introgression can get easily confused since these terms are often mentioned together or even interchangeably in some evolutionary

contexts. Introgression will always be preceded by hybridization, while hybridization is not necessarily followed by introgression. (Although Anderson [1949] considers the repeated back crossing as one of the commonest side consequences of natural hybridization, and Heiser (1973) states that hybridization will often be followed by introgression.) Introgression and hybridization are clearly different processes, but in practice it seems hard to make a distinction between the two, especially when discussing the evolutionary consequences. Many reviews and other studies (e.g. Anderson, 1949; Rieseberg & Wendel, 1993; Rieseberg & Soltis, 1991; Rieseberg & Brunsfeld, 1992; Arnold, 1992, 1997) use hybridization, introgression and introgressive hybridization interchangeably.

Notwithstanding the existing range of definitions as listed in Table 2.2, most papers dealing with introgression do not present a specific definition. Arnold (1992) refers to the definition of Anderson (1949) while Anderson & Hubricht (1938) and also Rieseberg & Carney (1998) mention a slightly modified definition based on these definitions. Rieseberg & Wendel (1993) make a distinction between permanent incorporation of genes and transient gene flow in hybrid swarms and incorporated this view in their definition of introgression (see Table 2.2). Dowling & Secor (1997), reviewing on hybridization and introgression in animals, also use this definition of Rieseberg & Wendel (1993).

Other hybrid categories

In addition to the above-mentioned definitions other hybrid categories have been defined, for instance focussing on the age of hybrids or on the amount of evolution in hybrids, in e.g. Rieseberg & Ellstrand (1993) and McDade (1995). Rieseberg & Ellstrand (1993), in their review on markers in plant hybrids, make a distinction between first generation hybrids, later generation hybrids and hybrid species, without further defining these categories. McDade (1995) reviews different patterns of character distribution in plant hybrids and defines two categories of hybrids: primary and derived. She delimits primary hybrid as hybrids that are in more or less the same genetic and phenotypic condition as when they were initially formed. Derived hybrids are specified as having undergone considerable evolutionary change since they originated. In this view primary hybrids are probably easier to detect than derived hybrids, since the latter category can be as different from their parents as other distinct species. Hybrid individuals and populations (F1) plus allopolyploids and stabilized hybrid lineage's (that have undergone little change) fall into the category of primary hybrids. Most (homoploid) hybrid species can probably be categorized as derived hybrids (McDade, 1995).

This view on primary hybrids can be compared with the definition of hybrid taxa given by Dowling & Secor (1997) (see above). In their definition they emphasize the need of historical stability, which implies that the mosaic of characters inherited from independent lineages is retained in the population. These delimitations can be very useful since it makes a very clear distinction between categories of hybrids. Nevertheless, they are not applied in any other studies (as far as I know). In addition, the term "primary hybrid" can be confusing as it is also used to indicate first generation (F1) hybrids (see for instance Rieseberg & Ellstrand, 1993).

It is clear that different studies use the term hybrid in various contexts. While some workers only consider the difference between the rank of the parents as an important distinction, others look into more detail to the process of hybridization, thereby for example differentiating between introgression on one hand and hybridization with instant isolation of the hybrid from its parents on the other hand. Most of the hybridization definitions summarized in Table 2.1 only differ in the parental rank and do not consider aspects of the process, the reproductive stages, whether or not introgression is involved, the patterns occurring in the hybrid etc. It is useful to summarize the plethora of concepts and definitions of hybridization and related processes in a conceptual framework, see Chapter 3, Fig. 1.

Conclusions

Although many definitions of the term hybrid have been suggested, in practice these definitions are seldomly applied in studies on hybrids. The simple use of the term hybrid as "a crossing between unlike taxa" is the most practical definition since it does not need any knowledge of the age, history or reproductive mode of the hybrid. In my opinion this is a good application of the term hybrid, as long as additional information is given. It is good to state whether the object of study is a first generation hybrid or an older hybrid, especially because this implies that some hybrid detection methods might be more successful than others. For instance, trying to infer the patterns in a F1 hybrid can be very important and interesting from a process-related view, but might not contribute to the knowledge on the behaviour of molecular characters of later generation hybrids or ancient species.

In studies on processes of hybridization where introgression can play a role, some use the term hybridization in the title or introduction of their study and further elaborate on possible influence of introgression. However, other studies start with introducing the term introgression (e.g. in Hardig et al., 2000) while they later use the term hybridization or mention that they investigate processes of "hybridization or introgression". This can be confusing, since the use of the different terms suggests that

the formation mode of the hybrids is known, while this is often not the case. In such cases it is probably better to just use the term hybridization as a neutral statement to introduce the subject and later investigate processes such as introgression.

The same is true for polyploid processes. Some studies introduce a polyploid species of putative hybrid origin as an allopolyploid (e.g. in *Erythronium,* Allen & Soltis, 2003), while others safely define the putative hybrid as polyploid (e.g. in *Viburnum,* Winkworth & Donoghue, 2004). Again, this might be confusing, since this difference suggests that the first example has more evidence than the latter, while this does not need to be the case.

In conclusion, the most important terms have been adequately defined and the introduction of any more terms will only make it more complicated and unnecessarily confusing. As long as it is clear on what ground a hybrid is defined as a hybrid, on what evidence a polyploid is called an allopolyploid, what the (presumable) age of the hybrid is, etc., it will be easier to compare the results of different studies and to use this knowledge to infer general conclusions.

2.2. EVOLUTIONARY SIGNIFICANCE OF HYBRIDIZATION IN PLANTS

Changing views through time

Early views on plant hybridization

Hybridization has been known since antiquity to occur in both animals and plants (Zirkle, 1935), but scientific studies on plant hybrids only started in the middle of the eighteenth century (by e.g. Linnaeus, 1760; Kölreuter, 1761-1766). One of the first to note that hybrids could form new species was Linnaeus (e.g. in 1760). Although his work mostly adhered to religious principles and to the constancy of species, illustrated by his often quoted remarks "we count as many species as different forms were created in the beginning", later he wrote that "species are the work of time" indicating that he might accept that species change through time. In his essay to proof the sexual nature of reproduction in flowering plants (Linnaeus, 1760, see 2.1) he wrote that "there can be no doubt that these are all new species produced by hybrid generation" (cited in Roberts, 1929).

However, these pre-Darwinian "evolutionary" thoughts on new species were not shared by most of his contemporaries. Kölreuter, who was a contemporary of Linnaeus, refuted these ideas on new species, and proved this with experiments on hybrids that reverted back to one of the parents, a process we now call introgression.

Kölreuter also stated that he had seen no proof of the occurrence of hybridization in nature (Stafleu, 1971).

During the 18th and 19th century many important workers on plant hybridization e.g. Darwin, Laxton, Shireff, Mendel, Naudin and Focke agreed with Kölreuter on the lack of constancy of hybrids (Rieseberg, 1997) and did not assign any important evolutionary role to hybrids. For instance, Naudin (1864, cited in Roberts, 1929) in his paper on variation in hybrid plants stated that "in this overlapping of the characters of the two different species, one does not see anything new appear" and concluded later from his crossing experiments with *Petunia* that "the hybrids have no constancy" (1861, cited in Roberts, 1929). Many workers conclude that hybrids are generally sterile and also Darwin (1859) discussed hybrids mostly in the context of the distinctiveness of species, describing the various degrees of sterility in hybrids. Although Darwin remarked that "the crossing of forms only slightly different is favourable to the vigour and fertility of their offspring" he did not reflect on a possible positive or long-term evolutionary role of hybrids.

At the end of the 18th and early 19th century the interest in hybridization in plants mostly concerned artificial hybrids, e.g. Knight (1906) and Herbert (1819, 1847), who were a fruit breeder and a practical plant hybridizer, respectively (Roberts, 1929). In the early 19th century, however, more botanists reported about spontaneous (i.e. natural) hybrids (for example Focke (1881)). In his overview "Die Pflanzen-Mischlinge" (1881), Focke reported not only on many experimental hybrids, but also on natural plant hybrids observed in the wild. He also reflected upon characteristics of hybrids and remarked that some genera or groups of species are "more inclined than others to enter into hybrid combinations" (Focke 1881, cited in Roberts, 1929).

Notwithstanding the lack of discussion on the evolutionary role of hybridization during this period, there were some speculations concerning the possible establishment of hybrid species. For instance Naudin (1863), despite of his remarks on the return of hybrids to parental forms, does also suggest that hybrid characters may become fixed in later generations, recognizing the potential role of hybridization in the process of evolutionary change leading to species formation (as discussed in Rieseberg & Carney, 1998). Kerner (1894-1895) expanded these ideas, considering the role of habitat in the speciation processes. He postulated that the limiting factor for success of a hybrid to become a new species must be the environment. Based on examples from *Rhododendron, Salvia* and *Nuphar*, he concluded that hybrid plants can only establish themselves in open and favourable habitats that are not occupied by the parents. Although he restricted his discussions to fertile hybrids, his remarks (on open habitats as a necessary condition for the establishment of hybrid species) significantly

contributed to later writings on hybrid speciation (Rieseberg, 1997; Grant, 1981; Rieseberg & Carney, 1998). He influenced for example the ideas of Templeton (1981) about the role of ecological divergence in later models of hybrid speciation. This model recognizes the important roles of selection and ecological divergence causing a hybrid to become stabilized and reproductively isolated from the parental species (see also the section on hybrid speciation below). In this view, hybrid segregates will often diverge ecologically from both parents and the availability of open habitat niches can play an important role in facilitating their establishments.

A very significant contribution to this discussion on hybrid speciation came from the discovery of polyploidy, following the exploration of plant cytogenetics in the beginning of this century (Grant, 1981). Winge (1917) was the first to postulate that a new species can arise from a hybrid following chromosome doubling. This hypothesis was soon confirmed by artificial hybridization experiments in for instance *Raphanus-Brassica* (Karpechenko, 1927) and *Galeopsis* (Müntzing, 1930). Allopolyploidy is now considered to be very common and widespread in vascular plants (Grant, 1981) and speciation via polyploidy is up to the present day recognized as a very important pathway of speciation in plants (Soltis & Soltis, 1993; Otto & Whitton, 2000). Many other example studies have focussed on the importance of allopolyploidy in speciation processes, e.g. in *Geinae* (Smedmark et al., 2003), *Spartina* (Ainouche et al., 2004), *Centaurium* (Mansion et al., 2005), *Houstonia* (Church & Taylor, 2005), *Achillea* (Guo et al., 2006) and *Cardamine* (Lihova et al., 2006).

Progression was also made in understanding hybrid speciation in the absence of polyploidy. Müntzing (1930) studied the genetics of hybrids in the genus *Galeopsis* and proposed the concept that chromosomal rearrangements in hybrids could lead to the formation of a new hybrid species, reproductively isolated from both parental species.

These mechanisms to describe homoploid hybrid speciation were further explored in theoretical and experimental studies, for instance by Grant (1958) who introduced the term recombination speciation. See also the section below on speciation mechanisms and Rieseberg (1997) for a comprehensive overview of studies related to this mode of hybrid speciation.

In the early 20th century more studies came to emphasize the possibly important role of hybridization in creating evolutionary novelty. Lotsy (1916) in his book "Evolution by means of natural hybridization" was the first to suggest a major evolutionary role of hybridization, and by the middle of the 20th century many experimental and conceptual studies involved the possible consequences of hybridization (e.g. Anderson, 1949; Anderson & Hubricht, 1938; Anderson &

Stebbins, 1954; Stebbins, 1950, 1959; Heiser, 1949). Most of these workers do not restrict their discussions to plant hybrids, e.g. Stebbins (1959) also discussed animal hybrids and evolutionary consequences of hybridization in animals. However, most of them are botanists and most of the considerations are related to examples from plant hybrids, especially in angiosperms.

Anderson and his co-workers investigated several examples of introgression in plants and stated that introgressive hybridization could potentially play an important role in the evolution of certain species, due to the "enrichment of variation in the participating species" (Anderson, 1949, Anderson & Hubricht, 1938). He also emphasizes the connection between hybridization and habitat, arguing that in certain disturbed habitats introgressive hybridization must have played an important role in the evolution of some plant species (e.g. Anderson, 1949; Anderson & Stebbins, 1954). In this view hybrid genotypes can possess novel adaptations that allow the invasions of new habitats not utilized previously by either parent. Heiser (1949) also stressed the role of introgression and enumerated several probable cases of introgression in higher plants. He concludes from this overview that it is still too early to evaluate the precise role of hybridization in evolution, but that it certainly does play a role.

Grant investigated the origin of new species from hybrids in *Gilia* (1966) and later devoted a large part of his book on plant speciation to evolutionary consequences of hybridization, describing several possible ways of hybrid stabilization (Grant, 1981). Many others have contributed to the discussion, with examples from studies on ecological isolation in putative hybrid species in *Delphinium* (Lewis & Epling, 1959), *Ophrys* (Stebbins & Ferlan, 1956) and on the stabilization of hybrid derivatives by pollination in *Penstemon* (Straw, 1955).

Hybridization in animals

In contrast to studies on plant hybrids, a different approach was taken by evolutionary biologists studying animal taxa (e.g. Dobzhansky 1951; Mayr, 1942, 1963). While botanists such as Anderson, Heiser, Stebbins and Grant considered natural hybridization as an important evolutionary process, zoological workers held different views. In general, they considered natural hybridization only of importance as a possible mechanism that could lead to the final development of barriers to reproduction between species (Dobzhansky, 1951; Mayr, 1942). One of the major opponents of a significant evolutionary role for hybridization in animals is Mayr, who summarized after having discussed several hybridization studies: "Instead of furthering speciation, that is the establishment of discontinuities, hybridization has, in all these

cases, accomplished just the opposite" (Mayr, 1942). And in 1963 he stated again that no evidence has been found for a major evolutionary role of hybrids in higher animals (as opposed to invertebrates) because hybrids are rarely found, are in most cases sterile or "produce genotypes of inferior viability that are eliminated by natural selection" (Mayr, 1963). He considered the genetic variability resulting from introgressive events negligible compared to the variability originating from mutation and regular gene flow from conspecific populations.

Another difference can be detected between studies on animal and plant hybrids. Many such zoological studies are process-orientated, investigating microevolutionary processes in hybrid zones, reflecting the main interest in the role of hybridization in the process of speciation (e.g. Hewitt, 1988, 2001; Harrison, 1990; Barton & Hewitt, 1985). In contrast, in plants emphasis has mostly been on the systematic implications of hybridization (i.e. pattern-orientated) and few studies at the population-level focussing on the process of natural hybridization have been performed (some examples are e.g. introgression in willow hybrid zones (Hardig et al., 2000), population-level studies in *Phlox* (Levin, 1967; Ferguson & Levin, 1999) and other experimental, microevolutionary studies, e.g. Arnold et al. (1990) and Rieseberg & Carney (1998).

One of the major causes for these differences in perspective is the less frequent observation of hybrid formation in animals than in plants (Levin, 1979). Also, the main mechanism of hybrid speciation, polyploidy, is considered to occur far rarer in animals than in plants. However, several examples of recent and ancient examples can be found throughout the animal kingdom with e.g. many examples of insect and fish species (Otto & Whitton, 2000) and based on a recent survey of natural interspecific hybridization studies, Mallet (2005) estimated that 10% of animal species is involved in hybridization.

Despite the negative attitudes towards the role of hybridizization in animals, many other studies did assign a significant evolutionary role to hybridization (e.g. Dobzhansky, 1951; White, 1954) and there are several experimental examples, for instance the study on the Australian fruit fly *Dacus tryoni* (Lewontin & Birch, 1966), indicating that introgressive hybridization could lead to new adaptations. For example, several studies on birds indicated that introgressive hybridization has been important in avian evolution (Short, 1972; Grant & Grant, 1992) and recent reviews indicate the evolutionary significance of hybrid speciation in animals and/or plants (Mallet, 2005, 2007; Schwarz, 2005) and the possible important role of hybridization in facilitating adaptive radiation (Seehausen, 2004).

Present views on plant hybridization

At present, there are still different points of view on the evolutionary significance of hybridization. At one extreme, hybridization is considered to be of little importance (see e.g. Schemske, 2000), while on the other hand the view is held that hybridization is a significant evolutionary force that creates evolutionary novelty (e.g. Arnold, 1992, 1997; Rieseberg, 1997). In the last decade, a number of studies have focussed on this creative role of hybridization in evolution with two major evolutionary consequences, i.e. the transfer and origin of new adaptations and the origin of new species (Arnold, 2004).

Of course, this also depends on the definition of hybridization (see also the discussion above and the conceptual framework (Fig. 3.1)). In many studies it is not explicitly stated what kind of hybrid definition is used. Whether hybrids are seen as constant hybrid species or as F1 hybrid individuals or what the taxonomic status of the parents is, is often not mentioned. It is important, however, to make such distinctions in discussions on the evolutionary significance of hybrids.

Introgression versus hybridization

As discussed above, introgression (the repeated backcrossing of hybrids to one or both parents) cannot be separated from hybridization. The origin of new adaptations is often related to introgression (e.g. Anderson, 1949; Heiser, 1973) and many have discussed the impact of the enrichment of new adaptations under the term introgressive hybridization or introgression. Anderson (1949) states that introgressive hybridization is responsible for much of the current genetic variability found in extant species. Several studies have been performed, e.g. in *Iris* and *Helianthus*, where the transfer of genetic material caused the involved species to differ in ecological preferences, thereby promoting the spread of these forms into new habitats (Arnold, 2004). Another example where introgression was found to be of evolutionary importance is in oak (*Quercus*) where gene flow among different sympatric species proved far more frequent than that between distant conspecific populations (Whittemore & Schaal, 1991).

One of the long-term consequences of hybridization is the origin of an entirely new species originating from hybrid individuals or populations. Of course, a new species may also arise from the origin of novel adaptations through introgression and it is often difficult to distinguish the different origins (Arnold, 2004). However, many plant systematists emphasize more specifically the creative role of hybridization in producing new lineages (new hybrid species) as one of the most important evolutionary consequences of hybridization (Grant, 1981; Rieseberg & Wendel, 1993).

Reviews on the role of hybridization (e.g. Rieseberg, 1997; Arnold, 1997) also emphasize this role.

Hybrid speciation was defined by Grant (1981) as "the origin of a new species directly from a natural hybrid" (see Table 2.1 on the different hybrid definitions). Grant distinguished this term from "hybrid race formation" which he defined as a situation in which a "new race, formed by hybridization processes diverges to the species level in geographical isolation from the original ancestral race" (Grant, 1981). Although he emphasized the origin of a new hybrid species directly in isolation, in practice it seems impossible not to include introgressive events. Many authors recognize the inevitability of including introgression in this definition since especially the origin of homoploid hybrid species is likely to involve backcrosses when the F1 hybrids are almost sterile. (Rieseberg, 1997; Rieseberg & Carney, 1998).

Moreover, it will be impossible to distinguish between the processes of backcrossing on one hand and sibcrossing or inbreeding on the other hand. In one extreme case, an F1-hybrid evolves further in complete isolation from either parent without any backcross to either parent. Another possibility is the situation where two parental taxa exchange genes without producing any new taxa and a third extreme could be a hybrid that crosses back freely during several generations, which can even result in the amalgamation of the species involved. The first process would be called hybrid speciation, the second introgression and the third e.g. "hybrid swarm formation and merging of species". However, if a hybrid crosses back once and then evolves further in isolation to a new lineage, one would probably still call this a hybrid species. But what about a hybrid that crosses back five times? Also, different gradients of gene flow between the parents may occur and different individual hybrids can have different multiple origins, but still form one new hybrid population. All these processes can be seen as extremes of a continuum of processes in hybridizing populations and/or species (Abbott, 1992).

Hybrid speciation and its mechanisms

In this overview I will focus on long-term consequences of hybridization in plants and the formation of new taxa as a result of hybridization, rather than just the introgression of genetic characters by hybridization. As it does not seem feasible to hold on to Grant's strict definition of hybrid speciation (see above: "the origin of a new species directly from a natural hybrid"), I will use the term hybrid speciation here in a broader perspective. I modified the definition of Grant (1981) and use hybrid speciation here as "the process through which a new species originates from a crossing between two different species, where the new species becomes reproductively isolated from its

parental species", regardless of whether introgression is of influence or not. Following this definition I will describe the different mechanisms suggested by several authors to be responsible for the stabilization of hybrid derivatives.

In hybrid speciation the hybrid is often sterile or semi-sterile and needs to restore its fertility in isolation from the parents, i.e. it needs to become "stabilized" (Abott, 1992; Grant, 1981). Grant formulated one of the most comprehensive overviews of several methods of this stabilization process. Following Grant, a first division can be made between asexual and sexual processes.

Asexual processes

As Grant described, there are several ways of asexual propagation of a sterile hybrid. One way is by apomixis, such as vegetative propagation and seed formation without fertilization (agamospermy). A recent example is described for the *Ranunculus cassubicus* complex (Paun et al., 2006) where apomixis follows hybridization to establish the divergent hybrid genotypes. While this reproduction mode is generally seen as an evolutionary dead-end because of the lack of genetic variation and possible degeneration due to increasing mutational load (van Dijk, 2003), there are several ways for a "sterile hybrid" to generate more diversity. For example, triploid hybrid individuals can sometimes form triploid bridges (Ramsey & Schemske, 1998). In addition, van Dijk (2003) shows that apomictic lineages can cross with sexuals to generate new apomictic clones.

Another interesting case is the multiple hybrid formation in *Arabis*, where most of the putative hybrid individuals are sterile and produce sterile pollen, but there is molecular variation due to the multiple origin of the hybrid population and the existence of a few (fertile) diploid hybrid individuals (Koch et al., 2003).

The biological meaning and importance of this mode of speciation is not clear. As Grant (1981) pointed out, asexual reproduction of the hybrid can lead to a new taxonomic species if the new hybrid possesses a new combination of characters, distinct from the parents and if it can spread by asexual means. However, when applying the commonly used biological species concept, this asexual mode of speciation can never contribute to the formation of new species, since no reproductive isolation mechanisms are involved (Rieseberg, 1997). Many reviews on plant hybrids do not discuss the role of asexual modes extensively (e.g. Rieseberg (1997) on homoploid hybridization and Rieseberg & Carney (1998) on plant hybridization) while other studies explicitly include asexual mechanisms of producing new taxa of hybrid origin (e.g. Dowling & Secor (1997) on the role of hybridization in animals and Sastad (2004) on polyploid speciaton in bryophytes).

Asexual reproduction is common in mosses and ferns and therefore hybrid speciation in these groups is likely to involve different asexual mechanisms or stages. The aberrant modes of reproduction in bryophytes provide opportunities for additional mechanisms of polyploid formation in bryophytes, such as apospory: the regeneration of a diploid gametophyte from sporophyte tissue (Sastad, 2005). In the polyploid complex of the fern *Asplenium*, allotetraploid fertile intermediates are formed as well as sterile diploid or triploid hybrid ones (Wagner, 1954). Another example is described in a study on the allopolyploid complex in *Polystichum* (Perrie et al., 2003). It was found that in addition to aborted spores, also a sexually outcrossing component existed in the breeding systems of the hybrids. Vogel et al. (1999) describe the processes of polyploidy formation in ferns, evaluating the involvement of multiple origins in several groups.

Formation of hybrid species through either asexual or sexual processes cannot be completely separated, since species can have different periods of asexual or sexual reproduction. For example a semi-sterile hybrid that mainly propagates vegetatively can become fertile after crossing back to one of the parental species. Also, some of the stabilization methods of hybrids (Grant, 1981) include some specific aberrant genetic system that cannot be classified into asexual or sexual mechanisms: examples are permanent translocation heterozygosity (where chromosome rings are formed at meiosis due to heterozygosity for successive translocations) or unbalanced polyploidy (caused by the combination of different sets of parental chromosomes in one plant).

Sexual processes

Most attention has been given to processes involving sexual mechanisms, since only these mechanisms can give rise to new biological species, following the biological species definition emphasizing reproductive isolation (Grant, 1981; Rieseberg 1997). The different processes where sexuality and interbreeding persist throughout the formative stages of new taxa by hybrid origin can be classified into three categories:

1. Homoploid hybrid speciation by postmating barriers (recombinational speciation)
2. Homoploid hybrid speciation by premating barriers
3. Allopolyploid speciation.

Ad 1. The process of hybrid speciation by postmating barriers involves the development of chromosomal sterility barriers between the hybrid species and its parent and can be summarized as follows. The two parental species need to have chromosomal differences that serve as sterility barriers to isolate the two species (these

differences are also called chromosomal rearrangements). The F1 hybrids of these species will be sterile or semi-sterile, but the chromosomal rearrangements in the progeny of these hybrids produce new homozygous recombinant types. Some of the F2 hybrids will be sterile, some will have the same chromosomal rearrangements as one of the two parents (making them indistinguishable to the parental species) while some new recombinant types will be fertile within the hybrid species and sterile with both parents. The formation and establishment of these new types is called recombinational speciation (Grant, 1981; Rieseberg, 1997; Rieseberg & Wendel, 1993). Müntzing (1930) was the first to propose the idea of recombination of chromosomal sterility factors to produce new fertile hybrids based on his work on *Galeopsis* and *Crepis* hybrids. Grant (1958) and Stebbins (1957) defined a model of recombinational speciation as described above. Later Templeton (1981) proposed a more general model attributing a more important role for selection (Rieseberg, 1997). The models have been verified by artificial synthesis of hybrids in comprehensive experiments involving for instance *Elymus* (Stebbins, 1957), *Gilia* (Grant, 1966) and *Nicotiana* (Smith & Daly, 1959). Furthermore, a few confirmed cases of recombinational speciation in the wild have been investigated. See the examples below in the section on the frequency of homoploid hybrids.

Ad 2. Premating isolating mechanisms include habitat, temporal or ethological barriers. Rieseberg (1997) mentions a few examples where premating barriers appeared to be important in the establishment and maintenance of a hybrid species e.g. in *Rhodondendron* (Kerner, 1894-1895) where ecological factors such as soil preferences and behaviour of pollinators might play a role. Here the hybrids differ in flower color and in soil preference, causing external isolation of the hybrids from the parental species (Grant, 1981). The putative hybrid between two closely related species of *Pinus* (Mirov, 1967) remains distinct due to its different ecological requirements. Other studies e.g. in *Ophrys* (Stebbins & Ferlan, 1956), *Delphinium* (Lewis & Epling, 1959) and *Alsophila* (Conant & Cooperdriver, 1980) all suggest the role of external ecological barriers as a factor in keeping the hybrid species isolated from their parents (Grant, 1981).

Ad 3. The term polyploidy was introduced by Winkler in 1916 and Winge (1917) was the first to associate polyploidy with interspecific hybridization. He suggested that chromosome number doubling in species hybrids could convert the hybrid to a fertile type with instant creation of a new constant species. This hypothesis was experimentally confirmed through the artificial synthesis of a tetraploid hybrid of *Nicotiana glutinosa* and *N. tabacum* (Clausen & Goodspeed, 1925) followed by

several other examples e.g. in *Raphanus-Brassica* (Karpechenko, 1927) and *Galeopsis* (Müntzing, 1930).

Allopolyploidy is now generally recognized as the main process of hybrid speciation and currently two main mechanisms of polyploid formation are described: (a) somatic doubling in mitosis and (b) sexual polyploidisation via non-reduction in meiosis.

a) Somatic polyploidisation concerns genomic doubling that involves a failure of cell division following mitotic division, resulting in polyploid tissues. If this takes place in the zygote or the early embryo, this generates complete polyploid new organisms (Grant, 1981; Ramsey & Schemske, 1998; Otto & Whitton, 2000). An example of somatic doubling is the allotetraploid *Primula kewensis* that arose by somatic doubling of particular flower branches on the sterile diploid hybrid (Newton & Pellew, 1929). Little is known about the frequency of somatic polyploidisation in plants and the effects of interspecific hybridization on its occurrence (Ramsey & Schemske, 1998; Carputo et al., 2003). Some authors conclude that its frequency is very low and that it is therefore an unimportant process in the production of polyploids (Harlan & de Wet, 1975) although it could be a pathway for stabilization in a completely sterile hybrid as e.g. *Primula kewensis* (Newton & Pellew, 1929).

b) Sexual polyploidisation: gametic non-reduction, involving a failure of cell division during meisois, resulting in diploid spores or gametes and the formation of tetraploid zygotes after their union (Grant, 1981). This can happen either directly from two parental species, or via the formation of unreduced gametes in the originally (diploid) hybrid (Ramsey & Schemske, 1998). A two-step mechanism including a triploid-bridge is also known as one of the possible pathways to produce sexual polyploids. Triploids are formed within a diploid population by the union of an unreduced and a normal (reduced) gamete and have often been observed in nature. Selfing of these triploid species or backcrossing of the triploids to the parental diploids can produce fertile tetraploids. In contrast to the common claim that triploids are sterile, the pathway via a triploid bridge seems a very plasusible pathway for the formation of both auto- and allopolyploids (Ramsey & Schemske, 1998).

Both somatic and sexual polyploidization lead to chromosome doubling, but the genetic consequences of the two modes are very different (Carputo et al., 2003). Somatic doubling transmits all the parental heterozygosity, but brings no additional heterozygosity at the same allele. It can transmit only two different alleles per locus, whereas sexual polyploidy can have a maximum of 4 different alleles per locus, making sexual polyploidization likely to result in species that are genetically much more variable (Carputo et al., 2003). The occurrence and impact of this process in the

evolution of natural polyploids is unknown, but it may have contributed to the success and diversification of many polyploid lineages in both plants and animals (Song et al., 1995). Due to their higher levels of heterozygosity in comparison with their diploid progenitors, polyploids are able to adapt faster than diploids and generally have a broader ecological tolerance (Otto & Whitton, 2000). In addition, they also have a relatively high genetic diversity caused by other sources of genetic variation such as recurrent polyploidization, genome re-shuffling (intra- and intergenomic re-arrangements) and gene-level changes (for instance concerted evolution and gene silencing) (e.g. Soltis & Soltis, 2000). The degree of genomic change is further influenced by cytoplasmic-nuclear interactions and transposable elements (Soltis & Soltis, 1999). All these different processes lead to the much larger variation within polyploids and can be an explanation for the great evolutionary success of polyploids (Soltis et al., 1992). A different view on the putative advantageous characteristics of polyploids is presented by Meyers & Levin (2006). They suggest that the abundance of polyploids may be mainly coincidental and does not require evolutionary advantages. In this view many plants have a polyploid mode of origin just because polyploidy is largely irreversible.

The term allopolyploidy refers to polyploids of hybrid origin as opposed to autopolyploids formed by doubling of genomes within a species (following the first classification by Kihara & Ono, 1926). Therefore, considering our interest in the evolutionary impact of hybridization through polyploidy, our focus should be on allopolyploids. However, in many examples it is not known whether a polyploid species has an autopolyploid or allopolyploid origin. The traditional view is that autopolyploids possess lower fertility and are much less common than allopolyploids (Stebbins, 1950; Grant, 1981) due to meiotic irregularities, caused by the irregular separation of the tetravalents during meiosis in autopolyploids (Soltis et al., 2004). Autopolyploids are thought to be recognizable by the formation of multivalents during meiosis, while allopolyploids form stable bivalents. However, Ramsey & Schemske (1998) found that these rules are often not valid with remarkably little difference among auto-and allotetraploids. There are indications that in comparison with allopolyploids autopolyploidy may not be as rare as once thought based on several examples of ancient autopolyploid events (Thompson & Lunaret, 1992) and e.g. estimations of triploid bridge pathways contributing to autopolyploid formation (Ramsey & Schemske, 1998). In addition to these studies, Ramsey & Schemske (2002) did not find evidence for lower fertility in autopolyploids in comparison with allopolyploids. Consequently, Otto & Whitton (2000), in their review on polyploid evolution, do not distinguish between the mode of origin of polyploidy and the mode of chromosomal segregation.

Accordingly, they do not refer to the distinction between allopolyploids and autopolyploids in their conclusions. Also, many case studies on polyploids do not specify the mode of origin simply because it is not known (e.g. polyploid clades within *Viburnum* (Winkworth & Donoghue, 2004)). This is especially observed in studies on ancient polyploids where it is almost impossible to determine the processes involved in the formation of these species.

Prevalence and importance of hybridization

So what is the contribution of hybridization to evolutionary processes? How common are hybrids in the wild and how often does hybridization lead to evolutionary change? Polyploidy is recognized to be widespread in plants, although estimates of frequency vary considerably in literature. This is also dependent on the method of estimation. Table 2.3 shows that estimated polyploid frequencies range from 30-35% (Stebbins, 1950) to 43-58% (Grant, 1981) to even 80% (Goldblatt, 1980). Most of these estimates were based on chromosome counts, where all plant species with a number higher than a certain value are considered to be of polyploidy origin.

Table 2.3. Estimates of frequency of polyploidy in angiosperms.

Publication	% polyploids	Criterion for assigning "polyploid" status
Stebbins, 1971	30-35%	Chromosome number is a multiple of a lower chromosome number within the same genus
Grant, 1981	47%	Chromosome number higher than n=14
Goldblatt. 1980	70-80%[a]	Chromosome number higher than n=9/10
Masterson, 1994	70%	Compare stomatal size fossil with extant taxa
Otto & Whitton, 2000	2-4%[b]	Distribution odd/even haploid chromosome numbers

[a] % of polyploids in monocots

[b] % of polyploids involved in speciation events

Masterson (1994) used another estimation method where differences in size of the stomata in fossils and extant taxa are used as an indication of polyploidy. Otto & Whitton (2000) also applied a completely different approach and estimated the amount of polyploids based on the distribution of haploid chromosome numbers. By analysing the pattern of even and odd haploid chromosome numbers they estimated that roughly 2-4 % of all speciation events in angiosperms involved polyploidy.

The proportion of polyploids is estimated to be even higher in ferns, for which Grant estimated that about 95% of all species have a polyploid origin, while Otto & Whitton (2000) mention a 7% frequency (see above for methods of estimation). Sastad

(2005) gives estimates of the extent of polyploidy in mosses based on different methods: the threshold method (with a haploid threshold number of n <9) gives an estimation of 84% of moss species with a polylpoid origin, while another approach based on a formula using the number of intrageneric polyploids and number of speciation events within genera gives an estimation of 6.4-18.6 % (Sastad, 2005). The high frequency of polyploids is often used as an indication for the prevalence of hybridization, because polyploidy can be seen as a process that restores sterility caused by hybridization. However, polyploid species do not necessarily have a hybrid origin as they can also be autopolyploid, see discussion above.

The frequency of homoploid hybrid species is even more difficult to establish. Many cases of homoploid hybrids can only be detected with certainty after artificial crossing, detailed DNA analysis (such as FISH), biogeographical investigations, etc. These studies are highly laborious and consequently not many examples are known. In his comprehensive review of 1997, Rieseberg listed only 7 confirmed examples of homoploid species (in *Helianthus*, *Iris, Paeonia, Pinus* and *Stephanomeria).* However, he asserts that there are about 50 putative examples and more recent studies have suggested or even confirmed more cases of homoploid hybrid species e.g. in *Gossypium* (Wendel et al., 1995a), *Armeria* (Aguilar & Feliner, 2003), *Viburnum* (Donoghue et al., 2004), *Pleione* (Gravendeel et al., 2004), *Argyranthemum* (Borgen et al., 2003), *Stephanandra* (Oh & Potter, 2003), *Solanum* (Hawkes, 1990) and *Hippophae* (Sun et al., 2003).

Although some studies described hybridization in angiosperms as ubiquitous and uniform, this view was not shared by Ellstrand et al. (1996) in a comprehensive review on the frequency of "spontaneous" (interspecific) hybrids. They examined five floras and counted all spontaneous interspecific hybrids and hybrid species (including allopolyploids). The main conclusion was that spontaneous hybridization is not ubiquitous among plant families. It was found to be restricted to certain plant families (between 16-34% of all families have at least one reported hybrid) and plant genera (6-16% of the genera have one or more reported hybrids). Unfortunately, they did not distinguish between "spontaneous" interspecific hybrids and hybrid species, making it difficult to distinguish between the frequency of hybrid individuals and long term consequences of hybridization, i.e. the stabilization of hybrids leading to a new species. Another study by Mallet (2005) on hybridization rates was based on the number of species involved in hybridization, with an estimated rate of 25% in vascular plants, based on UK data. Allopolyploids and other "stable species" of hybrid origin were excluded and these counts do also not give an indication about "long-term" consequences.

It seems more reasonable to look at the "result of hybridization" and the number of allopolyploid species and homoploid hybrid species. Only if hybridization events lead to a new stable evolutionary lineage they are important in evolution (Arnold, 1997), and hence a high frequency of (spontaneous) hybrids does not necessarily imply an important role. In a review on the extent of introgression in plants Rieseberg & Wendel (1993) listed 165 proposed cases of introgression and concluded that introgression plays a major role in the evolution of species. They based this conclusion on the large number of confirmed cases of introgression having contributed to genetic diversity and to the origin of new plant species. However, only if the process causes a faster evolution rate or novel directions it can be related to evolutionary importance.

Although they found evidence that occurrence of polyploidy may have extensively influenced the tempo and mode of evolution, Otto & Whitton (2000) could not find convincing evidence that polyploidization itself has a significant effect on patterns and rates of diversification. For example, they found a small positive correlation between the amount of polyploidy and species richness (at genus level). This can also be explained, however, by the observation that older lineages often have more species and at the same time more polyploids, merely by its older age. It would be more useful to study sister genera with different ploidy levels, but these are hard to find. Moreover, it is very hard to say whether new traits in duplicate genes of polyploids would also have evolved in the absence of gene duplication. To prove an adaptive role of these gene duplications, multiple independent transitions to polyploids would be needed to assess whether increased rates of speciation follow polyploid events (Otto & Whitton, 2000).

While there are some studies on specific taxa where hybridization played a major role in adaptive evolution (e.g. in *Helianthus*, Rieseberg et al., 2003), the role of hybridization in generating new plant species in general remains elusive. However, there are many recent studies where hybrids are included or hybrid speciation investigated, for instance recent studies focussing on the role of hybridization and polyploidization within the Brassicaceae (summarized in Marhold & Lihova, 2006). Additionally, in the next decades new technologies within plant evolutionary biology will become widely available, and create for instance opportunities to study speciation at the level of genome and transcriptome (Hegarty & Hiscock, 2005). These new molecular approaches will facilitate the study of genetic processes in hybrid speciation, and will provide more insights and hopefully help us understand how important the evolutionary role of hybrids actually is and has been.

chapter 3

RECONSTRUCTING PATTERNS OF RETICULATE EVOLUTION IN ANGIOSPERMS: WHAT CAN WE DO?

Bastienne Vriesendorp & Freek T. Bakker

Taxon 54; 593-604, 2005

Hybridization is thought to be an important phenomenon in angiosperm evolution and it has been suggested that a majority of all plant species may be derived from past hybridization events (e.g. Raven, 1976; Stebbins, 1959; Grant, 1981; Arnold, 1997). In addition, there is an increasing interest in the reconstruction of reticulate patterns (e.g. Linder & Rieseberg, 2004), with increased emphasis on the need to explore multiple independent markers to investigate the origin of putative hybrid species (e.g. Hamzeh & Dayanandan, 2004; Koontz & al., 2004 and other examples listed in Table 3.1). Several reviews have recently been published on the process of hybridization itself or on issues indirectly related to it: Hegarty & Hiscock (2005) and Zhou & al. (2005) present an overview of molecular techniques as well as criteria for distinguishing hybrid speciation; Gross & Rieseberg (2005) evaluate the ecological genetics of homoploid hybrid speciation, and Seehausen (2004) reviewed the possible role of hybridization in adaptive radiation. Mallet (2005) presented several hybrid examples in plants and animals to discuss the evolutionary significance of hybridization. Also, many studies have been published in the past several years on the incidence and role of (allo)polyploidy in evolution (e.g. in Soltis & Soltis, 1993, 2000; Soltis & al., 2004; Ramsey & Schemske, 1998; Otto & Whitton 2000 and Crawford & Mort, 2003).

In addition to (species-level) hybridization, other (genome-level or molecular) evolutionary processes such as recombination, gene conversion or horizontal gene transfer can confound the phylogenetic signal in the data to such an extent that it may become non-treelike, and phylogenetic methods are not appropriate for analysis. It is best to check prior to phylogenetic analysis whether the data are non-treelike, and if so, then apply network methods to represent it (Bryant & Moulton 2004).

Nevertheless, the pages of botanical systematic journals are still remarkably devoid of examples of reticulate patterns, and plant species-level relationships are predominantly depicted as trees. The question can be asked whether this is because there are no suitable tools available for detection or whether the problem is merely ignored. In this paper we will explore the current practice of dealing with hybrid terminals in published phylogenetic studies, briefly describe a selection of network-producing methods currently available, as well as discuss future possibilities of reconstructing reticulate patterns in angiosperm evolution.

Many recently published studies report the occurrence of plant hybrids in several plant genera and families, both at the polyploid and homoploid level (e.g. *Spartina* [Poaceae] in Ainouche & al., 2004; *Actinidia* [Actinidiaceae] in Chat & al., 2004; *Glycine* [Fabaceae] in Doyle & al., 2004; *Phoenix* [Arecaceae] in Gonzalez-Perez & al., 2004; *Pleione* [Orchidaceae] in Gravendeel & al., 2004; *Gagea* [Liliaceae] in

Peterson & al., 2004). Ellstrand & al. (1996) surveyed frequency and taxonomic distribution of spontaneous hybridization in vascular plants in five major floras. They concluded that most hybrids are concentrated in particular families such as Poaceae, Cyperaceae and Rosaceae, and several genera within these families account for most of the hybrid species encountered (Ellstrand & al., 1996). Some life-history characteristics seemed to be associated with hybridizing taxa such as perennial habit, asexual reproductive modes and outcrossing breeding system (Ellstrand & al., 1996; Rieseberg, 1997; Wisseman & Ritz, 2005). However, it is unclear whether the observed uneven distribution is due to intrinsic (biological) differences of the lineages involved (such as breeding system or ecological preferences), or to extrinsic factors such as extreme habitat or ecological transitions (Gross & Rieseberg, 2005; Rieseberg, 1997), or distribution pattern (the extent of sympatry with other species). Also, sampling bias could be a factor with the number of reports on hybrids influenced by the systematic attention given to particular taxa (Ellstrand & al., 1996).

When hybrid species are included in phylogenetic analyses they can affect the overall tree topology. Remarkably few studies on the behaviour of hybrids in cladistic analyses have been published. The landmark studies of McDade using artificial hybrids of *Aphelandra* (Acanthaceae) are often cited to indicate that a hybrid is not expected to disrupt phylogeny reconstruction unless the hybridization event is between divergent lineages (McDade, 1990, 1992). However, this is based on a data set consisting of morphological markers that are mostly intermediate in state for the hybrid, causing it to be placed at a basal position relative to the most derived parent. In contrast, molecular characters do not express intermediacy but can display apomorphies of both parents simultaneously (i.e. polymorphic sites or mosaic sequences) that may cause the hybrid to be placed proximate to the most derived parent. Many such apomorphic characters shared between hybrid and parents could cause long-branch attraction (McDade, 1995). Biparentally-inherited markers expressing additivity will possibly influence tree topology (loss of resolution), tree length (increase or decrease depending on treating additivity as polymorphism or uncertainty, respectively; see Kornet & Turner, 1999), or support analysis (Simmons, 2001). Many molecular phylogenetic studies use multiple markers with different modes of inheritance (i.e. nuclear and organelle). In fact, Seehausen (2004) uses the ensuing "cytonuclear discordance" as evidence for ancestral hybridization preceding evolutionary radiation.

Phylogenetic studies are sometimes conducted excluding putative hybrids in order to avoid (a priori) expected disruptive effects on the analyses (e.g. *Cardamine* [Brassicaceae], Marhold & al.,2004 and *Calopogon* [Orchidaceae], Goldman & al.,

Table 3.1. List of representative phylogenetic studies in angiosperms which include putative hybrids.

Genus	References	Species/ acc.[1]	Ploidy level of putative hybrid	Markers used	External evidence on hybrid status[2]
Achillea	Guo & al., 2004	63/82	polyploid	nrDNA ITS; cpDNA trnL-F	morph; AFLP
Actinidia	Chat & al., 2004	40/79	mix	mtDNA nad; cpDNA matK, psbC-trnS, rbcL, trnL-F	--
Amelanchier	Campbell & al., 1997	19/26	polyploid	nrDNA ITS	morph; pl
Anacamptis	Bateman & Hollingsworth, 2004	3/4	diploid?	morphology; nrDNA ITS; cpDNA trnL-F; RFLP	geo
Arabis	Koch & al., 2003	3/402	triploid	nrDNA ITS; chromosome counts	morph
Armeria	Aguilar & Feliner, 2003	72/131	diploid	nrDNA ITS	morph; geo; DNA
Calopogon	Goldman & al., 2004	5/56	polyploid	cpDNA restriction analysis; nrDNA ITS, AFLP; chromosome counts	--
Cardamine	Marhold & al., 2004	17/36	diploid	nrDNA ITS; AFLP	--
Cardamine	Lihova & al., 2004	22/22	tetraploid	nrDNA ITS; cpDNA trnL-F	pl
Ceanothus	Hardig & al., 2002	4/23	diploid	nrDNA ITS; cpDNA matK; allozymes; morphology	morph
Cicer	Shan & al., 2005	9/146	diploid	AFLP	morph
Dactylorhiza	Shipunov & al., 2004	9/125	mix	cpDNA trnL-F, trnS-G, nrDNA ITS; morphology	--
Delphinium	Koontz & al., 2004	30/30	diploid	nrDNA ITS; cpDNA trnL-F	morph; cross
Dendrochilum	Barkman & Simpson, 2002	22/22	mix	nrDNA ITS; cpDNA accD	morph; geo
Elymus	Mason-Gamer, 2004	33/45	hexaploid	cpDNA rpoA, trnT-L; nDNA GBSSI	pl
Elymus	Helfgott & Mason-Gamer, 2004	27/27	tetraploid	nDNA pepC	iso
Erythronium	Allen & al., 2003	24/24	tetraploid	nrDNA ITS; cpDNA matK	morph; iso
Fagopyrum	Nishimoto & al., 2003	15/15	tetraploid	nDNA Flo/Lfy, AG; cpDNA rbcL-accD, trnK, trnC-rpoB	--
Gagea	Peterson & al., 2004	7/32	diploid?	cpDNA psbA-trnH & trnL-F; nrDNA ITS; morphology	morph
Hippophae	Sun & al., 2002	15/15	diploid	nrDNA ITS	morph; geo; DNA
Hordeum	Petersen & Seberg, 2004	28/30	tetraploid	nrDNA DMC1, EF-G; cpDNA rbcL	pl; iso

Table 3.1. continued.

Genus	References	Species/ acc.[1]	Ploidy level of putative hybrid	Markers used	External evidence on hybrid status[2]
Lepidium	Mummenhoff & al., 2004	56/56	polyploid	cpDNA trnT-L, trnL-F; nrDNA ITS	morph
Mimulus	Beardsley & al., 2004	18/18	diploid?	nrDNA ITS, ETS; cpDNA trnL-F	morph
Miscanthus	Hodkinson & al., 2002	3/5	triploid	nrDNA ITS; AFLP; cpDNA; FISH	pl
Mitella	Okuyama & al., 2005	12/66	diploid[3]	nrDNA ITS & ETS; cpDNA matK, trnL-F	--
Nicotiana	Chase & al., 2003	66/70	polyploid	nrDNA ITS; GISH; cpDNA matK	pl, morph
Paeonia	Sang & al., 1995	33/45	mix	nrDNA ITS	--
Paeonia	Sang & al., 1997	32/37	tetraploid	cpDNA matK; nrDNA ITS	pl
Paeonia	Sang & Zhang, 1999	12/12	tetraploid	nDNA Adh1	pl
Pleione	Gravendeel & al, 2004	20/20	?	morphology; nrDNA ITS; cpDNA trnT-L, trnL-F, matK	morph; geo
Populus	Hamzeh & Dayanandan, 2004	21/21	?	cpDNA trnL-F; nrDNA ITS	RFLP
Ranunculus	Hörandl & al., 2005	c. 200/c. 200	polyploid	nrDNA ITS	morph; pl; cross
Sphagnum[4]	Shaw & al., 2005	31/136	mix	nrDNA ITS; nDNA Leafy/Flo; cpDNA trnL-F; RAPD	iso
Stephanandra	Oh & Potter, 2003	9/17	diploid	nDNA Leafy; nrDNA ITS; cpDNA trnL-F, trnD-Y-E-T, matK-trnK	--
Stylosanthes	Vanderstappen & al., 2002	28/40	tetraploid	STS[5]; nrDNA ITS; cpDNA trnL intron	pl; morph
Tarasa	Tate and Simpson., 2003	27/27	polyploid	cpDNA psbA-trnH, trnT-L, matK-trnK; nrDNA ITS	--
Viburnum	Donoghue & al., 2004	42/43	diploid	cpDNA trnK; nrDNA ITS	geo
Viburnum	Winkworth & Donoghue, 2004	41/41	polyploid	nDNA GBSSI	--
Zaluzianskya	Archibald & al., 2005	23/28	?	nrDNA ITS; cpDNA rpl16, trnL-F	--

[1] Number of ingroup species (including hybrids) /accessions used

[2] Evidence is sometimes inferred from the publications, i.e. not stated explicitly by the authors; morph = morphology; pl = ploidy level; cross= crossing experiments; geo= biogeography; iso=isozymes; DNA= "DNA evidence" (not specified)

[3] Evidence of introgression between taxa; no specific hybrid taxon is identified.

[4] Bryophyta

[5] STS: nuclear sequence-tagged site PCR

2004) or after determination of incongruence between different gene data sets (e.g. *Gaura*, [Onagraceae], Hoggard & al., 2004 and *Pleione*, [Orchidaceae], Gravendeel & al., 2004). In addition, several authors have analysed their data both including and excluding the putative hybrid, in order to investigate its influence on phylogenetic reconstruction. The effects of hybrid exclusion from nrDNA ITS data sets was investigated in *Achillea* (Asteraceae) by Guo & al., 2004; *Armeria* (Plumbaginaceae) by Aguilar & Feliner (2003); *Delphinium* (Ranunculaceae) by Koontz & al. (2004); *Hippophae* (Elaeagnaceae) by Sun & al. (2002); and *Nicotiana* (Solanaceae) by Chase & al. (2003). While the *Delphinium* and *Nicotiana* studies recorded little effect of hybrid exclusion on the analysis, the two other studies found fewer most parsimonious trees with a higher consistency index and a higher resolution in the analysis upon hybrid exclusion. In *Bikinia* (Fabaceae) exclusion of a putative hybrid caused an increase in jackknife support values for both clades containing the parental species, from 68% to 93% and from less than 50% to 77%, using AFLP data (Wieringa & Guhl, 2005). The authors argue that this taxon jackknifing approach could possibly be used as a standard tool to trace undetected hybrids.

The effect of hybrid exclusion in a combined analysis of nrDNA ITS and chloroplast DNA RFLPs was investigated in *Calopogon* (Goldman & al., 2004). No effect of removal of the putative hybrid (inferred from its ploidy level) on the combined analyses was found. In contrast, Hoggard & al (2004) studied two tetraploid species of *Gaura* (Onagraceae) and found a disruptive effect on tree topology of a putative hybrid with distant parents, while no effect was seen for another hybrid with "close" parents.

Cytonuclear incongruencies have confirmed several hypotheses of suspected hybrids, for example, *Anacamptis* (Orchidaceae) by Bateman & Hollingsworth (2004); *Delphinium* (Ranunculaceae) by Koontz & al. (2004); and *Dendrochilum* (Orchidaceae) by Barkman & Simpson (2002). Comparison of discordant phylogenetic trees from independent data sets has even revealed new unexpected cases of possible hybridization (e.g. *Braya* [Brassicaceae], Warwick & al., 2004; *Stephanandra* [Rosaceae], Oh & Potter, 2003; and *Viburnum* [Adoxaceae], Donoghue & al., 2004). In addition, this approach appears promising in phylogeography (see Comes & Abbott, 2001; Franzke & al., 2004; Lorenz-Lemke & al., 2005).

Of course, incongruent phylogenetic patterns within a data set or between data sets can have causes other than the hybrid origin of one or more of the species involved. Such causes may include incomplete lineage sorting, that is, the persistence and retention of ancestral polymorphisms through multiple speciation events (e. g. Avise, 2000; Comes & Abbott, 2001; Andreasen & Baldwin, 2003; Goldman & al.,

2004), homoplasy and taxonomic sampling error (Wendel & Doyle, 1998). Therefore, hybridization should not be a "standard" interpretation when incongruencies are found. Other causes should be considered carefully because the incongruent pattern alone can never be an indicator of hybrid status.

Many examples of hybrid detection involve investigation of the additivity of nucleotides at single positions (polymorphic sites) of rDNA ITS sequences (e.g. Gravendeel & al., 2004; Koontz & al., 2004; Marhold & al., 2004; Peterson & al., 2004; Sun & al., 2003; Warwick & al., 2004). An example of intraspecific ITS additivity can be found in *Clausia aprica* (Brassicaceae) where accessions of an intermediate group showed additivity, possibly indicating hybridization (Franzke & al., 2004).

Furthermore, hybrid origin and relationship to putative parents, when not extinct, can be explored in more detail using several different markers. Morphology or patterns of geographical distribution can provide valuable additional evidence for a hybrid origin (Hughes & Harris, 1998; Bateman & Hollingsworth, 2004; Peterson & al., 2004; Shan & al., 2005). Additionally, karyological evidence, such as chromosome counts, C-values, and GISH or FISH patterns can discriminate between parental genome donors and the hybrid relationships (Hodkinson & al., 2002; Borgen & al., 2003; Chase & al., 2003; Bures & al., 2004; Harper & al. 2004; Pires & al., 2004; Tel-Zur & al., 2004). Another line of evidence for hybrid status can be found using analyses of fragment-length polymorphisms (e.g. RFLP) or the currently more often employed PCR-based markers (e.g. AFLP, ISSR, RAPD, or PCR-RFLP). For example, an additive pattern of AFLPs and the lack of unique bands confirmed the hybrid status of a species of *Mangifera* (Teo & al., 2002). Kiew & al. (2003) used AFLP data to test hybrid origin in several taxa (*Begonia, Mangifera, Nepenthes* and *Lausium*) and these data permitted the reconstruction of relations with the putative parents. Despite many examples where AFLP data are considered useful in phylogenetic studies (e.g. Kardolus & al., 1998; El-Rabey & al., 2002; Spooner & al., 2005), the application of these data (and similar single-locus markers) in infrageneric studies requires caution. One major concern involves the difficulty in assessing homology between the co-migrating fragments of more distant taxa (El-Rabey & al., 2002; but see Crawford & Mort, 2004). Therefore, the general value of the use of these markers in assessing hybrid origin remains questionable as only the successful cases tend to get published (but see Krauss & Hopper, 2001, who report that high genetic variability made it difficult to distinguish between different hybrid scenarios). More insight is needed into the "behaviour" of ALFPs, and to this end, simulation (*in silico* AFLP, see Koopman & Gort, 2004) may become increasingly important as more

complete genome sequences become available (Antonov, 2002). Recent studies of hybrids in angiosperm phylogenies are listed in Table 3.1, with the different hybrid detection markers used.

Ideally, additional studies, such as crossing experiments, need to be conducted to support any hybrid hypothesis. Artificial crossing experiments permit investigating the possibility of crossing of the putative parents, and also comparison of the character pattern (either molecular or morphological) in progeny of controlled crosses with that of putative hybrids. For instance, experimental crosses have been used to investigate morphology and fertility in *Solanum* (Clausen & Spooner, 1998); to compare nrDNA ITS sequences between artificial and natural hybrids in *Begonia* (Chiang & al., 2001); to determine the maternal donor of F1 hybrids in *Phlox* (Ferguson & al., 1999); and to study genomic changes in synthetic polyploids of *Brassica* (Song & al., 1995). Even more extensive studies have been performed in sunflower hybrid species, where the genomic structure of a newly formed hybrid was compared with that of ancient hybrids to study the process of diploid hybrid speciation in *Helianthus* (Rieseberg & al., 1996). In later studies, adaptive quantitative trait loci (QTL) were compared to investigate ecological divergence and adaptive genetic variation of the hybrids (Lexer & al., 2004; Rieseberg & al., 2003). While such genomic evidence can be regarded as the best and most direct evidence for documenting the hybrid nature of a species, as well as allowing assessment of the actual mechanisms involved, such data will probably never be available for most groups on a routine basis.

HYBRID DEFINITIONS

The range of possible characteristic hybrid patterns listed above (e.g. additivity of AFLP bands, polymorphic nucleotides, incongruence between gene trees, intermediate morphology, etc.), may well not apply to each hybrid plant species as not all will "behave" in the same way. Rieseberg & Ellstrand (1993) investigated chemical, morphological and molecular characters in hybrid plants and found that hybrids can display a range of characteristics at both the morphological and molecular level. This ranges from closely resembling one parent to complete intermediacy between the parentals, and in some cases to the formation of a completely new character. The authors emphasised the unpredictable nature of character expression in hybrids, hence preventing hybrid detection based on a specific "hybrid character syndrome". Many other studies corroborate these findings with different character patterns found in different hybrids (Table 3.1). To our knowledge, no recent review on patterns in hybrid characters has been performed, comparable to the list of Rieseberg & Ellstrand (1993).

The other reason that no general pattern in hybrids can be inferred from published studies lies in inconsistent terminology. For example, Rieseberg and Ellstrand (1993) discriminate between "first generation hybrids", "later generation hybrids" and "hybrid species". According to the authors, the latter category is the most difficult to detect, since hybrid species are more prone to display many new and extreme characters, while F_1 hybrids will probably more often show a blend of characters of the parental species. McDade (1995) reviewed character patterns in hybrids and introduced the terms "primary hybrids" ("with simple histories and little change since origination") and "derived hybrids" (with "considerable evolutionary change since origination"). She used these categories to indicate that the amount of evolutionary change will probably define whether hybrids can be dealt with in systematics. "Primary hybrids" are the only category where we can expect to understand the behaviour of their characters (McDade, 1995).

Additionally, the term "hybrid" is used for a wide variety of entities (McDade, 1995), often without reference to important factors that must be considered, such as age, ploidal level and parental phylogenetic distance. Most systematic studies dealing with hybrids do not explicitly state what hybrid definition is used, but simply assume that a hybrid is a cross between different species, or define it as "interspecific" or "hybrid between species". There are, however, some studies that explicitly refer to the age of the hybrid or "stability" of the hybrid individuals. For example Bures & al. (2004) specify that they include (sterile) F_1 hybrids in their study of *Cirsium*; Koontz & al (2001) and Goldman & al., (2004) both discuss the possibility of ancient hybridization in respectively *Delphinium* and *Calopogon*, but these are exceptions.

The term "hybridization" is rarely specified, instead it is assumed that every worker knows what is meant, but it is important to note that several definitions exist. The most often used one is by Harrison (1993): "interbreeding of individuals from two populations, or groups of populations, which are distinguishable on the basis of one or more heritable characters." This definition does not require any consideration of species concepts, but most workers use the "standard" definition of a hybrid "resulting from crossing between different species". In an attempt to clarify matters, we include here a conceptual framework in which various hybrid definitions are logically arranged according to factors and scales that are of importance in hybrid formation, and hence in character evolution (Fig. 3.1). The two main axes here are "age of the hybrid" (whether it is a newly formed (F_1) hybrid or a more ancient and established hybrid species), and the taxonomic level of the hybrid's parents. The latter is further subdivided into relevant mechanism(s) involved during or after the process of

hybridization, such as the amount of introgression and change in ploidy level. As McDade (1995) noted, the pattern of character state transmissions in hybrids and the amount of evolutionary change in characters are important for possible detection and behaviour of hybrid terminals in phylogenetic studies. Therefore, we include these factors here as well, and nest them within the different time scales.

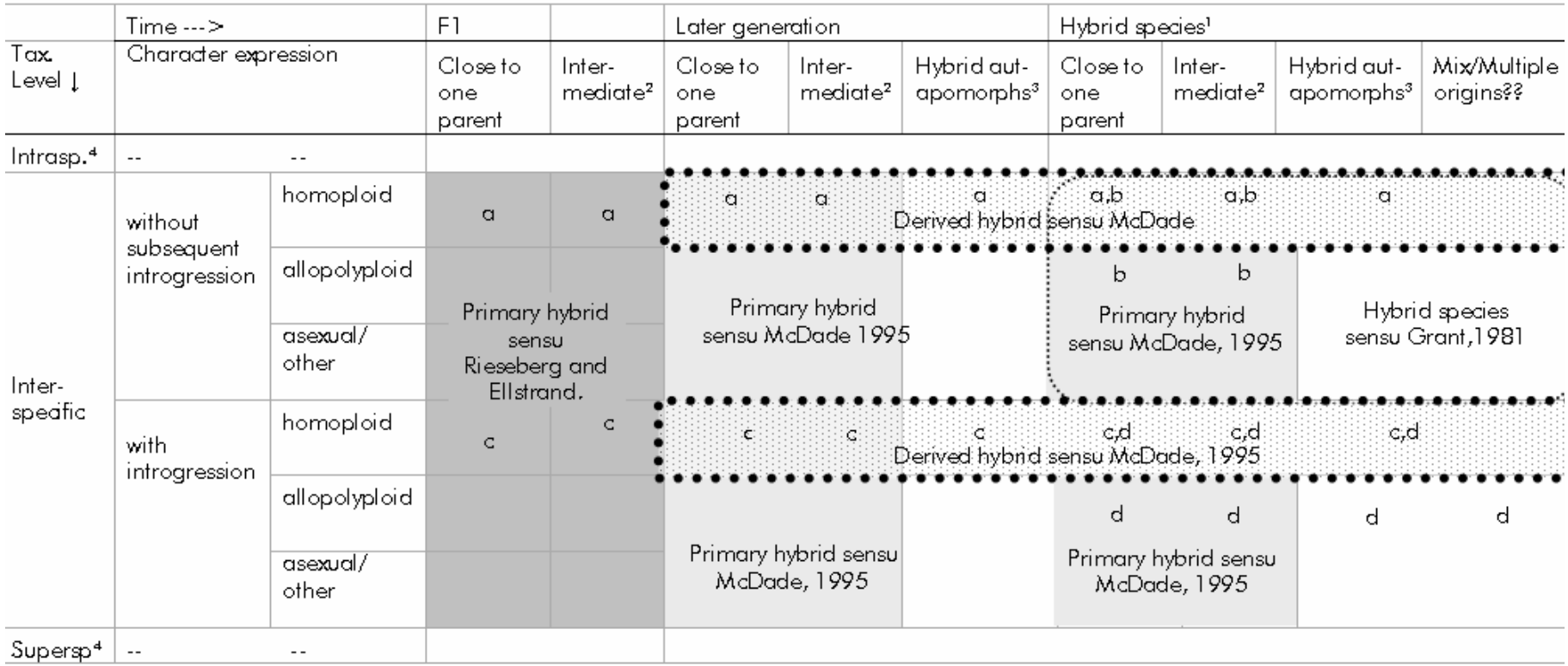
REPRESENTING HYBRIDS IN PHYLOGENETIC ANALYSIS: VISUALISATION OF PATTERNS

As mentioned above, hybrids are sometimes excluded from phylogenetic analyses, often after incongruence testing among multiple data sets, or because the hybrid is expected to have a disruptive effect on the tree topology (e.g. Marhold & al., 2004). This approach intuitively makes sense because trees cannot depict hybrids and tree reconstruction could be confounded by their inclusion, with a polytomy a likely result (but see below). Moreover, when using packages such as Mesquite (Maddison & Maddison, 2004) or MacClade (Maddison & Maddison, 2005) for optimisation of characters, resolved trees are usually required as input. However, it would be a waste of potentially important information if the hybrid sequences were not used in such analyses. In addition, inclusion of the hybrid might not have a disruptive effect on the trees after all (see Chase & al., 2003; Guo & al., 2004; Koontz & al., 2004).

As outlined above, one solution to this problem could be to conduct analyses that both include and exclude putative hybrids and present both results (as in Sun & al., 2002; Aguilar & Feliner, 2003; Chase & al., 2003; Guo & al., 2004; Koontz & al., 2004; Wieringa & Guhl, 2005). An alternative, and perhaps better approach might be to represent hybrid relationships directly in a network. This can be done by hand for relatively clear hybrid relationships (Sang & al., 1995; Hardig & al., 2002). However, for more complex situations this is not feasible and depicting reticulate evolutionary patterns (and all possible relationships) in one network would be a desirable feature in a computer package.

In analogy to the gene tree/species tree problem (Maddison, 1997) however, the question must be addressed of what is actually represented in such reticulated

Fig. 3.1. How to be(come) a hybrid? Conceptual framework of commonly-used hybrid terminology with exemplar studies indicated.



[1] Stabilised lineage/ancient hybrid
[2] Can be either mosaic or additive characters
[3] "New" characters
[4] Does not apply as intra- and superspecific hybridization falls outside standard hybrid terminology

[a] *Helianthus* (Rieseberg et al., 1996, 2003)
[b] *Paeonia* (Sang et al., 1995)
[c] *Iris* (Arnold, 1993, 1997; Arnold et al., 1990, 1991)
[d] *Gossypium* (Wendel et al. 1991, 1995a, 1995b; Cronn et al., 2003

networks: character conflict or relationships among organisms (or species). For instance, Bryant & Moulton (2002) characterise networks as “a representation of the data rather than a phylogenetic inference", and to "indicate whether or not the data is substantially tree-like". Holland & al. (2005) on the other hand describe their Consensus Network (see below) method as one “that generalizes the notion of consensus trees to allow conflicting evolutionary hypotheses to be displayed within a network”.

While one would certainly like to distinguish between sources of phylogenetic tree incongruence, it is unlikely that the cause of the conflict can be inferred by analysing the pattern alone. Both homoplasy/sampling artifacts and hybridization (or other evolutionary processes) can give the same (sometimes incongruent) patterns. Also, hybridization need not necessarily result in a clear reticulated pattern of evolution. The data can not indicate any incongruence, for instance, when using uniparentally inherited markers or markers that show strong gene conversion. In addition, processes during or after hybridization (such as repeated secondary contact, as for instance under a post-glacial refugium expansion scenario), can make the actual split “messy”, and hence the relationships more complex and difficult to resolve. Nevertheless, Seehausen (2004) argues that “adaptive radiation” (or at least “functional diversification”) can be facilitated by interspecific hybridization and that such patterns can be clearly reconstructed.

Probably the best way of distinguishing between the different above-mentioned causes for the observed phylogenetic incongruences is therefore to use additional data or evidence of hybridization from other sources such as morphology, genomics and karyology. Yet, in spite of the objections outlined above, a network can be used as a starting point for investigating relationships. Whether or not hybridization is the cause, it is desirable to display the source of conflict. One way to do this is to visualise the character incongruences in a network, where a “hybrid” or “problematic terminal” can be connected to more than one other terminal or internode.

Unfortunately, although some programs exist that can deal with population-level data and possibly hybridization events (see below), no method is available that can be considered the perfect “hybrid interpreter”. The only way to make progress with this problem is to seek methods to deal with complex hybrid terminals using simulation and experimental data. The common practice of leaving suspected taxa out of the analysis to avoid confounding effects on phylogenetic reconstruction will not stimulate further progress. Below we explore some of the methods currently available and infer possible solutions and suggest some future research directions.

CURRENT TOOLS FOR REPRESENTING INCONGRUENT PHYLOGENETIC PATTERNS SIMULTANEOUSLY

Posada & Crandall (2001a) listed a range of methods and software for network estimation that can possibly "take into account population-level phenomena and allow for persistent ancestral nodes, multifurcations and reticulations". Examples are the method of statistical parsimony (as implemented in the package *TCS*, Clement & al., 2000), *SplitsTree* (Huson & Bryant, 2004), and *Network* (Bandelt & al., 1995, 1999) but so far, these methods have not been used frequently in published studies of angiosperm species phylogenies. In Table 3.2 we list currently available and accessible methods aiming at network reconstruction. Some of these programs are character-based, such as the *Median Network* and *Median-Joining network* approach of Bandelt (1995, 1999), but most other methods are distance based. Generally, programs can be distinguished by whether they are based on an algorithmic approach or use an optimality criterion. Most network methods are based on an algorithmic approach and do not explore alternative solutions. However, the *Median Network* approach displays all parsimonious solutions in one network, and it is not immediately clear whether an algorithm or criterion-based approach applies here.

Table 3.2. List of selected currently available network reconstruction packages.

Package	Reference	Network reconstruction method	Input data
Network[a]	Bandelt & al., 1995	Median networks[1]	Binary characters
		Median-joining networks[2]	Multistate characters
Spectronet[b]	Huber & al., 2002	Median networks[1]	Binary characters
Arlequin[c]	Schneider & al., 2000	Molecular-variance parsimony[3]	Multistate characters/ haplotype frequencies
SplitsTree[d]	Huson & Bryant, 2004	Split decomposition[4]	Multistate characters or distances[g]
		NeighborNet[5]	Multistate characters or distances[g]
		Consensus networks[6]	Trees [h]
T-rex[e]	Makarenkov, 2001	Reticulogram reconstruction[7]	Distances
TCS[f]	Clement & al., 2000	Statistical parsimony[8]	Multistate characters/ haplotype frequencies

[a] http://www.fluxus-technology.com/
[b] http://awcmee.massey.ac.nz/spectronet/index.html
[c] http:// lgb.unige.ch/arlequin
[d] http:// www-ab.informatik.uni-tuebingen.de/software/ jsplits/welcome.html
[e] http://www.labunix.uqam.ca/~makarenv/trex.html
[f] http:// darwin.uvigo.es/software/tcs.html
[g] Input data are multistate characters or distances, analysis is based on distances
[h] Input data are trees, analysis is based on splits of these trees

[1] Bandelt & al., 1995
[2] Bandelt & al., 1999
[3] Excoffier & Smouse, 1994
[4] Bandelt & Dress, 1992
[5] Bryant & Moulton, 2002, 2004
[6] Holland & Moulton, 2003
[7] Makarenkov & Legendre, 2004
[8] Templeton & al., 1992

Most of the programs listed in Posada and Crandall (2001a) aim at population-level data and there are several recent examples of their application. For example, multiple origins in *Glycine tomentella* were investigated by Rauscher & al. (2004) using *TCS*; representation of phylogeographic relationships in dolphins was reconstructed using *TCS, Network, Arlequin* and *SplitsTree* (Cassens & al., 2003); and patterns of genetic diversity were explored in *Scaevola plumieri* (Goodeniaceae) using *SplitsTree* (Barker & al., 2003). However, there are few examples of applications of using network programs at the angiosperm species-level. For instance, hybrid relationships in *Opuntia* were investigated with the *Median Network* approach of *Spectronet* (Griffith, 2003) and split decomposition (with the program *SplitsTree*) was used to study species radiation and reticulate relationships in *Ranunculus* (Lockhart & al., 2001; Hörandl & al., 2005). Also, *Median-Joining* networks (Bandelt & al., 1999) have been used for detailed analyses of introgression and hybridization zones between two species of *Populus* (Lexer & al., 2005) and *Passiflora* (Lorenz-Lemke & al., 2005).

In addition to the methods outlined above, there are several methods aimed specifically at detecting recombination (reviewed by Posada & Crandall, 2001b, 2002; Posada & al., 2002; Posada, 2002). These methods can only be used to test whether or not recombination is likely to be present in the data and do not display reticulate relationships. For instance, the new package TOPALi (see http://www.bioss.ac.uk/~iainm/topali/; Milne & al., 2004) is one of the available recombination detection programs that has several methods implemented to automatically identify recombinant sequences within DNA multiple alignments. One of these methods works by sliding a window along a sequence alignment, and measuring the discrepancy between the trees suggested by the first and second halves of the window, using distance matrix methods. If we could use these programs to "correct" phylogenetic data sets prior to phylogenetic analysis by scanning and removing recombined regions this could prove highly useful.

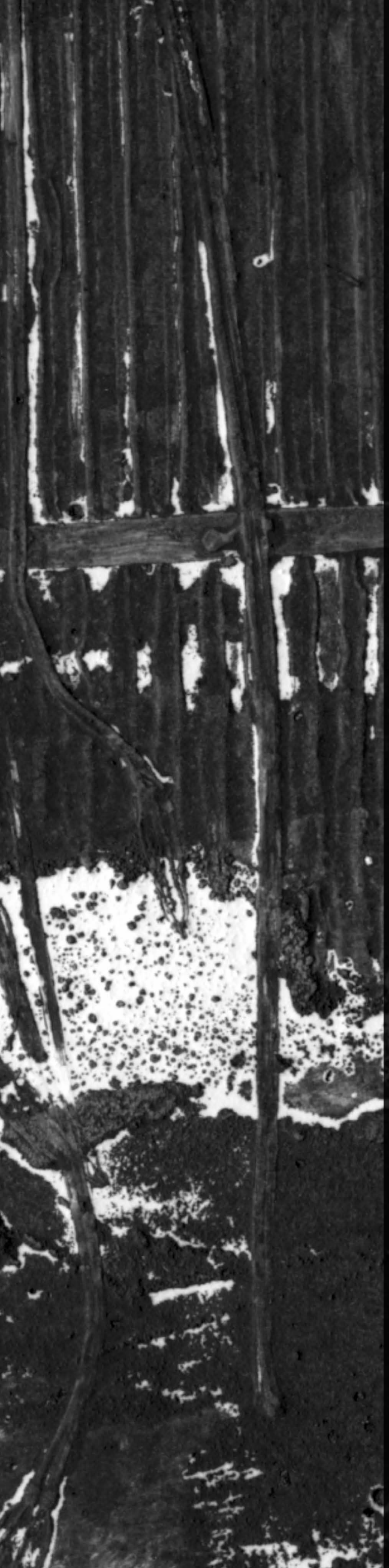
Ideally, one would like to test the performance of network reconstruction methods using simulated data as has been done for several other phylogenetic methods (e.g. Suzuki & al., 2002; Douady & al., 2003; Hall, 2005). However, since many network reconstruction programs do not have a batch mode, simulation can become a cumbersome enterprise (pers. obs. and L. Nakhleh, pers. comm.). More importantly, in network simulations it is not clear what test statistic to use when comparing networks to a simulated model network. Measurements such as the partition metric (Robinson & Foulds, 1981) as used in many simulation studies (e.g. Leitner & al., 1996; Zwickl & Hillis, 2002; Piontkivska, 2004), are not available for a

network. Nakhleh & al., (2003) tested the performance of *SpNet* and *SplitsTree*, using a modified version of the Robinson-Fould metric specifically designed and implemented for their experiments, but not suitable for wider use.

In another study, we will focus on the performance of selected network reconstruction methods using "real" published data (Vriesendorp & Bakker, in prep.). Here, we would like to note that despite the wealth of examples of hybrids in phylogenetic studies (Table 3.1), the successful use of network reconstruction seems to be restricted to the population level. This is understandable since phylogenetic DNA sequence data sets usually contain far higher levels of variation than what most methods are designed to deal with. Moreover, the consistency and relative performance of these methods at the species level is not well understood, as outlined above.

In conclusion, the question remains whether a network representation can give a meaningful representation of species relationships, and whether such an analysis provides added value with respect to tree analysis. It remains to be seen whether network representation can be a good alternative to placement of the hybrid by hand on a phylogenetic tree (e.g. Sang & al., 1995; Whitehouse, 2002). However, when networks are not interpreted as displaying evolutionary relationships, but instead represent character conflict in the data, these packages prove very useful. For instance, uncovering data ambiguity using NeighborNet or Consensus networks in a way that (consensus) trees cannot, is a valuable addition to our phylogenetic tool palette, providing new insights in the analysis of data structure.

Many network reconstruction methods are based on a combined approach and explore incongruencies within a combined data set (Posada & Crandall, 2001a; Linder & Rieseberg, 2004). Exploring and summarising separate data sets (or gene trees) could have preference above combining all data "a priori" and conducting simultaneous analysis. In their review, Linder & Rieseberg (2004) stress the importance of using multiple independent markers in the reconstruction of patterns of reticulate evolution in plants. The best approach appears to be to combine as many independent gene trees as possible into a species tree and infer hybrid relationships from there. This was done by Nakhleh & al. (2005) using their package RIATA-HGT which enables inference of horizontal gene transfer events based on analysis of incongruence among species and gene trees.



chapter 4

MOSAIC DNA SEQUENCES AND THEIR EFFECT ON PHYLOGENETIC TREE RECONSTRUCTION: SIMULATIONS INVOLVING RECOMBINATION AND HYBRIDIZATION

Bastienne Vriesendorp & Freek T. Bakker

Submitted for publication in Plant Systematics and Evolution

ABSTRACT

The effect of mosaic sequences on phylogenetic reconstruction is potentially disruptive, but poorly studied explicitly. We use simulated and published data from species-level angiosperm studies to investigate the effect of recombination on phylogenetic reconstruction. In addition to recombination, a scenario including hybridization and extinction was applied where instead of exchanging clades an extra lineage was inserted and/or parental lineages were removed.

Bayesian inference and Jackknife resampling-based topologies inferred from the simulated data sets were compared with the model tree topology using partition metric (*PM)* based quantitative accuracy measures. We also present a *PM* based measure for assessing tree topology conflict. Accuracy of tree topology recovery was significantly affected at 50% recombination between recent-divergent lineages, while the trees based on other levels and scenarios of recombination were less affected. An extra hybrid lineage, or extinction of lineages, only had a small effect on accuracy. Choice of tree topology/simulated model combination affected accuracy scores, with remarkably high conflict and *PM* values for the trees simulated under high rate heterogeneity ($\alpha=0.09$) compared with a low value ($\alpha=2.62$).

Finally, recombination detection methods were applied to our data sets to infer the simulated recombination events. With highly variable data sets, containing low rate heterogeneity ($\alpha>2$), detection of true recombination events was more accurate, although the amount of false positives was also relatively high. Reciprocal events provided more correctly identified events than non-reciprocal recombination.

Phylogenetic data sets could be "corrected" for these events prior to phylogenetic analysis, but no major improvement on *PM* values can be found compared with a "naive" approach.

KEY WORDS

Phylogenetic reconstruction, Hybridization, Recombination, Reticulate evolution, Tree comparison, Mosaic sequences

INTRODUCTION

In recent literature the notion has become prevalent that evolutionary reticulation can be a significant factor in clade proliferation and hence should be taken into account in phylogeny reconstruction (Spring 2003; Linder & Rieseberg, 2004; Hypsa, 2006; McBreen & Lockhart, 2006; Willis et al., 2006). While some authors focus on reticulation from a biological perspective, emphasizing underlying processes and

ecological importance of hybridization (Seehausen, 2004; Mallet, 2005), others have addressed more conceptual questions such as "do networks represent conflict in data or evolutionary patterns?" (Linder & Rieseberg, 2004; Huson et al., 2005; Vriesendorp & Bakker, 2005/Chapter 3; Huson & Bryant, 2006; McBreen & Lockhart, 2006). In addition, analytical questions such as i) what is the accuracy of network and phylogeny reconstruction (Cassens et al., 2003; Nakhleh et al., 2003; Morrison, 2005) and ii) how can a separate gene tree approach help in reconstructing hybrid speciation (Linder & Rieseberg, 2004), have been the subject of increased attention in the past decade. In this paper we focus on one particular analytical aspect of reticulation in the context of angiosperm phylogenetics, namely the influence of reticulation on accuracy of DNA sequence-based phylogeny reconstruction.

Because most phylogeny reconstruction algorithms assume hierarchical structure among species, problems can arise when processes such as recombination, hybridization or the sorting of ancestral polymorphisms generate reticulate relationships at the sequence-level (Wendel & Doyle, 1998; Posada & Crandall, 2001b; Linder & Rieseberg, 2004). This can result in terminals of a mosaic nature, where different parts of the alignment have different evolutionary histories, analogous to the gene tree/species tree problem (e.g. Maddison, 1997). The effect of including such mosaic terminals in phylogenetic tree reconstruction is not well known, due to a scarcity of experimental data. Simulation allows addressing this question and, in addition, may allow relating effects on tree topology to particular causes. For example genetic distance between the mosaic terminal and its parents, sequence variation, or length of sequences may correlate with phylogeny reconstruction.

Despite the above, only a few simulation studies have actually dealt with mosaic terminals in phylogeny reconstruction. Wiens (1998) investigated the effects of combining data sets generated under different genealogies in phylogenetic analyses allowing only bifurcations. As expected, he found that combining data sets can result in poor estimates of the underlying "true" trees when phylogenies had different histories, whereas these estimates improved when the gene genealogies agreed. Schierup & Hein (2000a, b) studied the effect of recombination on estimates of population genetic parameters, using simulations under a coalescent perspective. They found that ignoring recombination can lead to overestimation of branch length, underestimation of the age of the most recent common ancestor of the sequences and incorrect rejection of the molecular clock (Schierup & Hein, 2000a, b). Recombination is an important factor in population-level studies (e.g. Posada et al., 2002; Hedderson & Nowell, 2006; Houliston & Olson, 2006), but rarely addressed in species-level phylogenetic analyses. However, given the uncertainty of perceived angiosperm

taxonomic rank, i.e. whether the population or species level is actually being targeted, population-level processes such as recombination could have a major influence in such cases as well.

Posada & Crandall (2002) explored the effect of recombination on accuracy of phylogeny estimation based on recombinant sequences created by evolving the data along two different 8-taxon tree topologies. The authors tested the effect of recombination events on the accuracy of tree reconstruction based on these simulated data, measured as the probability of recovering the overall tree topology or of clades therein. Ruths & Nakhleh (2005) extended these simulations by using a 20-taxon tree. In addition, they assessed the effect of both single and multiple subsequent recombination events on tree topology, measured by partition metric (*PM*, Robinson & Foulds, 1981) distances between true and simulated trees. The *PM* between two trees is simply the number of "edges" (=internal branches) they have in common. Ruths & Nakhleh (2005) used the *PM* as a relative distance, i.e. difference in edges proportional to the total number of edges in the tree topologies.

Here, we study the effect of recombination on tree reconstruction using large, published phylogenetic tree topologies. We use "recombination" here in an operational sense to simulate patterns of cyto-nuclear incongruence often seen in angiosperm species-level studies, and not in a genetic sense, i.e. simulating within-locus exchange with two cross-over points. We consider different scenarios of recombination, as well as a scenario of "hybridization" where hybrid lineages are included in phylogenetic trees along with both parents. Furthermore, we assess the effect of the absence of one or both parental lineages on the accuracy of tree reconstruction, as this may well reflect actual angiosperm evolution. For instance, possible extinct parental lineages were described in e.g. *Eleusine* (Poaceae) (Neves et al., 2005), *Rosa* (Rosaceae) (Ritz et al., 2005), and *Cardamine* (Brassicaceae) (Lihova et al., 2006).

In cases where recombination at angiosperm species-level has been investigated (e.g. Barkman & Simpson, 2002; Koch et al., 2003; Beardsley et al., 2004; Devos et al., 2005; Howarth & Baum, 2005; Poke et al., 2006), the authors concluded that it could have possibly lead to "mosaic" sequences. Therefore, ideally, pre-phylogenetic analysis routine could include scanning and fixing of possible recombination events in the data, prior to subsequent phylogenetic analysis. We compare this approach to phylogenetic analysis with the more naive approach of knowing recombination is present but ignoring it, using a selection of published angiosperm data sets. In addition to this, we include testing of the accuracy of selected

recombination detection tools, using our simulated recombinant DNA sequence data sets.

MATERIALS AND METHODS

Simulation

Our procedure roughly follows Posada & Crandall (2002), but using larger and 'real' published angiosperm tree topologies, and improved output measures in order to achieve a more realistic perspective on the phenomenon. Two contiguous parts of a DNA sequence alignment were simulated along two tree topologies, that differed in the placement of a "recombinant" lineage, see Fig. 4.1 and below. Subsequently, simulated data sets were analyzed phylogenetically and resulting tree topologies compared with one of the two model topologies. All simulated sequences are generated using Seq-Gen version 1.3.2 (Rambaut & Grassly, 1997).

Tree topologies and branch lengths used for the simulations were based on published angiosperm species-level phylogenetic studies. Because only contemporary branches should be used as recombinant lineages, we chose to use ultrametric trees for our simulations. We are aware that ultrametrization could introduce artifacts (as ultrametric trees are rarely found in 'real' angiosperm species-level phylogenetic studies), and we tested for this by comparing ultrametric and non-ultrametric tree topologies with respect to resulting tree topologies, see below. Trees were made ultrametric using nonparametric rate smoothing (NPRS) as implemented in TreeEdit (Rambaut & Charleston, 2001), using the default NPRS rooting variant of "across root"; other rooting methods included in that package gave no significant differences in resulting (average) branch lengths. The ultrametric "model tree" (T_M in Fig. 4.1) was used as reference tree to compare the consensus trees inferred from the simulated data sets (T_C in Fig. 4.1) against. In order to use T_M as input tree for Seq-Gen, polytomies had to be arbitrarily resolved, using zero-length branches, as the program can only take fully-resolved trees. Subsequently, tree branches were moved or interchanged to create the different recombination events using Mesquite version 1.12 (Maddison & Maddison, 2006). A non-reciprocal recombination event implied a movement of the recombinant branch to another lineage, while reciprocal recombination involved two branches interchanging their position in the tree (see Fig. 4.1-I, T_{Rn} and T_{Rr}). Three possible recombination events were used (similar to Posada & Crandall, 2002): "ancient", between ancient lineages (A); "recent-divergent", between divergent lineages (RD), and "recent-close", between closely related lineages (RC), see Fig. 4.2.

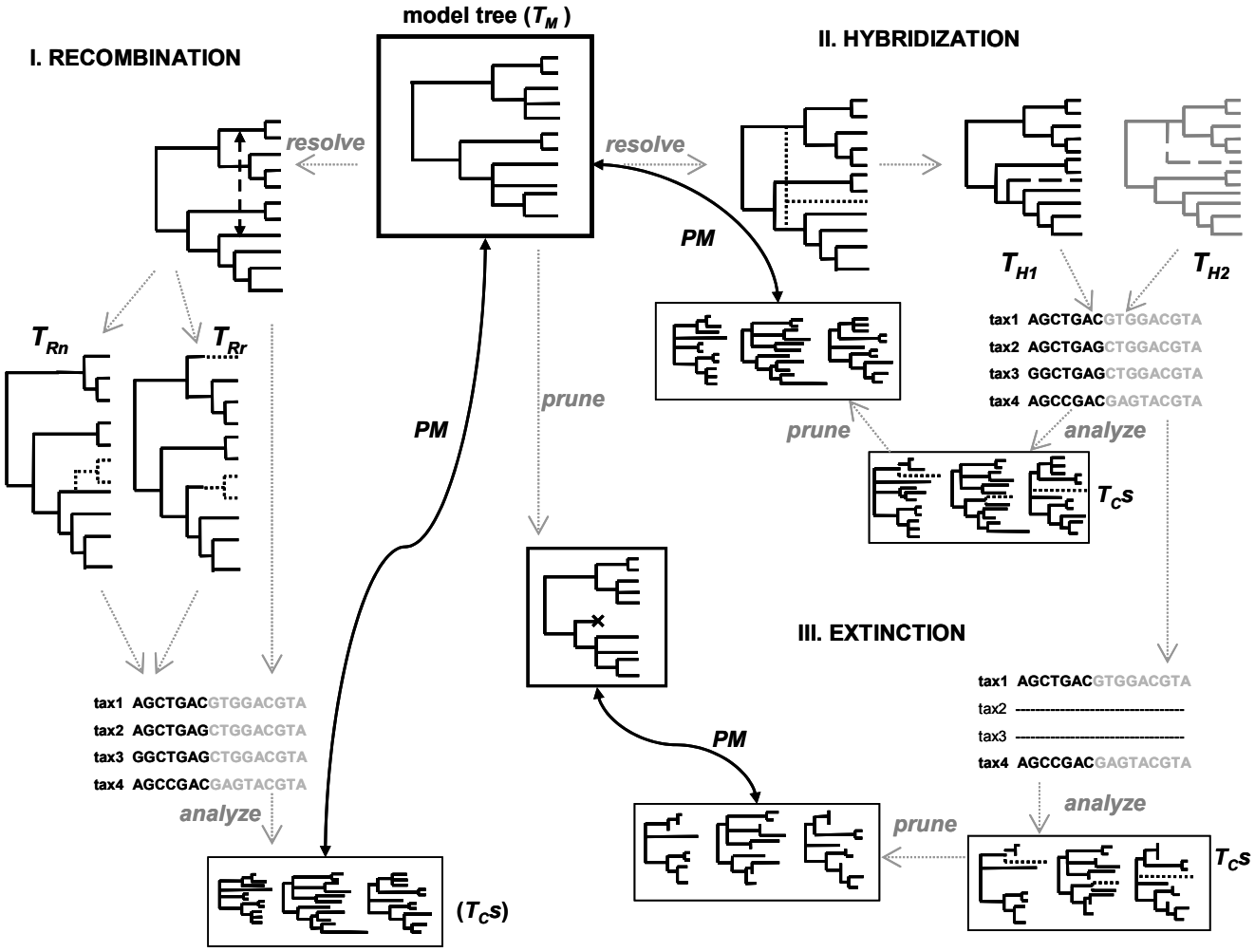


Fig. 4.1. Simulation set-up. Different simulation routes are shown, with on the left the recombination scenario (I), and on the right the hybridization (II) and extinction scenario (III). Dashed arrows indicate direction of flow-scheme. From the model tree T_M the unresolved nodes are arbitrarily resolved ("resolve") with subsequent modification of the input tree for the simulations. This results in two recombinant trees for the recombination (I) scenario, with non-reciprocal (T_{Rn}) or reciprocal (T_{Rr}) exchange of lineages. Simulation of DNA sequence data is based on two input trees, a recombinant tree (either T_{Rn} or T_{Rn}) and the (resolved) T_M. In the hybridization (II) scenario an extra lineage is inserted in the resolved T_M resulting in two trees, T_{H1} and T_{H2}, both input for the simulations. In the extinction scenario (III) parental taxa were removed from the same simulated data sets used in (II). Bayesian and Jackknife resampling analyses ("analyze") were performed on the simulated data sets and the consensus trees (T_C's) of all replicates were compared to T_M using the partition metric *(PM).* In (II) and (III) the hybrid lineages (indicated by dashed lines in the T_C's) were pruned from the consensus trees ("prune"). In addition, in the extinction scenario (III), the parental lineages were removed from the T_M before the *PM* comparison with the T_C's.

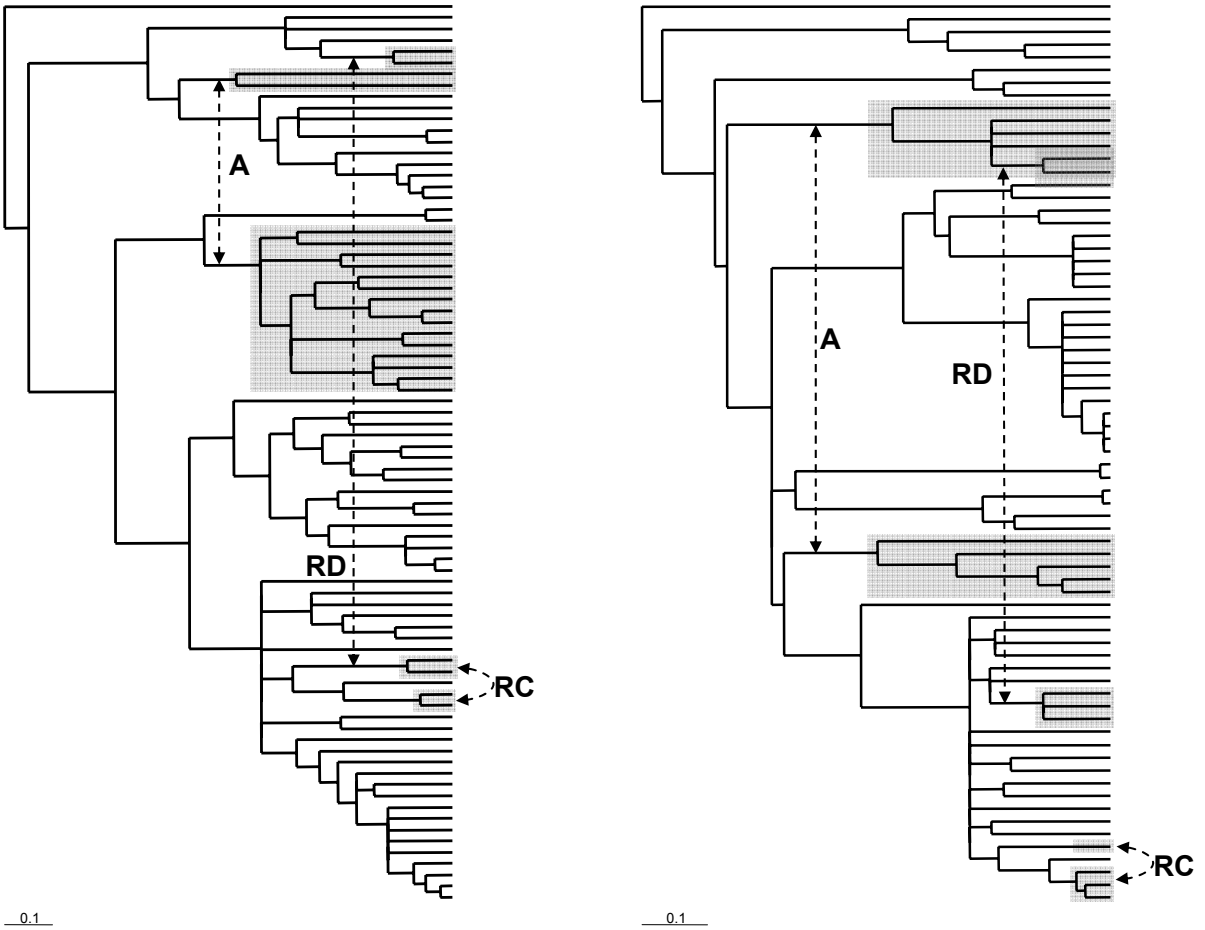


Fig. 4.2. Model tree topologies, $T_{M\ Pelargonium}$ from Bakker et al. (2004) and $T_{M\ Celtidoids}$ from van Velzen & Bakker (in prep.). Tree topologies are ultrametricized majority-rule Bayesian consensus trees. Dashed lines and shaded areas indicate the branches and terminals involved in the ancient (A) recent-close (RC) and recent-divergent (RD) recombination and hybridization scenarios.

For the data simulation, contiguous parts of the DNA sequence alignment (of a total of 1000bp) were simulated along two tree topologies: the "input tree" (arbitrarily resolved T_M) and the recombinant trees. The simulated recombination breakpoints were at 10%, 25%, 50% of the sequence length (the null-situation, R=0%, was also included in all tests). Thirty replicate data sets were created for each recombination scenario and breakpoint value. Models of nucleotide substitution were based on values reported in studies used (Bakker et al., 2004; van Velzen & Bakker, in prep., and see Table 4.1).

Table 4.1. Summary of model settings in simulation studies of Posada & Crandall (2002), Ruths & Nakhleh (2005) and this study.

Parameters	Posada & Crandall, 2002	Ruths & Nakhleh, 2005	This Study	
			Pelargonium	*Celtidoids*
#taxa	8	20	80	71
Colless's index [a]	0-1	0.26	0.12	0.18
Substitution model	HKY	GTR [c]	GTR [d]	GTR [e]
Ti/Tv ratio	2	-	-	-
Base frequencies				
A	0.1	0.18	0.21	0.33
C	0.2	0.33	0.30	0.16
G	0.3	0.26`	0.28	0.19
T	0.4	0.23	0.21	0.32
Substitution rates				
A↔C	-	3.297	0.988	1.206
A↔G	-	12.55	2.153	1.696
A↔T	-	1.167	1.812	0.404
C↔G	-	2.060	0.424	0.934
C↔T	-	13.01	5.388	2.132
G↔T	-	1.00	1.00	1.00
α [b]	-	0.82	0.3	0.09/ 2.62
% of invariant sites	-	0.545	22.5	25.1
Branch length scaling factor	0.3 & 0.6	0.1, 0.3 & 0.6	n.a. [f]	n.a. [f]
% of informative sites	?	?	52	25/ 74

[a] Colless's Imbalance statistic for tree asymmetry (Colless, 1982), normalized by maximum asymmetry
[b] Shape parameter of gamma distribution
[c] GTR + Γ model settings of Zwickl & Hillis (2002): based on ML estimation for 12SrRNA and cnr1 genes (from study on origin placental mammals, Murphy et al., 2001)
[d] Model settings of Bakker et al. (2004)
[e] Model settings of van Velzen & Bakker (in prep.)
[f] The branch lengths from the published tree topologies are used

The Seq-Gen output was input in subsequent phylogenetic analyses, including Bayesian inference and Jackknife resampling (JR). Bayesian MCMC analysis was performed using MrBayes 3.1.2 (Huelsenbeck & Ronquist, 2001; Ronquist & Huelsenbeck, 2003) with settings for the best-fit model (nst=6, rates=invgamma) selected by applying the hLRT criterion in MrModeltest 2.2 (Nylander, 2004). Two sets of 4 MCMC chains were run for 4 million generations or until the standard deviation of their split frequencies was below 0.01. Trees were sampled every 100th generation, and summarizing trees was performed using the MrBayes default 50% majority-rule, with the first 25% of samples discarded as burn-in.

Jackknife resampling searches were carried out using PAUP* 4.0b10 (Swofford, 2002) with the following settings: 37% of characters deleted with Jac emulation, 1000

replicates, TBR branch-swapping, 10 random-addition sequence replicates, 1 tree held at each step during stepwise addition, saving 1 tree per replicate (following Freudenstein et al., 2004). For comparison, we also performed NJ analyses (with 1000 bootstrap replicates) for some of our recombinant data sets in order to assess how a distance-based method performs.

Jackknife resampling of all replicate data sets was performed on a teen-node Sun V60x servers cluster (dual 2.8 GHz Xeon CPUs, 3GB RAM, Gigabit Ethernet) running Sun Grid Engine 6.0 on SuSE Linux Enterprise Server 9. MrBayes analyses were run on multiple Pentium IV Windows machines, using a Condor cluster (http://www.cs.wisc.edu/condor). File-handling was performed using the relevant Condor commands.

Measuring success

In order to assess the effect of recombination on tree topology recovery we need a measure that can estimate topological differences. Computation of metric values based on tripartitions and quartets can become too complex and inefficient for large-sized tree topologies (Vriesendorp, pers. obs.). Therefore, we used the less computationally intensive partition metric (*PM)* (Robinson & Foulds, 1981), as implemented in TREEDIST in the PHYLIP package version 3.*6* (Felsenstein, 2005), for comparing estimated tree topology T_C (either the Jackknife consensus or Bayesian consensus trees) with T_M (see Fig. 4.1). *PM* values are then corrected for the baseline values of the *PM* in a scenario without recombination (see Table 4.2).

Table 4.2. *PM* baseline values of null-situation, i.e. no recombination (R=0%).

R=0%	Bayesian	JR
$T_{M\ Pelargonium}$	8.5	14.4
$T_{M\ Celtidoids}$ [a]	15.5	20.2
$T_{M\ Celtidoids}$ [b]	9.3	12.8

[a] simulated using α=0.09
[b] simulated using α=2.62

We also used multidimensional scaling (MDS) as implemented in the Tree Set Visualization Module (Amenta & Klingner, 2002) of the Mesquite software to explore the estimated phylogenetic trees by visualization of tree space. MDS represents the (*PM*-based) distances between trees in such a way that the distortion between the true distance between pairs of trees and the screen distance is minimized, using a stress function (Hillis et al., 2005).

PM has been extended to include branch length information as well as the Branch Score (*BS*) distance (Kuhner & Felsenstein, 1994). *BS* is based on the sum of squares of differences between the branch lengths of branches in common between two trees. *BS* is implemented as the "branch score" in TREEDIST. While we included this measure in addition to *PM* for the Bayesian trees, we did not apply this procedure to the Jackknife resampling analysis because we felt branch lengths to be less meaningful in a Jackknife resampling approach.

Tree topological difference, i.e. the presence of splits in one tree and absence in another, can be due either to conflicting nodes or merely to a difference in resolution between the trees. The model trees (T_M) used in this study contained several polytomies, probably causing the *PM* to represent mainly resolution differences between T_C and T_M, rather than actual conflict. In order to filter out conflict from resolution differences, we opted for establishing whether or not trees are in conflict, using a procedure involving consensus trees, see Fig. 4.3 and below.

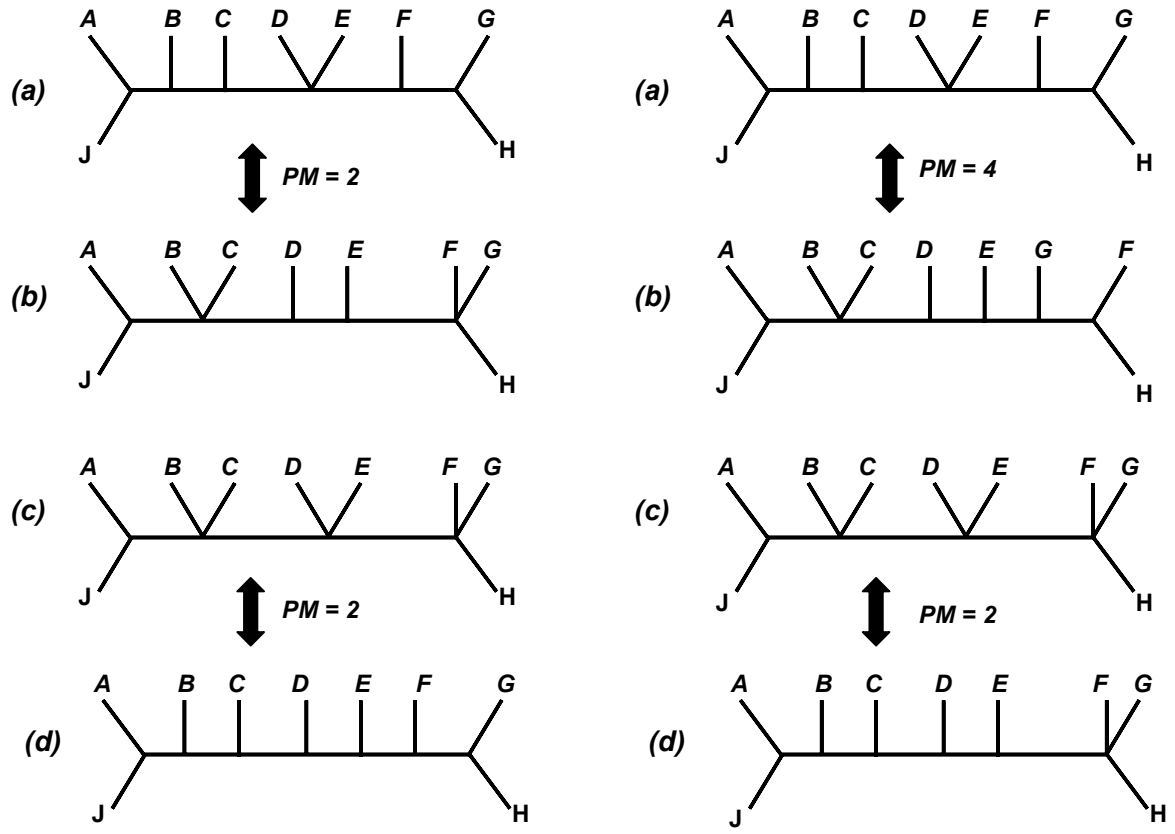


Fig. 4.3. Qualitative conflict measurement: trees in a non-conflict (left) or conflict situation (right). Shown are T_M(a), T_C (b), Ts_{CM} (c) and Tss_{CM} (d). Note the difference in representation of unresolved groups in these trees as compared with the rectangular representation elsewhere in this paper, i.e. the common node of taxa D and E in (a) is unresolved.

We first took T_C plus T_M and computed their strict (Ts_{CM}) as well as semi strict (Tss_{CM}) consensus tree topologies. As a strict consensus allows *identical* components only, Ts_{CM} represents all *resolved* nodes in common between T_M and T_C, whereas Tss_{CM} represents *all* nodes in common, both resolved and unresolved. In addition to the *PM* between T_M and T_C for each replicate data set, we also calculated the *PM* between Ts_{CM} and Tss_{CM} which we name here *PM**. If there is no topological conflict between these trees, then *PM* = *PM** and *PM* values represent solely the differences in tree resolution. If there is actual conflict between the trees, then *PM* > *PM**, and Δ_{PM}, i.e. *PM-PM**, therefore represents the difference in conflict in tree topology only.

PM values resulting from the different recombination scenarios (A, RC, RD) used were analyzed by a two-way ANOVA against a null situation with no recombination (R=0%). Least significance difference (LSD) testing was applied to test for significance of differences between the means of *PM*'s per scenario over the 30 replicates.

Hybrid scenarios

The hybrid scenarios included here used the same model tree topologies and parameter settings as under the recombination scenario (see Fig.1-II and Table 4.1). After resolving T_M we constructed two input trees from it by adding an extra ("hybrid") lineage in a different position on either tree, following the A, RC and RD scenarios (T_{H1} and T_{H2} in Fig. 4.1). Each scenario resulted in two different tree topologies, inserting the hybrid as sister to one or the other parental lineage. DNA sequences were simulated over these two trees at equal rate (500 bp per tree topology) using Seq-Gen as described above. Subsequently, Bayesian tree inference and Jackknife resampling was performed as described above. Before evaluating topological incongruence with T_M, hybrid lineages were pruned from the simulated consensus trees to enable a fair comparison.

In addition, we included a scenario with missing (extinct) parental lineages. These taxa were removed from the data sets according to four different scenarios of extinction: i) no extinction, ii) one or iii) the other parental lineage excluded and iv) both parental lineages excluded (see Fig. 4.1-III). The rest of the procedure was as above, except for an extra step of pruning the extinct parental lineages from T_M.

PM values resulting from the comparison of (pruned) T_M with the (pruned) T_C resulting from the hybridization scenarios were analyzed by one-way ANOVA to test for significance of means of *PM* difference.

Model trees

The first model tree, $T_{M\ Pelargonium}$, was based on 79 nrDNA ITS sequences of *Pelargonium* (Geraniaceae) published in Bakker et al. (2004) and reanalyzed here (see Fig. 4.2 and Table 4.1). The outgroup used in that study, *Monsonia ciliata* (Geraniaceae), was on such a long ITS branch that it was expected to disturb ultrametricizing, and hence the resulting mean branch length of the tree. Therefore, in order to avoid this artifact it was replaced by an artificial outgroup.

The second model tree topology, $T_{M\ Celtidoids}$ was based on a Bayesian analysis of combined rbcL and trnL-F DNA sequence data of 118 taxa including Celtidoids, Cannabaceae, Urticaceae, Moraceae and related taxa in the so-called urticoid Rosids (Van Velzen & Bakker, in prep.). We reduced their 50% majority rule Bayesian consensus tree to their subclade of 70 taxa ("Celtidoids + Cannabaceae"). One outgroup was randomly selected from the sister clade of Urticaceae (*Boehmeria calophleba)*. Selected lineages involved in the hybridization and recombination scenarios are depicted in Fig. 4.2. Simulation parameter settings for both the rbcL and trnLF partitions were based on Van Velzen & Bakker (in prep.) Two different shape parameters for the gamma rate distribution (α=0.09 for the *rbcL* and α=2.62 for the *trnLF* partition) were used as test parameters in the simulations. We used α as test parameter, because it is considered to be the most important DNA sequence model parameter. In the simulation studies of Lemmon & Moriarty (2004) for instance, a significant impact was shown of ignoring this variable on true trees recovery, examining several cases of model misspecification. The other parameter values were almost identical among the two partitions. These values were averaged over the partitions and used as fixed values throughout the simulations.

Calibrating the simulations

Simulating too many phylogenetically informative sites in the data may confound reliable tree reconstruction, and hence result in poor recovery of "true" tree topology (i.e. increase in *PM* between T_M and T_C). As an example, based on a 17 taxon tree we found an optimum of 40% phylogenetically informative sites as measured by minimizing *PM* between model and inferred tree. Scaling factors directly influence simulated branch lengths and therefore result in a varying amount of simulated informative sites. The effect of scaling (i.e. informative sites) on *PM* values and branch scores was also investigated on one of the model trees, $T_{M\ Pelargonium}$, see Fig. 4.4. No significant increase in *PM* values between T_M and T_C was found compared to the default value of branch length scaling factor 1. Therefore, in all further simulations

branch lengths were used as they appear on the published tree topology without additional scaling, but ultrametricized.

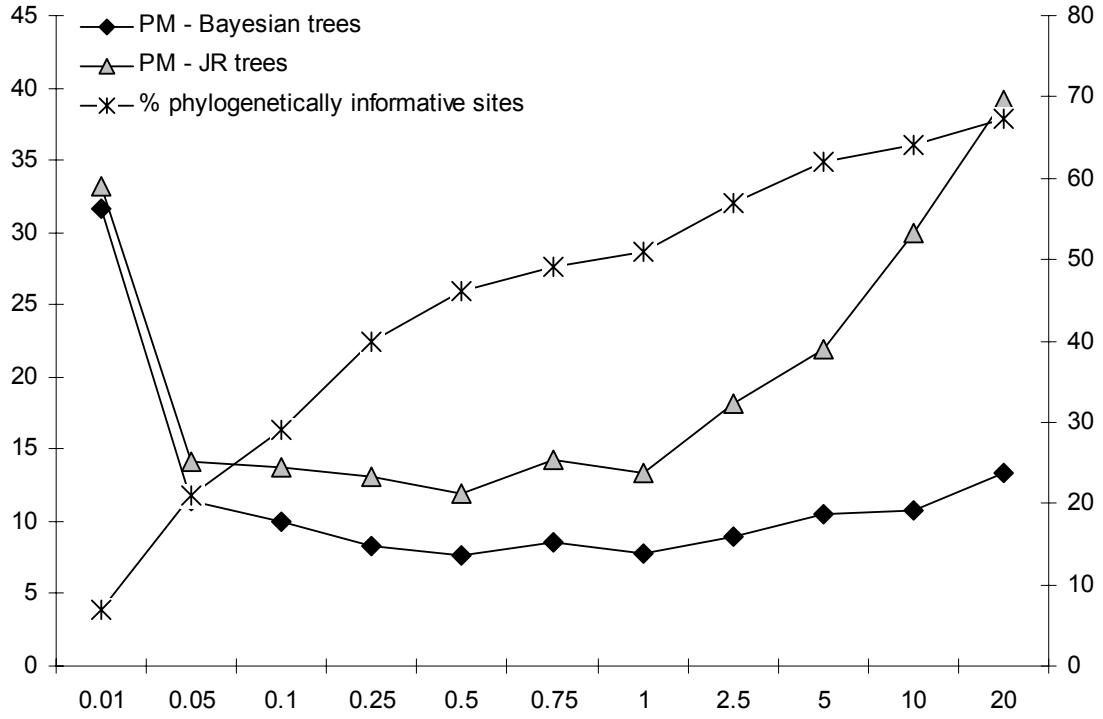


Fig. 4.4. The effect of scaling factor (i.e. phylogenetically informative sites) on tree topology using the first model tree $T_{M\ Pelargonium}$. Nine different scaling factors are explored in the range from 0.01 to 20. Partition metric (*PM)* values between T_M and Jackknife resampling (JR) and Bayesian consensus topologies (left scale) and % informative sites (right scale) are presented as average values (over 10 independent analyses from 10 replicate data sets) and shown against the scaling factors.

We tested for the effect of simulated sequence length on recovery of true tree topology (results not shown). No significant improvement was found using 5000bp, and therefore all simulations were performed using 1000bp.

As outlined above, ultrametrization could introduce an artifact in our method because such branch lengths may in reality never be achieved because of limitations of character sampling. In order to assess this possible effect, we used the tree topology before and after ultrametrization as input in Seq-Gen. The resulting data sets (100 replicates) were analyzed with MrBayes, with subsequent checking for effect on *PM* with $T_{M.}$ There was a significant difference (one-way ANOVA; F=7.81, p=0.0057), with the non-ultrametric tree resulting in a lower average value of the *PM* (6.82 ±0.40) compared with the ultrametric tree as input (*PM* of 7.81 ±0.52).

Recombination detection

Many different methods exist that test for recombination; see for an overview and testing of several of them Posada & Crandall (2001a), Wiuf et al. (2001), Posada et al. (2002) and Posada (2002). As part of our objective of testing an "informed" versus "naive" approach to phylogenetic analysis of recombinant DNA sequence data sets, our simulated DNA sequence alignments provide an excellent opportunity to test how efficiently these methods pick up the recombination breakpoints. We used the package RDP2 (Martin et al., 2005), which contains non-parametric methods for identifying recombinant and parental sequences, as well as for estimating breakpoint positions in the sequences. The implemented methods in the RDP2 package enable fast automated analysis of large alignments (up to 300 sequences containing 13,000 sites). We used the following 5 recombination detection algorithms as implemented in the RDP2 package. RDP (Martin & Rybicki, 2000) and Bootscan (Salminen et al., 1995) are both topology-based, e.g. they compare (NJ or UPGMA) trees derived from either side of a window sliding through the data set, and pick up shifts in topology, based on triplet scanning (RDP) or bootstrap replicates of the complete alignment (Bootscan). We also use character-based methods GENECONV (Padidam et al., 1999), MaxChi (Smith 1992) and Chimaera (Posada & Crandall, 2001a) that compare taxon-douplets or triplets to examine the sequences either for a significant clustering of substitutions or for a fit to an expected statistical distribution to induce recombination points. Three replicate data sets for all recombination and hybridization scenarios for both the $T_{M\ \textit{Pelargonium}}$ and $T_{M\ \textit{Celtidoids}}$ were tested, using NJ where possible. Various settings were explored, however, as most methods only allowed one- or two-parameter models, or in some cases could not be run optimally, we consider our test results to be not conclusive.

Correcting phylogenetic data sets?

When recombination effects on true tree recovery are encountered, we investigated possible solutions to this problem. The data available in this study provide us with an opportunity to test whether it would be sensible to discard part of the recombinant sequence, i.e. remove traces of recombination, prior to phylogenetic analysis. The alternative would be to analyze the data set *as is*, including the recombinant patterns, i.e. follow a "naive" approach. In order to compare both approaches we used 6 published composite cyto-nuclear angiosperm DNA sequence data sets. They all comprise topological incongruence between trees inferred from the separate nuclear and plastid DNA sequence data sets. Aligned DNA sequence data sets were provided directly by the authors (Barkman & Simpson, 2002; Beardsley et al., 2004; Hamzeh &

Dayanandan, 2004), or obtained from Treebase (www.treebase.org) (Oh & Potter, 2003; Donoghue et al., 2004; Petersen & Seberg, 2004). In the naive approach, we inferred trees from the combined and separate data sets using MrBayes, with subsequent comparison of their topologies. In the corrected (or "informed") approach, however, part of the hybrid sequence (either from the first or second data set) is effectively removed from the data, i.e. positions transformed to question marks, prior to phylogenetic analysis.

RESULTS

The baseline-corrected *PM* values for comparison with the Bayesian T_C's are presented in Fig. 4.5. Results from the Jackknife resampling analyses were almost similar and are not shown here. Some of the tests show a small but significant effect at 25% recombination for the scenario A, but at 50% recombination highly significant effects can be seen throughout. The highest impact of recombination on *PM* values can be found under the scenario RD for all model tree topologies and also under the scenario A for $T_{M\ Pelargonium}$. Visual comparison of the phylogenetic trees by MDS (not shown) clearly showed the T_C's from scenario RD as a separate cluster of trees for $T_{M\ Pelargonium}$, $T_{M\ Celtidoids}$ with $\alpha=2.62$, and to a smaller extent for $T_{M\ Celtidoids}$ with $\alpha=0.09$. T_C's from scenario A did only cluster as a separate group for $T_{M\ Pelargonium}$, on the opposite site from the cluster of RD-based trees. Results from Bayesian and Jackknife resampling analyses are similar with respect to the different scenarios used, but the resampling analyses show higher *PM* values overall. MDS corroborated this by showing a clustering of Bayesian trees much closer to the T_M compared with the JR trees. The values for the branch scores (only calculated for Bayesian T_C's) did not show any significant difference between the scenarios and are not further discussed.

The amount of conflict for simulated trees, measured as the average Δ_{PM}(*PM* – *PM**) of the consensus trees (T_C's) is given in Fig. 4.6. In addition, the total number of conflicting trees is presented here. Both measures suggest a discrepancy between resampling and model-based analysis, showing different levels of conflict for the different model trees. For the $T_{M\ Pelargonium}$ the majority of simulated trees from Jackknife resampling analyses is in conflict with the model tree and shows higher Δ_{PM}, even at low levels of recombination. However, the $T_{M\ Celtidoids}$ results indicate that trees inferred from simulated data with a high α-value (2.62) present a more mixed situation, and a low α of 0.09 resulted in less conflicting JR trees with lower Δ_{PM} values for all scenarios. It is noteworthy to add that JR trees are in general less resolved than trees inferred with MrBayes, see Table 4.3 where the resolution of simulated trees is given, measured as

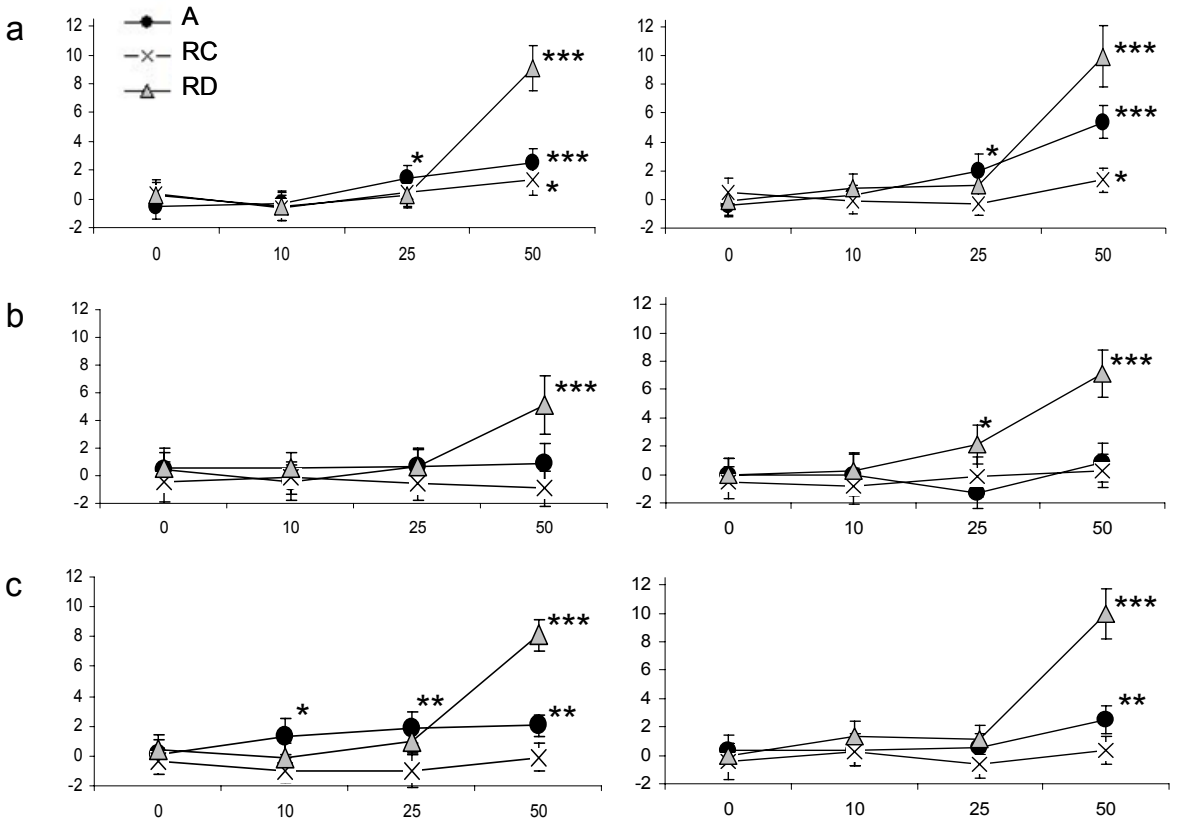


Fig. 4.5. Partition metric *(PM)* values T_C*'s* plotted against simulated recombination proportion for both non-reciprocal (left) and reciprocal (right) recombination as average values (over 30 replicates). Presented are the Bayesian T_C*'s* inferred from DNA sequence data sets simulated over the model tree topologies $T_{M\ Pelargonium}$ (a) and $T_{M\ Celtidoids}$ with α =0.09 (b) and α =2.62 (c). Recombination scenarios used were ancient (A), recent-close (RC) and recent-divergent (RD). *significant at p=0.05; **significant at p=0.01; ***significant at p=0.001

the number of splits in the T_C as a percentage of the maximum number of splits in the completely resolved tree (averaged over all replicates T_C 's). Using MDS, the Bayesian T_C's for $T_{M\ Celtidoids}$ with α = 0.09 showed a distribution which was more scattered than the JR trees. This is a nice illustration of the higher conflict in Bayesian trees despite their lower *PM* values, for this particular model tree topology. Indeed, in Fig. 4.6 the $T_{M\ Celtidoids}$ with a low α-value of 0.09 shows lower Δ_{PM} values resulting from JR resampling analyses compared with the Bayesian trees.

Differences between reciprocal and non-reciprocal exchange were small for Bayesian T_C's and negligible under resampling (data not shown) and are not further discussed here.

In addition to Jackknife and Bayesian analysis the $T_{M\ Pelargonium}$ was also analyzed using NJ. This resulted in higher *PM* values than in the Bayesian analysis, and, interestingly, lower values than in Jackknife resampling analyses (baseline value was 12.02).

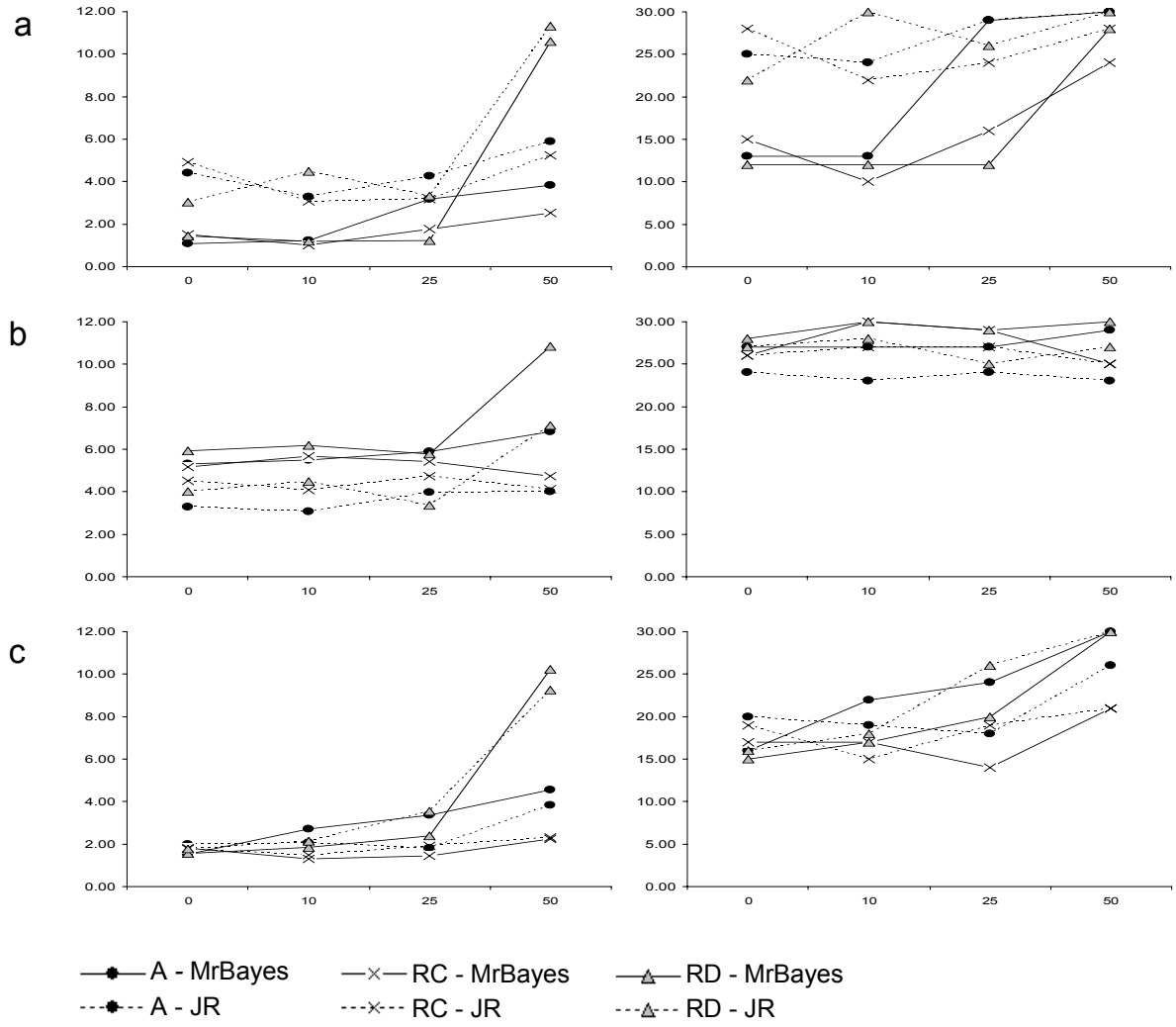


Fig. 4.6. Conflict values (left) and number of conflicting trees (right) for $T_{M\ Pelargonium}$ (a) and $T_{M\ Celtidoids}$ with α =0.09 (b) and α =2.62 (c), for Bayesian and Jackknife resampling (JR) analyses, non-reciprocal recombination. Conflict values (Δ_{PM}) are averaged over all 30 replicates. Number of conflicting trees is presented as the total amount of consensus trees that show any topological conflict with the model tree.

Table 4.3. Resolution of simulated trees, as a percentage of the maximum number of splits, averaged over all replicates.

%resolution	Bayesian	JR
$T_{M\ Pelargonium}$	87	81
$T_{M\ Celtidoids}$ [a]	62	52
$T_{M\ Celtidoids}$ [b]	71	72

[a] simulated using α=0.09
[b] simulated using α=2.62

Hybridization

The effect of hybridization and extinction on *PM* values of the consensus trees is summarized in Table 4.4. Overall no clear pattern can be seen, although extinction apparently had an effect in the A scenario in $T_{M\ Pelargonium}$ and the RD scenario for $T_{M\ Celtidoids}$ with an α-value of 2.62. Overall, the Bayesian topologies appear to have been more influenced by the hybridization scenarios than those inferred under JR.

Table 4.4. Effect of hybridization and extinction on tree topology measured as significant increase in *PM* values against baseline values (see Table 4.2) for Bayesian and JR analyses. Further explanation see text. *significant at p =0.05; **significant at p=0.01

	Bayesian			JR		
	A	RC	RD	A	RC	RD
$T_{M\ Pelargonium}$						
hybrid	*	-	-	-	-	-
hybrid, ex1	-	-	-	**	-	-
hybrid, ex2	**	-	-	-	-	-
hybrid, ex12	**	-	-	-	-	-
$T_{M\ Celtidoids}$ [a]						
hybrid	-	-	-	**	-	-
hybrid, ex1	-	-	-	-	-	-
hybrid, ex2	-	-	-	-	-	-
hybrid, ex12	-	-	-	-	-	-
$T_{M\ Celtidoids}$ [b]						
hybrid	-	-	-	-	-	-
hybrid, ex1	-	-	**	-	-	-
hybrid, ex2	-	-	*	-	-	-
hybrid, ex12	-	-	**	-	-	-

[a] simulated using α=0.09
[b] simulated using α=2.62

Detection of recombination

Recombination detection methods are usually applied to population-level data where several sequences (or alleles) represent different populations. The data used in this study contain simulated recombination resulting from (ancient) exchanging of lineages of species, where the different time scales prevent comparison with the population-level approach. Precision of locating the breakpoints in our simulated sequences depends on several factors, such as the window and step size as used in the topology-

based methods RDP and Bootscan (here, we only used default settings). Another factor influencing the success of recombination detection is the percentage variable sites in the DNA sequence alignment, because detection of recombination events depends on available informative sites. The three test-situations ($T_{M\ Pelargonium}$ and $T_{M\ Celtidoids}$ with α-values of 0.09 and 2.62) contained respectively 52%, 25% and 74% phylogenetically informative positions. This amount of variation should be sufficient to approximate the location of the 'true' breakpoints. We assigned the detected breakpoint as "correctly identified" (or "one-side correct") if the value of the breakpoint falls within a range of 50bp of the "true" breakpoint.

Table 4.5. Detected recombination breakpoints for $T_{M\ Pelargonium}$ and $T_{M\ Celtidoids}$. Detected recombination breakpoints are scored as "correctly" or "one-side correctly" if the range of beginning or ending points in the alignment are within 50 bp of the real breakpoints.

Simulated data set	Correct	One side correct	False positives
$T_{M\ Pelargonium}$			
RC, reciprocal	-	-	-
RD, non-reciprocal	1 [d]	2 [c]	-
RD, reciprocal	5 [c]	3 [c]	-
Hybrid RD	2 [c f]	4 [e]	1 [d]
$T_{M\ Celtidoids}$ [a]			
RC, reciprocal	-	-	1 [c]
RD, non-reciprocal	2 [c]	-	-
RD, reciprocal	2 [c]	1 [c]	-
Hybrid RD	-	-	-
$T_{M\ Celtidoids}$ [b]			
RC, reciprocal	-	-	-
RD, non-reciprocal	-	6 [d]	8 [d]
RD, reciprocal	5 [d]	5 [d]	2 [f]
Hybrid RD	6 [d]	7 [d]	2 [d]

[a] simulated using α=0.09

[b] simulated using α=2.62

[c] Detected by RDP and Bootscan

[d] Detected by RDP, Bootscan and Geneconv

[e] Detected by RDP, Bootscan and Chimaera/Geneconv

[f] Detected by RDP and Geneconv

In Table 4.5 breakpoints are listed that were identified by more than 1 of the selected 5 methods. Recombination breakpoints are added up per scenario over the 3 replicate data sets from all breakpoint proportions (10%, 25% and 50%). With the exception of a few correctly identified breakpoints (detected by RDP method only) (see Table 4.5), all other breakpoints detected by one method only, are out of the range of the true breakpoint. They occur randomly in all data sets, regardless of the amount of recombination or of the recombination scenario used. These "single-method" breakpoints appear to be a result of noise in the data and are therefore ignored in the rest of the analysis.

Breakpoints are only detected in the RD scenarios, where the sequences are different enough to be recognized as having a different ancestry. As expected, reciprocal recombination is easier to detect than non-reciprocal exchange of lineages, since both lineages contain signals of recombination. There are clear differences between data sets based on the different model tree topologies. The data sets with the highest average percentage of phylogenetically informative sites (from $T_{M\ Celtidoids}$ with $\alpha=2.62$) result in the highest amount of correctly inferred breakpoints. While the amount of false positives is also higher in this situation, most "misidentified" events are in the vicinity of the "true" breakpoints and the recombination test can still be used to indicate the occurrence of recombination.

Since a recombination event cannot be directly linked to a specific parental or daughter sequence (see above), we used these methods to identify and locate recombination events, and we did not infer the (putative) parental or daughter sequences. Putative parents or daughters indeed never accurately reflected the underlying 'true' scenario of recombination. Instead, related sequences are indicated as equally likely parents together with the "true" parental sequences. Apparently, all these related sequences contribute to the efficiency of detection of recombination events. This is especially obvious in the hybridization scenario (RD) where the results are independent of the "extinction of one or both parental lineages". The signal of "hybridization" is well enough maintained in the remaining related sequences to make identification of events just as possible without the parental lineages.

RDP and Bootscan identify most of the correctly detected breakpoints, sometimes in combination with Geneconv. Based on our simulated data we conclude that detection of events is most likely for ancient exchange of lineages in data sets with a high amount of variable sites. We consider the use of RDP2 to be a useful and fast check of DNA sequence data sets prior to phylogenetic analysis.

Naive versus corrected analyses

The results of our tests on the effects of ignoring any recombination (i.e. combining all data) versus "correcting" the data set (removing part of the hybrid sequence) is given in Table 4.6. No significant effect on average *PM* values is found. However, if only the *PM* against one of the data sets is considered (i.e. the recovery of either one of the "underlying trees"), the corrected approach gives better results: the separate *PM* values show a slight decrease after removing part of the sequence for almost all example studies (exceptions are *Dendrochilum* against tree 1 (*PM*1) and *Viburnum* against tree2 (*PM*2)). Therefore, we conclude that it would be worthwhile to explore correction of phylogenetic data sets using recombination detection wherever possible and in any case where recombination is suspected.

DISCUSSION

Recombination and phylogeny reconstruction

Recombination has potentially a severe influence on phylogenetic analyses, but how robust are our tree building methods actually to recombination? In many recombination studies the (potential) recombinant sequences are discovered and analyzed by using different clones and several individuals per species. The impact of recombination on phylogeny reconstruction in an angiosperm species-level context, using one hybrid terminal represented by only one individual (one accession) is more difficult to predict and not many example studies exist. In earlier simulation studies it was shown that 50% recombination resulted in divergent topologies from the model topology under a scenario of ancient or recent-divergent recombination (Posada & Crandall, 2002). In this study we have validated these findings for much larger tree topologies based on real angiosperm DNA sequence data. The success measure based on the *PM* is more suitable for larger tree topologies and we introduced a modification of *PM* (Δ_{PM}) that serves as a conflict measure to deal with the high percentage of unresolved nodes in these tree topologies. We also investigated differences in performance of model versus resampling analysis and explored different values of the shape parameter of gamma distribution.

Recombination was considered here as "any event that causes incongruence among trees", using the definition of Ruths & Nakhleh (2005). Recombination as such defined is comparable to the situation of a hybrid sequence with concatenated sequences from different markers (different genes or regions). Several phylogenetic studies, including those involving putative hybrids, have shown that the hybrid terminal is placed on different positions depending on the data set, for example in *Clausia*

(Brassicaceae) (Franzke et al., 2004), *Dendrochilum* (Orchidaceae) (Barkman & Simpson, 2002), *Mimulus* (Phrymaceae) (Beardsley et al., 2004), *Populus* (Salicaceae) (Hamzeh & Dayanandan, 2004) and *Scaevola* (Goodeniaceae) (Howarth & Baum, 2005). Most of these studies do not combine the data sets because their authors expect incongruence and a disturbing effect on the resulting combined tree topology. A study on 30 species of *Gaura* (Onagraceae) (Hoggard et al., 2004) is one of the few examples where the effect of including a mosaic terminal (putative hybrid) on phylogeny reconstruction was investigated. One putative hybrid with closely related parents caused no disruption, while the inclusion of a second putative hybrid, from distantly related taxa, disrupted the branching order within parts of the topology. This corresponds with the outcomes of our simulation studies where the effect of recombination in the scenario RC was almost negligible compared to the other scenarios. Although our model trees have different base *PM* values, due to different properties of the tree and data sets, the relative effects of the scenarios are almost similar and do indicate the confounding effect of recombinant lineages on tree topology reconstruction.

Main effect at 50% recombination?

The main effect of increasing *PM* values under different scenarios of recombination can be seen at a recombination breakpoint value of 50%. While most scenarios show this pattern, there is no specific property of the 50-50 recombination breakpoint that causes a decrease in tree topology recovery. Rather, it is a continuum of increase in *PM* value with the increase in the proportion of recombination. In some test situations an effect can be seen at 25%, and even 10% recombination can lead to a significant increase in *PM*. Re-analysis of two randomly selected scenarios (non-reciprocal A [$T_{M\ Pelargonium}$] and RD [$T_{M\ Celtidoids}$ with α=0.09]) with smaller steps (5% recombination breakpoints) indicated a linear increase in *PM* values. This contrasts with the non-linear distributions of *PM* values depicted in Fig. 4.5, but is probably due to stochastic processes using 30 replicate data sets only.

An interesting result is the difference in tree topology recovery for the two values of the gamma shape parameters α for $T_{M\ Celtidoids}$. While the effect of recombination is similar for both values, the average values of *PM* for both model- and character-based analysis are much higher in the simulated trees based on $\alpha = 2.62$, compared to simulations using $\alpha = 0.09$. A possible explanation can be a randomization of variable characters in the data sets (with on average 74% informative and 26% invariable sites).

Model-based versus Jackknife resampling approach

Throughout this simulation study, the model-based analyses seem to outperform Jackknifing resampling using *PM* values. This is partly due to a lower resolution in the Jackknife T_C's, illustrated by the lower (average) amount of splits compared with the Bayesian T_C's (see Table 4.3). However, as can be seen in Fig. 4.6, the proportion of conflicting trees can also be higher in resampling trees (based on $T_{M\ Pelargonium}$ and scenario RC and RD of $T_{M\ Celtidoids}$ with α=2.62). Accordingly, the differences are due to less resolution as well as more conflict in the trees based on Jackknife resampling.

PM as measure of success

The *PM* is easy to compute, but highly sensitive to all differences between trees (e.g. Felsenstein, 2004). Indeed, values for the "baseline" (no recombination or hybridization events) are relatively high compared to the increase in *PM* values. In addition, the values for the *PM between* the simulated trees are lower on average than those for the resulting trees against the model tree (results not shown). The most likely explanation is the better resolution of the simulated trees, i.e. splits that do occur in these consensus trees cannot be found in the model trees. These splits add up to the value of the *PM*, while the *PM* does not distinguish between conflicting splits or splits that are absent in one of the trees (i.e. unresolved).

Although a measure based on conflicting splits instead of number of splits seems more informative, it is not straightforward what exact measure of conflict is suitable and feasible for analyzing these simulated data sets. The conflict measure used in this study showed a correlation between the amount of conflicting trees and percentages of recombination. However, JR and Bayesian analyses show no consistent results concerning conflict values, and the level of conflicting trees was already high at R=0%. Especially the $T_{M\ Celtidoids}$ with α=0.09 revealed a high percentage of conflicting trees, at all levels of recombination.

Summarizing, measures based on splits do not give us a complete picture about the distortion of phylogeny reconstruction caused by recombination. A measure based on quartets, triplets or for instance (percentage) clade recovery could possibly give a better indication of the recovery of phylogenetic relationships, but were not available in a suitable implementation for our data sets.

Recommendations

How severe is the effect of ignoring “recombination” on actual situations of phylogeny estimation? And what can we do to solve this? Posada & Crandall (2002), Ruths & Nakhleh (2005) and this present study show that in most situations the “non-

recombined" tree topology is still recovered, but only for specific ("extreme") scenarios of recombination and with specific conditions of recombination. That is, if a low percentage of the combined data set (up to about 25%) comes from the recombined tree, the "non-recombined" tree topology is still recovered. If the proportion of the recombined data is almost equal to the amount of non-recombining sites, recovery of one of the tree topologies is not straightforward anymore and the recovered tree will be different from both model trees (measured both as higher *PM* values as well as more conflict). Therefore, if the goal of phylogeny reconstruction is to recover at least one of the underlying tree topologies, the data can be analyzed in a combined way, as long as the proportion of the recombination does not approach 50%. However, if the goal is to recover all underlying evolutionary histories, these trees will never be recovered, simply because tree estimation does only allow recovering bifurcating trees and does not represent conflicting nodes. The only way to represent all conflicting signals in a phylogeny is to generate separate trees for the separate parts or to use network methods, such as NeighborNet, Split Decomposition or Consensus Networks, as implemented in SplitsTree (version 4.6, Huson & Bryant, 2006).

Furthermore, not all the scenarios of recombination result in deviating tree topologies. The RD scenario had a significant effect on *PM* values in all situations. The A scenario, however, only affected the analyses based on $T_{M\ Pelargonium}$ and $T_{M\ Celtidoids}$ with $\alpha=2.62$, probably due to the high amount of informative sites in both data sets (52 and 74, respectively). Admittedly, these percentages are high and most angiosperm studies will show much lower percentages of variable sites, so this situation is probably an exceptional case. Nevertheless, the effect of 50% recombination still holds for almost all situations.

As described earlier, the use of "recombination" can both refer to "real" recombination (e.g. recombination between paired chromosomes, creating chimeric alleles) or referring to the combination of incongruent separate data sets, i.e. mimicking cyto-nuclear incongruence. In the latter situation, separate analyses are recommended to check for possible recombination events, or to estimate tree topologies for the separate parts to analyze possible incongruence. However, if there is just one data set, the possible recombination breakpoints are not known. We used here RDP2, to detect these recombination points for some of our data. In highly variable data sets, most of the recombination events were detected (for the RD scenario), regardless of proportion of recombination. These same data sets resulted in a significant impact on tree topology recovery. However, some data sets with a lower amount of recombination under scenario A did also result in significantly different tree

topologies, but are not detected here. So, only "worst cases" are likely to be detected using a program such as RDP2.

In the case of detection of a breakpoint, the "recombined" sequences can be removed from the data set in order to have a "clean" data set without taxa that cause incongruence. However, this practice of removing "hybrids" will not provide any insight in the evolutionary history of the "hybrid" and its relatives. It depends on what kind of information is needed. If the "summary of relationships" is of primary interest, the recombination can be seen as noise. Then, the data set only needs to be adjusted (analyzed separately) if the breakpoint divides the alignment in a part higher than 25%. Sequences can also be "corrected" prior to analysis (i.e. still combining all data, but removing part of the hybrid sequence) to explore the possibility of possible improvement. Using published data sets, either combining or correcting the data, did not affect *PM* values.

As a concluding remark we can say that in a naive approach – analyzing the data combined, without checking or correcting for recombination – the resulting trees will only significantly differ from one of the underlying tree topologies in extreme cases such as in a RD scenario. If there is cause for suspecting recombination different data sets from different marker regions should be analyzed separately instead of combined.

ACKNOWLEDGEMENTS

We are grateful to Mark Fiers for help with the simulations and providing analysis space on the computer cluster. We thank Jack Leunissen for helpful suggestions and use of Unix-computer cluster. Timo van der Niet and Fred van Eeuwijk provided helpful suggestions for improvement of the manuscript.

chapter 5

EXPLORING NETWORK METHODS PERFORMANCE USING ANGIOSPERM SPECIES-LEVEL DNA SEQUENCE DATA SETS

Bastienne Vriesendorp & Freek T. Bakker

Submitted for publication in Molecular Phylogenetics and Evolution

ABSTRACT

Network reconstruction methods are rarely used in phylogenetic studies at species-level, and also studies testing their relative performance are scarce. We use simulated and published composite DNA sequence data from species-level angiosperm studies to study the performance of a selection of current network methods. All data sets comprised a "hybrid" terminal that is incongruently placed in phylogenies derived from the separate data sets. Recovery of this "hybrid" in the reconstructed networks is compared against our expectation of hybrid position in the underlying data. We also compare rates of false positives and false negatives between the retrieved networks to test for recovery of resolution among all included taxa. Here, we demonstrate that while some network methods do place the hybrid in the "expected" position, most network methods produce graphs with either too much (Median Networks and NeighborNet) or too little (Parsimony Splits and Split Decomposition) resolution, thereby complicating interpretation of phylogenetic relationships. While network methods are used frequently in population studies, at species-level the higher levels of variation probably make it difficult to visualize the relationships in a meaningful way. Future developments including likelihood approaches and incorporation of models to deal with other kind of data (e.g. genomic) could improve the ability of network methods to represent organism-level phylogenetic relationships.

KEY WORDS

Network reconstruction, Hybridization, Recombination, Reticulate evolution

INTRODUCTION

Reticulation, i.e. the occurrence of hybridization and introgression, has always been thought to have played an important role in the evolution of plants (e.g. Stebbins, 1959; Raven, 1976; Grant, 1981). During the last decade, reticulate evolution has become more and more an important topic in the systematic and general biological literature, with several reviews published on the occurrence and/or evolutionary consequences of hybridization (e.g. Arnold, 1992, 1997; Rieseberg, 1997, Rieseberg & Carney, 1998; Ellstrand et al., 1996, Hegarty & Hiscock, 2005). While most of these reviews have been written from a plant perspective, hybridization and introgression is increasingly considered an important evolutionary mechanism in the animal kingdom too (e.g. Dowling & Secor, 1997; Seehausen, 2004; Schwarz et al., 2005; Willis et al., 2006; Mallet, 2007). Other studies focus on the detection of

angiosperm hybrids in a phylogenetic context (e.g. Mansion et al., 2005; Gravendeel et al., 2004, Chat et al., 2004; Wendel et al., 1995a).

The standard representation of evolutionary relationships at species-level, i.e. bifurcating trees, will fail to describe evolutionary relationships when they are actually networks. However, methods to reconstruct reticulate patterns are still not widely used (Vriesendorp & Bakker, 2005/Chapter 3, Linder & Rieseberg, 2004). This is in contrast with population-level studies where relationships are commonly represented in a network, e.g. as haplotype networks (recent examples in angiosperms are e.g. Ikeda & Setoguchi, 2007; Stracke et al., 2007, Shih et al., 2007; Chung et al., 2007) or minimum spanning networks (e.g. in Liao et al., 2007; Paula & Leonardo, 2006). Some studies have used phylogenetic networks to represent evolutionary relationships between plant species: e.g. Lockhart et al. (2001) used Split Decomposition (Bandelt & Dress, 1992) for constructing splits graphs to analyze hybridization in alpine buttercups (*Ranunculus*, Ranunculaceae), and Hörandl et al. (2005) applied the same method to other clades within the Ranunculaceae. NeighborNet (Bryant & Moulton, 2002, 2004) has been used to explore evolutionary relationships within *Heliosperma* (Caryophyllaceae) (Frajman & Oxelman, 2007) and among 40 individuals representing 7 species of *Rosa* (Rosaceae) (Joly & Bruneau, 2006). However, in general, most systematic papers to date only contain tree-like representations.

There are different ways to use networks in evolutionary studies (e.g. Morrison, 2005; Huson & Bryant, 2006), i.e. either to represent organism-level or character-level relationships. Based on directed acyclic graphs reticulate organismal-level evolutionary relationships can be described and hence, such networks would represent therefore "true phylogenetic networks" in the sense of Huson & Bryant (2006). In order to incorporate processes such as introgression, hybridization, recombination or horizontal gene transfer, mathematical models are needed to be able to represent the evolutionary relationships (e.g. Xu, 2000; Hallet & Lagergren, 2001; Addario-Berry et al., 2003; Moret et al., 2004; Nakhleh et al., 2005b).

Another way of using networks in phylogenetic studies is to visualize incongruence of phylogenetic signals, or represent character conflict (e.g. Holland et al., 2005; Huber & Moulton, 2005). Networks in such a context might give an *indication* of the amount of organism-level hybridization, but do not necessarily strive after presenting reticulate relationships. More network branches just indicate more instances of site pattern incompatibility in the DNA sequence alignment, making hybridization hypotheses more likely.

Several algorithms have been developed, either to reconstruct reticulate organism-level relationships in acyclic graphs (e.g. Hallet & Lagergren, 2001; Nakhleh et al., 2003, 2005b) or to represent character-level networks (e.g. Bandelt & Dress, 1992; Bryant & Moulton, 2002). Posada & Crandall (2001b) gave a comprehensive review of different network programs and packages and their use for analyzing intraspecific gene genealogies. Network methods can be characterized based on input data (i.e. pair wise distances, characters, or trees) or, ideally, on optimality criterion (see Table 5.1), although the latter is only possible in a minority of methods. NeighborNet (NN), Split Decomposition (SD) Median Network (MN), Parsimony Splits (PS) and Consensus Networks (CN) (see Table 5.1 for references) are all based on the mathematical concept of a split, i.e. bipartition of the data.

Table 5.1 Selection of network reconstruction methods. Abbreviations of methods further used in this study are given in brackets.

Method	Reference	Input data	Optimality criterion
Median networks(MN)[a]	Bandelt et al., 1995	characters	?[b]
NeighborNet (NN)[c]	Bryant & Moulton,2002, 2004	distances	n.a. (algorithmic approach)
Split Decomposition (SD)[c]	Bandelt & Dress, 1992	distances	n.a. (algorithmic approach)
Parsimony Splits (PS)[c]	Bandelt & Dress, 1993	characters	n.a. (algorithmic approach)
Consensus Networks (CN)[c]	Holland & Moulton, 2003	trees	n.a. (algorithmic approach)
Molecular Variance Parsimony (MVP)[d]	Excoffier & Smouse, 1994	characters	?[b]
Statistical Parsimony(SP)[e]	Templeton et al., 1992	distances	n.a. (algorithmic approach)

[a] implemented in e.g. Network (Bandelt et al., 1995; www. fluxus-engineering.com), Spectronet (Huber et al., 2002) SplitsTree (Huson & Bryant, 2006)
[b] all parsimonious solutions are presented in one network, but no optimality criterion is used to search tree or network topology space
[c] implemented in SplitsTree (Huson & Bryant, 2006)
[d] implemented in MINSPNET (http://cmpg.unibe.ch/software.htm)
[e] implemented in TCS (Clement et al., 2000)

Splits are inferred either from a distance matrix, characters or from phylogenetic trees. NN will only identify splits that can be represented in a plane (i.e. in two dimensions), using circular ordening. SD and PS use a "weak compatibility" criterion to identify incompatible splits, either based on the distances (SD) or the raw data (PS) and combines these splits into one graph. CN can visualize incongruence between gene trees when they include the same set of terminals, but for comparing trees with non-identical sets of terminals, supernetwork approaches can be used, see e.g. Huson et al. (2004) and Holland et al. (2007). CN summarizes splits of all trees, using a threshold to control the complexity of the network. In the character-based

Median Network (MN) approach all observable splits are represented, and ancestral nodes are inferred in order to represent both extant and median (ancestral) sequences. Another character-based method is Molecular Variance Parsimony (MVP) (Excoffier & Smouse, 1994 and see Table 5.1) which, like MN, summarizes all possible solutions, i.e. all most parsimonious trees or all mimimum-spanning trees, into a single graph (Bandelt, 1995; Excoffier & Smouse, 1994). Statistical Parsimony (SP) combines a character-based approach using a "parsimony connection limit" to determine what subset of taxa will be connected together, with subsequently an algorithmic approach to construct the network (Templeton et al., 1992). Actually, most of the methods listed here reconstruct networks using an algorithmic approach; at least no optimality criterion is applicable to choose between different network solutions, or the use of the optimality criterion is not immediately clear (i.e. with MN and MVP).

Testing the comparative performance of network methods with "real" data sets has only been described in a few publications; Morrison (2005) and Cassens et al. (2003, 2005) compared a range of methods using empirical and/or simulated data. However, these studies were mainly limited to population-level settings and associated levels of DNA sequence divergence (Posada & Crandall, 2001b). DNA sequence variation at angiosperm species-level could prevent efficient analysis using these network methods (Vriesendorp & Bakker, 2005/Chapter 3), as higher levels of variation can cause nodes to collapse (SD) or produce an indecipherably large amount of reticulations (e.g. MN and NN).

Ideally, we would like to test the performance of network reconstruction methods using simulated data as is done for phylogenetic tree methods (e.g. Bayesian methods in Suzuki et al., 2002 and Douady et al., 2003, or comparing Bayesian, MP, ML and NJ in Hall, 2005). However, since many network reconstruction programs do not have a batch mode, simulation can become a cumbersome enterprise (pers. obs. and L. Nakhleh, pers. comm.). More importantly, it is not clear what test statistic to use as "succes measure", i.e. how to compare retrieved networks to the simulated model network. Measurements such as the Partition Metric (Robinson & Foulds, 1981) are used in many simulation studies (e.g. Leitner et al., 1996; Zwickl & Hillis, 2002; Piontkivska, 2004), but cannot be used for networks because its original description applies to bifurcating trees only.

Nakhleh et al. (2003) tested the performance of distance-based *Split Decomposition* using a modified version of the Partition Metric *(PM),* which sums up the number of taxa bipartitions between trees. Bipartitions can be divided into false negatives *(FN)*, i.e. present in the simulated but absent in the reconstructed network,

and false positives *(FP)*, i.e. present in the reconstructed but not the simulated network. Nakhleh et al. take the average of *FN* and *FP* (instead of adding these values) and they measure *FN* as: (| *C(T')* - *C(T)*|) / (*n*-3) and *FP* as: (| *C(T)* - *C(T')*|) / (*n*-3), where *C* is the set of splits of the model tree *(T)* or reconstructed tree *(T'),* and n is the number of terminal taxa. Other studies used a tripartition-based distance to infer distances between phylogenetic networks (Linder et al., 2003; Moret et al., 2004; Nguyen et al., 2005). This metric is an extension of the *PM*. First, all terminals in a network are partitioned into sets of tri-partitions and subsequently these sets of tripartitions are compared between different networks. When the network is actually a tree, this distance reduces to the standard *PM*.

Several studies focus on the trees induced from a network instead of the network itself in order to express network distance. Known metrics, such as the SPR-distance (e.g. Baroni et al., 2006) proceed by calculating the minimum number of rooted subtree prune and regraft operations required to transform a tree *T* into another tree. However, in addition to the problem of finding trees induced from the networks (e.g. Gusfield & Bansal, 2005; Baroni et al., 2006) it has been shown that the SPR distance can only be solved exactly if there are few differences between the trees, i.e. if their SPR distance is small, e.g. < 20 for trees with up to 1000 leaves (Bordewich & Semple, 2004; Baroni et al., 2005).

Therefore, in this study we test the relative performance of network reconstruction methods against topological hypotheses of expected reticulation, i.e. the position of the hybrid terminal in a network, using angiosperm species-level DNA sequence data sets including (putative) hybrid terminals for which ample (external) evidence on their hybrid origin is present. Both an organism- and character-level approach is followed, because we test network methods that display character conflict, while using data sets including "actual" hybrids. Using a range of sequence divergence levels "typical" for angiosperm species-level systematic studies, as well as different (putative) parent-hybrid species scenarios, we want to test relative performance of network reconstruction methods. In addition, we apply some of the methods to data sets simulated on 71- and 80-taxon tree topologies for verification of main conclusions.

Note that we use the term "hybrid terminal" or "hybrid" throughout this paper in an operational sense, and not explicitly as an indication of evidence for organism-level hybrid ancestry.

MATERIALS AND METHODS

Data sets

We used composite DNA sequence data sets from published angiosperm phylogenetic studies with (external) evidence for the presence of a hybrid terminal, and preferably encompassing a wide range of tree size, sequence variation as well as parental distances. The selected data sets are composite, i.e. they comprise separate data sets which, after their separate analysis, appeared incongruent concerning the phylogenetic position of the hybrid terminal. (Note that it cannot be ruled out that other hybrids occur in the data sets selected.)

In addition to these "real" data sets, we used simulated data sets generated for another study (Chapter 4). These simulations involved generating DNA sequence data over angiosperm phylogenetic tree topologies comprising 71 and 80 terminals, based on different models of DNA sequence evolution (see Table 5.2).

Table 5.2 Input tree topologies for the three simulated data sets (see Chapter 4). Simulated data sets obtained using a GTR + Γ model of sequence evolution. Shape parameter of gamma distribution and % phylogenetically informative sites are listed (averaged over 30 replicates).

Input-tree	α	% invariable. sites	% inform. sites
Pelargonium [a]	*0.3*	*22.5*	*52*
Celtidoids with α=0.09 [b]	*0.09*	*25.1*	*25*
Celtidoids with α=2.62 [b]	*2.62*	*25.1*	*74*

[a] Model tree topology from Bakker et al. (2004)
[b] Model tree topology from van Velzen & Bakker (in prep.)

Tree analyses

Most published studies selected had used maximum parsimony (MP) for phylogenetic analysis, in most cases only for the separate data sets. Using Bayesian inference, we re-analysed all data sets, separate and combined, in order to assess how the incongruence between data sets would resolve. This resulted in separate trees T_1 and T_2 and T_{1+2} for the combined data (see Fig. 5.1). Bayesian MCMC analysis was performed using MrBayes 3.1.2 (Huelsenbeck & Ronquist, 2001; Ronquist & Huelsenbeck, 2003) with the model of sequence evolution selected by applying the hLRT criterion in MrModeltest 2.2 (Nylander, 2004). Two sets of 4 MCMC chains were run for 4 million generations or until the standard deviation of their split frequencies was below 0.01. If necessary, some data sets were run with lower

temperature (T = 0.05) because this resulted in better mixing of the chain. Trees were sampled every 100th generation, and summarizing trees was performed using the MrBayes default 50% majority-rule, with the first 25% of samples discarded as burn-in.

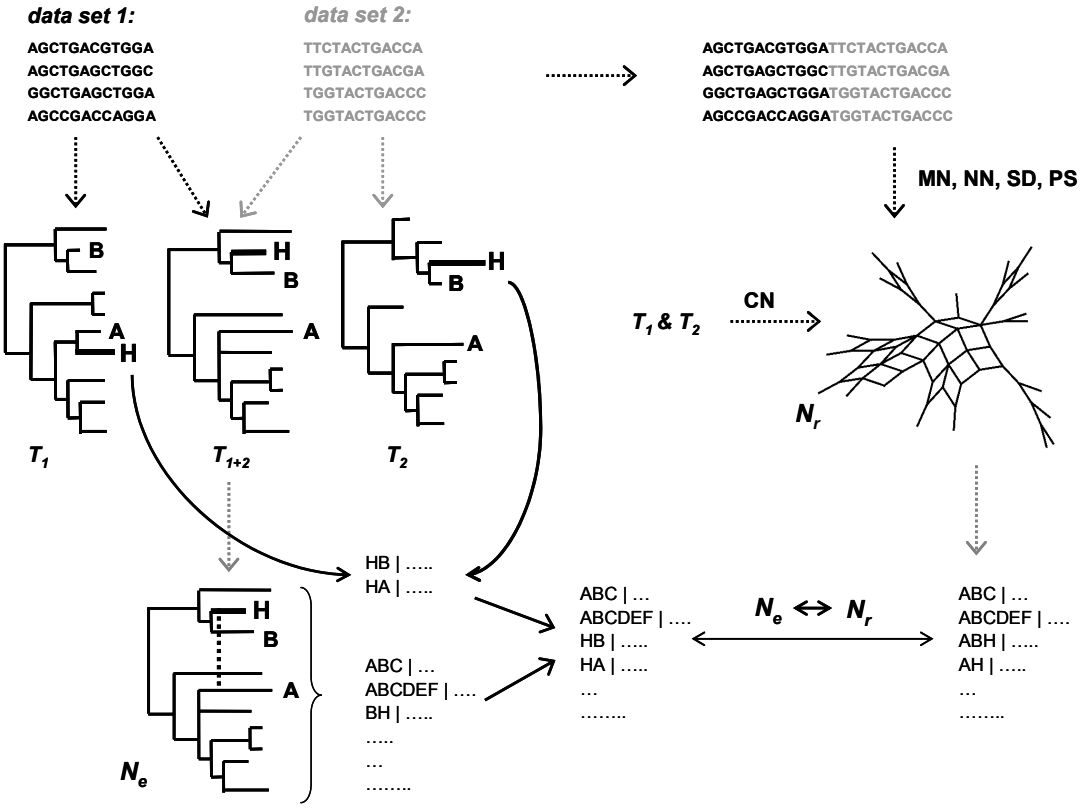


Fig. 5.1 Performance testing of network methods using published data sets.
Based on two (or three) different data sets, separate *(T_1* and *T_2)* and combined *(T_{1+2})* MrBayes analysis are performed. The splits in the expected network *(N_e)* are based on the splits in the combined network plus the splits that define the hybrid position as sister to the different parents in T_1 and T_2. Input for network analyses are the concatenated data sets (MN, NN, SD and Ps) or the separate trees T_1 and T_2 (CN). All splits from the reconstructed network (N_r) are compared against all splits from N_e. to establish hybrid position and values of *FN* and *FP*.

Network methods

All network analyses were performed using SplitsTree v.4.6 (Huson & Bryant, 2006). Five methods from that package were chosen (see also Table 5.1). For the distance-based approach we used NeighborNet (NN) and Split Decomposition (SD), for the character-based approach we used Median Network (MN) and Parsimony Splits (PS), and for the tree-approach Consensus Networks (CN). For the latter the 50% majority

rule consensus trees T_1 and T_2 were used as input. Bootstrapping (1000 replicates) of the resulting networks was applied to all NeighborNet analyses by random resampling the DNA sequence alignments.

Performance testing

We used arbitrary criteria (based on splits) to assess performance of network methods against "biologically intuitive expectation", i.e. following the "gene tree = species tree" approach. We are aware that such an approach may appear to be inconsistent with the recommendation as described above, i.e. to visualize incongruence and not to describe organismal-level evolutionary history. Nevertheless, we feel that such an approach is sensible in this context and in the absence of formal success measures, which is why we choose the following arbitrary criteria for assessing success of method output (see also Fig. 5.1):

i) *Hybrid position in the reconstructed network.* The position of the hybrid in the reconstructed network (N_r see Fig. 5.1) is compared with the position in the "expected" network (N_e), where the hybrid terminal shows a sister-connection to both parents. Parents are appointed by selecting the terminals (or groups of terminals) that are sister to the hybrid in the separate trees (T_1 and T_2 see Fig. 5.1). The tree resulting from the composite analysis of the separate data sets (T_{1+2}) was transformed to represent N_e by including all connections of the hybrid to its parents, i.e. adding two extra splits representing connections to both parents or add just one extra split if the hybrid is placed as sister to one parent in the $T_{1+2.}$ The hybrid in N_r can be reconstructed in the expected position as indicated in N_e (i.e. it shows a sister-connection to both parents), or is either sister to one parent or has no (direct) connection to either. Although the reconstructed networks sometimes allow assessment of hybrid position by visual inspection, we established this criterion comparing lists of splits.

ii) *False negatives and positives (FN, FP). FN* are the splits present in the expected network (N_e) but absent in the reconstructed network (N_r), whereas *FP* are splits present in the reconstructed but not in the expected network. These two criteria give an indication of how much structure of the overall tree/network topology is recovered by the network method under scrutiny. Thus, low *FN* and *FP* indicate a high overall correspondence with N_e. Calculating *FN* and *FP* is based on comparing the list of splits from N_r with all splits contained in $N_{e.}$ The list of splits from N_e is derived from combining all splits in T_{1+2} with the (one or two) extra splits due to the hybrid-parent connection in the separate trees T_1 and T_2. The list of splits from N_r are taken directly from the SplitsTree output and compared against the list of expected splits (N_e) using

the CMPNETS package as implemented in the toolkit Phylonet 1.2 (Nakhleh et al., 2006).

iii) *Effect of hybrid exclusion on network structure.* Because a hybrid terminal possibly has a disruptive influence on reconstructing trees or networks, we tested for this effect by constructing networks (using our selected set of network methods) both with and without the hybrid terminal. Resulting networks were compared using *FN* and *FP* values as outlined above.

Consensus Networks (CN) is an efficient approach to summarize and visualize tree conflict (Holland et al., 2003, 2005). The question can be posed whether CN is actually an analytical tool ("network method") or "merely" a consensus tool. Testing of the CN method is not a meaningful approach here because it will return "whatever you feed it", i.e. it compiles a network out of all splits concerned. As our first performance criterion is based on tree topologies (i.e. T_1 and T_2) CN will always resolve the hybrid in the expected position. However, in general CN is a valuable tool to display hybrid relationships and simultaneously check for incongruencies in non-hybrid parts of the tree. Therefore, we included CN here to visualize and compare the resulting consensus networks and their *FP* and *FN* levels against the other methods.

The test procedure using the simulated data differs from the "real" composite data sets in that the underlying T_1 and T_2 are identical, apart from the placement of the hybrid terminal (no difference in "rest-topology"). Thus, split calculation is much more straightforward with this effectively single topology. In order to assess the recovery of the overall tree topology of the reconstructed network N_r compared to the expected network N_e, the list of splits in the N_e is determined by taking all splits from both trees T_1 and T_2.

RESULTS

We identified 6 available data sets that complied with the restrictions. Table 5.3 lists these data sets and some of their characteristics. The associated aligned DNA sequence data sets were provided directly by the authors (Barkman & Simpson, 2002; Beardsley et al., 2004; Hamzeh & Dayanandan, 2004), or obtained from Treebase (www.treebase.org) (Oh & Potter, 2003; Donoghue et al., 2004; Petersen & Seberg, 2004). To simplify the analysis and to more accurately compare the performance of the network methods, we adjusted one data set (i.e. removed 2 taxa) to comprise an equal amount of sequences. The selected composite data sets ranged from conserved to variable (2-13 % phylogenetically informative sites in the nrDNA and 1-4% for the plastid partitions, see Table 5.3). The *Dendrochilum* and

Viburnum data sets were most variable concerning the overall (average) nrDNA variation, whereas the *Populus* data set shows the lowest variation in terms of % informative sites. The position of the hybrid in the N_r for all methods is listed in Table 5.4, including the placement in T_{1+2}. MN and NN resolved the hybrid in 2 out of 6 data sets in the expected position, in contrast to PS (1 out of 6) and SD (none, i.e. unresolved or sister position). Also, SD resulted in many largely unresolved trees.

Table 5.3 Data sets used.

Genus	Reference	hybrid sp./ total sp. [a]	Par. Dist. (% K2P)	Marker		Average nDNA var. (% p.i.) [b]
				nDNA (% p.i.) [b]	cpDNA (% p.i.) [b]	
Dendrochilum (Orchidaceae)	*Barkman & Simpson, 2002*	1/ 22	10	ITS (13)	accD (1)	13
Hordeum (Poaceae)	*Petersen & Seberg, 2004*	2/ 28	48	DMC (6) EF-G (8)	rbcL (1)	7
Mimulus (Phrymaceae)	*Beardsley et al., 2004*	1/ 18	23	ITS (8) ETS (14)	trnL-F (3)	11
Populus (Salicaceae)	*Hamzeh & Dayanandan, 2004*	1/ 21	49	ITS (2)	trnL-F (3)	2
Stephanandra (Rosaceae)	*Oh & Potter, 2003*	3/ 9	3/9	ITS (8) LFY (6)	trnL-F (2)	6
Viburnum (Adoxaceae)	*Donoghue et al., 2004*	1/ 43	34	ITS (13)	trnK (4)	13

[a] Number of hybrid terminals / total terminals, excluding outgroups.
[b] Variation as % phylogenetically informative sites, excluding outgroups

Table 5.4. Position of the hybrid terminal in the N_r per network method against N_e and the position in T_{1+2}.

Data set	Hybrid position in MrBayes analysis	Hybrid position in N_r against N_e			
		MN	NN	PS	SD
Dendrochilum	unresolved	sister to P1	sister to P1	sister to P1	unresolved
Hordeum	sister to P2	sister to P2	sister to P2	sister to P2	sister to P2
Mimulus	sister to P1	as expected	unresolved	sister to P2	unresolved
Populus	unresolved	sister to P2	sister to P2	unresolved	unresolved
Stephanandra	unresolved	as expected	as expected	sister to P1	sister to P1
Viburnum	sister to P1	sister to P2	as expected	as expected	sister to P2

The number of false negatives *(FN)* and false positives *(FP)* per method for all 6 data sets are presented in Fig. 5.2. CN is included here for completeness and represents the amount of differences solely due to tree topology differences between combined and separate trees (i.e. rest-topology). The performance of CN, however, should not be compared against the other methods, for reasons outlined above.

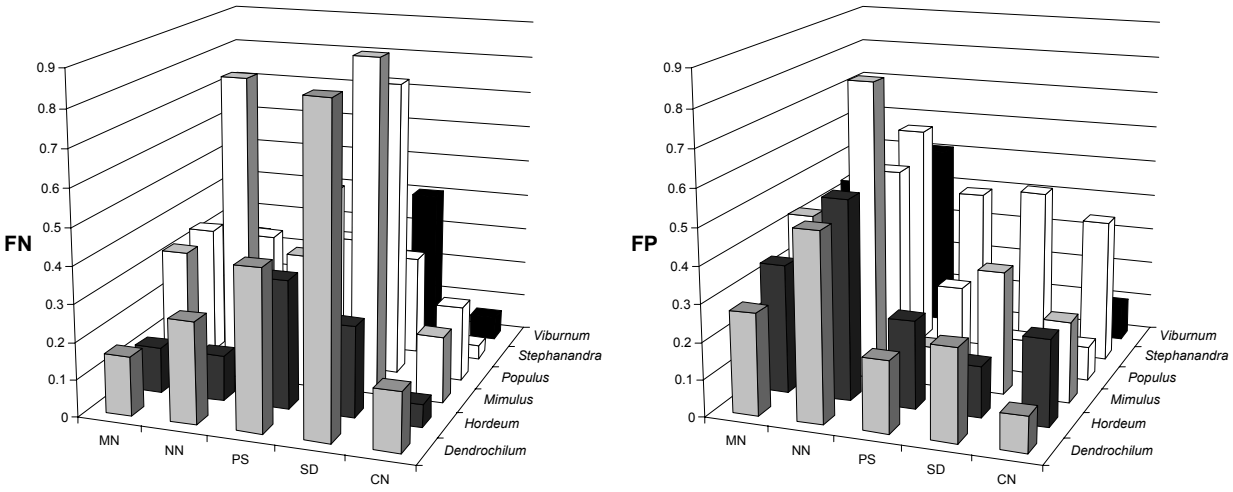


Fig. 5.2 Number of *FN* (A) and *FP* (B) per method for all 6 data sets.

FN values are significantly higher for SD and PS compared with MN and NN (except for the *Mimulus* data set). *FP* values are relatively higher for MN and NN. This is in correspondence with visual inspection of the graphs (not shown), where the NN and MN graphs show much more network structure than the SD and PS graphs. NN and MN networks are more likely to include the splits of the N_e, but also more "redundant" splits, as many more splits are reconstructed in these methods. SD and PS showed many unresolved nodes, resulting in a high number of "missing" splits. In some reconstructed graphs the topology of N_e is not recovered at all, resulting in high values of *FN*.

There were also clear differences among the different data sets, with relatively high values of *FN* for SD in the *Mimulus, Populus* and *Dendrochilum* data sets, compared with the other three data sets. The *Mimulus* data set also shows a high *FN* value for NN. Two data sets from different ends of the spectrum of sequence variation (*Populus* and *Dendrochilum*) behave almost similar in terms of *FN*, making it difficult to link sequence divergence to performance of network reconstruction. *FP*

values did not differ much among data sets, except for higher *FP* values for PS, SD and CN in the *Stephanandra* data set.
Using the simulated data sets the hybrid terminal resolved in the expected position (our criterion 1) in a few cases (results not shown). Only the *Pelargonium* data sets recovered the hybrid as sister to one parent or in the expected position. NN performed best, with placement of the hybrid as sister to one or expected in more than 50% of the data sets, while MN and PS recovered the hybrid as sister to one parent in 30% and 3% of the data sets respectively. SD did not recover the relationship between hybrid and either of the parents.

Values of the *FN* and *FP* values for all data sets are given in Fig. 5.3. Most methods show the same pattern for all 3 data sets. SD shows high values of *FN*, whereas *FP* values of 0 were found for PS and SD. In fact, most network graphs consisted of just one or two nodes, with all taxa collapsed.

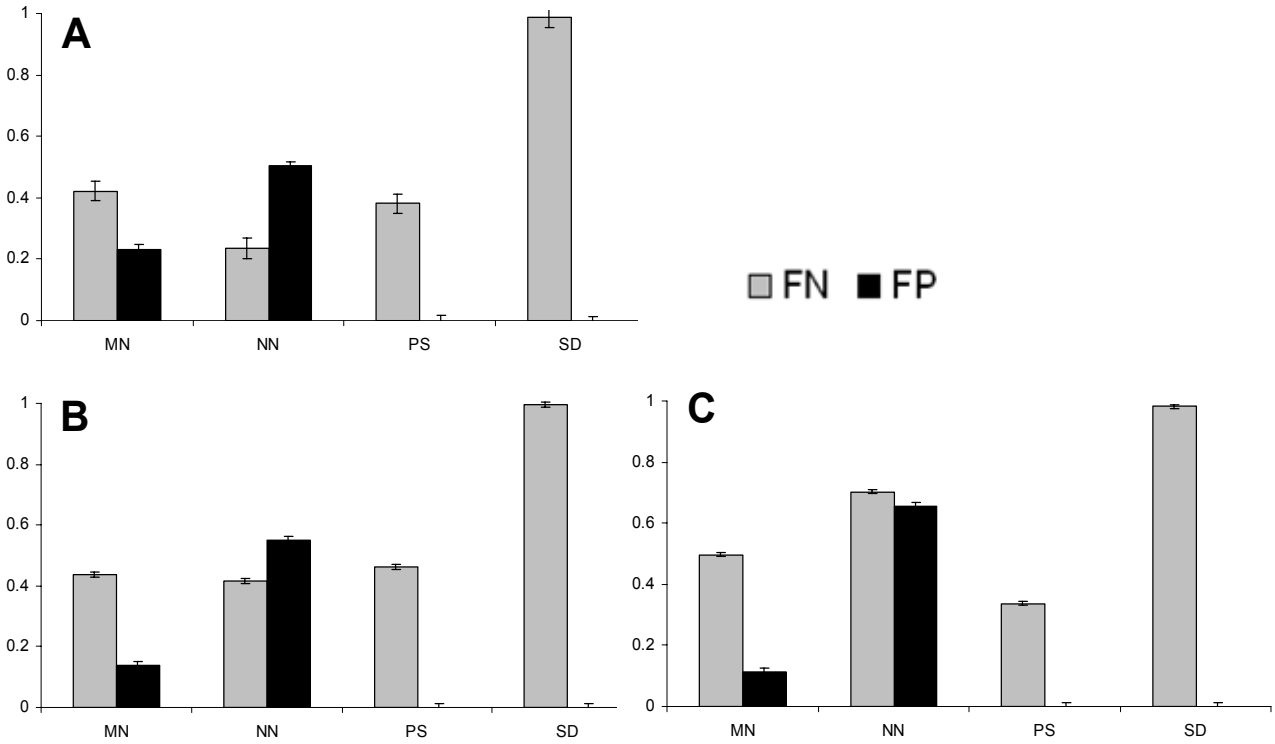


Fig. 5.3 Number of *FN* and *FP* per method for the three different simulated data sets based on tree topologies of Pelargonium (A) and Celtidoids with $\alpha = 0.09$ (B) and $\alpha = 2.62$ (C),averaged over all 30 replicate data sets. All values are also averaged over the three different scenarios of "hybridization" (see Chapter 4), because no significant differences in *FN* or *FP* per scenario were found.

Effect of removal of hybrid on tree and network analysis

In general, the effect of removal of the hybrid on (Bayesian) tree analysis did not have a high impact on tree recovery. After removing the hybrid from the *Dendrochilum* and *Mimulus* data set, posterior probabilities on a clade with one of the parents increased significantly (18% and 26%, respectively, data not shown). In the *Stephanandra* data set, one of the parental clades showed a (15%) *decrease* in posterior probability. In the latter case, this is probably due to a considerable reduction in the total number of taxa (due to the removal of 5 hybrid terminals), resulting in less resolution and lower confidence values in the analysis.

The effect of hybrid removal on *network* topology is presented in Fig. 5.4, as percentage decrease in *FN* and *FP* values. As expected, the *Viburnum* data set showed the smallest decrease, with only one hybrid terminal removed from a data set consisting of 43 taxa. *Dendrochilum*, *Hordeum*, *Mimulus* and *Populus* data sets all indicate a small decrease in *FP* values and a higher increase in *FN*. Apparently, the hybrid terminal impedes correct resolution of relationships among the other taxa. In the *Stephanandra* data set the decrease in *FN* is much less apparent upon removal of the hybrids, but here *FP* is strongly decreased. In this data set the 5 hybrid terminals caused many extra reticulations compared with the expected network, also illustrated by the high *FP* values in Fig. 5.2 for the *Stephanandra* data set. Per method, the strongest effect was measured for CN, the smallest for SD (Fig. 5.4), probably due to the low number of resolved relationships in the splits graphs.

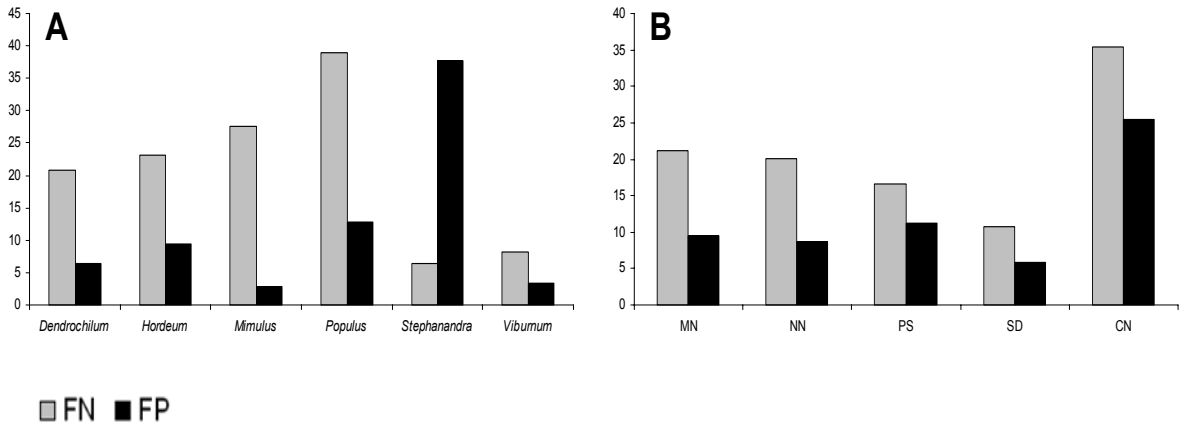


Fig. 5.4 Effect of the removal of the hybrid terminal on percentage decrease in *FN* and *FP* values per data set (left), averaged over all methods and per method (right), averaged over all data sets.

DISCUSSION

Use of network methods

Phylogenetic networks can visualize potentially competing signals from alternative phylogenetic signals obtained from bifurcating phylogenetic trees or data sets (Huber & Moulton, 2005; Winkworth et al., 2005). Here, we tested some of the available methods to see whether they are able to visualize ("detect") the hybrid terminal and whether they provide a useful alternative in displaying evolutionary organism-level relationships, using arbitrary topological criteria. Thus measured the methods tested showed clear differences in performance.

Split Decomposition never displayed the hybrid in the expected position and in many cases showed unresolved relationships among taxa. In general, SD may be of limited use in presenting reticulate relationships due to a conservative criterion for branch length selection in its default setting, as also implemented in SplitsTree implementation, i.e. the smallest value is chosen for the length of the split. This causes only well-supported splits to be represented and therefore a splits graph inferred by SD will only display the strongest signals of incompatibility (Perrie et al., 2003; Winkworth et al., 2005).
Its derived variant, Parsimony Splits, also shows many "missing connections" and overall low resolution. In our performance tests this resulted, just as with SD, in networks without reticulations connecting the hybrid and one parent to other lineages.

In the NeighborNet and Median Network analyses reticulations between the hybrid and other lineages can often be observed, resulting in the expected position (i.e. the first criterion) more often than in the other methods. However, in general NN displays many redundant connections, with reticulations among almost all clades, as is illustrated by the high *FP* values (see Fig. 5.2). Bootstrap support of the edges (internal branches) does not provide any evidence for a particularly stronger signal of reticulation between the hybrid and the other lineages than the average "reticulation-noise" in the graph. Although some of the reconstructed networks represented the correct relationships of the hybrid, the high numbers of false positives makes interpretation of the graph difficult.

It is not straightforward to assess the best performing network method. The *FN* and *FP* criteria give an indication of correspondence of N_e with N_r, but are not necessarily informative of hybrid relationships, i.e. a network with a high *FN* score can still include the "correct" hybrid connection, while a high *FP* score does not guarantee that the network includes the "correct" hybrid connection. These values,

however, do give an indication of the presence or absence of a network structure. Additionally, the hybrid position itself may not be a reliable test criterion. In MN and NN analyses, the hybrid is often placed in the expected position, while the other relationships in the graph are undecipherable. In general, it can be said that SD and PS display little resolution with our data sets, while NN and to a lesser extent MN contain too much (i.e. additional internal branches).

However, the results also appear to depend on size and level of DNA sequence divergence in the separate input data sets. Our simulated data sets resulted, in general, in networks with hardly any resolution (many polytomies) using PS and SD, or fully illegible graphs with reticulations among almost all taxa in NN and MN. These data sets include a large amount of taxa (71 or 80), and are highly variable, with the percentage of phylogenetically informative sites ranging from 25 – 74. Whether it is the number of taxa or divergence level per se that affect performance is not clear from our results. Clearly, all network methods in our test performed much better with the "real" data sets with a maximum of 45 taxa and lower values of DNA sequence variation. While sequence divergence and parental distance will no doubt influence network reconstruction performance, we could not pinpoint specific values to optimal performance of the different test criteria.

Networks are clearly useful as a visualization tool when the underlying evolutionary history of the data is complex and non-bifurcating. Moreover, even in case the underlying history is treelike, the history can be difficult to present as a single tree due to processes such as parallel evolution, model heterogeneity or sampling error (Huber & Moulton, 2005; Winkworth et al., 2005; Huson & Bryant, 2006). Network methods can also simply serve as a way to make full use of all data and to visualize the uncertainty in phylogenetic reconstruction. Even without the suspicion of reticulate evolutionary history, these methods can be used as additional tools to analyse the data from a different perspective, namely investigating the uncertainty of the data. A program such as SplitsTree could serve as a tool investigating the uncertainty of the data in addition to analytical and summarizing tools. It might even be used before any tree analyses is done, to check how tree-like the data is, and subsequently assessing whether it is justified to use tree-like phylogenetic analysis. However, at the moment no objective criterion exists to check whether data is tree-like "enough", but if available in the future, this could be a highly useful check prior to any phylogenetic analysis.

There is a striking lack of literature on relative performance testing of networks at species-level compared with literature describing the tree approach. At

the same time, in population-level studies networks are used frequently and apparently without problems, illustrating the division between the scientific community of population geneticists and systematists. Besides a "traditional" view that network methods belong to the population-level world, a "fear for phenetics" may be another obstacle for the use of networks in phylogenetic studies at species-level, because of the emphasis of most network methods on the use of distances.

It would be interesting to see what happens in a model-based approach, incorporating explicit models of DNA substitution to reconstruct networks, but it is not immediately clear how to apply models to a network. For instance, it is not straightforward how to locate substitutions in a cluster of splits because parallel splits, and the ones perpendicular to them, should undergo the same substitutions. Additionally, a single split (i.e. a node) or a cluster of splits cannot be directly linked to a substitution event, since the internal nodes do not necessarily represent ancestors (e.g. Bryant & Moulton, 2004). Therefore, optimizing parameter values and calculating likelihood values for a given data set and a set of parallel splits, may be unfeasible. It might be that parallel splits need to be collapsed first in order to pursue the calculations. Some studies have addressed this problem (e.g. Strimmer & Moulton, 2000, 2001) and suggested an approach where phylogenetic networks first need to be rooted and converted into directed acyclic graphs (DAG's) and then into a Bayesian network to be able to infer parameter values using a probability distribution involving all nodes of the network. Subsequently Monte Carlo simulations are used to approximate the likelihood values. This approach allows comparison among competing networks or tree topologies and the example data set showed for instance better likelihood scores for a network over the competing tree (Strimmer & Moulton, 2000). DNA sequences (from HTLV viruses) were used as example data, but it is also possible to extend the approach to other kinds of data (e.g. amino acid sequences). In its current implementation, however, likelihood calculations are not feasible for large data sets and "in the general case, it is not all clear how networks should be unambiguously rooted or how prior probabilities should be assigned to parent nodes" (Strimmer & Moulton, 2000). Until now, no further applications of this method using actual data sets have been published.

Other model-based approaches to organism-level network construction are described in e.g. Jin et al. (2007) who calculate LnL of the data under a network perspective by combining likelihood values of its decomposed trees ("bi-components"). Improvement in likelihood scores of the sequence data is investigated upon introduction of HGT events ("reticulation edges") to obtain an optimal phylogenetic network. This looks a promising direction, with future research plans

that include developing more computationally efficient implementations to analyze large data sets and implement different models (Jin et al., 2007).

Implications for future hybrid studies

In this age of genomics, the input data for phylogenetic studies will rapidly change (e.g. Brown, 1996). With the availability of complete genomes, much larger data sets can be compared instead of using a limited number of genes (Boore, 2006; Clark, 2006). In addition, other genomic data, such as gene arrangements, presence/absence of genes and positions of short interspersed elements (SINEs) will become available to serve as phylogenetic markers (Boore, 2006). Modelling may become prohibitively complicated, which is why returning to a parsimony approach is being advocated by some (e.g. Albert et al., pers. comm.).

This will also have implications for the use of networks in phylogenetics. Genomic data have probably more signal and can give especially more conclusive answers about hybridization (Linder & Rieseberg, 2004). With the analysis of just a few genes, it is more difficult to discern between conflict in the data due to e.g. sampling error and "actual" conflict due to underlying conflicting evolutionary histories of the genes involved. However, using genomic data, noise will be less of a problem, especially with the essentially homoplasy-free markers such as SINEs (Shedlock & Okada, 2000). These kinds of markers are also suitable to detect gene flow between species (Deragon & Zhang, 2006). If "real" reticulate events such as hybridization are concerned, the information of multiple unlinked genes will probably provide more statistical power to determine the conflicting signals than with the use of just a few genes (e.g. Sang & Zhong, 2000; Linder & Rieseberg, 2004; Huson et al., 2005). Therefore, the need for modelling this kind of data, if possible at all, will probably be less because signal in the data may be strong enough already.

chapter 6

AFLPs AND HYBRID DETECTION – A CASE STUDY OF SOLANUM

B. Vriesendorp, M. Vrielink-van Ginkel, R.G. van den Berg

INTRODUCTION

During the last decade, reticulate evolution has increasingly become an important topic, with several recent reviews that have been published on the occurrence and evolutionary consequences of hybridization (e.g. Rieseberg, 1997, Rieseberg & Carney, 1998; Ellstrand et al., 1996, Hegarty & Hiscock, 2005). In addition to these general publications indicating the importance of hybridization, many recent angiosperm studies focussed on the detection of hybrids in a phylogenetic context, e.g. studies in *Capsella* (Slotte et al., 2006), *Rosmarinus* (Rosello et al., 2006), *Mercurialis* (Obbard et al., 2006) and the Brassicaceae (Marhold & Lihova, 2006). Most molecular-based methodologies to investigate hybrid relationships involve DNA sequence data, but AFLP has proven to be another potentially powerful marker system to do this.

There are many examples of studies concerning hybridization where AFLPs have been used, most of them at population-level, but also some at species-level. We focus here on the latter. Some examples from recent studies in angiosperms using AFLPs are listed in Table 6.1. The term hybrid is often used loosely and can imply a whole range of entities, e.g. F1 hybrids or established ancient species, and there can be a range of different parental distances (see Chapter 2 and 3 of this thesis for a summary of terms and a conceptual framework for the terminology). Here, we consider only (putative) hybrids between different species, but without further limitations to their nature (e.g. F1 or ancient).

AFLP is a DNA fingerprinting technique based on the selective amplification of fragments resulting from a total digest of genomic DNA (Vos et al., 1995). This genome-wide mapping provides signals from both parental lineages, in contrast to sequencing of one or a few genes, where often just one of the parental contributions is analyzed. Whereas these parental genes are not distributed evenly over the genome, AFLP markers are in theory well dispersed. This makes AFLPs potentially suitable markers for the molecular analysis and detection of hybrids.

Most studies on AFLP markers in a hybrid between different species focus on a detailed investigation of many accessions of the involved species, looking for additivity of parental bands. Additivity of AFLP-markers is based on the recognition of species-specific bands in the hybrids, where "species-specific" is defined in different ways, e.g. as bands that are exclusive to a species and occur in at least 90% of the individuals of one or more of its populations (Guo et al., 2006).

Studies that investigate patterns of species-specific bands involve only a few species (see Table 6.1). These studies reveal a high percentage of species-specific bands and only a few unique bands in the hybrid (Hodkinson et al., 2002; Divakaran

et al., 2006) or even complete additivity where all bands in the hybrids are shared with one or both parents (e.g. Teo et al., 2002; Chauhan et al.; 2004 Guo et al., 2006).

Table 6.1. Selection of angiosperm studies where AFLP analysis was used to resolve hybrid relationships.

Genus	References	Species/ /acc.[a]	AFLP-analysis[b]	Ploidy level
Achillea	Guo et al., 2006	7 / 169	band freq.	polyploid
Calopogon	Goldman et al., 2004	7 / 60	PCO; phenetic	polyploid
Conostylis	Krauss & Hopper, 2001	5 / 36	phenetic: MP: MDS	polyploid
Dactylorhiza	Hedren et al., 2001	23 / 108	phenetic; PCO	polyploid
Mangifera	Teo et al., 2002	3 / 20	band freq.; phenetic	diploid (?)
Mentha	Gobert et al., 2002	8 / 62	phenetic; MP; PCO	mixed
Miscanthus	Hodkinson et al., 2002	3 / 10	phenetic; PCO	allotriploid
Onopordum	O'Hanlon et al., 1999	10 / 80	band freq; PCO; NMDS	diploid
Pistacia	Kafkas, 2006	11 / 35	phenetic, PCO	diploid (?)
Polylepis	Schmidt-Lebuhn, 2006	29 / 48	PCO; phenetic; MP	diploid (?)
Populus[c]	Chauhan et al., 2004	3 / 31	band freq; PCO; phenetic	diploid (?)
Salix	Beismann et al., 1997	3 / 26	phenetic, PCO	diploid (?)
sum	Lara-Cabrera & Spooner, 2004	19 / 63	phenetic; MP	diploid
Trillium	Kubota et al., 2006	9/65	PCO	polyploid
Utricularia	Kameyama et al., 2005	3/26	band freq; PCO	diploid (?)
Vanilla[c]	Divakaran et al, 2006	3/16	band freq; phenetic	diploid (?)
Vasconcellea	Van Droogenbroeck et al., 2006	5/61	phenetic: PCO;band freq	diploid (?)

[a] Number of species (including hybrids) / number of accessions included in the study.
[b] AFLP-analyses is based on different categories; "band freq" = analysis of additivity of species-specific bands in the putative hybrid; "PCO" = principal coordinates analysis; "(N)MDS" = (Non-metric) mulitdimensional scaling; "MP" = maximum parsimony analysis; "phenetic" = UPGMA and/or NJ analyis
[c] Hybrid is result of artificial crossing

When more than a few species are included the most commonly used methods to analyze the AFLP data are clustering (mostly UPGMA or NJ) and ordination using PCO or PCA plots. The results of cluster analyses or ordination often put the hybrids either close to one of its parents or in an intermediate position. Examples of studies using PCO where hybrids indeed cluster as expected in between the parental taxa are those on *Mentha* (Gobert et al., 2002), *Dactylorhiza* (Hedren et al., 2001) and *Trillium* (Kubota et al., 2006). In UPGMA or NJ analysis three main hybrid positions can be found. Hybrids cluster either close to one of the putative parents (Kafkas, 2006; Teo et al., 2002), different hybrid accessions cluster to different parents (e.g. Divakaran et al., 2006; Gobert et al., 2002), or hybrids take a more basal position in the tree (e.g. Kardolus et al., 1998; Krauss & Hopper, 2001).

Another option is to use AFLP data in a phylogenetic (e.g. MP or ML) approach, but only few studies have been performed using this type of analysis. Since in theory AFLPs are representatives of the whole nuclear genome, it is expected that a hybrid has AFLP bands from both parents. This could resolve in either an intermediate position or in inclusion in the clade with the parent with which it shares most synapomorphic characters (McDade, 1995). Lara-Cabrera & Spooner (2004) investigated the placement of the putative hybrid species *Solanum x michoacanum* and *Solanum x sambucinum* using AFLP and found different results for cladistic and phenetic analyses. While the hybrids did not group with either of the parental groups in the MP tree, the phenetic analysis resulted in clustering with one of the putative parental species (*S. pinnatisectum* for both hybrids). In *Polylepis* (Schmidt-Lebuhn et al., 2006) the hybrid was resolved within the clade that included the parental species and in *Mentha* (Gobert et al., 2002) the hybrid appeared as sister to one of its (putative) progenitors. These last two studies showed identical results for both phenetic and MP analyses.

Additional analyses could include removing the hybrid from the data set to investigate its effect on tree topology and support values. As a hybrid individual could potentially disturb the tree topology, it is expected that removing it results in higher topological resolution or higher bootstrap values. A few studies based on DNA sequence data indeed resulted in a higher resolution (e.g. Aguilar & Feliner, 2003; Sun et al., 2002. This was also done for AFLP data in *Polylepis* (Schmidt-Lebuhn et al., 2006), with a loss of resolution in the clade including the putative hybrid and its parents upon exclusion of hybrid. Unfortunately, to our knowledge no other studies including AFLPs in hybrid analyses have performed these kind of explicit testing.

Finally, analysis of the AFLP data set using network reconstruction tools can be an additional approach to reconstruct hybrid relationships. To our knowledge, only one published study on network analyses of AFLPs at species-level exists. This study concerns fern species where 4 different lineages within an allopolyploid species complex are analyzed using Split Decomposition (Perrie et al., 2003).

Objectives

In this study we will investigate the suitability of AFLPs to provide markers that enable hybrid detection. The objectives of this study are to (i) explore patterns of AFLP additivity and (ii) investigate behaviour of AFLPs in phylogenetic and network analyses with respect to the hybrid individual and its parents. We use a putative ancient hybrid and a re-synthesized F1 hybrid from a crossing of the same putative parental species. All material belongs to *Solanum* L. sect. *Petota* Dumort.

Hawkes (1990) defined *S. x michoacanum* (Bitter) Rydb. as "definitely a natural hybrid" of *S. bulbocastanum* Dunal and *S. pinnatisectum* Dunal. Both species occur within the same geographical region and *S. x michoacanum* is morphologically intermediate between the two species (e.g. Spooner et al., 2004). Based on this hypothesis, artificial crossings between *S. pinnatisectum* x *S. bulbocastanum* were performed to produce an interspecific F1 hybrid. For the additivity analysis we use this F1 hybrid and its parental species. We also use another artificial F1 hybrid within the Mexican group, between *S. pinnatisectum* and *S. brachistotrichium* (Bitter) Rydb., where the parental accessions are known, but not the exact individuals. These F1 hybrids were produced in a project to test the EBN (endosperm balance number) hypothesis (Koopman, unpublished).

Table 6.2a. Accessions of Solanum species used in hybrid crossings and additivity analyses.

Species	Number of accessions			
	Ind. level	Acc. level	Species level	Total
S. pinnatisectum (PNT)				
CGN_17745	1		5	5
CGN_17743[a]		3	3	3
CGN_2650 (204-4)[b]			1	1
CGN_507 (374-2)[b]			1	1
S. bulbocastanum (BLB)				
CGN_22367	1		5	5
CGN_4072 (515-1)[b]			1	1
CGN_1171 (517-1)[b]			1	1
F1 Hybrid (PNT x BLB)[c]	4			4
S. brachistotrichium (BST)				
CGN_17681[a]		3		3
F1 Hybrid (PNT x BST)[a]		3		3
S. x michoacanum (MCH)				
CGN_2573 (185-2)[b]			1	1
CGN_3005 (279-3)[b]			1	1
Herbarium[d]			1	1
Total number	6	9	20	30

[a] Material available from Koopman (unpublished)

[b] Number in brackets is CBSG accession (and clone) number

[c] F1 hybrid is crossed between *S. pinnatisectum* (CGN 17745) as maternal parent and *S. bulbocastanum* (CGN 22367) as paternal parent.

[d] Herbarium specimem, Rivera-Pena et al. 903

METHODS

Plant material

In total 30 samples were used for the additivity analysis, including the species *S. pinnatisectum* (PNT), *S. bulbocastanum* (BLB), *S. brachistotrichium* (BST), the F1 hybrids between PNTxBLB and PNTxBST and the ancient hybrid *S. x michoacanum* (MCH), see Table 6.2a. This part of the study involved analysis at three different levels, each with different selections of samples:
i) Individual level (F1 of PNT & BLB): 4 F1 hybrids, 1 PNT (maternal parent), 1 BLB (paternal parent) (6 samples).
ii) Accession level (F1 of PNT & BST): 3 F1 hybrids, 3 PNT individuals (from the accession of the maternal parent), 3 BST individuals (from the accession of the paternal parent) (9 samples).
iii) Species level (Ancient hybrid species and its parental species): 3 ancient hybrids MCH, 10 PNT and 7 BLB individuals (from different accessions) (20 samples).
Plants were grown from seeds from the CGN (Centre for Genetic Resources), see Table 6.2a. Flowers of *S. pinnatisectum* and *S. bulbocastanum* were emasculated followed by hand-pollination. This resulted in only one successful crossing between *S. pinnatisectum* (female) and *S. bulbocastanum* (male), with 4 fruits. Seeds from 1 fruit were collected and grown, resulting in 4 F1 hybrid plants which were used to collect fresh leaf material for molecular analysis.

S. pinnatisectum, *S. brachistotrichium* and its F1 hybrid (accession-level) were grown using seeds available from earlier crossing experiments (Koopman, unpublished). As further representatives of the ancient hybrid, plant material from 1 herbarium specimen (Rivera-Pena et al. 903 – WAG) and extracted DNA from two CGN accessions were used, see Table 6.2a. An AFLP analysis was performed on the 30 samples using two EcoRI/MseI AFLP primer combinations, E35/M48 and E32/M49.

To study the behaviour of AFLPs in phenetic and phylogenetic analyses we used an existing data set containing cpDNA sequences and AFLPs of the group of Mexican diploid *Solanum* species generated within the CBSG (Centre for Biosystem Genomics) program (Jacobs et al., submitted). The group of Mexican diploid species consists of 14 recognized species plus two putative ancient hybrid species, *S. x michoacanum* and *S. x sambucinum*. Most species were included with several accessions, resulting in a data set of 39 accessions and 1 outgroup, *S. palustre* (Table 6.2b). The sequences from chloroplast regions *trnL-F* and *psbA-trnH* were combined in one data set. This data set consists of 2371 nucleotides with 36 (1.5%) phylogenetically informative

Table 6.2b. Mexican diploid Solanum accessions used in this study (data from Jacobs et al., submitted)

Species	CBSG[a]	Collector	Gene Bank[b]
S. bulbocastanum Dunal (BLB)	331-5	HAW 1595	CGN_17693

S. bulbocastanum (BLB)	330-4	HAW 1593	CGN_21306
S. bulbocastanum (BLB)	330-5	HAW 1593	CGN_21306
S. bulbocastanum (BLB)	524-4	SMHV 7043	CGN_21364
S. brachistotrichium (Bitter) Rydb. (BST)	117-4	GRA 347 x 348	PI_255529
S. clarum Correll (CLR)	568-1	SMHV 7011	PI_604052
S. clarum (CLR)	52-1	HAW 1833	PI_275202
S. cardiophyllum Lindl (CPH)	336-1	GRA 118 x 207	CGN_18325
S. cardiophyllum (CPH)	336-9	GRA 118 x 207	CGN_18325
S. cardiophyllum (CPH)	337-3	HAW 1059 x HHLs 1729	CGN_18326
S. cardiophyllum (CPH)	541-2	HAW 1010 x 1032	CGN_22387
S. cardiophyllum (CPH)	539-5	BGRC 55227	BGRC_55227
S. cardiophyllum (CPH)	542-3	WRF 1274	CGN_17697
S. ehrenbergii (Bitter) Rydb. (EHR)	155-5	HAW 1100	PI_186548
S. ehrenbergii (EHR)	153-3	HAW 1095	PI_184765
S. ehrenbergii (EHR)	154-2	HAW 1097	PI_184767
S. jamesii Torr. (JAM)	268-1	CPC BPC 463	CPC_7510
S. jamesii (JAM)	355-7	GRA 381 x 386	CGN_18349
S. jamesii (JAM)	672-2	CPC 3850 x 5847	BGRC_53860
S. lesteri (LES)	20-5	SHGRF 4155	PI_558434
S. lesteri (LES)	21-5	SHGRF 4177	PI_558435
S. x michoacanum (Bitter) Rydb. (MCH)	185-1	HHLs 1541 x 1517	GLKS_2346
S. morelliforme Bitter & Muench (MRL)	187-4	HAW 1805	PI_275222
S. morelliforme (MRL)	78-1	RSSV 946	PI_619119
S. morelliforme (MRL)	79-1	TRHRG 178	PI_545720
S. nayaritense (Bitter) Rydb. (NYR)	27-41	TRHRG 234a	PI_545820
S. nayaritense (NYR)	26-5	TRHRG 225	PI_545825
S. polyadenium Greenm. (PLD)	377-2	UGN 1766	CGN_17746
S. polyadenium (PLD)	377-3	UGN 1766	CGN_17746
S. polyadenium (PLD)	376-2	EBS 51	CGN_17749
S. pinnatisectum Dunal (PNT)	375-9	ROC S- 44	CGN_17743
S. pinnatisectum (PNT)	374-3	HAW 1505	CGN_17745
S. pinnatisectum (PNT)	778-5	HAW 1455	CGN_23011
S. pinnatisectum (PNT)	204-5	HAW 1041	GLKS_1586
S. pinnatisectum (PNT)	880-10	GLK 124. 6	CGN_18335
S. x sambucinum Rydb. (SAM)	92-1	ROD 2563	PI_595478
S. trifidum Correll (TRF)	881-1	GRA 301 x 244	CGN_22722
S. tarnii Hawkes et Hjert. (TRN)	228-2	TRHRG 62	PI_498046
S. tarnii (TRN)	40-2	HJT 7365	PI_570642
S. palustre Poepp. (PLS)	284-2	CPC BPC 1030	CPC_7034

[a] Accession number from CBSG (Centre for Biosystem Genomics) project (Jacobs et al., submitted)

[b] Accesion numbers from different Gene Banks:

CGN = Centre for Genetic Resources, the Netherlands

PI = US Potato gene bank ("Plant Introduction" number)

BGRC = Braunschweig Genetic Resources Collection

CPC = Commonwealth potato collection

GLKS= Gros Lusewitzer Kartoffel-Sortimente

characters. The AFLP data set was based on the same primer-combinations used in the first part (see above) and contained 202 informative bands.

Data analysis

1. Additivity of AFLP bands

Species-specific bands are defined here as bands that are present in at least one of the accessions of a specific species and completely absent in all accessions of the other species. Shared bands between 2 species are bands that occur in both species in at least one of the accessions and are completely absent in all other species.

2. Behaviour of AFLPs in phylogenetic analyses

Ordination analysis

A principal component analysis (PCA) was performed based on the correlation of the AFLPs and a principal coordinates analysis (PCO) based on the similarity matrix, both using the NTSYS-pc software package version 2.2 (Rohlf, 2004).

Phylogenetic and phenetic analysis

Maximum Parsimony (MP) analyses were conducted using PAUP* 4.0b10 (Swofford, 2002), Unix version. Heuristic search of shortest trees was performed using 1000 replicates of random addition sequences with 1 tree held at each step during stepwise addition and TBR branch swapping. The robustness of the resulting trees was estimated by bootstrapping (1000 replicates) using TBR branch swapping. Parsimony jackknife support was calculated using PAUP* with the following settings: 37% of characters deleted with Jac emulation, 1000 replicates, full heuristic search criterion, TBR branch swapping, 10 random-addition sequence replicates, 1 tree held at each step during stepwise addition, saving 1 tree per replicate (following Freudenstein et al., 2004). For the NJ analysis distance matrices were calculated using Nei & Li's (1979) genetic distance, with bootstrap support for the clusters obtained by using 1000 replicates, using PAUP*.

A Bayesian MCMC analysis was performed using MrBayes 3.1.2 (Huelsenbeck & Ronquist 2001; Ronquist & Huelsenbeck 2003) with the parameters for the model of sequence evolution selected by applying the hLRT criterion in MrModeltest 2.2 (Nylander, 2004). For the AFLP-data set no model was selected, due to the lack of such a model for these kind of markers (see also Brouat et al., 2004). Default settings were used (standard model, 2 state and equal rates) except for the lset coding to "no absencesites". Two sets of 4 MCMC chains were run for 4 million generations or until the standard deviation of split frequencies was below 0.01. MrBayes settings for the best-fit model were selected by hLRT in MrModeltest 2.2. All three data sets were run with lower temperature (of 0.05) because this resulted in good mixing of the chain. Sampling and summarizing of trees was performed following the default settings.

To estimate the effect of potential distortion of the putative hybrid terminal on all tree analyses, *S. x michoacanum* or one of the parental taxa were removed from the data set, checking the influences on tree topology and bootstrap values or posterior probabilities.

All network analyses were performed using SplitsTree v.4.6 (Huson & Bryant 2006). Five methods were chosen: the distance methods NeighborNet (NN) and Split Decomposition (SD), the character-based methods Median Network (MN) and Parsimony Splits (PS), and the tree-based method Consensus Networks (CN). The chloroplast sequences and the AFLP data were analyzed both separately and in a combined data set. For the tree-approach (Consensus Networks) 1 tree (majority rule consensus tree) per data set was included to study conflict between the different consensus trees. In addition, bootstrap replicate trees from the separate analyses were used as input to investigate incongruencies within the separate data sets. Character transformation is only applicable for the distance methods (NN and SD). For AFLP no character transformation was used (i.e. the default setting of p-distance was applied) since there is no model of AFLP character evolution available. Selection of the model parameters in the chloroplast data was estimated by ModelTest, see above. Bootstrapping (1000 replicates) of the resulting networks was applied to all the NeighborNet analyses.

RESULTS

1. Additivity of AFLP bands

The AFLP primer combinations E32/M49 and E35/M48 generated 85 and 77 polymorphic bands respectively for the complete data set of 30 samples (Table 6.2a). The bands ranged in size from 48bp to 480 bp. Band frequencies from the separate primer-combinations were almost identical and for further analysis the results from these two data sets were combined. A summary of the distribution of all AFLP bands among the hybrids and parents is given in Table 6.3. At all three (individual, accession and species) levels species-specific bands for the parents (or parental accessions) could be indicated. At the individual and accession level a high proportion of these bands are shared by the F1 hybrid, while at the species level a much lower percentage of the species-specific bands can be found in the putative ancient hybrid species.

Table 6.3. Distribution of AFLP bands for each of the three different testing-levels and different sample sizes. The individual level includes 6 samples (1xP1, 1xP2 and 4xF1); the accession level 9 samples (3xP1, 3xP2 and 3xF1) and at species level 20 samples are included (7xP1, 10xP2 and 3xMCH).

Level	Total	Shared AFLP-bands						
		all	P1 + P2	P1 unique	P1 + H	P2 unique	P2 + H	H unique
Individual:								
P1=PNT, P2=BLB, H=F1 hybrid	131	35	3	2	44	2	33	12
Accession:								
P1=PNT, P2=BST, H=F1 hybrid	140	72	4	2	26	8	21	7
Species:								
P1=BLB, P2=PNT, H=MCH (ancient hybrid)	172	81	10	19	26	11	21	4

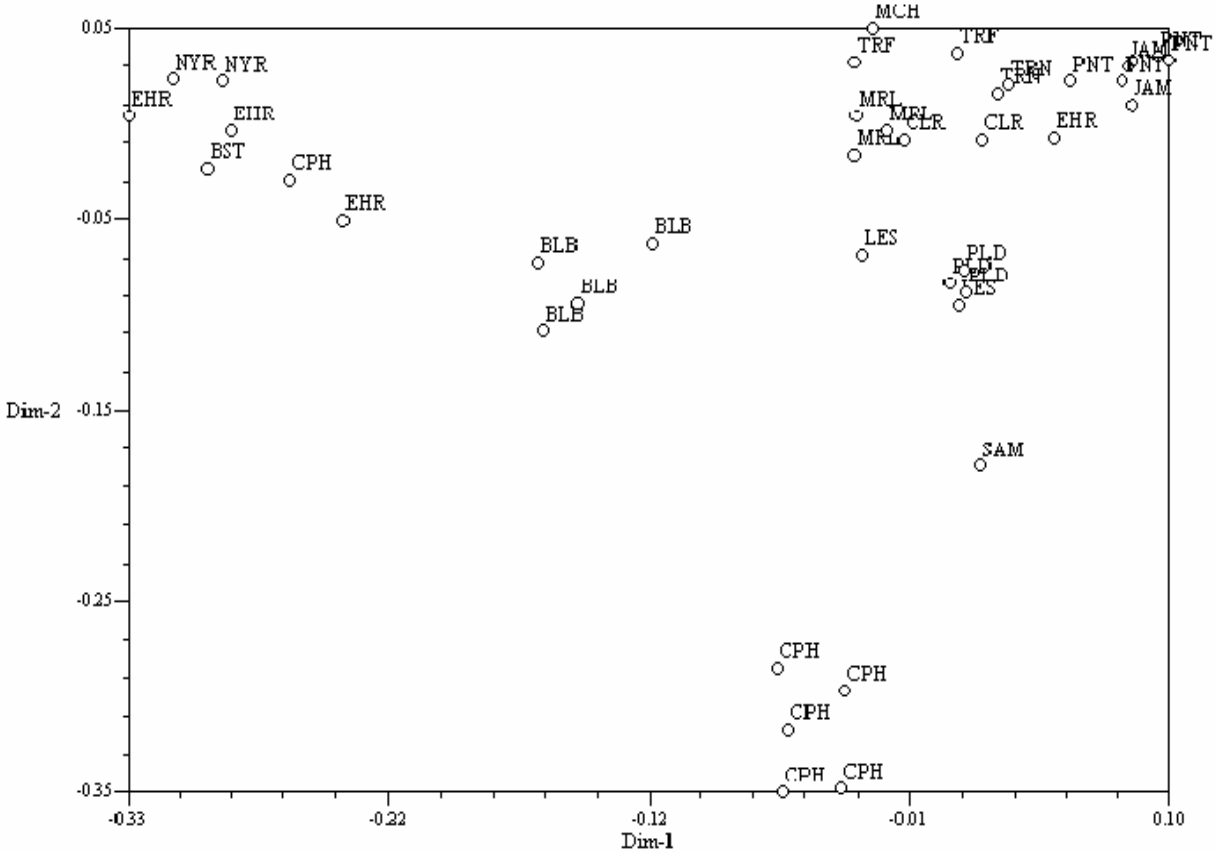


Fig. 6.1a. PCA of AFLPs Mexican diploid *Solanum* species. See Table 6.2a for abbreviations of species names.

2. Behaviour of AFLPs in phylogenetic and cluster analyses

The plots from PCA and PCO are shown in Fig. 6.1a and b. The proportions of total explained variance along the first and second axes were 47.2 % and 8% in the PCA analysis and 15.6% and 12.1% in the PCO analysis. In both graphs, no clear intermediate position of the hybrid (MCH) can be seen between the putative parental

species (*S. pinnatisectum* and *S. bulbocastanum*). The putative ancient species (*S. x michoacanum*) clusters with *S. trifidum* in a group with *S. pinnatisectum* but there is no connection with the other parent, *S. bulbocastanum*, which is placed far away from *S. x michoacanum.*

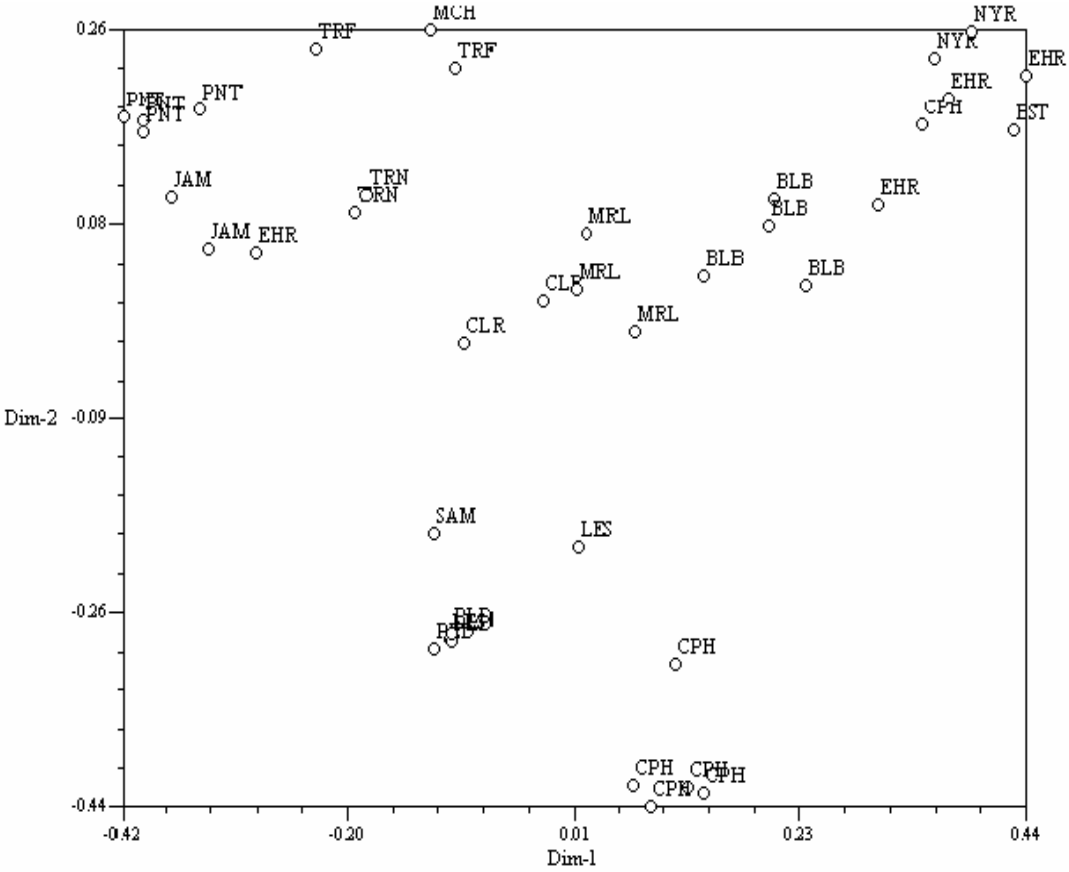


Fig. 6.1b. PCO of AFLPs Mexican diploid *Solanum* species. See Table 6.2a for abbreviations of species names.

Tree analysis

The NJ, MP (Jackknife and heuristic search) and Bayesian analyses of chloroplast DNA sequences, AFLP and the combined data all resulted in almost identical trees. Here, we only show the (50%) majority rule consensus trees recovered by the Bayesian analysis. The trees resulting from the separate and combined analyses of the chloroplast and AFLP data are represented in Fig. 6.2a-c. The tree from the chloroplast data (Fig. 6.2a) resolved the putative ancient hybrid (*S. x michoacanum*) in a basal position to a clade which includes *S. jamesii, S. nayaritense*, *S. ehrenbergii*, *S. morelliforme*, *S. clarum*, *S. polyadenium* and one of the putative parental species, *S. pinnatisectum*. In the separate AFLP analysis (Fig. 6.2b) and the combined analysis (Fig. 6.2c) *S. x michoacanum* is placed as sister to one accession of *S. trifidum*, having 2 accessions

of *S. tarnii* as closest relatives. This group of four is placed as sister to a group containing *S. jamesii, S. ehrenbergii* and *S. pinnatisectum*.

Effect of exclusion of taxa

Excluding the hybrid (*S. x michoacanum*) or one of the putative parents (*S. pinnatisectum* or *S. bulbocastanum*) from the data set had no effect on tree topology and only a minor effect on posterior probabilities in the MrBayes trees. (Posterior probability values on most clades remain unchanged, except for an increase of 0.55 to 0.56 on the *S. bulbocastanum* clade after removal of the hybrid or *S. pinnatisectum*.)

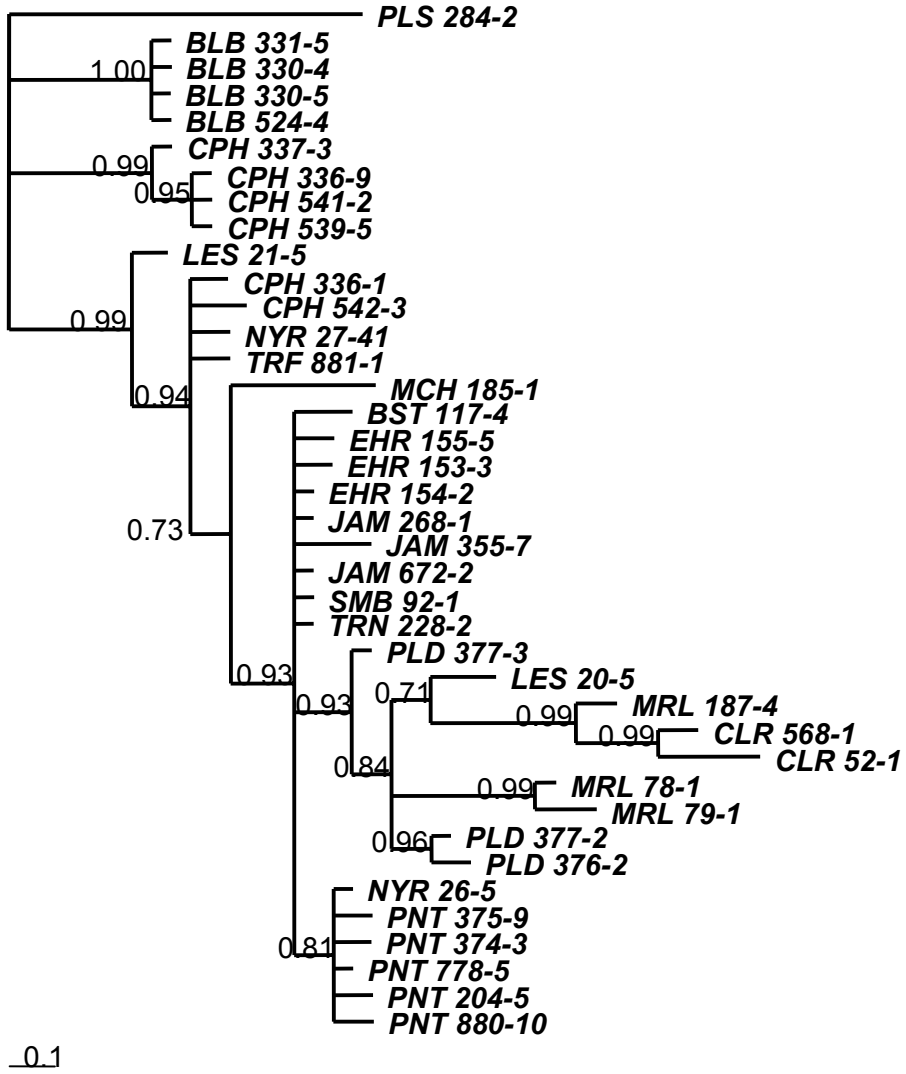


Fig. 6.2a. 50% majority rule consensus of the trees from Bayesian analysis based on trnL-F and psbA_trnH sequences.

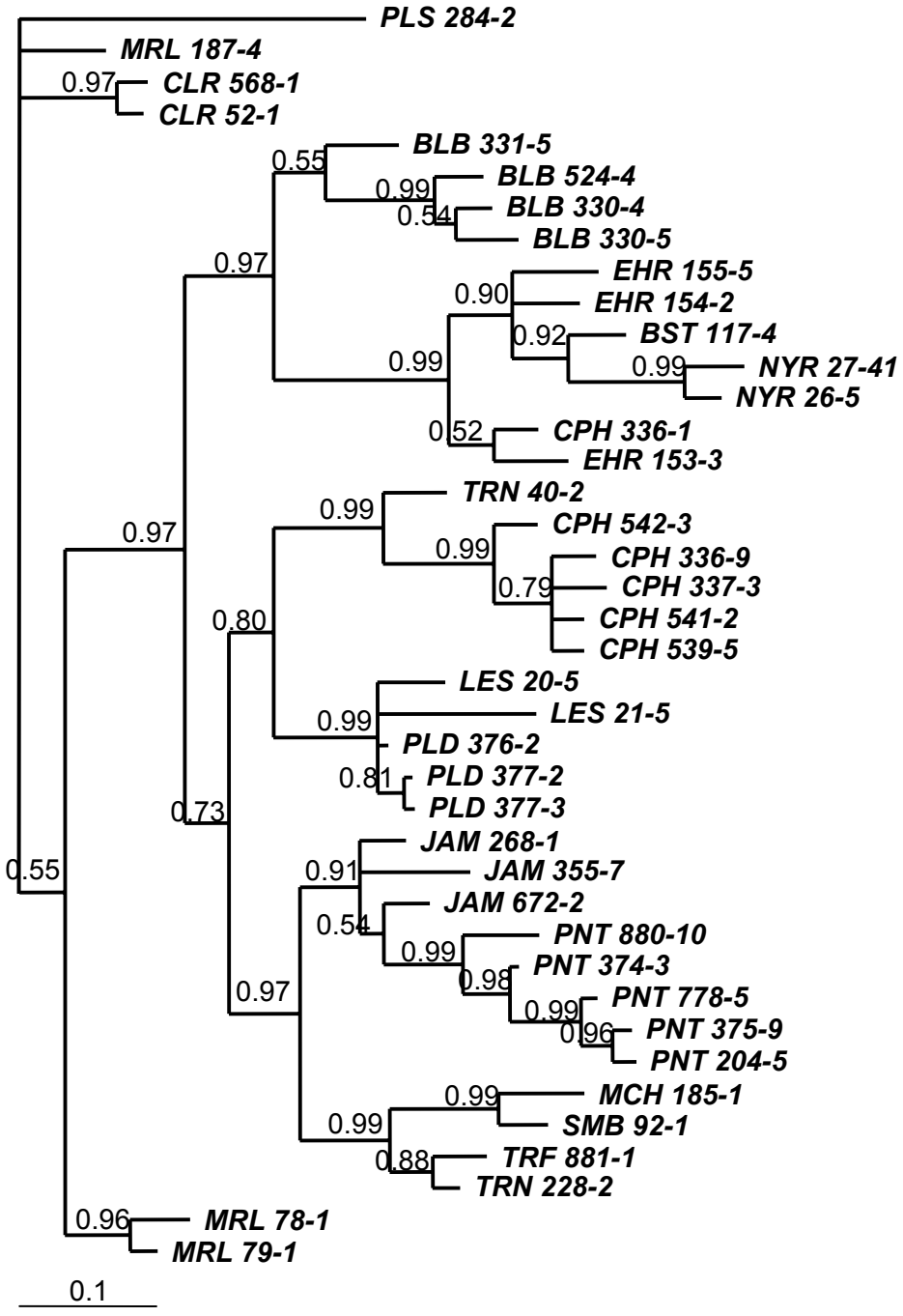


Fig. 6.2b. 50% majority rule consensus of the trees from Bayesian analysis based on AFLP bands.

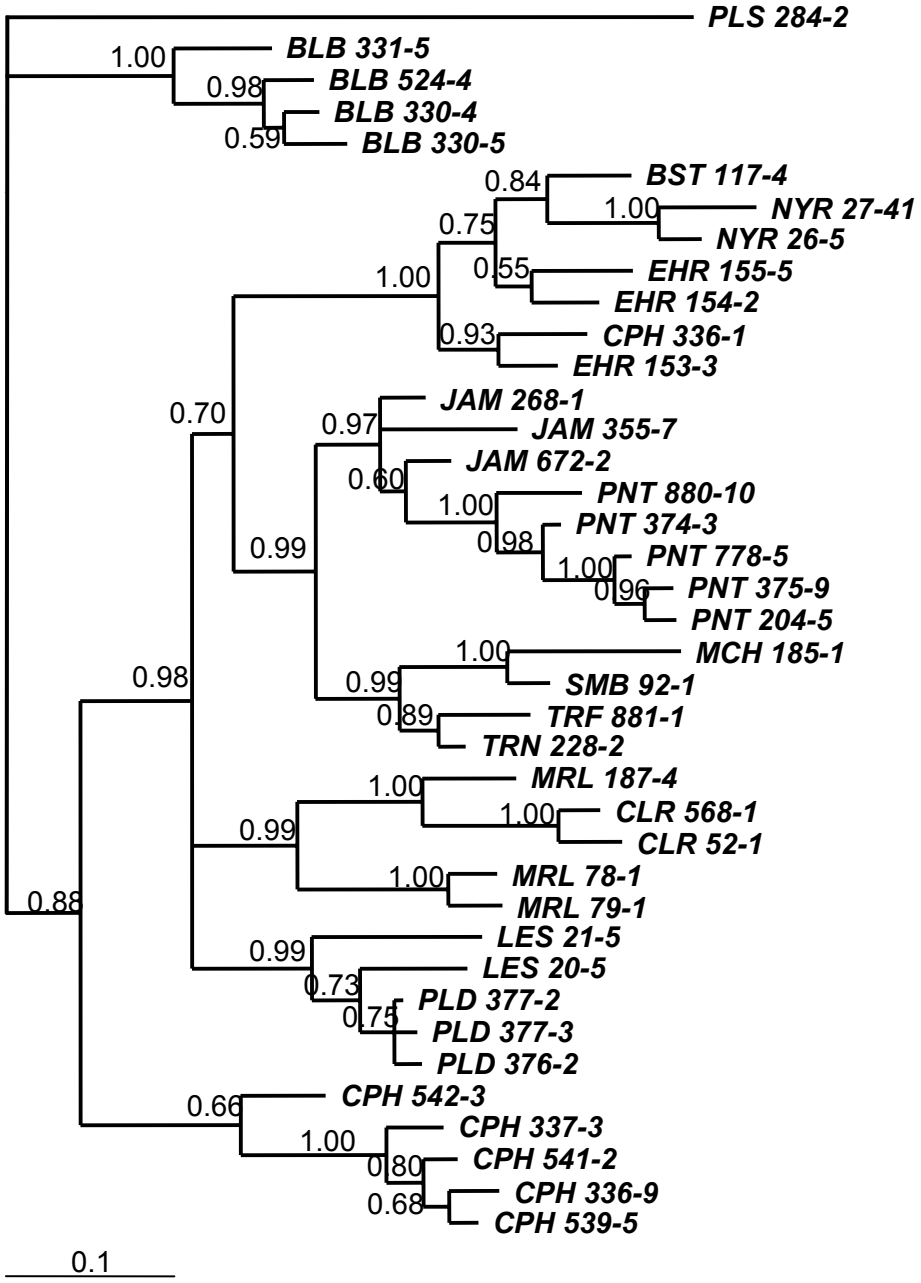


Fig. 6.2c. 50% majority rule consensus of the trees from Bayesian analysis based on combined data from chloroplast DNA sequences and AFLPs.

Network analysis

Only the results for NN and CN are discussed here. SD and PS showed too little resolution in their resulting networks and MN presented too many reticulations (additional internal branches) and made the figures and relationships in the graph undecipherable.

For NN the separate AFLP analysis is given in Fig. 6.3a. This graph was almost identical to the results from the combined analysis, and the separate chloroplast

analyses resulted in little structure, therefore these two graphs are not shown. The results largely resemble the tree analyses. *S. x michoacanum* is placed as sister to *S. trifidum* in the network, with high bootstrap support values. Connections to the other species also concur with relationships in the bifurcating trees. In a network analysis of a "true distinguishable hybrid" it would be expected that the hybrid shows connections to both parents, but no direct split that connects *S. x michoacanum* to *S. bulbocastanum* can be found here.

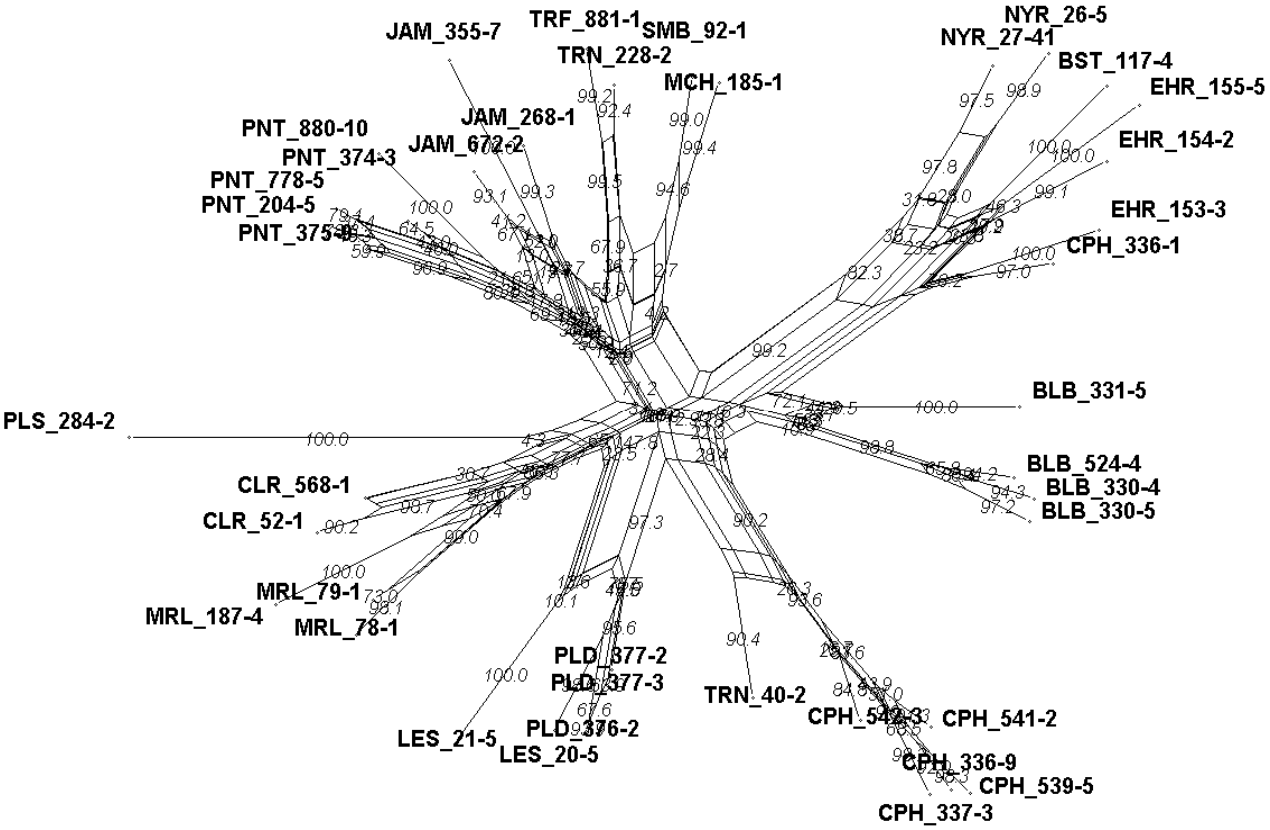


Fig 6.3a Network analysis - NeighborNet of AFLP data. Bootstrap values (1000 replicates) are indicated on the edges.

The CN analysis inferred conflicting branches between the three consensus trees from the separate analyses and displays the conflict between these trees, see Fig. 6.3b. Since there is no conflict in the position of *S. x michoacanum*, there are also no reticulations between *S. x michoacanum* and other species in the resulting CN graph. Another objective was to investigate the possible conflicting signals within the AFLP data combining all bootstrap trees (8000) in the CN analysis, see Fig. 6.3c. At the basal position where different clades come together some reticulations can be found.

In addition, within clades additional network branches are displayed. However, this leads to no indication of conflict in the placement of *S. x michoacanum*. While there are clearly reticulations in this graph, *S. x michoacanum* resolves in a bifurcating manner to *S. trifidum* and both *S. tarnii* accessions.

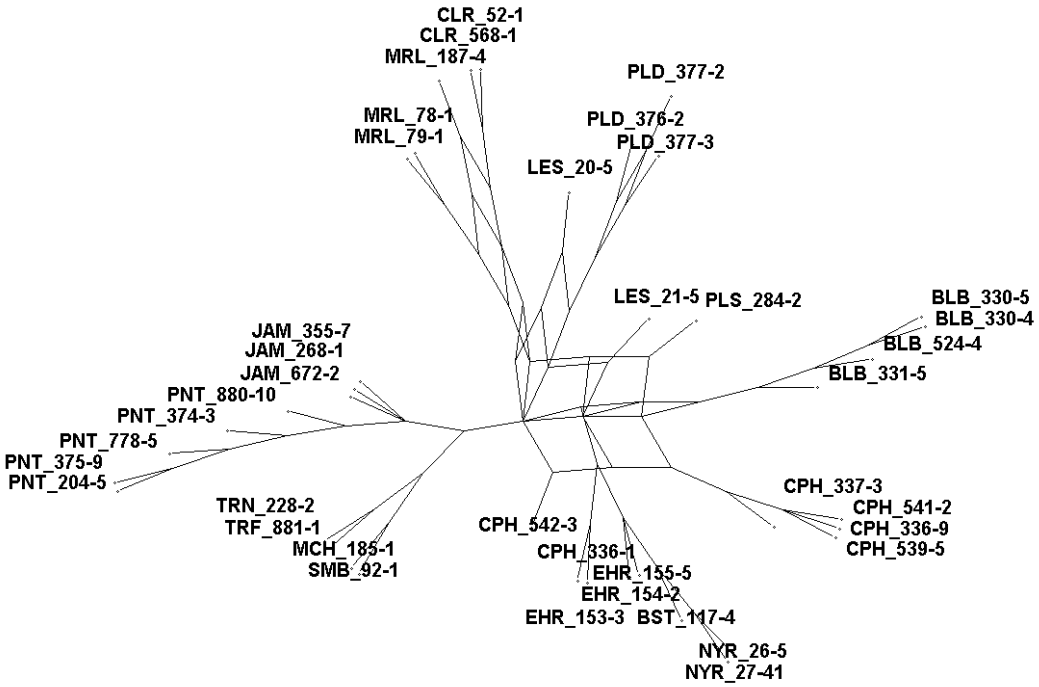


Fig. 6.3b Network analysis - Consensus network of the (50% majority rule) consensus trees from Bayesian analysis based on chloroplast DNA sequences, AFLPs and combined data.

DISCUSSION

Several studies have reported on the usefulness of AFLP to resolve relationships between parents and hybrids such as Beismann et al. (1997), Gobert et al. (2002), Guo et al.(2006) and Chauhan et al. (2004). However, most of these "success stories" are dealing with clear hypotheses of just a few putative parental species and their hybrids or even just the two parents and their hybrid. In addition, most of the putative hybrids in studies where AFLP performed well are F1 hybrids, instead of more ancient hybrid taxa.

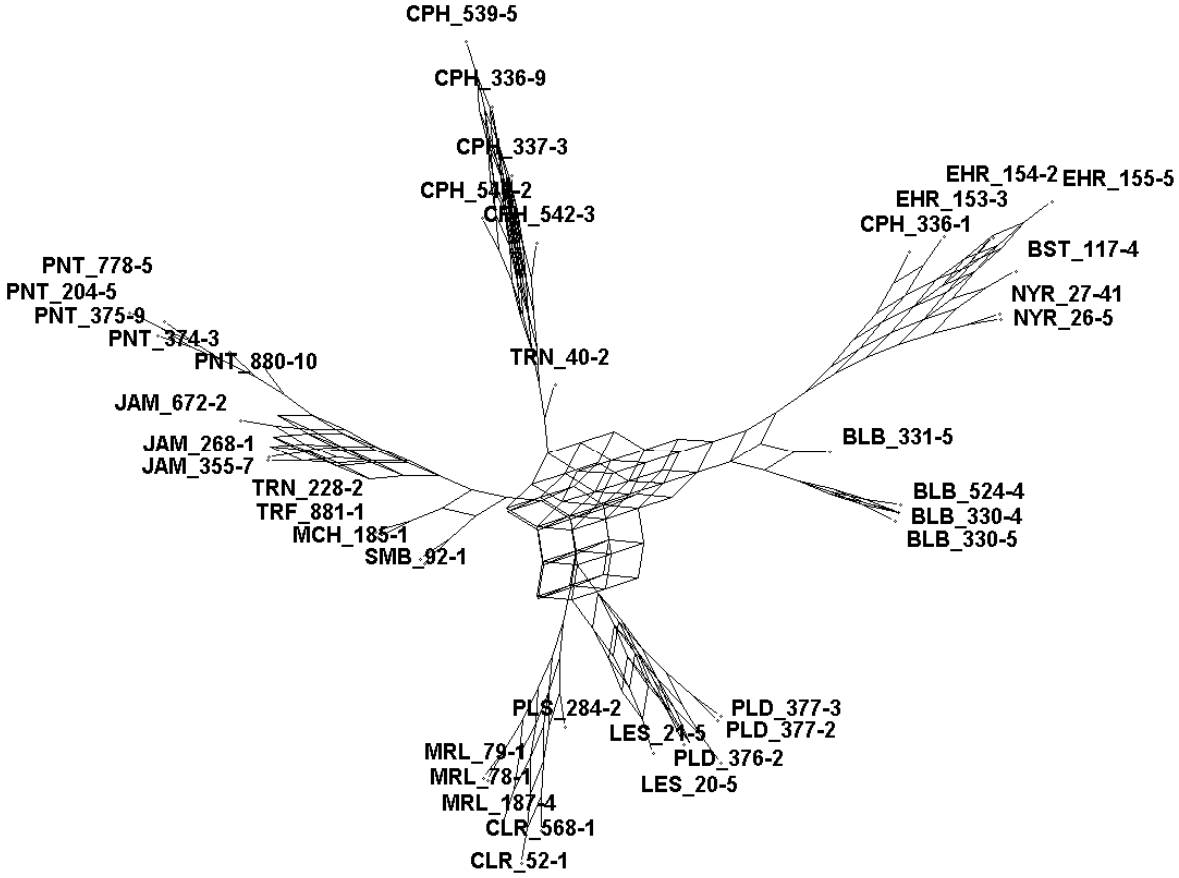


Fig 6.3c Network analysis - Consensus network of 8000 bootstrap trees from AFLP data. The threshold of splits to display is set at 10%, to reduce complexity of the graph.

In the studies cited above, interpretation of results was mostly based on the additivity of AFLP bands, the grouping in PCO-plots with the hybrids placed in between the two parents, or positioning of the hybrid to one of the two putative parents in UPGMA trees. In the present study, we could also recover hybridity when the included taxa were restricted to the parental lineages and the hybrid. Clearly, the proportion of shared hybrid-parental bands is less in the species-level test with the ancient hybrid when compared to the pattern in F1 hybrids (Table 6.3). It seems paradoxal that in this study, the number of unique AFLP-bands in both F1 hybrids (at individual and accession level) is higher than in the ancient hybrid. However, this is probably an artifact of differences in sample size. For the first two levels 6 and 9 samples were used, respectively, while the third level included 20 samples. With many accessions per species it is less likely to find species-specific bands, defined by the presence in all accessions of one species and complete absence in all accessions of the other.

In this test including just the parental species and the hybrid, additive bands can still be inferred, even at the species level. However, if more species would have been included, species-specific bands would probably not have been present. This is supported by the fact that in the group of Mexican diploid species, no species-specific bands for the (putative) parental taxa could be found (data not shown). This lack of species-specific AFLP markers for the same group of species was also described by Lara-Cabrera & Spooner (2004).

Interestingly, the definition of species-specific bands differs among authors. For instance, Zhou et al. (2005) distinguishes between common bands and "monomorphic species-specific bands", the latter defined as being present in 100% of the samples of a species and absent in all other samples. In a study on *Achillea* (Guo et al., 2006) with 7 species and 169 accessions, AFLP-bands were called species-specific when they occur with frequencies of 90% or more in at least one of its populations. Van Droogenbroeck et al. (2006), considers markers as species-specific if present in at least 95% of the individuals of one putative parent and absent in at least 95% of the other putative parent species. Applying a less stringent definition could have resulted in the presence of some species-specific bands in our data set. However, we could not find straight-forward criteria that resulted in clearly defined species-specific bands. For instance, defining species-specific bands as bands that occur in all accessions of one species and in just a few accessions in other species did not result in a higher number of these bands. Selecting only the accessions of the putative parental species resulted in 22 species-specific bands of *S. pinnatisectum* (compared to *S. bulbocastanum* only*)* and 46 species-specific bands of *S. bulbocastanum* (compared to *S. pinnatisectum*). *S. x michoacanum* shared 7 and 9 of these bands with *S. pinnatisectum* and *S. bulbocastanum* respectively (data retrieved from AFLP data set including all Mexican diploids, data not shown).

However, the studies mentioned above used a much broader population-wide sampling and in this study with a few accessions per species, the used definition seems justifiable.

In general, there are several drawbacks concerning AFLPs, such as their dominant and anonymous character, which possibly make them less suitable for phylogenetic analysis. Koopman (2005), however, reviewed this topic elaborately and demonstrated the presence of phylogenetic signal in AFLP data. Other studies also report on the presence of phylogenetic signal, mostly based on similarity of the results between AFLP and DNA sequence data sets (e.g. Schmidt-Lebuhn et al., 2006; Tremetsberger et al., 2006). In the present study, the patterns of the phenetic and phylogenetic analyses are

comparable and not in contrast with the (less resolved) chloroplast tree topology. No differences in hybrid position and tree topology structure in general can be seen between the different analyses. Therefore, our study provides further support for the presence of phylogenetic signal in AFLP data.

The use of AFLPs, as a marker system across the genome, may also circumvent problems involving single genes, like introgression, lineage sorting and paralogy (Simmons et al., 2007). However, with hybridization, it is expected that not just a single gene is involved, but instead many genes from both parents are expected to be distributed over the genome. Therefore, a signal of hybridization is expected to be revealed by AFLP markers and this marker system can possibly even be used to detect (unexpected) hybrids. The presence of high numbers of species-specific bands in the several studies (e.g. see Table 6.1) illustrated the usefulness of AFLPs to detect signals form both parents.

However, in a broader phylogenetic context, i.e. including more taxa in the analysis, this may be less straightforward, as also indicated by our results where we could not find species-specific bands in the larger data set. With the increase in the use of AFLPs for phylogenetic inference (Simmons et al., 2007), it is expected that more studies will be published where hybridity is explored using phylogenetic position in AFLP trees. These studies might provide evidence to clarify whether AFLPs are suitable to detect hybrids within a context of larger data sets.

Bayesian analysis of AFLPs

There are not many examples of Bayesian analyses using AFLPs. One of the few examples is in *Leonardoxa* (Brouat et al., 2004), where a probability approach resulted in a tree with exactly the same topology of the deeper nodes as in a NJ analysis. In our study, Bayesian analysis also results in a consensus tree with identical tree topology as compared to the MP and phenetic analyses. The results from all different analyses closely resemble each other and the signal in the data set seems to be strong and without conflict. It is possible that other data sets, with a less "strong signal", could have led to different results when Bayesian analysis is applied. In such a case the choice of a good model of AFLP character evolution would also become essential, to account for the differences in probability of gain or loss of bands (see also Brouat et al., 2004). Recent advancement in modelling AFLP marker evolution and its application in Bayesian approach has been made by Luo et al., (2007). In this study a likelihood model was developed based on the explicit underlying genetic mechanisms, taking information such as marker fragment lengths into account. The computational costs of the current implementation of this method are still prohibitively high. However,

this might be improved in future versions, and this method seems to be a promising start for a Bayesian approach to phylogenetic inference from AFLP marker data.

Network reconstruction and AFLPs

Several programs have the capacity to display incongruency in the data, possibly linked to reticulate evolution. Using AFLP data, ideally a hybrid terminal has part of the AFLP markers grouped to one parent and about the same proportion of markers to the other parent. Subsequently, these conflicting signals can be presented as a network with connecting branches between both parental terminals to the hybrid. However, in this particular "ideal" situation it is expected to see conflict among the various analyses, either between the chloroplast and AFLP trees or among the AFLP trees themselves. This conflict should surface as less topological resolution among the hybrid, parents and/or closely related species. Furthermore, upon exclusion of the hybrid an effect on tree topology or clade support is expected, such as lower support values for the different clades. This, however, was not the case in our study. Apparently a clear signal was present in the data set itself, with no (or non-discernible) conflict. This resulted in network analyses without clear reticulations connecting the lineage of the hybrid and one parent to other lineages. In the NeighborNet analyses reticulations between the hybrid and other lineages can be observed. However, NeighborNet in general displays many extra reticulations, with all clades displaying reticulations to other lineages. The bootstrap support of the edges (internal branches) does not provide any evidence for a particularly stronger signal of reticulation between hybrid and the other lineages than the average "reticulation-noise" in the graph.

Network reconstruction should, in theory, be able to display the conflict even when it is not visible in the data set or cannot be inferred from tree analysis. Unfortunately, few other studies have used network methods at the species-level, but Perrie et al. (2003) used Split Decomposition to resolve relationships among allopolyploid fern species with AFLP data as input. The reconstructed splits graphs displayed separate evolutionary lineages and did not implicate an allopolyploid origin of any taxon. However, the other evidence for an allopolyploid origin was convincing, i.e. intermediate morphology, indicating that a hybrid terminal does not necessarily lead to a conflicting signal that can be displayed in the split graph. It should be noted here, that not all conflicts are displayed in a split graph. Split Decomposition is a conservative method and only recovers branches that are relatively well supported (Perrie et al., 2003), so in our case it means that the signal to indicate separate lineages is much stronger than a signal of data conflict.

Network methods can be useful as additional tools for visualization of conflict and may

assist in formulating hypotheses. However, network methods alone are insufficient to reconstruct evolutionary relationships between hybrids and their parents and additional evidence is necessary to draw final conclusions.

Evidence for ancient hybrid origin of S. x michoacanum?

In general the position of *S. x michoacanum* in the analyses does not point towards a hybrid origin. No conflicting positions were found in the different analyses from cpDNA and AFLP and no (strong) within-data set conflict is present. This can be concluded from the effect of hybrid removal on tree resolution and bootstrap values and the position in the network analyses, with no reticulations connecting the lineage of *MCH -TRF-TRN* to other lineages. In addition, the absence of species-specific bands does not provide evidence for a hybrid ancestry in *S. x michoacanum* in general. The position of the putative ancient hybrid *S. x michoacanum* is almost identical in all clustering and phylogenetic analyses. It resolves within the group including one of the putative parents, *S. pinnatisectum*, but as sister to one of the *S. trifidum* accessions. This is in contrast with the results from Lara-Cabrera & Spooner (2004) based on an AFLP analysis of the same group of Mexican diploid species. In their study *S. x michoacanum* falls within a clade consisting of *S. lesteri* and *S. polyadenium* in the MP analysis and it clusters with *S. pinnatisectum* in UPGMA. This is clearly different from our results, but both analyses do indicate *S. pinnatisectum* as closely related and therefore a likely parent.

S. x michoacanum was always assumed to be a hybrid between *S. pinnatisectum* and *S. bulbocastanum* based on geographical (and morphological) evidence (Hawkes (1990), Graham & Dionne (1961) and Correll (1962)). Graham & Dionne (1961) performed crossing experiments where artificial crosses between *S. pinnatisectum* and *S. bulbocastanum* resulted in F1 and F2. Correll (1962) stated that "there is no question in my mind that *S. x michoacanum* is a naturally produced hybrid of the above two species". This statement is based on the observation of products of artificial crosses between *S. pinnatisectum* and *S. bulbocastanum* that are highly similar to known *S. x michoacanum*. While this is indeed a clear indication of possible parentage of the two involved species, it does not rule out the possibility of other parental combinations. Many species in this complex show similarity in several characteristics. Therefore, it seems likely that more than one possible crossing combination can result in hybrids that closely resemble *S. x michoacanum*. This could possibly have resulted in different varieties of this hybrid (resulting from different crosses) all identified under the same name "*S. x michoacanum*". We assume that the hypothesis of *S. x michoacanum* as a hybrid between *S. bulbocastanum* and *S.*

pinnatisectum is probably generally accepted because it seems likely, without further thorough investigations of alternatives. Alternative hypotheses including one of the other related species are also possible. Based on the analysis of molecular data in our study, *S. trifidum* could be a possible alternative to S*. pinnatisectum* as one of the involved parents. Morphological characteristics fit the hypothesis just as well as *S. pinnatisectum* and the geographical distribution shows overlap with the region of all species involved. Also, the crossings of Graham & Dionne (1961) showed that *S. trifidum* x *S. bulbocastanum* resulted in a viable F1.

While the inclusion of *S. trifidum* as a potential parent would fit the results of cluster and phylogenetic analyses better, the hybrid hypothesis still involves *S. bulbocastanum* (or a similar parent) as the other parent. However, no indication of parentage of this species can be inferred from the results we presented. Possible explanations are the selection of more markers similar to one parent by chance or maybe the hybrids have backcrossed to one of the parents, resulting in a higher percentage of corresponding markers to this parent. The results here also correspond to other studies (e.g. Slotte et al., 2006; Lee et al., 2005) where the phylogenetic position does not indicate hybridity, but where there is evidence from other sources. In these examples, a low resolution limited clear inferences of hybrid ancestry. This also illustrates the notion of McDade (1995) that phylogenetic position alone is not sufficient to indicate hybrid ancestry. Conclusions about hybrid nature should therefore always be based on several lines of evidence and a single marker system can never be sufficient to corroborate or reject hybrid hypothesis.

ACKNOWLEDGEMENTS

We would like to thank Henk de Leeuw and other people from the Botanical Gardens for the assistance with the experimental work. Also, we are very grateful for the help of Theo Damen to create the only viable hybrid crossing.

chapter 7

TOWARDS SPECIES TREES

Bastienne Vriesendorp & Freek T. Bakker

Submitted for publication in Taxon

ABSTRACT

How to infer species trees? In DNA-based systematics the focus of the last decades has been on how to convert nucleotides to gene trees, but development of methods on how to infer species trees from gene trees are lagging behind. The focus of this chapter is on the use of gene trees for inferring organism-level relationships (species trees) and different methods to reconcile gene trees are discussed. Most of these implement a parsimony approach to solving gene tree conflict (*tree noise*) minimizing specific reticulation events, such as hybridization, introgression, horizontal gene transfer, gene duplication/loss or incomplete lineage sorting. There is clearly a need for process modelling, enabling concerted testing and comparison of different processes in reconciling gene tree conflict in a probabilistic approach. In a Bayesian approach to species tree estimation, tree noise can be optimized keeping in mind that gene trees are in fact all samples from the same species tree, in stead of assuming all gene trees are independent samples. Although methods such as BEST implement this approach, it is so far not possible to model different underlying (reticulate) processes simultaneously. We suggest such model parameters and their possible use in the future.

INTRODUCTION

How to represent gene trees in a meaningful way, i.e. can incongruent gene trees be resolved in a single tree to represent organismal relationships? Obviously, reconstructed phylogenetic trees are often used to test hypotheses about evolutionary processes or relationships between organisms (e.g. Baldauf, 2003) or, for instance, to draw conclusions about the evolution of specific traits (e.g. tunicate tuber formation in Cape *Pelargonium*, Bakker et al., 2005). It is therefore assumed that the connections in a phylogenetic tree represent relationships between organisms. However, if individual gene trees are topologically incongruent, such representation is not straightforward and tracking any species phylogeny may become blurred. This "gene tree/species tree problem" (e.g. Maddison, 1997) has received a great deal of attention in the systematic literature of the nineties of the last century (e.g. Page & Charleston, 1998; Blake et al., 1999; Katz, 1999; Sang & Zhang, 1999) and is considered by some the *bête noire* of molecular systematics.

With the wealth of available gene trees on our way, chances to display organismal evolutionary relationships look promising. First of all, increased numbers of gene trees will give collectively a stronger signal of phylogenetic history and will override possible "phylogenetic noise" appearing in just a few gene trees. This is, however, dependent on the phylogenetic or taxonomic level, for instance the

para/monophyly of "the gymnosperms" proved more difficult to resolve as more DNA sequence data sets from different genomic compartments were added (Burleigh & Matthews, 2004, but see Hajibabei et al., 2006). Also, determining prokaryotic "phylogenetic" relationships (the "ring of life", Rivera & Lake, 2004) is complicated using genome sequences as many genes are known to have been transferred horizontally (Martin & Embley, 2004; Embley & Martin, 2006). Nevertheless, combining as many as possible gene trees or data sets is expected to converge to what could be considered an organismal tree (e.g. Edwards et al., 2007) and in any case constitutes *de facto* the most highly corroborated hypothesis of evolutionary relationships (e.g. Sang and Zhong, 2000; Rokas et al., 2003; Linder and Rieseberg, 2004; Delsuc et al., 2005; Huson et al., 2005). A nice example is the study of Rokas et al. (2003) who used a data set of 106 genes for phylogenetic analyses of eight yeast species and concluded that more than 20 genes were necessary to support "the species tree", i.e. the single tree yielded from the analysis of all concatenated genes with 100% bootstrap values on all branches. Increased population-level sampling, as done by for instance Koch et al. (2005) using approximately 750 individuals to represent several species, will be instrumental for accurate "species-level" phylogeny reconstruction. Among loci, coalescents can be variable and with only one locus and one gene sampled per species, the coalescent tree should be used as the gene tree. Only with multiple copies per species it is possible to decide whether the source of discrepancy between gene genealogy and species tree is caused by coalescent phenomena or other causes, such as reticulation processes. In order to do this, explicit modelling of effective population size N_e and "branch length in generations" t was proposed by Felsenstein (2004), i.e. $t/2N_e$ and probability of obtaining the correct species tree topology are positively related. A large N_e increases maintenance of polymorphic ancestral alleles, and hence incomplete lineage sorting artefacts, whereas ancestral polymorphism is quickly lost in small populations.

Incongruence between gene trees can be taken as evidence of reticulate evolution between organisms (but see Faith & Trueman, 2001, 2002), enabling distinction between *character* noise (i.e. the character support for clades in separate gene trees) and *tree* noise (i.e. topological incongruence among gene trees). "Reticulation processes" that may cause gene trees to be incongruent are e.g. hybridization (H), introgression, (I), horizontal gene transfer (HGT), gene duplication/loss (GDL) or incomplete lineage sorting (ILS), i.e. when genes fail to coalesce before species diverge (looking backwards in time).

Reticulate evolution and its implications for phylogeny reconstruction has become a prevalent topic in recent systematic literature (e.g. Linder & Rieseberg,

2004; Hypsa, 2006; McBreen & Lockhart, 2006; Willis et al., 2006). Phylogeny reconstruction usually results in bi- or multifurcating trees that may be interpreted as a visualization of character-state evolution among ancestors and descendants, or even as an actual phylogenetic hypothesis (e.g. Page & Holmes, 1998: 23). With reconstructed networks this is usually not as straightforward: extra nodes and edges are inserted in a network to display (all) character (or tree) conflict in a highly economized way. However, not all conflict between individual characters will actually represent organism-level evolutionary events. Therefore, it is not clear what internal network nodes and edges represent (Bryant & Moulton, 2004; Huson & Bryant, 2006; Chapter 3, 5) and to what extent the gene tree/species tree problem is paralleled in a gene network/species network problem.

Many studies have focused on the reconstruction of networks that do actually represent an "organism-level" phylogenetic history (e.g. Hein, 1990, 1993; Xu, 2000; Hallet & Lagergren, 2001; Addario-Berry et al., 2003; Moret et al., 2004; Gusfield, 2005; Nakhleh et al., 2005a, 2005b). This is described below in more detail. Most of these studies use a directed acyclic graph (DAG) to portray organism-level relationships, i.e. where one or more reticulation processes are considered to better explain the pattern against the data at hand. In this case, reticulation edges are considered to represent the outcome of reticulation processes. Ideally, process models describing them would be included using an optimality criterion framework in order to assess alternative event scenarios against, for instance, some likelihood score of the data. However, most methods reconstruct the networks in an algorithmic approach preventing the use of such a likelihood framework. Therefore, DAG-based approaches are useful as a starting-point for indicating reticulation processes between organisms instead of actual reconstruction of these relationships.

DAG-based approaches have not been used frequently in phylogenetic studies (Morrison, 2005). The main obstacle for using them is usually a practical one: only applicable to small data sets, limited number or types of data allowed or only few reticulations allowed, which prohibits the frequent use of these algorithms in the sense that restrictions in sampling design may seem unrealistic or impractical. Besides, most of the methods are not implemented in a computer package; the algorithms themselves are often available, but not directly applicable to the data, which can be an obstacle for their use.

DATA

Regardless of what method is used, "good" input data is the prerequisite for inferring correct relationships, both in terms of quantity (more genes), but also in terms of quality. Homoplasy-free markers for example, or DNA sequence data that show no saturation will probably be powerful in phylogenetic reconstruction. As indicated by several authors (e.g. Sang & Zhong, 2000; Linder & Rieseberg, 2004; Huson et al., 2005) the use of multiple unlinked genes can be an important method for estimating the extent of reticulation in organismal-level relationships. A promising development is the availability of genomic data including whole-genome sequences (see http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome). Increased numbers of independently evolving gene sequences as well as new and different types of data, e.g. short interspersed elements (SINEs), gene order, or the presence/absence of genes (e.g. Brown, 1996; Boore, 2006; Clark, 2006) are all highly promising as phylogenetic markers. SINEs, for instance, are essentially homoplasy-free (Shedlock & Okada, 2000) and will also be highly suitable to detect gene flow between species (Deragon & Zhang, 2006) because with the billions of possible states (in contrast to four states for nucleotide sequences) shared states are highly likely to indicate shared ancestry (i.e. gene flow) instead of caused by coincidence (Delsuc et al., 2005). These data will not only facilitate reconstruction of contested relationships (Boore, 2006), but will also be a great stimulus for the analysis and inference of reticulate patterns and hence, possibly hybridization processes. In addition, information about genome duplications or rearrangements can be highly valuable too to infer or investigate hybridization processes (Lowe et al., 2004).

Having good data available is one thing, using them well is another. In the nineties of the last century several studies have described the discrepancy between a "total evidence" versus "separate" approach to analyzing different data sets (e.g. Dequieroz et al., 1995). Basically, the "total evidence" school advocated the concatenation of all data sets into one supermatrix to use all the evidence in a combined approach, e.g. in Driskell et al. (2004), whereas the "separate" school proposed to analyze separate gene partitions individually and combine the resulting trees, for instance into a consensus tree or even into a supertree (e.g. Binanda-Emons et al., 2002). Nowadays, inference of evolutionary trees from genomic data can be done using both approaches and can produce similar results (see Delsuc et al., 2005). Methods that are based on rare genomic events (such as SINEs), gene order or gene content will reconstruct phylogenetic trees based on presence/absence data or distances matrices. Although it is expected that new genomic markers will reveal strong

phylogenetic signals, the only caveat could be how to model change in these different kinds of data (see below) for use in a statistical or Bayesian framework.

EVIDENCE

As most methods use gene trees as input, the question may be asked what the data actually represents when using such tree topologies. Is it right to use derived "meta-data" (gene trees) as evidence for hypothesis-testing instead of the primary data itself (e.g. DNA nucleotides)? Nevertheless, it is common practice to use gene trees as evidence for testing hypotheses on evolutionary trends, ancestral relationships, etc. Moreover, due to the inherent historical (i.e. not experimental) nature of biological systematic hypotheses they have sometimes been claimed not to be part of hypothetico-deductive science. Faith (2006) in his short communication "Science and philosophy for molecular systematics: Which is the cart and which is the horse?" addresses the question whether corroboration (in a true Popperian sense) is applicable to biological systematic hypotheses, in terms of evidence (*e*), hypotheses (*h*) and background knowledge (*b*). He points out the close relation of Popper's logical corroboration formula with Bayes' theorem in which $p(h, eb)$ represents posterior probability. Faith & Trueman (2001, 2002) asserted that Bayesian posterior probabilities of phylogenetic hypotheses, i.e. *goodness of fit* of data to hypothesis, can be treated as Popperian evidence statements subject to further corroboration. The authors show, however, that strong support may not always result in corroboration.

GENE TREE/SPECIES TREE SOLUTIONS

Approaches to reconciling gene trees into a species tree have already been proposed by Takahata in 1989 and Slowinsky & Page (1999). In general, methods are based on the assumption of events that could have caused the topological conflict, i.e. topological conflict is interpreted in terms of an event of choice (i.e. HGT, H, ILS, etc.). Subsequently, the number of events is minimized to select from alternative solutions, in a true optimization criterion approach in order to explain the set of gene trees. This reflects Occam's razor principle, already formulated in the 14th century, which states that the best approach in science is that "wherein the number of assumptions required to explain observations is minimized" (Schuh, 2000). When the nature of the underlying processes is known, event minimization is straightforward. However, we would like to be able to compare the relative importance of GDL with processes such as H or HGT to search for the best explanation of tree incongruence, especially in a probabilistic framework. However, it is generally not accepted to assign weights to these processes as this is regarded highly subjective (pers. comm. T. van der Niet).

One solution to explain incongruence between a species tree and its constituting gene trees is to reconcile gene trees in terms of duplication and loss of the actual genes involved, as introduced by Goodman et al. (1979) and formalized by Page et al. (1994). This concept is used frequently in gene orthology analysis and implemented in e.g. GeneTree (Page, 1998) and Notung (Durand et al., 2005). In addition to these procedures that are mainly applicable to rooted trees, other methods for unrooted gene trees have also been designed (Gorecki & Tiuryn, 2007). Ideally, branch length information should be included to use as evidence to be able to distinguishing between H and ILS (as e.g. suggested by Holder et al., 2001), but unfortunately programs such as GeneTree or Notung do not accommodate such information. Even worse, they can only have fully resolved trees as input, treating the (arbitrarily) resolved branches equal to the other well-supported branches.

Other approaches to reconciling gene trees in a species tree are based on the assumption of horizontal gene transfer (HGT) events, Efficient algorithms have been described for individual cases, for instance based on approximating SPR distances between trees (e.g. Hein, 1990, 1993; Hallet & Lagergren, 2001), where one SPR operation between different trees can be inferred to represent one HGT event. Addario-Berry et al. (2003) implemented the algorithm of Hallett & Lagergren (2001) in their package *Lattrans*, which calculates the minimum number of SPR operations (i.e. transfers) between trees. Some limitations of these algorithms are that gene trees can only be analyzed pairwise, only few HGT events can be inferred, or solutions are restricted to independent reticulations, i.e. "galled" networks (see Gusfield, 2005) - whereas many data sets will include reticulations that will interfere with other reticulated regions. Also, these methods are restricted to strictly bifurcating trees, which will increase the differences between trees and lead to "false positive" conflicts (see also Chapter 4).

MacLeod et al. (2005) and Beiko & Hamilton (2006) build further on Addario-Berry's approach (2003) also based on an algorithm that approaches the SPR distance between trees. These methods (implemented in the program *Horizstory* and the *EEEP* algorithm respectively) allow for multiple trees to be compared against a reference tree, but also impose restrictions on the input trees, e.g. limited to relatively similar trees and a specified number of SPR operations. Applications can mainly be found in studies aimed at detection of transferred genes in prokaryotes (e.g. Liu et al., 2006; Susko et al., 2006; Poptsovka & Gogarten, 2007) and may be less applicable to the representation of evolutionary relationships.

One of the network reconstruction tools developed by Nakhleh and co-workers (e.g. Nakhleh et al., 2003, 2005a, 2005b; Moret et al., 2004), involved optimizing

scenarios of HGT in order to reconcile conflicting gene trees, and is implemented in the method RIATA-HGT (Nakhleh et al., 2005a), which is included in the package Phylonet 1.2 (Nakhleh et al., 2006). Claimed as the first algorithm to solve the *general case* (i.e. without restrictions) of HGT (Nakleh et al., 2005a), their algorithm involves a relaxation of the criterion of always finding the minimum number of HGT events. This results in suboptimal solutions, making this method a heuristic search tool. RIATA-HGT displays multiple optimal solutions (of equal length), but this implicates that using input gene trees with many data uncertainty may result in prohibitively high numbers of possible solutions (Than et al., 2006). Most of these HGT and GDL event-minimizing methods assume a reference (species) tree and the gene trees, i.e. they resolve the different gene trees *on* the species tree. In practice, however, there are several gene trees, without prior preference as to one of them being the species tree. Results can therefore dangerously depend on assignments of species tree, but this problem has now been elegantly addressed using a Bayesian approach by Edwards et al. (2007).

While the process of HGT among angiosperm species is not common (but see for a dramatic exception the example of HGT in mtDNA of *Amborella* [Bergthorsson et al., 2004]), it is related to H or I in terms of underlying mechanisms (i.e. exchange of genetic material) and resulting reticulate patterns. A model that "realistically" describes HGT events may therefore be able to represent H or I events too.

Another solution to present reticulate evolutionary relationship at organism-level is described by Huson et al. (2005). The authors present an algorithm for drawing "reticulate networks" that they describe as an "explicit representation of evolutionary history" with "internal nodes corresponding to hypothetical ancestors" and edges that represent "lineages of descent or reticulate events such as hybridization, horizontal gene transfer, or recombination" (Huson & Bryant, 2006). The starting point for this algorithm is a Consensus Network (CN) (Holland & Moulton, 2003), which presents a summary of all splits that occur in the input trees. Or, alternatively, a Super Consensus Network (Huson et al., 2004; Holland et al., 2007) can be used, where trees with non-identical sets of terminals can be analyzed. A reticulate network is derived by analyzing each "netted component" individually using an algorithm to find the most parsimonious solution to reconcile all trees with a minimum of reticulations. This minimization-problem is related to the reconciliation problem as described above and, for the general case, computationally intractable. However, when the algorithm is restricted to independent reticulations (i.e. "galled" networks) it is computationally tractable. Netted regions that can be explained by reticulations are resolved and drawn "simplified" and otherwise these regions remain netted (Huson et al., 2005).

A main drawback of this method is that the algorithm is thus restricted to "galled networks", because most species-level data sets may contain many "netted regions" that are intertwined instead of separate. With complex data sets, errors on the trees may cause the netted regions to be too complicated to resolve, i.e. additional incompatible splits will interfere with netted regions that "describe the hybridization events" (Huson et al., 2005 and own observation). As always with methods that use phylogenetic trees as input, be it for network reconstruction (CN), character optimization (SIMMAP, Bollback, 2006) or historical biogeographic reconstruction (DIVA, Ronquist, 1996), ideally, phylogenetic uncertainty will be taken into account. In the approach of Huson et al. (2005) it appears that character support for input trees is not considered in constructing the networks. Therefore, this approach may lead to suboptimal solutions. Moreover, the algorithm results in only one solution for the inference of a network, obscuring equally good solutions.

Apart from HGT, H or GDL discussed above, obviously, gene tree incongruence can also be caused by incomplete lineage sorting (ILS), i.e. the presence of ancestral polymorphisms. Although it is complicated to distinguish between the different processes (e.g. Sang & Zhong, 2000) several studies have suggested procedures to differentiate at least H from ILS. Sang & Zhong (2000) described a test statistic, where the authors assume that the time of divergence (or time to coalescence) for multiple genes is nearly equal in case of H (or HGT), but differs significantly if ILS is the underlying process. This makes sense because under ILS, different alleles coalesce at different time periods, while a H event occurs simultaneously for all involved genes. Holder et al. (2001) criticize Sang and Zhong in implicitly assuming that variance in coalescence times of genes is negligible, except in cases of ILS. However, Holder et al. state that coalescence times will vary among genes, regardless of whether ILS is a problem. The statistic can therefore not reliably discriminate between H and ILS and only detect difference in coalescence times between two genes. This difference will especially be strong when genes from different compartments are compared, as effective population sizes N_e of organellar genomes are smaller than that of nuclear genomes (in diploid species). Holder et al. (2001) recommend using a method aimed at distinguishing between H and ILS that evaluates the probabilities of all gene trees (given a species tree), and the likelihood of the data given each gene tree, or posterior probabilities in a Bayesian approach. In such an approach models could be used that take more information into account, such as branch lengths or information on geographical distance between taxa (see "where are the models" below).

Huson et al. (2005) present a statistical test for distinguishing between reticulations due to H and those due to ILS or to tree-estimation error. They describe a

statistic that is applicable when the number of gene trees is *large*. However, sampling a large number of gene trees would in general make it easier to distinguish between ILS or H, even without such a statistical test: ILS should appear as random noise, i.e. as a random set of incongruent trees, while in the case of H a more consistent signal of incongruence is expected, i.e. high frequencies of a particular incongruent node (e.g. Holder et al., 2001).

Other studies consider H more likely than ILS because the hypothesis of one H event is more parsimonious than two independent hypotheses required to explain ILS. The latter implies the maintenance of duplicated genes (or alleles) at multiple unlinked loci in particular species, together with a loss in other species (Sang & Zhang, 1999; Oh & Potter, 2003). However, if one species has for instance a much larger effective population size it is more likely to maintain polymorphic ancestral alleles. This underlines the importance of modelling N_e, μ, reticulation processes and species relationships in the analysis.

There are also approaches to distinguishing H from ILS where different alleles per gene (population-level data) are analyzed. A monophyletic clustering of all alleles per gene indicates evidence for H, whereas ILS is expected to cause monophyletic patterns of alleles in only one of the gene trees. Church & Taylor (2005) for instance used two different (cpDNA and nrDNA) gene trees based on 74 populations and covering 17 species of *Houstonia* (Rubiaceae). They claim to have obtained support for a hybridization hypothesis as underlying cause, because both the nuclear and chloroplast data show a pattern of shared haplotypes. A different pattern in the cpDNA haplotype tree versus the ITS-tree would have suggested ILS as underlying cause. Howarth & Baum (2005) analyzed several alleles for rDNA ITS, as well as intron regions of nuclear encoded genes LFY, NIA and G3DPH from 20 populations for seven species of *Scaevola* (Goodeniaceae). Based on the distribution of alleles, they infer hybrid speciation as the most likely event to explain the reticulate patterns, which was consistent with morphological, ecological, and geographic information. Just as mentioned above, if H is the underlying cause, the alleles per gene are expected to show a monophyletic clustering. Other evidence for H is a pattern of clustering where different genes cluster with different parents. On the other hand, if different copies (alleles) are distributed over the different parents, this is an indication for ILS.

To summarize, the issue of distinguishing between H and ILS is not straightforward to settle. As Machado et al (2002) recognized, the distinction between the two processes could be too difficult to resolve analytically. An alternative approach to this problem is detailed population genetic analyses of intra- and interspecific variation across multiple loci. The amount of gene flow can then be tested based on patterns of linkage disequilibrium, allowing the assessment of the relative importance of gene flow and natural selection during species divergence (Machado et al., 2002). Since this test discerns whether patterns of shared variation can be best explained by gene flow or isolation with lineage sorting, this is analogous to distinguishing between H and ILS. Several more studies followed this approach (e.g. Stadler et al., 2005; Mazzoni et al., 2006) where population-level data was used to infer introgression.

WHERE ARE THE MODELS?

Most methods discussed so far are in principle parsimony-based, i.e. minimizing individual, "non-modeled" reticulation events. As we have seen in the tree building literature ever since Huelsenbeck et al. (1993), the application of models to describe processes relevant to the change in characters used to infer our patterns, has been instrumental in phylogenetic reconstruction. Hence, a model-based approach is considered the *bees knees* of DNA sequence-based phylogenetic reconstruction (e.g. Lewis, 2001). Especially the Bayesian approach, allowing uncertainty about model parameters and their values to be expressed as prior density probability distributions (e.g. Felsenstein, 2004) and allowing our hypotheses to be expressed in terms of posterior probabilities *given the data*, has proven to be an important and logically attractive approach. Bayesian inference is already common in other areas of both biological (ecology, physiology) and extra-biological sciences (varying from engineering to economics and has now firmly arrived in biological systematics as well.

In an ideal situation, we would like to be able to distinguish between different processes or errors in explaining gene tree conflict, and represent the inferred species relationships in a single graph - which will be a network in case of H, I or HGT. Preferably, such an assessment would include the use of models to incorporate parameters other than just topological differences. Parameters such as distance between taxa could be used to assign likelihoods to different processes, as a successful transfer or H event is for instance less likely between phylogenetically distant species. Different models would need to be formulated describing the above mentioned "reticulation processes" such as HGT, ILS, H, DLG, etc. Model parameters for HGT for instance, would include length of hypothesized transferred genes (where smaller genes are assigned higher prior probabilities than long genes) and the taxonomic groups

involved (e.g. within bacteria HGT is known to be rampant and could be assigned higher priors than within most eukaryotic organisms). Other possible model parameters could involve geographical information (e.g. increase in geographical distance will decrease likelihood of HGT), or whether or not multiple transfer events are likely along the same lineage or specific circumstances. Factors such as branch length will also be important to assign likelihood to ILS or H events.

The power of Bayesian approach is the incorporation of complex statistical models, optimizing a set of model parameters, such as topology, branch lengths and substitution parameters simultaneously, while expressing uncertainty about parameter "behaviour" as prior probability density distributions and ultimately arriving at posterior probabilities for hypotheses given the data (e.g. Huelsenbeck et al., 2002; Felsenstein, 2004; Nylander et al., 2004). Distinguishing ILS from H could be computationally feasible using Markov Chains in a Bayesian framework, as long as suitable models can be formulated. Of course, modelling needs to be incorporated in computationally tractable algorithms and whether convergence and mixing in MCMCMC is achieved will always be of concern (e.g. Hillis et al., 2005; Beiko et al., 2006). Because different models can probably not be applied simultaneously to the same data set, it is worthwhile investigating whether an approach involving "evolutionary process model testing" in a hierarchical way, for instance using a hierarchical likelihood-ratio test (hLRT; Huelsenbeck & Crandall, 1997; Posada & Crandall, 2001c), can achieve selecting the most probable scenario of evolution. A hierarchical criterion such as hLRT, however, is not likely to be used, because process models are expected to be non-nested. Therefore, criteria such as the Akaike information criterion (AIC; Akaike, 1973) or the Bayesian information criterion (BIC; Schwartz, 1978) will probably be more appropriate.

Future developments that can include evaluating different models simultaneously would be very useful, as is for instance foreseen in MrBayes 4 (Ronquist, pers. comm.). Such an approach will include "model jumping", i.e. the relative assessment of different models in efficiency of explaining the data. Model jumping in an organismal-level reticulation optimization context would involve jumping from a HGT model to an ILS model etc., whilst tracking the likelihood of the data.

Edwards et al. (2007) proposed an elegant solution to deal with several gene trees in a Bayesian framework. Obviously, gene trees are all "taken" from the same "species tree" ultimately, and are therefore not to be considered independent estimates of it. This method (Bayesian Estimation of Species Trees, BEST, see also Liu & Pearl, 2007) incorporates a joint prior probability distribution on gene trees and coalescent time across loci, and estimates the joint posterior distribution $K(\boldsymbol{G}|D)$ of

gene trees from DNA sequences for each locus. This is done using a "joint prior", which specifies the joint probability of gene trees and coalescent times for across loci. Coalescent theory is used to model the probability distribution of gene trees given an approximation of the species tree, constrained by the condition that all divergences of species pairs must occur after the genes coalesce, i.e. gene trees track the species tree. $K(\boldsymbol{G}|D)$ is then used to approximate the posterior of the species tree under coalescent theory (Edwards et al., 2007; Liu & Pearl, 2007). The authors apply "importance sampling" to further select from this posterior distribution. Applied to the Rokas et al. (2003) data set BEST achieved more statistical power than the concatenated approach of Rokas et al. (based on ML and MP): eight instead of 20 genes were necessary to estimate the correct species tree with >0.95 confidence. Incorporation of processes such as hybridization, gene flow and lateral gene transfer into a more general model is not yet possible in BEST and may not be feasible due to a dramatic increase in both number of parameters and computation time (Liu & Pearl, 2007) but would be an important improvement (Edwards et al., 2007).

CONCLUSION

In general, no specific algorithm or method is able to accommodate processes underlying incongruence between gene trees (recombination, HGT, H, ILS or GDL) in inferring actual reticulate organismal relationships. Although some methods do not provide a general solution to pinpoint the underlying process, they at least give an idea about how probable a certain cause is. For instance a HGT algorithm that results into many HGT events may be an indication that there are (also) other underlying causes for the conflict. Naturally, additional biological evidence (e.g. geographical patterns, morphology, etc.) could already be an indication of the nature of the underlying processes and serve as a starting point for further analyses.

One important underlying cause for gene tree incongruence can be "character noise", i.e. either stochastic error due to taxonomic/character sampling artifacts, or systematic error in the data. For instance, variable rates across lineages and heterogeneous base compositions can cause strong systematic bias that will not fade with using more genes (Phillips et al., 2004). The implementation of models in a Bayesian approach will help in reducing these sources of error.

An issue related to this is that most methods to infer reticulate events (HGT, ILS or H) are based on "ideal" input trees, i.e. perfectly reconstructed tree topologies that only differ in the placement of the reticulate taxa. Nakhleh et al. (2005b) describes a situation that involves tree errors, implementing "strict consensus calculations" of sets of trees instead of handling individual gene trees, resulting in a solution that can only

be used for situations with "one reticulation event" only. Than et al. (2006) also describe the effect of error in inferred trees on the estimation of HGT-events and they suggest removing poorly supported edges prior to analysis. Also methods could be improved, e.g. future implementations should be designed to handle non-binary trees (and allow for elimination of statistical error by collapsing branches) and should search for all most parsimonious solutions to HGT minimization (Than et al., 2006).

A promising perspective is the development of model-based methods in a Bayesian context, such as the BEST method (Edwards et al., 2007; Liu & Pearl, 2007). These model-based methods are good starting points for implementing reticulate processes in the future, to be able to infer organism-level relationships based on underlying gene trees, even if complicated reticulate processes have played a role.

chapter 8

GENERAL DISCUSSION

The goal of this thesis was to elucidate patterns in both character data and phylogenetic trees, in order to detect reticulation events and discuss the consequences of reticulate evolution for phylogenetic reconstruction. While reticulate evolution comprises several different processes at different levels, the main focus here was at species-level reticulation, with examples and data originating from angiosperms.

Reticulate patterns

The general view is that reticulate evolution can possibly have a disruptive effect on phylogenetic reconstruction. Chapter 3 described the general practice of dealing with hybrids, where putative hybrids are often left out of the analysis. However, the potential effect of hybrids depends on their underlying character pattern, which not necessarily consists of conflicting signals in the data. Using morphological data, hybrids are expected to show intermediacy, although a whole range of variation appears possible (ranging from unique or extreme characters within the hybrid or the hybrid being similar to one parent (Chapter 3; McDade, 1992, 1995)). Using molecular data, there are two main expected character patterns that cause problems in phylogeny reconstruction: a pattern of incongruence or a pattern of additivity. The emphasis of this thesis was on the first category, incongruent parts of DNA sequence data that can both have a disruptive effect on phylogenetic reconstruction (investigated using simulation studies in Chapter 4) and can be suitable to detect reticulate evolution and possibly even depict hybrid terminals in a network (Chapter 3, 5, 7).

The additivity of e.g. rDNA ITS copies to detect or investigate reticulate events has also been successfully applied, with several examples where nucleotide additivity in ITS sequences supported hypotheses of hybrid origin, see for instance a recent study of Shi et al. (2006). Several other examples listed in Chapter 3 illustrate the usefulness of these markers as well. In this thesis however, these types of markers have received less attention, because analysis of ITS-additivity is mostly done by visual inspection of the data set instead of running an analysis program. Therefore it is less appropriate for simulation experiments. In addition, most network methods used here are not (yet) suitable to deal with polymorphic sites and just treat ambiguous character states as uncertainty.

Markers such as AFLPs are also expected to show additivity in hybrids, as explored in Chapter 6. AFLPs and other genomic-wide markers (such as gene order, presence/absence of genes, or SINEs) will become more widely available in the near future. These markers may replace the use of just a few incongruent genes to detect reticulate evolution by a genomic approach and will probably be more "reliable", especially if appropriate models can be assigned to the behaviour of these types of data (see Chapter 5, 7).

The most important conclusion that one may draw from this study and recent literature is that hybridization leaves no unique "definite" signature that can always be identified as long as you use the right methods or gather enough data. Reticulate processes may cause many different patterns and lead to different results. As reviewed in Chapter 2, hybridization alone can follow different pathways, with possible involvement of e.g. introgression or backcrossing, or the multiple formation of hybrid lineages can occur, etc. All these various processes can result in different character patterns in the hybrid and therefore leave variable "signatures" in the data.

As can be concluded from our example of *Solanum* (Chapter 6), even with AFLPs that are assumed to give a large amount of genome-wide markers, the signal of both parental histories may simply not be there anymore.

Network solutions

Most network methods are originally used and/or designed for population-level data. This makes sense, because conspecific individuals can be expected to cross with each other and this type of data will always include reticulate relationships. However, at the species-level, where relationships are expected to be linear and bifurcating, the (exceptional) reticulate events may possibly be represented in a network too.

The present study is performed on the boundary of population-level and species-level, testing population-level methods using species-level data – with often only one individual per "species". This may not be the best approach. Even for straightforward species-level phylogenies without hybridization it is always recommended to include more individuals per species. Since we know that some hybrids can have multiple independent origins (e.g. Chapter 2 and the statement of Soltis & Soltis [1993, 1999] that "recurrent formation of polyploidy species is the rule, rather than the exception") multiple individuals per "species" are a must.

Moreover, most network methods can be "too sensitive", representing all character-conflict as reticulations, which makes sense in a population-level setting where the substitution rate between individuals is low and incongruence may indeed indicate organism-level reticulation. However, at species-level, levels of variation can be much higher and many incongruent positions in the data will just represent character-level conflict (i.e. due to high levels of substitution or sampling error). Hybridization may occur as an occasional event between two terminals represented by a reticulate pattern, but the other terminals in the phylogeny are preferably still better represented by multi- or bifurcating lineages.

While the simulation experiments and network testing (Chapter 4 & 5) are necessarily restricted to character-level patterns, in the last Chapter (7) the link between reticulate patterns and organism-level relationships is described and

suggestions are made on how to filter character-level noise from underlying organism-level reticulate processes. One approach would be a method that presents a "median"' solution – i.e. not display all but only the most pronounced conflict. But then, this can already be done using existing network methods – e.g. in Consensus Networks (Holland & Moulton, 2003) a threshold can be applied to display only the largest conflict. This method, however, can also be seen as a summarizing tool rather than an actual analysis method. Furthermore, approaches such as "reticulate networks" have been suggested where the netted regions can be simplified to yield the important signals, potentially indicating the "organism-level" relationships (see Chapter 7, Huson et al., 2005). In these cases you would expect the hybridization event to cause more conflict than a "base level" of noise, but this is not necessarily happening. In Chapter 7 other options are also discussed, e.g. using models in a Bayesian context that could lead to a solution to the problem of resolving incongruent gene trees in a species tree.

Consequences for phylogeny reconstruction

It is often considered general knowledge that recombination can severely influence phylogenetic inference and several recent phylogenetic studies exclude the hybrid prior to an analysis. However, published studies that explicitly target this problem in an experimental approach are rare. The famous studies of McDade (1990, 1992, 1995) have thoroughly investigated this influence, using artificial hybrids. However, her studies dealt with morphological characters, where morphological intermediacy often does not lead to a disruptive result but often places the hybrid in a basal position in the tree. In Chapter 3 several "ad hoc" examples in phylogenetic studies are presented where the effect of including hybrid terminals in a molecular data set is investigated. In some cases hybrids may cause disruptive effects, but this is not an inevitable outcome. Simulation studies can make it possible to draw conclusions about the effect of the general case of "recombinant terminals" on phylogenetic reconstruction, as we have done for large tree topologies (Chapter 4) in addition to earlier studies using an 8-taxon example tree (Posada & Crandall, 2002) and a 20-taxon tree (Ruths & Nakhleh, 2005).

The main conclusion is that most recombinant (hybrid) terminals do not disrupt the tree topology much. This is good news, if hybrid terminals are unknowingly included in a data set and the main goal of the phylogeny reconstruction is to provide a general overview of relationships. But this also means that hybrid terminals are often not detected, unless the exchange of material is between distantly related species and the extent of mixing is about 50%.

Taxonomic implications

In the case of reticulate evolution at organism-level it is of course also the question; what to do next? Besides the problem of detection and the potentially disturbing effects of reticulate events on tree reconstruction, if we assume that we can infer the correct underlying relationships (and maybe even represent them in a phylogenetic network or by hand on a tree), what kind of status do these reticulate lineages have?

Should species that exchange genetic material still be considered separate species or maybe a complex of species? If there is no new lineage, but just introgression, i.e. exchange between lineages, what would be the taxonomic status of these lineages? And what about a "hybrid" lineage that eventuates from a mixing of lineages and forms a new lineage, is such a lineage assigned to a new separate species? It is clear that we enter into the realm of the "species problem" and species concepts (e.g. Hey, 2006; Stamos, 2003), a partly philosophical discussion which is not the scope of this study.

First of all, reticulate patterns, detected or not, should only be considered important if they cause long-term consequences. Only self-sustaining hybrid lineages may be considered "new species" and assigned a specific status (e.g. Zhou et al., 2005). Otherwise, they can just be assumed "ad hoc" F1 individuals that are just exceptions or recurrently formed in a hybrid zone. These individuals will not lead to new evolutionary lineages and therefore do not deserve a specific status. The occurrence of F1 hybrids may indicate that there is introgression, but not necessarily a merger of species nor the creation of a new hybrid species. However, it will always be difficult to distinguish introgressed individuals from real new hybrid species because introgression, introgressive hybridization or hybridization cannot be separated on clearly distinctive criteria (Chapter 2). Hybrids in phylogenetic analyses can be either seen as temporary individual exceptions (and therefore left out from final taxonomic delimitations), or as a representative of a new stabilized form or "hybrid species" (i.e. self-sustainable), and therefore cannot be ignored.

Morphologically intermediacy, additivity of specific markers, or incongruence between different genes is often used as indication or corroboration of hybrid ancestry (see examples in Chapter 3). However, the observation that the hybrid consists of material from both parents is not sufficient to assign a status of hybrid species, as stressed by Zhou et al. (2005). They also suggest ways to determine the status of putative hybrids (F1, F2 or more advanced), proposing molecular criteria to determine the extent of admixture between parental genomes based on the segregation patterns of AFLP markers. However, these kind of tests can only be performed when more individuals per species (lineage) are included, which is probably a limitation in many phylogenetic studies. In phylogenetics often only one representative of a hybrid is

included in the study, which may not be enough to make statements about "stability" of lineages or about how widespread they are.

In addition to this, it can be valuable to design criteria to assess how deep in the phylogeny the reticulation events can still be traced, e.g. reticulation between ancient lineages might be assigned more importance than recent reticulation events. In practice, however, the specific status of the hybrid individual is not mentioned frequently. Many studies do not elaborate on degree of independency or age (contemporary or long-term etc.) of the hybrid, probably because it is often not known (Chapter 3). It would be good if studies that include reticulate events would always perform additional detailed studies to investigate the reticulate events, but this is of course not feasible. However, regardless of whether anything definite can be said about the hybrid, at least it should be made more explicit what is known or unknown about the status of the hybrid individuals.

And what are the implications for classification? Most systematists strive after a natural classification containing only monophyletic groups (Judd et al., 2002), but reticulate events cause a disruption of the tree structure such that strict monophyletic classification becomes impossible or at least much more difficult (Sosef, 1997). Especially examples of reticulation between clearly distinct lineages show the disruptive nature of reticulations. For instance, a recombination event that involved species between different sections (e.g. Poke et al., 2006) causes these sections to be no longer monophyletic. Besides the problem of how to deal with the parental entities (the sections), of course, the question is also what to do with the hybrid lineage. If you want to classify the hybrid lineage as a distinct species (after designating the stability and evolutionary significance of this entity, see above) in a monophyletic group, it would have descended from two lineages (instead of one which is required for monophyly). However, some have argued lately that the monophyletic system in general would be unsuitable for the classification of nature (Sosef, 1997; Brummitt, 2002; Hörandl, 2006).

Yet, with the intractability of reticulate problems this is probably not a relevant discussion within this thesis, where priorities lie with reconstruction of the patterns and the different ways of classifying are of secondary importance.

Related to this issue is how to represent reticulate evolution in a graph that can depict organism-level relationships. Hypothetically, in a phylogenetic network the internal nodes and edges may represent character-state evolution among ancestors and descendants. In such a network the multi– or bifurcating lineages represent the splitting into new lineages, and "horizontal" edges represent genetic transfer or the mixing of lineages into new lineages (i.e. hybrid lineages), see Chapter 7. At this

moment, no program or algorithm is available to lead to such a phylogenetic network, and because of the complexity of reticulate processes probably no method will ever be capable of do this unambiguously. However, the new (genomic) data, the improvements in computer capacity, developments of new models, etc. will render it possible to analyze large amounts of data and to give us indications about likelihoods of (reticulate) events. In addition, new model-based approaches to infer species trees from gene trees are currently under development and seem very promising. Eventually this will all provide information that helps us distinguishing between different scenarios and making it possible to be more confident about underlying causes. Hopefully, this will also prevent studies in the future from ignoring hybrids or from having to remove reticulate terminals from their analysis.

LITERATURE CITED

Abbott, R.J., 1992. Plant invasions, interspecific hybridization and the evolution of new plant taxa. Trends Ecol. Evol. 7, 401-405.

Addario-Berry, L., Hallet, M.T., Lagergren, J., 2003. Toward identifying lateral gene transfer events. Proceedings of the Proc. Eighth Pacific Symp. Biocomputing (PSB '03), pp. 279-290.

Aguilar, J.F., Feliner, G.N., 2003. Additive polymorphisms and reticulation in an ITS phylogeny of thrifts (*Armeria*, Plumbaginaceae). Mol. Phylogenet. Evol. 28, 430-447.

Ainouche, M.L., Baumel, A., Salmon, A., Yannic, G., 2004. Hybridization, polyploidy and speciation in *Spartina* (Poaceae). New Phytol. 161, 165-172.

Akaike, H., 1973. Information theory as an extension of the maximum likelihood principle. In: Petrov, B.N., Csaki, F. (Eds.), *Second International Symposium on Information Theory*, Akademiai Kiado, Budapest, pp. 267–281.

Allen, G.A., Soltis, D.E., Soltis, P.S., 2003. Phylogeny and biogeography of *Erythronium* (Liliaceae) inferred from chloroplast matK and nuclear rDNA ITS sequences. Syst. Bot. 28, 512-523.

Amenta, N., Klingner, J., 2002. Case study: Visualizing sets of evolutionary trees. 8th IEEE Symposium on Information Visualization 2002, 71-74

Anderson, E., 1949. Introgressive hybridization. John Wiley and Sons, New York.

Anderson, E., Hubricht, L., 1938. Hybridization in *Tradescantia*. III. The evidence for introgressive hybridization. Am. J. Bot. 25, 396-402.

Anderson, E., Stebbins, G.L., 1954. Hybridization as an evolutionary stimulus. Evolution. 8, 378-388.

Andreasen, K., Baldwin, B.G., 2003. Nuclear ribosomnal DNA sequence polymorphism and hybridization in checker mallows (*Sidalcea*, Malvaceae). Mol. Phylogenet. Evol. 29, 563-581.

Antonov, A.S., 2002. Genomics and genosystematics. Russ. J. Genet. 38, 622-627.

Archibald, J.K., Mort, M.E., Wolfe, A.D., 2005. Phylogenetic relationships within *Zaluzianskya* (Scrophulariaceae s.s., tribe Manuleeae): Classification based on DNA sequences from multiple genomes and implications for character evolution and biogeography. Syst. Bot. 30, 196-215.

Arnold, M.L., 1992. Natural Hybridization as an Evolutionary Process. Annu. Rev. Ecol. Syst . 23, 237-261.

Arnold, M.L., 1993. *Iris nelsonii* (Iridaceae) - Origin and genetic composition of a homoploid hybrid species. Am. J. Bot. 80, 577-583.

Arnold, M.L., 1997. Natural hybridization and evolution. Oxford University Press, Oxford.

Arnold, M.L., 2004. Transfer and origin of adaptations through natural hybridization: Were Anderson and Stebbins right? Plant Cell. 16, 562-570.

Arnold, M.L., Bouck, A.C., Cornman, R.S., 2004. Verne Grant and Louisiana Irises: Is there anything new under the sun? New Phytol. 161, 143-149.

Arnold, M.L., Buckner, C.M., Robinson, J.J., 1991. Pollen-Mediated introgression and hybrid speciation in Louisiana Irises. Proc. Natl. Acad. Sci. USA. 88, 1398-1402.

Arnold, M.L., Hamrick, J.L., Bennett, B.D., 1990. Allozyme variation in Louisiana Irises - a test for introgression and hybrid speciation. Heredity. 65, 297-306.

Arnold, M.L., Hodges, S.A., 1995. The fitness of hybrids - reply. Trends Ecol. Evol. 10, 289-289.

Arnold, M.L., Robinson, J.J., Buckner, C.M., Bennet, B.D., 1992. Pollen dispersal and interspecific gene flow in Louisiana Irises. Heredity. 68, 399-404.

Avise, J.C., 2000. Phylogeography: The history and formation of species. . Harvard University Press., Cambridge, MA.

Bakker, F.T., Culham, A., Hettiarachi, P., Touloumenidou, T., Gibby, M., 2004. Phylogeny of *Pelargonium* (Geraniaceae) based on DNA sequences from three genomes. Taxon. 53, 17-28.

Bakker, F.T., Marais, E.M., Culham, A., Gibby, M., 2005. Nested radiation in Cape *Pelargonium.* . In: Bakker, F.T., Chatrou, L.W., Gravendeel, B., Pelser, P. (Eds), *Plant Species-level Systematics: New perspectives on pattern & process,* Regnum Vegetabile 143, Gantner Verlag, Liechtenstein, pp. 75-100

Baldauf, S.L., 2003. The deep roots of eukaryotes. Science. 300, 1703-1706.

Bandelt, H.J., Dress, A., 1992. Split decomposition: A new and useful approach to phylogenetic analysis of distance data. Mol. Phylogenet. Evol. 1, 242-252.

Bandelt, H.J., Dress, A.W.M., 1993. A relational approach to Split Decomposition. In: O. Opitz, O., Lausen, B., Klar, R. (Eds.), Information and Classification, Springer Berlin, pp.123-131.

Bandelt, H.J., Forster, P., Rohl, A., 1999. Median-joining networks for inferring intraspecific phylogenies. Mol. Biol. Evol. 16, 37-48.

Bandelt, H.J., Forster, P., Sykes, B.C., Richards, M.B., 1995. Mitochondrial portraits of human-populations using Median Networks. Genetics. 141, 743-753.

Barker, N.P., Harman, K.T., Ripley, B.S., Bond, J., 2003. The genetic diversity of *Scaevola plumieri* (Goodeniaceae), an indigenous dune coloniser, as revealed by Inter Simple Sequence Repeat (ISSR) fingerprinting. S. Afr. J. Bot. 68, 532-541.

Barkman, T.J., Simpson, B.B., 2002. Hybrid origin and parentage of *Dendrochilum acuiferum* (Orchidaceae) inferred in a phylogenetic context using nuclear and plastid DNA sequence data. Syst. Bot. 27, 209-220.

Baroni, M., Grunewald, S., Moulton, V., Semple, C., 2005. Bounding the number of hybridisation events for a consistent evolutionary history. J. Math. Biol.. 51, 171-182.

Baroni, M., Semple, C., Steel, M., 2006. Hybrids in real time. Syst. Biol. 55, 46-56.

Baroni, M., Semple, C., Steel, M.A., 2004. A framework for representing reticulate evolution. Ann. Combin. 8, 391-408.

Barton, N.H., Hewitt, G.M., 1985. Analysis of hybrid zones. Annu. Rev. Ecol. Syst . 16, 113-148.

Bateman, R.M., Hollingsworth, P.M., 2004. Morphological and molecular investigation of the parentage and maternity of *Anacamptis xalbuferensis (A-fragrans x A-robusta),* a new hybrid orchid from Mallorca, Spain. Taxon. 53, 43-54.

Beardsley, P.M., Schoenig, S.E., Whittall, J.B., Olmstead, R.G., 2004. Patterns of evolution in Western North American *Mimulus* (Phrymaceae). Am. J. Bot. 91, 474-489.

Beiko, R.G., Hamilton, N., 2006. Phylogenetic identification of lateral genetic transfer events. BMC Evol. Biol. 6.

Beiko, R.G., Keith, J.M., Harlow, T.J., Ragan, M.A., 2006. Searching for convergence in phylogenetic Markov chain Monte Carlo. Syst. Biol. 55, 553-565.

Beismann, H., Barker, J.H.A., Karp, A., Speck, T., 1997. AFLP analysis sheds light on distribution of two *Salix* species and their hybrid along a natural gradient. Mol. Ecol. 6, 989-993.

Bergthorsson, U., Richardson, A.O., Young, G.J., Goertzen, L.R., Palmer, J.D., 2004. Massive horizontal transfer of mitochondrial genes from diverse land plant donors to the basal angiosperm *Amborella*. Proc. Natl. Acad. Sci. USA. 101, 17747-17752.

Bigelow, R.S., 1965. Hybrid Zones and Reproductive Isolation. Evolution. 19, 449-458.

Bininda-Emonds, O.R.P., Gittleman, J.L., Steel, M.A., 2002. The (Super)tree of life: Procedures, problems, and prospects. Annu. Rev. Ecol. Syst . 33, 265-289.

Blake, N.K., Lehfeldt, B.R., Lavin, M., Talbert, L.E., 1999. Phylogenetic reconstruction based on low copy DNA sequence data in an allopolyploid: The B genome of wheat. Genome. 42, 351-360.

Bollback, J.P., 2006. SIMMAP: Stochastic character mapping of discrete traits on phylogenies. BMC Bioinformatics. 7.

Boore, J.L., 2006. The use of genome-level characters for phylogenetic reconstruction. Trends Ecol. Evol. 21, 439-446.

Bordewich, M., Semple, C., 2004. On the computational complexity of the rooted subtree prune and regraft distance. Ann. Combin. 8, 409-423.

Borgen, L., Leitch, I., Santos-Guerra, A., 2003. Genome organization in diploid hybrid species of *Argyranthemum* (Asteraceae) in the Canary Islands. Bot. J. Linn. Soc. 141, 491-501.

Brouat, C., McKey, D., Douzery, E.J.P., 2004. Differentiation in a geographical mosaic of plants coevolving with ants: phylogeny of the *Leonardoxa africana* complex (Fabaceae : Caesalpinioideae) using amplified fragment length polymorphism markers. Mol. Ecol. 13, 1157-1171.

Brown, J.R., 1996. Preparing for the flood: Evolutionary biology in the age of genomics. Trends Ecol. Evol. 11, 510-513.

Brummitt, R.K., 2002. How to chop up a tree. Taxon. 51, 31-41.

Bryant, D., Moulton, V., 2002. NeighborNet: An agglomerative method for the construction of planar phylogenetic networks. Algorithms in Bioinformatics, Proceedings, vol.2452, pp. 375-391.

Bryant, D., Moulton, V., 2004. Neighbor-Net: An agglomerative method for the construction of phylogenetic networks. Mol. Biol. Evol. 21, 255-265.

Bures, P., Wang, Y.F., Horova, L., Suda, J., 2004. Genome size variation in Central European species of *Cirsium* (Compositae) and their natural hybrids. Ann. Bot. 94, 353-363.

Burke, J.M., Arnold, M.L., 2001. Genetics and the fitness of hybrids. Annu. Rev. Genet. . 35, 31-52.

Burleigh, J.G., Mathews, S., 2004. Phylogenetic signal in nucleotide data from seed plants: Implications for resolving the seed plant tree of life. Am. J. Bot. 91, 1599-1613.

Camerarius, R.J., 1694. Epistola ad M.B. Valentini de sexu plantarum. In Ostwald's Klassiker der exakten Naturwissenschaften, No. 105, 1899 (Leipzig: Verlag von Wilhelm Engelmann).

Campbell, C.S., Wojciechowski, M.F., Baldwin, B.G., Alice, L.A., Donoghue, M.J., 1997. Persistent nuclear ribosomal DNA sequence polymorphism in the *Amelanchier* agamic complex (Rosaceae). Mol. Biol. Evol. 14, 81-90.

Carputo, D., Frusciante, L., Peloquin, S.J., 2003. The role of 2n gametes and endosperm balance number in the origin and evolution of polyploids in the tuber-bearing Solanums. Genetics. 163, 287-294.

Cassens, I., Mardulyn, P., Milinkovitch, M.C., 2005. Evaluating intraspecific "Network" construction methods using simulated sequence data: Do existing algorithms outperform the global maximum parsimony approach? Syst. Biol. 54, 363-372.

Cassens, I., Van Waerebeek, K., Best, P.B., Crespo, E.A., Reyes, J., Milinkovitch, M.C., 2003. The phylogeography of dusky dolphins (*Lagenorhynchus obscurus*): a critical examination of network methods and rooting procedures. Mol. Ecol. 12, 1781-1792.

Chase, M.W., Knapp, S., Cox, A.V., Clarkson, J.J., Butsko, Y., Joseph, J., Savolainen, V., Parokonny, A.S., 2003. Molecular systematics, GISH and the origin of hybrid taxa in *Nicotiana* (Solanaceae). Ann. Bot. 92, 107-127.

Chat, J., Jauregui, B., Petit, R.J., Nadot, S., 2004. Reticulate evolution in kiwifruit (*Actinidia,* Actinidiaceae) identified by comparing their maternal and paternal phylogenies. Am. J. Bot. 91, 736-747.

Chauhan, N., Negi, M.S., Sabharwal, V., Khurana, D.K., Lakshmikumaran, M., 2004. Screening interspecific hybrids of *Populus (P-ciliata x maximowiczii)* using AFLP markers. Theor. Appl. Genet. 108, 951-957.

Chiang, T.Y., Hong, K.H., Peng, C.I., 2001. Experimental hybridization reveals biased inheritance of the internal transcribed spacer of the nuclear ribosomal DNA in *Begonia x taipieiensis*. J. Plant Res. 114, 343-351.

Chung, J.D., Lin, T.P., Chen, Y.L., Cheng, Y.P., Hwang, S.Y., 2007. Phylogeographic study reveals the origin and evolutionary history of a *Rhododendron* species complex in Taiwan. Mol. Phylogenet. Evol. 42, 14-24.

Church, S.A., Taylor, D.R., 2005. Speciation and hybridization among *Houstonia* (Rubiaceae) species: The influence of polyploidy on reticulate evolution. Am. J. Bot. 92, 1372-1380.

Clark, A.G., 2006. Genomics of the evolutionary process. Trends Ecol. Evol. 21, 316-321.

Clausen, A.M., Spooner, D.M., 1998. Molecular support for the hybrid origin of the wild potato species *Solanum x rechei*. Crop Science. 38, 858-865.

Clausen, R.E., Goodspeed, T.H., 1925. Interspecific hybridization in *Nicotiana*. II. A tetraploid *gllutinosa-tabacum* hybrid, and experimental verification of Winge's hypothesis. Genetics. 10, 279-284.

Clement, M., Posada, D., Crandall, K.A., 2000. TCS: a computer program to estimate gene genealogies. Mol. Ecol. 9, 1657-1659.

Colless, D.H., 1982. Review of Phylogenetics: the Theory and Practice of Phylogenetic Systematics. Syst Zool 31, 100-104

Comes, H.P., Abbott, R.J., 2001. Molecular phylogeography, reticulation, and lineage sorting in Mediterranean *Senecio* sect. Senecio (Asteraceae). Evolution. 55, 1943-1962.

Conant, D.S., Cooperdriver, G., 1980. Autogamous Allohomoploidy in *Alsophila* and *Nephelea* (Cyatheaceae) - a new hypothesis for speciation in homoploid homosporous ferns. Am. J. Bot. 67, 1269-1288.

Correll, D.S., 1962. The potato and its wild relatives: Section Tuberarium of the genus *Solanum*. Texas Research Foundation. Renner, Texas.

Crawford, D.J., Mort, M.E., 2004. Single-locus molecular markers for inferring relationships at lower taxonomic levels: observations and comments. Taxon. 53, 631-635.

Cronn, R., Small, R.L., Haselkorn, T., Wendel, J.F., 2003. Cryptic repeated genomic recombination during speciation in *Gossypium gossypioides*. Evolution. 57, 2475-2489.

Darlington, C.D., 1937. What is a hybrid? J. Hered. 28, 308.

Darwin, C., 1859. On the Origin of Species by means of natural selection. John Murray, Ltd., London.

Delsuc, F., Brinkmann, H., Philippe, H., 2005. Phylogenomics and the reconstruction of the tree of life. Nat. Rev. Genet. 6, 361-375.

Dequeiroz, A., Donoghue, M.J., Kim, J., 1995. Separate versus combined analysis of phylogenetic evidence. Annu. Rev. Ecol. Syst . 26, 657-681.

Deragon, J.M., Zhang, X.Y., 2006. Short interspersed elements (SINEs) in plants: Origin, classification, and use as phylogenetic markers. Syst. Biol. 55, 949-956.

Devos, N., Oh, S.H., Raspe, O., Jacquemart, A.L., Manos, P.S., 2005. Nuclear ribosomal DNA sequence variation and evolution of spotted marsh-orchids (*Dactylorhiza maculata* group). Mol. Phylogenet. Evol. 36, 568-580.

Divakaran, M., Babu, K.N., Ravindran, P.N., Peter, K.V., 2006. Interspecific hybridization in vanilla and molecular characterization of hybrids and selfed progenies using RAPD and AFLP markers. Sci. Hortic. 108, 414-422.

Dobzhansky, T., 1951. Genetics and the origin of species. Colombia University Press, New York, New York.

Donoghue, M.J., Baldwin, B.G., Li, J.H., Winkworth, R.C., 2004. *Viburnum* phylogeny based on chloroplast trnK intron and nuclear ribosomal ITS DNA sequences. Syst. Bot. 29, 188-198.

Douady, C.J., Delsuc, F., Boucher, Y., Doolittle, W.F., Douzery, E.J.P., 2003. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. Mol. Biol. Evol. 20, 248-254.

Dowling, T.E., Secor, C.L., 1997. The role of hybridization and introgression in the diversification of animals. Annu. Rev. Ecol. Syst . 28, 593-619.

Doyle, J.J., Doyle, J.L., Rauscher, J.T., Brown, A.H.D., 2004. Diploid and polyploid reticulate evolution throughout the history of the perennial soybeans (*Glycine* subgenus Glycine). New Phytol. 161, 121-132.

Driskell, A.C., Ane, C., Burleigh, J.G., McMahon, M.M., O'Meara, B.C., Sanderson, M.J., 2004. Prospects for building the tree of life from large sequence databases. Science. 306, 1172-1174.

Durand, D., Halldorsson, B.V., Vernot, B., 2006. A hybrid micro-macroevolutionary approach to gene tree reconstruction. J. Comput. Biol. 13, 320-335.

Edwards, S.V., Liu, L., Pearl, D.K., 2007. High-resolution species trees without concatenation. Proc. Natl. Acad. Sci. USA. 104, 5936-5941.

Ellstrand, N.C., 2003. Current knowledge of gene flow in plants: implications for transgene flow. - Biological Sciences. 358, 1163-1170.

Ellstrand, N.C., Whitkus, R., Rieseberg, L.H., 1996. Distribution of spontaneous plant hybrids. Proc. Natl. Acad. Sci. USA. 93, 5090-5093.

El-Rabey, H.A., Badr, A., Schafer-Pregl, R., Martin, W., Salamini, F., 2002. Speciation and species separation in *Hordeum L.* (Poaceae) resolved by discontinuous molecular markers. Plant Biology. 4, 567-575.

Embley, T.M., Martin, W., 2006. Eukaryotic evolution, changes and challenges. Nature. 440, 623-630.

Excoffier, L., Smouse, P.E., 1994. Using allele frequencies and geographic subdivision to reconstruct gene trees within a species - Molecular Variance Parsimony. Genetics. 136, 343-359.

Faith, D.P., 2006. Science and philosophy for molecular systematics: Which is the cart and which is the horse? Mol. Phylogenet. Evol. 38, 553-557.

Faith, D.P., Trueman, J.W.H., 2001. Towards an inclusive philosophy for phylogenetic inference. Syst. Biol. 50, 331-350.

Faith, D.P., Trueman, J.W., 2002. Reverend Bayes reports back to Popper. Science Online (15 February), <http://www.sciencemag.org/cgi/eletters/294/5550/2310#401>

Felsenstein, J., 2004. Inferring Phylogenies. Sinauer Associates, Sunderland, MA.

Felsenstein, J., 2005. PHYLIP (Phylogeny Inference Package) version 3.6. *Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.*

Ferguson, C.J., Levin, D.A., Jansen, R.K., 1999. Natural hybridization between an outcrossing and a selfing *Phlox* (Polemoniaceae): the maternal species of F-1 hybrids. Plant Syst. Evol. 218, 153-158.

Focke, W.O., 1881. Die Pflanzen-Mischlinge. Berlin, Borntraeger.

Frajman, B., Oxelman, B., 2007. Reticulate phylogenetics and phytogeographical structure of *Heliosperma* (Sileneae, Caryophyllaceae) inferred from chloroplast and nuclear DNA sequences. Mol. Phylogenet. Evol. 43, 140.

Franzke, A., Hurka, H., Janssen, D., Neuffer, B., Friesen, N., Markov, M., Mummenhoff, K., 2004. Molecular signals for Late Tertiary Early Quaternary range splits of an Eurasian steppe plant: *Clausia aprica* (Brassicaceae). Mol. Ecol. 13, 2789-2795.

Freudenstein, J.V., van den Berg, C., Goldman, D.H., Kores, P.J., Molvray, M., Chase, M.W., 2004. An expanded plastid DNA phylogeny of Orchidaceae and analysis of jackknife branch support strategy. Am. J. Bot. 91, 149-157.

Gärtner, C.F.v., 1827. Notice sur des expériences concernant la fécondation de quelques végétaux. Annales des Sciences Naturelles. 10, 113-148.

Gobert, V., Moja, S., Colson, M., Taberlet, P., 2002. Hybridization in the section *Mentha* (Lamiaceae) inferred from AFLP markers. Am. J. Bot. 89, 2017-2023.

Godron, D.A., 1844. De l'hybridité dans les végétaux, Nancy.

Goldblatt, P., 1980. Polyploidy in angiosperms: monocotyledons. In: Lewis, W.H. (Ed.), Polyploidy: Biological Relevance. Plenum, New York, pp. 219-240.

Goldman, D.H., Jansen, R.K., Van Den Berg, C., Leitch, I.J., Fay, M.F., Chase, M.W., 2004. Molecular and cytological examination of *Calopogon* (Orchidaceae, Epidendroideae): Circumscription, phylogeny, polyploidy, and possible hybrid speciation. Am. J. Bot. 91, 707-723.

Gonzalez-Perez, M.A., Caujape-Castells, J., Sosa, P.A., 2004. Molecular evidence of hybridisation between the endemic *Phoenix canariensis* and the widespread *P. dactylifera* with Random Amplified Polymorphic DNA (RAPD) markers. Plant Syst. Evol. 247, 165-175.

Goodman, M., Czelusniak, J., Moore, G.W., Romeroherrera, A.E., Matsuda, G., 1979. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. Syst. Zool. 28, 132-163.

Gorecki, P., Tiuryn, J., 2007. URec: a system for unrooted reconciliation. Bioinformatics. 23, 511-512.

Graham, M.K., Dione, L.A., 1961. Crossability relationships of certain diploid Mexican *Solanum* species. Can. J. Genet. Cytol. 3: 121-127.

Grant, P.R., Grant, B.R., 1992. Hybridization of bird species. Science. 256, 193-197.

Grant, V., 1958. The regulation of recombination in plants. Cold Spring Harbor Symp Quant Biol. 23, 337-363.

Grant, V., 1966. Origin of a new species of *Gilia* in a hybridization experiment. Genetics. 54, 1189-&.

Grant, V., 1981. Plant Speciation. Colombia Univ. Press, New York.

Gravendeel, B., Eurlings, M.C.M., van den Berg, C., Cribb, P.J., 2004. Phylogeny of *Pleione* (Orchidaceae) and parentage analysis of its wild hybrids based on plastid and nuclear ribosomal ITS sequences and morphological data. Syst. Bot. 29, 50-63.

Griffith, M.P., 2003. Using molecular evidence to elucidate reticulate evolution in *Opuntia* (Cactaceae). Madrono, a West Am. J. Bot. 50, 162-169.

Gross, B.L., Rieseberg, L.H., 2005. The ecological genetics of homoploid hybrid speciation. J. Hered. 96, 241-252.

Guo, Y.P., Ehrendorfer, F., Samuel, R., 2004. Phylogeny and systematics of *Achillea* (Asteraceae-Anthemideae) inferred from nrITS and plastid trnL-F DNA sequences. Taxon. 53, 657-672.

Guo, Y.P., Vogl, C., Van Loo, M., Ehrendorfer, F., 2006. Hybrid origin and differentiation of two tetraploid *Achillea* species in East Asia: molecular, morphological and ecogeographical evidence. Mol. Ecol. 15, 133-144.

Gusfield, D., 2005. Optimal, efficient reconstruction of root-unknown phylogenetic networks with constrained and structured recombination. J. Comp. Syst. Sci. 70, 381-398.

Gusfield, D., Bansal, V., 2005. A fundamental decomposition theory for phylogenetic networks and incompatible characters. Research in Computational Molecular Biology, Proceedings, vol.3500, pp. 217-232.

Hajibabaei, M., Xia, J.N., Drouin, G., 2006. Seed plant phylogeny: Gnetophytes are derived conifers and a sister group to Pinaceae. Mol. Phylogenet. Evol. 40, 208-217.

Hall, B.G., 2005. Comparison of the accuracies of several phylogenetic methods using protein and DNA sequences. Mol. Biol. Evol. 22, 792-802.

Hallet, M.T., Lagergren, J., 2001. Efficient algorithm for lateral gene transfer problems. Proceedings of the RECOMB 2001, Montreal, Canada, pp. 149-155.

Hamilton, C.W., Reichard, S.H., 1992. Current practice in the use of subspecies, variety, and forma in the classification of wild plants. Taxon. 41, 485-498.

Hamzeh, M., Dayanandan, S., 2004. Phylogeny of *Populus* (Salicaceae) based on nucleotide sequences of chloroplast TRNT-TRNF region and nuclear rDNA. Am. J. Bot. 91, 1398-1408.

Hardig, T.M., Soltis, P.S., Soltis, D.E., 2000. Diversification of the North American shrub genus *Ceanothus* (Rhamnaceae): Conflicting phylogenies from nuclear ribosomal DNA and chloroplast DNA. Am. J. Bot. 87, 108-123.

Hardig, T.M., Soltis, P.S., Soltis, D.E., Hudson, R.B., 2002. Morphological and molecular analysis of putative hybrid speciation in *Ceanothus* (Rhamnaceae). Syst. Bot. 27, 734-746.

Harlan, J.R., DeWet, J.M.J., 1975. Winge,O and Prayer,A - Origins of polyploidy. Bot. Rev. 41, 361-390.

Harper, J.A., Thomas, I.D., Lovatt, J.A., Thomas, H.M., 2004. Physical mapping of rDNA sites in possible diploid progenitors of polyploid *Festuca* species. Plant Syst. Evol. 245, 163-168.

Harrison, R.G., 1990. Hybrid zones: windows on evolutionary process. In: Futuyma, D. and Antonovics, J. (Eds.), Oxford Surveys in Evolutionary biology, vol.7. Oxford University Press, Oxford, pp. 69-128.

Harrison, R.G., 1993. Hybrids and hybrid zones: historical perspective. In: Harrison, R. G. (Ed.), Hybrid zones and the evolutionary process. Oxford University Press, Oxford, pp. 3-12.

Hawkes, J.G., 1990. The potato: evolution, biodiversity and genetic resources. Belhaven Press, London.

Hedderson, T.A., Nowell, T.L., 2006. Phylogeography of *Homalothecium sericeum* (Hedw.) Br. Eur.; toward a reconstruction of glacial survival and postglacial migration. J. Bryol. 28, 283-292.

Hedren, M., Fay, M.F., Chase, M.W., 2001. Amplified fragment length polymorphisms (AFLP) reveal details of polyploid evolution in *Dactylorhiza* (Orchidaceae). Am. J. Bot. 88, 1868-1880.

Hegarty, M.J., Hiscock, S.J., 2005. Hybrid speciation in plants: new insights from molecular studies. New Phytol. 165, 411-423.

Hein, J., 1990. Reconstructing evolution of sequences subject to recombination using parsimony. Mathematical Biosciences. 98, 185-200.

Hein, J., 1993. A heuristic method to reconstruct the history of sequences subject to recombination. J. Mol. Evol. 36, 396-405.

Heiser, C.B., 1949. Natural hybridization with particular reference to introgression. Bot. Rev. 15, 645-687.

Heiser, C.B., 1973. Introgression reexamined. Bot. Rev. 39, 347-366.

Helfgott, D.M., Mason-Gamer, R.J., 2004. The evolution of North American *Elymus* (Triticeae, Poaceae) allotetraploids: Evidence from phosphoenolpyruvate carboxylase gene sequences. Syst. Bot. 29, 850-861.

Herbert, W., 1837. Amaryllidaceae; preceded by an attempt to arrange the Monocotyledonous orders, and folllowed by a treatise on cross-bred vegetables, and supplement. Ridgway, London, pp. 1-334.

Hewitt, G.M., 1988. Hybrid Zones - Natural laboratories for evolutionary studies. Trends Ecol. Evol. 3, 158-167.

Hewitt, G.M., 2001. Speciation, hybrid zones and phylogeography - or seeing genes in space and time. Mol. Ecol. 10, 537-549.

Hey, J., 2006. On the failure of modern species concepts. Trends Ecol. Evol. 21, 447-450.

Hillis, D.M., Heath, T.A., St John, K., 2005. Analysis and visualization of tree space. Syst. Biol. 54, 471-482.

Hodkinson, T.R., Chase, M.W., Takahashi, C., Leitch, I.J., Bennett, M.D., Renvoize, S.A., 2002. The use of DNA sequencing (ITS and trnL-F), AFLP, and fluorescent in situ hybridization to study allopolyploid *Miscanthus* (Poaceae). Am. J. Bot. 89, 279-286.

Hoggard, G.D., Kores, P.J., Molvray, M., Hoggard, R.K., 2004. The phylogeny of *Gaura* (Onagraceae) based on ITS, ETS, and trnl-F sequence data. Am. J. Bot. 91, 139-148.

Holder, M.T., Anderson, J.A., Holloway, A.K., 2001. Difficulties in detecting hybridization. Syst. Biol. 50, 978-982.

Holland, B., Conner, G., Huber, K., Moulton, V., 2007. Imputing supertrees and supernetworks from quartets. Syst. Biol. 56, 57-67.

Holland, B., Moulton, V., 2003. Consensus networks: A method for visualising incompatibilities in collections of trees. Algorithms in Bioinformatics, Proceedings, vol.2812, pp. 165-176.

Holland, B.R., Delsuc, F., Moulton, V., 2005. Visualizing conflicting evolutionary hypotheses in large collections of trees: Using consensus networks to study the origins of placentals and hexapods. Syst. Biol. 54, 66-76.

Horandl, E., 2006. Paraphyletic versus monophyletic taxa-evolutionary versus cladistic classifications. Taxon. 55, 564-570.

Horandl, E., Paun, O., Johansson, J.T., Lehnebach, C., Armstrong, T., Chen, L.X., Lockhart, P., 2005. Phylogenetic relationships and evolutionary traits in *Ranunculus* s.l. (Ranunculaceae) inferred from ITS sequence analysis. Mol. Phylogenet. Evol. 36, 305-327.

Houliston, G.J., Olson, M.S., 2006. Nonneutral evolution of organelle genes in *Silene vulgaris*. Genetics. 174, 1983-1994.

Howarth, D.G., Baum, D.A., 2005. Genealogical evidence of homoploid hybrid speciation in an adaptive radiation of *Scaevola* (goodeniaceae) in the Hawaiian Islands. Evolution. 59, 948-961.

Huber, K.T., Langton, M., Penny, D., Moulton, V., Hendy, M., 2002. Spectronet: a package for computing spectra and median networks. Appl Bioinformatics. 1, 159-161.

Huber, K.T., Moulton, V., 2005. Phylogenetic Networks. In: Gascuel, O. (Ed.), Mathematics of Evolution and Phylogeny. Oxford University Press, pp. 178-204.

Huelsenbeck, J.P., Crandall, K.A., 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. Annu. Rev. Ecol. Syst . 28, 437-466.

Huelsenbeck, J.P., Hillis, D.M., 1993. Success of phylogenetic methods in the 4-Taxon case. Syst. Biol. 42, 247-264.

Huelsenbeck, J.P., Larget, B., Miller, R.E., Ronquist, F., 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. Syst. Biol. 51, 673-688.

Huelsenbeck, J.P., Ronquist, F., 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 17, 754-755.

Hughes, C.E., Harris, S.A., 1998. A second spontaneous hybrid in the genus *Leucaena* (Leguminosae, Mimosoideae). Plant Syst. Evol. 212, 53-77.

Huson, D.H., Bryant, D., 2004. SplitsTree4.0 beta20. Distributed by the authors (http://www-ab.informatik.uni-tuebingen.de/software/jsplits/welcome.html).

Huson, D.H., Bryant, D., 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23, 254-267.

Huson, D.H., Dezulian, T., Klopper, T., Steel, M.A., 2004. Phylogenetic super-networks from partial trees. Algorithms in Bioinformatics, Proceedings, vol.3240, pp. 388-399.

Huson, D.H., Kloepper, T., Lockhart, P.J., Steel, M.A., 2005. Reconstruction of reticulate networks from gene trees. RECOMB 2005.

Hypsa, V., 2006. Parasite histories and novel phylogenetic tools: Alternative approaches to inferring parasite evolution from molecular markers. Int. J. Parasitol. 36, 141-155.

Ikeda, H., Setoguchi, H., 2007. Phylogeography and refugia of the Japanese endemic alpine plant, *Phyllodoce nipponica* Makino (Ericaceae). Journal of Biogeography. 34, 169-176.

Jacobs, M.J.M., Van den Berg, R.G., Vosman, B., submitted. Comparing chloroplast and AFLP data reveals a lack of phylogenetic resolution in *Solanum* section Petota. (submitted).

Jin, G.H., Nakhleh, L., Snir, S., Tuller, T., 2006. Maximum likelihood of phylogenetic networks. Bioinformatics. 22, 2604-2611.

Joly, S., Bruneau, A., 2006. Incorporating allelic variation for reconstructing the evolutionary history of organisms from multiple genes: An example from *Rosa* in North America. Syst. Biol. 55, 623-636.

Judd, W.S., Campbell, C.S., Kellogg, E.A., Stevens, P.F., Donoghue, M.J., 2002. Plant systematics: A phylogenetic approach. Sinauer, Sunderland, MA.

Kafkas, S., 2006. Phylogenetic analysis of the genus *Pistacia* by AFLP markers. Plant Syst. Evol. 262, 113-124.

Kameyama, Y., Toyama, M., Ohara, M., 2005. Hybrid origins and F-1 dominance in the freefloating sterile bladderwort, *Utricularia australis F. Australis* (Lentibulariaceae). Am. J. Bot. 92, 469-476.

Kardolus, J.P., van Eck, H.J., van den Berg, R.G., 1998. The potential of AFLPs in biosystematics: a first application in *Solanum* taxonomy (Solanaceae). Plant Syst. Evol. 210, 87-103.

Karpechenko, G.D., 1927. Polyploid hybrids of *Raphanus sativus* L. x *Brassica oleracea* L. Bull. Appl. Bot., Genet. Plant Breed. Ser. II. 17, 305-410.

Katz, L.A., 1999. The tangled web: Gene genealogies and the origin of eukaryotes. American Naturalist. 154, S137-S145.

Kerner, A., 1894-1895. The natural history of plants. London: Blackie & Son.

Kiew, R., Teo, L.L., Gan, Y.Y., 2003. Assessment of the hybrid status of some Malesian plants using Amplified Fragment Length Polymorphism. Telopea. 10, 225-233.

Kihara, H., Ono, T., 1926. Chromosomenzahlen und systematische Gruppierung der Rumex-Arten. Zeitschr. Zellforsch Mikrosk Anat. 4, 475-481.

Knight, T.A., 1806. Observations on the method of producing new and early fruits. Transactions, Horticultural Society of London. 1, 30-40.

Knight, T.A., 1809. On the comparative influence of male and female parents on their offspring. Transactions, Royal Society. 1, 392-399.

Koch, M.A., Dobes, C., Matschinger, M., Bleeker, W., Vogel, J., Kiefer, M., Mitchell-Olds, T., 2005. Evolution of the trnF(GAA) gene in *Arabidopsis* relatives and the Brassicaceae family: Monophyletic origin and subsequent diversification of a plastidic pseudogene. Mol. Biol. Evol. 22, 1032-1043.

Koch, M.A., Dobes, C., Mitchell-Olds, T., 2003. Multiple hybrid formation in natural populations: Concerted evolution of the internal transcribed spacer of nuclear ribosomal DNA (ITS) in north American *Arabis divaricarpa* (Brassicaceae). Mol. Biol. Evol. 20, 338-350.

Kölreuter, J.G., 1761-1766. Vorläufige Nachricht von einigen das geschlecht der pflanzen betreffenden versuchen und beobachtungen, nebst Forsetzungen 1, 2 und 3., Leibzig.

Koontz, J.A., Soltis, P.S., Brunsfeld, S.J., 2001. Genetic diversity and tests of the hybrid origin of the endangered yellow larkspur. Conservation Biology. 15, 1608-1618.

Koontz, J.A., Soltis, P.S., Soltis, D.E., 2004. Using phylogeny reconstruction to test hypotheses of hybrid origin in *Delphinium* section Diedropetala (Ranunculaceae). Syst. Bot. 29, 345-357.

Koopman, W.J.M., 2005. Phylogenetic signal in AFLP data sets. Syst. Biol. 54, 197-217.

Koopman, W.J.M., Gort, G., 2004. Significance tests and weighted values for AFLP similarities, based on Arabidopsis *in silico* AFLP fragment length distributions. Genetics. 167, 1915-1928.

Kornet, D.J., Turner, H., 1999. Coding polymorphism for phylogeny reconstruction. Syst. Biol. 48, 365-379.

Krauss, S.L., Hopper, S.D., 2001. From Dampier to DNA: the 300-year-old mystery of the identity and proposed allopolyploid origin of *Conostylis stylidioides* (Haemodoraceae). Aust. J. Bot. 49, 611-618.

Kubota, S., Kameyama, Y., Ohara, M., 2006. A reconsideration of relationships among Japanese *Trillium* species based on karyology and AFLP data. Plant Syst. Evol. 261, 129-137.

Kuhner, M.K., Felsenstein, J., 1994. Simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. 11, 459-468.

Lara-Cabrera, S.I., Spooner, D.M., 2004. Taxonomy of North and Central American diploid wild potato (*Solanum* sect. Petota) species: AFLP data. Plant Syst. Evol. 248, 129-142.

Lee, C., Kim, S.C., Lundy, K., Santos-Guerra, A., 2005. Chloroplast DNA phylogeny of the woody *Sonchus alliance* (Asteraceae: Sonchinae) in the Macaronesian Islands. Am. J. Bot. 92, 2072-2085.

Leitch, I.J., Bennett, M.D., 1997. Polyploidy in angiosperms. Trends Plant Sci. 2, 470-476.

Leitner, T., Escanilla, D., Franzen, C., Uhlen, M., Albert, J., 1996. Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. Proc. Natl. Acad. Sci. USA. 93, 10864-10869.

Lemmon, A.R., Moriarty, E.C., 2004. The importance of proper model assumption in Bayesian phylogenetics. Syst. Biol. 53, 265-277.

Levin, D.A., 1967. Hybridization between annual species of *Phlox* - population structure. Am. J. Bot. 54, 1122-1130.

Levin, D.A. (Ed.), 1979. Hybridization: an evolutionary perspective. Dowden, Hutchinson & Ross, Stroudsberg, PA, USA.

Lewis, P.O., 2001. Phylogenetic systematics turns over a new leaf. Trends Ecol. Evol. 16, 30-37.

Lewis, H., Epling, C., 1959. *Delphinium-Gypsophilum*, a diploid species of hybrid origin. Evolution. 13, 511-525.

Lewontin, R.C., Birch, L.C., 1966. Hybridization as a source of variation for adaptation to new environments. Evolution. 20, 315-336.

Lexer, C., Fay, M.F., Joseph, J.A., Nica, M.S., Heinze, B., 2005. Barrier to gene flow between two ecologically divergent Populus species, *P alba* (white poplar) and *P tremula* (European aspen): the role of ecology and life history in gene introgression. Mol. Ecol. 14, 1045-1057.

Lexer, C., Lai, Z., Rieseberg, L.H., 2004. Candidate gene polymorphisms associated with salt tolerance in wild sunflower hybrids: implications for the origin of *Helianthus paradoxus*, a diploid hybrid species. New Phytol. 161, 225-233.

Liao, P.C., Havanond, S., Huang, S., 2007. Phylogeography of *Ceriops tagal* (Rhizophoraceae) in Southeast Asia: the land barrier of the Malay Peninsula has caused population differentiation between the Indian Ocean and South China Sea. Conservation Genetics. 8, 89-98.

Lihova, J., Aguilar, J.F., Marhold, K., Feliner, G.N., 2004. Origin of the disjunct tetraploid *Cardamine amporitana* (Brassicaceae) assessed with nuclear and chloroplast DNA sequence data. Am. J. Bot. 91, 1231-1242.

Lihova, J., Shimizu, K.K., Marhold, K., 2006. Allopolyploid origin of *Cardamine asarifolia* (Brassicaceae): Incongruence between plastid and nuclear ribosomal DNA sequences solved by a single-copy nuclear gene. Mol. Phylogenet. Evol. 39, 759-786.

Linder, C.R., Moret, B.M.E., Nakhleh, L., Padolina, A., Sun, J., Tholse, A., Timme, R., Warnow, T., 2003. An error metric for phylogenetic networks. Technical Report TR–CS-2003–2026, University of New Mexico, Albuquerque, New Mexico.

Linder, C.R., Rieseberg, L.H., 2004. Reconstructing patterns of reticulate evolution in plants. Am. J. Bot. 91, 1700-1708.

Linné, C., 1760. Disquisitio de sexu plantarum, ab Academia Imperiali Scientiarum Petropolitana praemio ornata. Amoenitates Academicae. 10, 100-131.

Liu, J., Glazko, G., Mushegian, A., 2006. Protein repertoire of double-stranded DNA bacteriophages. Virus Research. 117, 68-80.

Liu, L., Pearl, D.K., 2007. High-resolution species trees without concatenation. Syst. Biol. 56, 504-514.

Lockhart, P.J., McLenachan, P.A., Harell, D., Glenny, D., Huson, D., Jensen, U., 2001. Phylogeny, radiation, and transoceanic dispersal of New Zealand alpine buttercups: Molecular evidence under split decomposition. Ann. Mo. Bot. Gard. 88, 458-477.

Lorenz-Lemke, A.P., Muschner, V.C., Bonatto, S.L., Cervi, A.C., Salzano, F.M., Freitas, L.B., 2005. Phylogeographic inferences concerning evolution of Brazilian *Passiflora actinia* and *P-elegans* (Passifloraceae) based on ITS (nrDNA) variation. Ann. Bot. 95, 799-806.

Lotsy, J.P., 1916. Evolution by means of natural hybridization. M. Nijhof, The Hague.

Lowe, A.J. Harris, S.A. and Ashton, P.A., 2004. Ecological genetics. Design, analysis and application. Blackwell Publishing, Oxford.

Luo R, Hipp, AL, Larget, B., 2007. A Bayesian model of AFLP marker evolution and phylogenetic inference. Stat. Appl. Genet. Mol. Biol. 6, art11.

Machado, C.A., Kliman, R.M., Markert, J.A., Hey, J., 2002. Inferring the history of speciation from multilocus DNA sequence data: The case of *Drosophila pseudoobscura* and close relatives. Mol. Biol. Evol. 19, 472-488.

MacLeod, D., Charlebois, R.L., Doolittle, F., Bapteste, E., 2005. Deduction of probable events of lateral gene transfer through comparison of phylogenetic trees by recursive consolidation and rearrangement. BMC Evol. Biol. 5.

Maddison, W.P., 1997. Gene trees in species trees. Syst. Biol. 46, 523-536.

Maddison, W.P., Maddison, D.R., 2004. Mesquite: a modular system for evolutionary analysis. Version 1.05 http://mesquiteproject.org.

Maddison, W.P., Maddison, D.R., 2005. MacClade version 4.07. Analysis of phylogeny and character evolution.). Sunderland, Massachusetts, Sinauer Associates.

Makarenkov, U., 2001. T-REX: reconstructing and visualizing phylogenetic trees and reticulation networks. Bioinformatics. 17, 664-668.

Makarenkov, V., Legendre, P., 2004. From a phylogenetic tree to a reticulated network. J. Comput. Biol. 11, 195-212.

Mallet, J., 2005. Hybridization as an invasion of the genorne. Trends Ecol. Evol. 20, 229-237.

Mallet, J., 2007. Hybrid speciation. Nature. 446, 279-283.

Mansion, G., Zeltner, L., Bretagnolle, F., 2005. Phylogenetic patterns and polyploid evolution within the Mediterranean genus Centaurium (Gentianaceae-Chironieae). Taxon. 54, 931-950.

Marhold, K., Lihova, J., 2006. Polyploidy, hybridization and reticulate evolution: lessons from the Brassicaceae. Plant Syst. Evol. 259, 143-174.

Marhold, K., Lihova, J., Perny, M., Bleeker, W., 2004. Comparative ITS and AFLP analysis of diploid *Cardamine* (Brassicaceae) taxa from closely related polyploid complexes. Ann. Bot. 93, 507-520.

Martin, D., Rybicki, E., 2000. RDP: detection of recombination amongst aligned sequences. Bioinformatics. 16, 562-563.

Martin, D.P., Williamson, C., Posada, D., 2005. RDP2: recombination detection and analysis from sequence alignments. Bioinformatics. 21, 260-262.

Martin, W., Embley, T.M., 2004. Evolutionary biology - Early evolution comes full circle. Nature. 431, 134-137.

Mason-Gamer, R.J., 2004. Reticulate evolution, introgression, and intertribal gene capture in an allohexaploid grass. Syst. Biol. 53, 25-37.

Masterson, J., 1994. Stomatal size in fossil plants - evidence for polyploidy in majority of angiosperms. Science. 264, 421-424.

Mayr, E., 1942. Systematics and the origin of species. Colombia University Press, New York, NY.

Mayr, E., 1963. Animal species and evolution. Belknap Press, Cambridge, MA.

Mazzoni, C.J., Souza, N.A., Andrade-Coelho, C., Kyriacou, C.P., Peixoto, A.A., 2006. Molecular polymorphism, differentiation and introgression in the period gene between *Lutzomyia intermedia* and *Lutzomyia whitmani*. BMC Evol. Biol. 6.

McBreen, K., Lockhart, P.J., 2006. Reconstructing reticulate evolutionary histories of plants. Trends Plant Sci. 11, 398-404.

McDade, L., 1990. Hybrids and phylogenetic systematics .1. Patterns of character expression in hybrids and their implications for cladistic analysis. Evolution. 44, 1685-1700.

McDade, L., 1995. Hybridization and phylogenetics. In: Hoch, P. C. and Stephenson, A. G. (Eds.), Experimental and molecular approaches to plant biosystematics. Missouri Botanical Garden, St. Louis, pp. 305-331.

McDade, L.A., 1992. Hybrids and phylogenetic systematics .2. The impact of hybrids on cladistic analysis. Evolution. 46, 1329-1346.

Meyers, L.A., Levin, D.A., 2006. On the abundance of polyploids in flowering plants. Evolution. 60, 1198-1206.

Miller, A.H., 1949. Some concepts of hybridization and intergradation in wild populations of birds. Auk. 66, 388-342.

Milne, I., Wright, F., Rowe, G., Marshall, D.F., Husmeier, D., McGuire, G., 2004. TOPALi: software for automatic identification of recombinant sequences within DNA multiple alignments. Bioinformatics. 20, 1806-1807.

Mineli, A., 1993. Biological systematics; the state of the art. Chapman & Hall, London.

Mirov, N.T., 1967. The genus *Pinus*, Ronald Press, New York.

Moret, B.M.E., Nakhleh, L., Warnow, T., Linder, C.R., Tholse, A., Padolina, A., Sun, J., Timme, R., 2004. Phylogenetic Networks: Modeling, Reconstructibility, and Accuracy. IEEE Transactions on computational Biology and Bioinformatics.

Morrison, D.A., 2005. Networks in phylogenetic analysis: new tools for population biology. Int. J. Parasitol. 35, 567-582.

Mummenhoff, K., Linder, P., Friesen, N., Bowman, J.L., Lee, J.Y., Franzke, A., 2004. Molecular evidence for bicontinental hybridogenous genomic constitution in *Lepidium* sensu stricto (Brassicaceae) species from Australia and New Zealand. Am. J. Bot. 91, 254-261.

Müntzing, A., 1930. Uber Chromosomenvermehrung in *Galeopsis*-Kreuzungen und ihre phylogenetische Bedeutung. Hereditas. 14, 153-172.

Nakhleh, L., Ruths, D., Than, C., 2006. "PhyloNet: A Phylogenetic Networks Toolkit." http://bioinfo.cs.rice.edu/phylonet.

Nakhleh, L., Ruths, D., Wang, L.S., 2005a. RIATA-HGT: A fast and accurate heuristic for reconstructing horizontal gene transfer. Computing and Combinatorics, Proceedings, vol.3595, pp. 84-93.

Nakhleh, L., Sun, J., Warnow, T., Linder, C.R., Moret, B.M.E., Tholse, A., 2003. Towards the development of computational tools for evaluating phylogenetic network reconstruction methods. Proceedings of the PSB '03, pp.

Nakhleh, L., Warnow, T., Linder, C.R., St John, K., 2005b. Reconstructing reticulate evolution in species - Theory and practice. J. Comput. Biol. 12, 796-811.

Naudin, C., 1861. Sur les plantes hybrides. Revue Horticole, 4me Série, pp. 396-399.

Naudin, C., 1863. Nouvelles recherches sur l'hybridité dans les végétaux. Annales des Sciences Naturelles, Botanique, 4me Série. 19, 180-203.

Naudin, C., 1864. De l'hybridité considérée comme cause de variabilité dans les végétaux. Comptes Rendus de l'Académie des Sciences, 59. 59, 837-845.

Nei, M., Li, W.H., 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA. 76, 5269-5273.

Neves, S.S., Swire-Clark, G., Hilu, K.W., Baird, W.V., 2005. Phylogeny of *Eleusine* (Poaceae : Chloridoideae) based on nuclear ITS and plastid trnT-trnF sequences. Mol. Phylogenet. Evol. 35, 395-419.

Newton, W.C.F., Pellew, C., 1929. *Primula kewensis* and its derivatives. J. Genet. 20, 405-467.

Nguyen, N.B., Nguyen, C.T., Sung, W.K., 2005. Fast algorithms for computing the tripartition-based distance between phylogenetic networks. Algorithms and Computation, vol.3827, pp. 402-411.

Nishimoto, Y., Ohnishi, O., Hasegawa, M., 2003. Topological incongruence between nuclear and chloroplast DNA trees suggesting hybridization in the urophyllum group of the genus *Fagopyrum* (Polygonaceae). Genes & Genetic Systems. 78, 139-153.

Nylander, J.A.A., 2004. MrModeltest v2. Program distributed by the author. Evolutionary Biology Centre, Uppsala University

Nylander, J.A.A., Ronquist, F., Huelsenbeck, J.P., Nieves-Aldrey, J.L., 2004. Bayesian phylogenetic analysis of combined data. Syst. Biol. 53, 47-67.

Obbard, D.J., Harris, S.A., Buggs, R.J.A., Pannell, J.R., 2006. Hybridization, polyploidy, and the evolution of sexual systems in *Mercurialis* (Euphorbiaceae). Evolution. 60, 1801-1815.

Oh, S.H., Potter, D., 2003. Phylogenetic utility of the second intron of LEAFY in *Neillia* and *Stephanandra* (Rosaceae) and implications for the origin of *Stephanandra*. Mol. Phylogenet. Evol. 29, 203-215.

O'Hanlon, P.C., Peakall, R., Briese, D.T., 1999. Amplified fragment length polymorphism (AFLP) reveals introgression in weedy *Onopordum* thistles: hybridization and invasion. Mol. Ecol. 8, 1239-1246.

Okuyama, Y., Fujii, N., Wakabayashi, M., Kawakita, A., Ito, M., Watanabe, M., Murakami, N., Kato, M., 2005. Nonuniform concerted evolution and chloroplast capture: Heterogeneity of observed introgression patterns in three molecular data partition phylogenies of Asian *Mitella* (Saxifragaceae). Mol. Biol. Evol. 22, 285-296.

Ostenfeld, C.H., 1928. The present state of knowledge on hybrids between species of flowering plants. J. R. Hort. Soc. 53, 31-44.

Otto, S.P., Whitton, J., 2000. Polyploid incidence and evolution. Annu. Rev. Genet. . 34, 401-437.

Padidam, M., Sawyer, S., Fauquet, C.M., 1999. Possible emergence of new geminiviruses by frequent recombination. Virology. 265, 218-225.

Page, R.D.M., 1994. Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. Syst. Biol. 43, 58-77.

Page, R.D.M., 1998. GeneTree: comparing gene and species phylogenies using reconciled trees. Bioinformatics. 14, 819-820.

Page, R.D.M., Charleston, M.A., 1998. Trees within trees: phylogeny and historical associations. Trends Ecol. Evol. 13, 356-359.

Page, R.D.M., Holmes, E.C., 1998. Molecular Evolution: A Phylogenetic Approach. Blackwell Science, Oxford.

Paula, M., Leonardo, G., 2006. Multiple ice-age refugia in a southern beech of South America as evidenced by chloroplast DNA markers. Conservation Genetics. 7, 591-603.

Paun, O., Stuessy, T.F., Horandl, E., 2006. The role of hybridization, polyploidization and glaciation in the origin and evolution of the apomictic *Ranunculus cassubicus* complex. New Phytol. 171, 223-236.

Perrie, L.R., Brownsey, P.J., Lockhart, P.J., Large, M.F., 2003. Evidence for an allopolyploid complex in New Zealand *Polystichum* (Dryopteridaceae). New Zeal. J. Bot. 41, 189-215.

Petersen, G., Seberg, O., 2004. On the origin of the tetraploid species *Hordeum capense* and *H-secalinum* (Poaceae). Syst. Bot. 29, 862-873.

Peterson, A., John, H., Koch, E., Peterson, J., 2004. A molecular phylogeny of the genus *Gagea* (Liliaceae) in Germany inferred from non-coding chloroplast and nuclear DNA sequences. Plant Syst. Evol. 245, 145-162.

Phillips, M.J., Delsuc, F., Penny, D., 2004. Genome-scale phylogeny and the detection of systematic biases. Mol. Biol. Evol. 21, 1455-1458.

Piontkivska, H., 2004. Efficiencies of maximum likelihood methods of phylogenetic inferences when different substitution models are used. Mol. Phylogenet. Evol. 31, 865-873.

Pires, J.C., Lim, K.Y., Kovarik, A., Matyasek, R., Boyd, A., Leitch, A.R., Leitch, I.J., Bennett, M.D., Soltis, P.S., Soltis, D.E., 2004. Molecular cytogenetic analysis of recently evolved *Tragopogon* (Asteraceae) allopolyploids reveal a karyotype that is additive of the diploid progenitors. Am. J. Bot. 91, 1022-1035.

Poke, F.S., Martin, D.P., Steane, D.A., Vaillancourt, R.E., Reid, J.B., 2006. The impact of intragenic recombination on phylogenetic reconstruction at the sectional level in *Eucalyptus* when using a single copy nuclear gene (cinnamoyl CoA reductase). Mol. Phylogenet. Evol. 39, 160-170.

Poptsova, M.S., Gogarten, J.P., 2007. The power of phylogenetic approaches to detect horizontally transferred genes. BMC Evol. Biol. 7.

Posada, D., 2002. Evaluation of methods for detecting recombination from DNA sequences: Empirical data. Mol. Biol. Evol. 19, 708-717.

Posada, D., Crandall, K.A., 2002. The effect of recombination on the accuracy of phylogeny estimation. J. Mol. Evol. 54, 396-402.

Posada, D., Crandall, K.A., 2001a. Evaluation of methods for detecting recombination from DNA sequences: Computer simulations. Proc. Natl. Acad. Sci. USA. 98, 13757-13762.

Posada, D., Crandall, K.A., 2001b. Intraspecific gene genealogies: trees grafting into networks. Trends Ecol. Evol. 16, 37-45.

Posada, D., Crandall, K.A., 2001c. Selecting the best-fit model of nucleotide substitution. Syst. Biol. 50, 580-601.

Posada, D., Crandall, K.A., Holmes, E.C., 2002. Recombination in evolutionary genomics. Annu. Rev. Genet. . 36, 75-97.

Rambaut, A., Charleston, M., 2001. TreeEdit v1.0a10. http://evolve.zoo.ox.ac.uk

Rambaut, A., Grassly, N.C., 1997. Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic frees. Computer Applications in the Biosciences. 13, 235-238.

Ramsey, J., Schemske, D.W., 1998. Pathways, mechanisms, and rates of polyploid formation in flowering plants. Annu. Rev. Ecol. Syst . 29, 467-501.

Ramsey, J., Schemske, D.W., 2002. Neopolyploidy in flowering plants. Annu. Rev. Ecol. Syst . 33, 589-639.

Rauscher, J.T., Doyle, J.J., Brown, A.H.D., 2004. Multiple origins and nrDNA internal transcribed spacer homeologue evolution in the *Glycine tomentella* (Leguminosae) allopolyploid complex. Genetics. 166, 987-998.

Raven, P.H., 1976. Systematics and Plant Population Biology. Syst. Bot. 1, 284-316.

Ridley, M., 2004. Evolution. Blackwell, Malden, MA.

Rieseberg, L.H., 1997. Hybrid origins of plant species. Annu. Rev. Ecol. Syst . 28, 359-389.

Rieseberg, L.H., Brunsfeld, S.J., 1992. Molecular evidence and plant introgression. In: Soltis, P. S. et al. (Eds.), Molecular Systematics of Plants, vol.I. Chapman & Hall, New York, pp. 151-176.

Rieseberg, L.H., Carney, S.E., 1998. Plant hybridization. New Phytol. 140, 599-624.

Rieseberg, L.H., Ellstrand, N.C., 1993. What can molecular and morphological markers tell us about plant hybridization. Cr. Rev. Plant Sci. 12, 213-241.

Rieseberg, L.H., Raymond, O., Rosenthal, D.M., Lai, Z., Livingstone, K., Nakazato, T., Durphy, J.L., Schwarzbach, A.E., Donovan, L.A., Lexer, C., 2003. Major ecological transitions in wild sunflowers facilitated by hybridization. Science. 301, 1211-1216.

Rieseberg, L.H., Sinervo, B., Linder, C.R., Ungerer, M.C., Arias, D.M., 1996. Role of gene interactions in hybrid speciation: Evidence from ancient and experimental hybrids. Science. 272, 741-745.

Rieseberg, L.H., Soltis, D.E., 1991. Phylogenetic consequences of cytoplasmic gene flow in plants. Evol. Trend. Plant. 5, 65-84.

Rieseberg, L.H., Wendel, J.F., 1993. Introgression and its consequences in plants. In: Harrison, R. G. (Ed.), Hybrid zones and the evolutionary process. Oxford University Press, Oxford, pp. 70-109.

Ritz, C.M., Schmuths, H., Wissemann, V., 2005. Evolution by reticulation: European dogroses originated by multiple hybridization across the genus *Rosa*. J. Hered. 96, 4-14.

Rivera, M.C., Lake, J.A., 2004. The ring of life provides evidence for a genome fusion origin of eukaryotes. Nature. 431, 152-155.

Roberts, H.F., 1929. Plant hybridization before Mendel. Princeton University Press, Princeton, NJ, USA.

Robinson, D.F., Foulds, L.R., 1981. Comparison of phylogenetic trees. Mathematical Biosciences. 53, 131-147.

Rohlf, F.J., 2004. NTSYSpc. Numerical taxonomy and multivariate analysis system. Version 2.2. Applied Biostatistics Inc., New York.

Rokas, A., Williams, B.L., King, N., Carroll, S.B., 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature. 425, 798-804.

Ronquist, F., 1996. DIVA version 1.1. Computer program and manual available by anonymous FTP from Uppsala University (ftp.uu.se or ftp.systbot.uu.se).

Ronquist, F., Huelsenbeck, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 19, 1572-1574.

Rossello, J.A., Cosin, R., Boscaiu, M., Vicente, O., Martinez, I., Soriano, P., 2006. Intragenomic diversity and phylogenetic systematics of wild rosemaries (*Rosmarinus officinalis L.* s.l., Lamiaceae) assessed by nuclear ribosomal DNA sequences (ITS). Plant Syst. Evol. 262, 1-12.

Ruths, D., Nakhleh, L., 2005. Recombination and phylogeny: effects and detection. Int J Bioinformatics Research and Applications. 1, 202-212.

Salminen, M.O., Carr, J.K., Burke, D.S., McCutchan, F.E., 1995. Identification of breakpoints in intergenotypic recombinants of Hiv Type-1 by bootscanning. Aids Research and Human Retroviruses. 11, 1423-1425.

Sang, T., Crawford, D.J., Stuessy, T.F., 1995. Documentation of reticulate evolution in Peonies (*Paeonia)* using internal transcribed spacer sequences of nuclear ribosomal DNA - Implications for biogeography and concerted evolution. Proc. Natl. Acad. Sci. USA. 92, 6813-6817.

Sang, T., Crawford, D.J., Stuessy, T.F., 1997. Chloroplast DNA phylogeny, reticulate evolution, and biogeography of *Paeonia* (Paeoniaceae). Am. J. Bot. 84, 1120-1136.

Sang, T., Zhang, D.M., 1999. Reconstructing hybrid speciation using sequences of low copy nuclear genes: Hybrid origins of five *Paeonia* species based on Adh gene phylogenies. Syst. Bot. 24, 148-163.

Sang, T., Zhong, Y., 2000. Testing hybridization hypotheses based on incongruent gene trees. Syst. Biol. 49, 422-434.

Såstad, S.M., 2005. Patterns and mechanisms of polyploid speciation in bryophytes. In: Bakker, F.T., Chatrou, L.W., Gravendeel, B., Pelser, P. (Eds), *Plant Species-level Systematics: New perspectives on pattern & process,* Regnum Vegetabile 143, Gantner Verlag, Liechtenstein, pp. 317-333.

Schemske, D.W., 2000. Understanding the origin of species. Evolution. 54,1069–1073

Schierup, M.H., Hein, J., 2000a. Consequences of recombination on traditional phylogenetic analysis. Genetics. 156, 879-891.

Schierup, M.H., Hein, J., 2000b. Recombination and the molecular clock. Mol. Biol. Evol. 17, 1578-1579.

Schmidt-Lebuhn, A.N., Kessler, M., Kumar, M., 2006. Promiscuity in the Andes: Species relationships in *Polylepis* (Rosaceae, Sanguisorbeae) based on AFLP and morphology. Syst. Bot. 31, 547-559.

Schneider, S., Roessli, D., Excoffier, L., 2000. Arlequin: A software for population genetics data analysis. Version 2.000. Genetics and Biometry Lab, Dept. of Anthropology, University of Geneva, Geneva.

Schuh, R.T., 2000. Biological Systematics: principles and applications. Ithaca,: Cornell University Press, New York.

Schwartz, G. 1978. Estimating the dimension of a model. Ann. Statist. 6:461–464.

Schwarz, D., Matta, B.M., Shakir-Botteri, N.L., McPheron, B.A., 2005. Host shift to an invasive plant triggers rapid animal hybrid speciation. Nature. 436, 546-549.

Seehausen, O., 2004. Hybridization and adaptive radiation. Trends Ecol. Evol. 19, 198-207.

Shaw, A.J., Cox, C.J., Boles, S.B., 2005. Phylogeny, species delimitation, and recombination in *Sphagnum* section Acutifolia. Syst. Bot. 30, 16-33.

Shedlock, A.M., Okada, N., 2000. SINE insertions: powerful tools for molecular systematics. Bioessays. 22, 148-160.

Shih, F.L., Hwang, S.Y., Cheng, Y.P., Lee, P.F., Lin, T.P., 2007. Uniform genetic diversity, low differentiation, and neutral evolution characterize contemporary refuge populations of Taiwan fir (*Abies kawakamii*, Pinaceae). Am. J. Bot. 94, 194-202.

Shipunov, A.B., Fay, M.F., Pillon, Y., Bateman, R.M., Chase, M.W., 2004. *Dactylorhiza* (Orchidaceae) in European Russia: Combined molecular and morphological analysis. Am. J. Bot. 91, 1419-1426.

Short, L.L., 1969. Taxonomic aspects of avian hybridization. Auk. 86, 84-105.

Short, L.L., 1972. Hybridization, taxonomy and avian evolution. Annals of the Missouri Botanical Garden. 59, 447-453.

Simmons, M.P., Zhang, L.B., Webb, C.T., Muller, K., 2007. A penalty of using anonymous dominant markers (AFLPs, ISSRs, and RAMS) for phylogenetic inference. Mol. Phylogenet. Evol. 42, 528-542.

Simmons, N.B., 2001. Misleading results from the use of ambiguity coding to score polymorphisms in higher-level taxa. Syst. Biol. 50, 613-620.

Slotte, T., Ceplitis, A., Neuffer, B., Hurka, H., Lascoux, M., 2006. Intrageneric phylogeny of *Capsella* (Brassicaceae) and the origin of the tetraploid *C-bursa-pastoris* based on chloroplast and nuclear DNA sequences. Am. J. Bot. 93, 1714-1724.

Slowinski, J.B., Page, R.D.M., 1999. How should species phylogenies be inferred from sequence data? Syst. Biol. 48, 814-825.

Smedmark, J.E.E., Eriksson, T., Evans, R.C., Campbell, C.S., 2003. Ancient allopolyploid speciation in *Geinae* (Rosaceae): Evidence from nuclear granule-bound starch synthase (GBSSI) gene sequences. Syst. Biol. 52, 374-385.

Smith, H.H., Daly, K., 1959. Discrete populations derived by interspecific hybridization and selection in *Nicotiana*. Evolution. 13, 476-487.

Smith, J.M., 1992. Analyzing the mosaic structure of genes. J. Mol. Evol. 34, 126-129.

Soltis, D.E., Soltis, P.S., 1993. Molecular data and the dynamic nature of polyploidy. Cr. Rev. Plant Sci. 12, 243-273.

Soltis, D.E., Soltis, P.S., 1999. Polyploidy: recurrent formation and genome evolution. Trends Ecol. Evol. 14, 348-352.

Soltis, D.E., Soltis, P.S., Tate, J.A., 2004. Advances in the study of polyploidy since Plant speciation. New Phytol. 161, 173-191.

Soltis, P.S., Doyle, J.J., Soltis, D.E., 1992. Molecular data and polyploid evolution in plants. In: Soltis, P. S. et al. (Eds.), Molecular Systematics of Plants, vol.I. Chapman & Hall, New York, pp. 177-201.

Soltis, P.S., Soltis, D.E., 2000. The role of genetic and genomic attributes in the success of polyploids. Proc. Natl. Acad. Sci. USA. 97, 7051-7057.

Song, K.M., Lu, P., Tang, K.L., Osborn, T.C., 1995. Rapid genome change in synthetic polyploids of *Brassica* and its implications for polyploid evolution. Proc. Natl. Acad. Sci. USA. 92, 7719-7723.

Sosef, M.S.M., 1997. Hierarchical models, reticulate evolution and the inevitability of paraphyletic supraspecific taxa. Taxon. 46, 75-85.

Spooner, D.M., Peralta, I.E., Knapp, S., 2005. Comparison of AFLPs with other markers for phylogenetic inference in wild tomatoes [*Solanum L.* section Lycopersicon (Mill.) Wettst.]. Taxon. 54, 43-61.

Spooner, D.M., Van den Berg, R.G., Rodrıguez, A., Bamberg, J., Hijmans, R.J., Lara-Cabrera, S.I., 2004. Wild potatoes (*Solanum* section Petota) of North and Central America. Syst. Bot. Monogr. 68: 1–209+9pl.

Spring, J., 2003. Major transitions in evolution by genome fusions: from prokaryotes to eukaryotes, metazoans, bilaterians and vertebrates. Journal of Structural and Functional Genomics. 3, 19-25.

Stace, C.A., 1989. Plant taxonomy and biosystematics. Edward Arnold, London.

Stadler, T., Roselius, K., Stephan, W., 2005. Genealogical footprints of speciation processes in wild tomatoes: Demography and evidence for historical gene flow. Evolution. 59, 1268-1279.

Stafleu, F., 1971. Linnaeus and the Linnaeans. Oosthoeck, Utrecht, Netherlands.

Stamos, D.N., 2003. The species problem: Biological species, ontology, and the metaphysics of biology. Lexington Books, Lanham.

Stebbins, G.L., 1950. Variation and evolution in plants. Colombia University Press, New York, New York.

Stebbins, G.L., 1957. The hybrid origin of microspecies in the *Elymus glaucus* complex. Cytologia Suppl. 36, 336-340.

Stebbins, G.L., 1959. The role of hybridization in evolution. Proc. Am. Philos. Soc . 103, 231-251.

Stebbins, G.L., 1971. Chromosome evolution in higher plants. Colombia University Press, New York, New York.

Stebbins, G.L., Ferlan, L., 1956. Population varibility, hybridization, and introgression in some species of *Ophrys.* Evolution. 10, 32-46.

Stracke, S., Presterl, T., Stein, N., Perovic, D., Ordon, F., Graner, A., 2007. Effects of introgression and recombination on haplotype structure and linkage disequilibrium surrounding a locus encoding *Bymovirus* resistance in barley. Genetics. 175, 805-817.

Straw, R.M., 1955. Hybridization, homogamy, and sympatric speciation. Evolution. 9, 441-444.

Strimmer, K., Moulton, V., 2000. Likelihood analysis of phylogenetic networks using directed graphical models. Mol. Biol. Evol. 17, 875-881.

Strimmer, K., Wiuf, C., Moulton, V., 2001. Recombination analysis using directed graphical models. Mol. Biol. Evol. 18, 97-99.

Sun, K., Chen, X., Ma, R., Li, C., Wang, Q., Ge, S., 2002. Molecular phylogenetics of *Hippophae L.* (Elaeagnaceae) based on the internal transcribed spacer (ITS) sequences of nrDNA. Plant Syst. Evol. 235, 121-134.

Sun, K., Ma, R.J., Chen, X.L., Li, C.B., Ge, S., 2003. Hybrid origin of the diploid species *Hippophae goniocarpa* evidenced by the internal transcribed spacers (ITS) of nuclear rDNA. Belg. J. Bot. 136, 91-96.

Susko, E., Leigh, J., Doolittle, W.F., Bapteste, E., 2006. Visualizing and assessing phylogenetic congruence of core gene sets: A case study of the gamma-proteobacteria. Mol. Biol. Evol. 23, 1119-1030.

Suzuki, Y., Glazko, G.V., Nei, M., 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. Proc. Natl. Acad. Sci. USA. 99, 16138-16143.

Swofford, D.L., 2002. PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods). version 4.0b10. Sinauer Associates, Sunderland, MA

Takahata, N., 1989. Gene genealogy in 3 related populations - consistency probability between gene and population trees. Genetics. 122, 957-966.

Tate, J.A., Simpson, B.B., 2003. Paraphyly of *Tarasa* (Malvaceae) and diverse origins of the polyploid species. Syst. Bot. 28, 723-737.

Tel-Zur, N., Abbo, S., Bar-Zvi, D., Mizrahi, Y., 2004. Genetic relationships among *Hylocereus* and *Selenicereus* vine cacti (Cactaceae): Evidence from hybridization and cytological studies. Ann. Bot. 94, 527-534.

Templeton, A.R., 1981. Mechanisms of speciation - a population genetic approach. Annu. Rev. Ecol. Syst . 12, 23-48.

Templeton, A.R., Crandall, K.A., Sing, C.F., 1992. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA-sequence data .3. Cladogram Estimation. Genetics. 132, 619-633.

Teo, L.L., Kiew, R., Set, O., Lee, S.K., Gan, Y.Y., 2002. Hybrid status of kuwini, *Mangifera odorata* Griff. (Anacardiaceae) verified by amplified fragment length polymorphism. Mol. Ecol. 11, 1465-1469.

Than, C., Ruths, D., Innan, H., Nakhleh, L., 2006. Identifiability issues in phylogeny-based detection of horizontal gene transfer. Comparative Genomics, Proceedings, vol.4205, pp. 215-229.

Thompson, J.D., Lumaret, R., 1992. The Evolutionary Dynamics of Polyploid Plants - Origins, Establishment and Persistence. Trends Ecol. Evol. 7, 302-307.

Tremetsberger, K., Stuessy, T.F., Kadlec, G., Urtubey, E., Baeza, C.M., Beck, S.G., Valdebenito, H.A., Ruas, C.D.F., Matzenbacher, N.I., 2006. AFLP phylogeny of South American species of *Hypochaeris* (Asteraceae, Lactuceae). Syst. Bot. 31, 610-626.

van Dijk, P.J., 2003. Ecological and evolutionary opportunities of apomixis: insights from *Taraxacum* and *Chondrilla*. Philos. T. Roy. Soc. B. 358, 1113-1120.

Van Droogenbroeck, B., Kyndt, T., Romeijn-Peeters, E., Van Thuyne, W., Goetghebeur, P., Romero-Motochi, J.P., Gheysen, G., 2006. Evidence of natural hybridization and introgression between *Vasconcellea* species (Caricaceae) from southern Ecuador revealed by chloroplast, mitochondrial and nuclear DNA markers. Ann. Bot. 97, 793-805.

Vander Stappen, J., Lopez, S.G., Davila, P., Volckaert, G., 2002. Molecular evidence for the hybrid origin of a new endemic species of *Stylosanthes* Sw. (Fabaceae) from the Mexican Yucatan Peninsula. Bot. J. Linn. Soc. 140, 1-13.

Van Velzen, R., Bakker, F.T., (in prep.) Molecular systematics of the Cannabaceae.

Vogel, J.C., Barrett, J.A., Rumsey, F.J., Gibby, M., 1999. Identifying multiple origins in polyploid homosporous pteridophytes. In: Hollingsworth, P. M. et al. (Eds.), Molecular Systematics and Plant Evolution. Taylor & Francis, London, pp. 101-117.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., Vandelee, T., Hornes, M., Frijters, A., Pot, J., Peleman, J., Kuiper, M., Zabeau, M., 1995. AFLP - a New Technique for DNA-Fingerprinting. Nucleic Acids Research. 23, 4407-4414.

Vriesendorp, B., Bakker, F.T., 2005. Reconstructing patterns of reticulate evolution in angiosperms: what can we do? Taxon. 54, 593-604.

Wagner, W.H., 1954. Reticulate evolution in the Appalachian *Aspleniums*. Evolution. 8, 103-118.

Wendel, J.F., Doyle, J.J., 1998. Phylogenetic incongruence: Window into genome history and molecular evolution. In: Soltis, P. S. et al. (Eds.), Molecular Systematics of Plants II. DNA Sequencing, vol.II. Kluwer Academic Publishers, Boston, MA, pp. 265-296.

Wendel, J.F., Schnabel, A., Seelanan, T., 1995b. Bidirectional interlocus concerted evolution following allopolyploid speciation in cotton *(Gossypium).* Proc. Natl. Acad. Sci. USA. 92, 280-284.

Wendel, J.F., Schnabel, A., Seelanan, T., 1995a. An unusual ribosomal DNA-Sequence from *Gossypium Gossypioides* reveals ancient, cryptic, intergenomic introgression. Mol. Phylogenet. Evol. 4, 298-313.

Wendel, J.F., Stewart, J.M., Rettig, J.H., 1991. Molecular evidence for homoploid reticulate evolution among Australian species of *Gossypium*. Evolution. 45, 694-711.

White, M.J.D., 1954. Animal cytology and evolution. Cambridge University Press, Cambridge.

Whitehouse, C.M., 2002. Systematics of the genus *Cliffortia* L. [Rosaceae]. In: Department of Botany). University of Cape Town.

Whittemore, A.T., Schaal, B.A., 1991. Interspecific gene flow in sympatric oaks. Proc. Natl. Acad. Sci. USA. 88, 2540-2544.

Wiens, J.J., 1998. Combining data sets with different phylogenetic histories. Syst. Biol. 47, 568-581.

Wieringa, J.J., Gühl, K., 2006. Species and generic delimitation in *Bikinia* and *Tetraberlinia* (Leguminosae-Caesalpinioideae) using ITS and AFLP. Proceedings of the XVIIth AETFAT International Congress, Kew, pp. 545-558.

Willis, B.L., van Oppen, M.J.H., Miller, D.J., Vollmer, S.V., Ayre, D.J., 2006. The role of hybridization in the evolution of reef corals. Annu. Rev. Ecol. Evol. Syst . 37, 489-517.

Winge, O., 1917. The chromosomes: their number and general importance. Compt Rend Trav Lab Carlsberg. 13, 131-275.

Winkler, H., 1916. Uber die experimentelle Erzeugung von Pflanzen mit abweichenden Chromosomenzahlen. Zeitschr f Bot. 8.

Winkworth, R.C., Bryant, D., Lockhart, P.J., Havell, D., Moulton, V., 2005. Biogeographic interpretation of splits graphs: Least squares optimization of branch lengths. Syst. Biol. 54, 56-65.

Winkworth, R.C., Donoghue, M.J., 2004. *Viburnum* phylogeny: evidence from the duplicated nuclear gene GBSSI. Mol. Phylogenet. Evol. 33, 109-126.

Wissemann, V., Ritz, C.M., 2005. The genus *Rosa* (Rosoideae, Rosaceae) revisited: molecular analysis of nrITS-1 and atpB-rbcL intergenic spacer (IGS) versus conventional taxonomy. Bot. J. Linn. Soc. 147, 275-290.

Wiuf, C., Christensen, T., Hein, J., 2001. A simulation study of the reliability of recombination detection methods. Mol. Biol. Evol. 18, 1929-1939.

Woodruff, D.S., 1973. Natural hybridization and hybrid zones. Syst. Zool. 22, 213-218.

Xu, S.Z., 2000. Phylogenetic analysis under reticulate evolution. Mol. Biol. Evol. 17, 897-907.

Zhou, R., Shi, S., Wu, C.-I., 2005. Molecular criteria for determining new hybrid species - An application to the *Sonneratia* hybrids. Mol. Phylogenet. Evol. 35, 595-601.

Zirkle, C., 1935. The beginnings of plant hybridization. University of Pennsylvania Press, Philadelphia, PA.

Zwickl, D.J., Hillis, D.M., 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst. Biol. 51, 588-598.

SUMMARY

In this thesis the phenomenon of reticulate evolution was explored, with the emphasis on consequences for phylogenetic reconstruction.

Patterns of reticulate evolution are a major topic in plant systematics and therefore Chapter 2 discusses on the evolutionary significance of reticulation within angiosperms, with the emphasis on hybridization and possible long-term consequences of hybrid species. First the history of hybrid-related studies is presented, together with an overview of terminology regarding hybridization, introgression and other related terms. Terminology can be highly confusing and it is recommended to include information about age or status of hybrid instead of using new complicated terminology. In this chapter also the evolutionary significance of hybridization is discussed, with an overview of possible processes that may be involved in hybrid speciation mechanisms and an evaluation about prevalence and frequency of hybrids in nature.

Chapter 3 describes the common practice of dealing with hybrids in angiosperm phylogenetic studies. Recent examples of studies including hybrids are given and possible consequences of putative hybrids for phylogenetic reconstruction are described. With just a few examples that include putative hybrids in their studies, no general conclusion can be given, except that inclusion might be disturbing in some cases. These studies showed again that hybrid terminology is often not clear, probably due to the many possible processes and factors that may be of influence during hybrid formation, illustrated in a conceptual framework. Also an overview of network reconstruction packages is presented and their usefulness as possible tools to represent reticulate patterns is discussed, although not used frequently at species-level.

The possible disruption of phylogenetic analysis is further explored in Chapter 4, which includes simulation studies to investigate the effect of hybrid terminals ("mosaic sequences") on resulting tree topologies. Effect on accuracy of tree topology recovery (using Bayesian inference and Jackknife resampling) was measured using the partition metric. From this study and earlier simulation studies it can be concluded that these mosaic sequences do not severely disturb phylogenetic reconstruction, unless in a few exceptional cases, such as parental taxa that are phylogenetically distant.

Chapter 5 evaluates the use of network reconstruction methods to visualize reticulate patterns. We tested how well the mosaic sequences (hybrid terminals) in the simulated data sets of Chapter 4 are visualized and we also used published DNA sequence data sets including hybrid terminals. Some methods presented the hybrid terminals in the "expected" position, connected to both parents. But no single method

did this for all example data sets. Most cases reveal that the network methods constructed too many "extra reticulations", probably due to the use of relatively variable data sets at species-level, while most methods were designed to deal with population-level data.

Besides an incongruent pattern, reticulation can also lead to additivity, as expected to be found in AFLP data that include hybrids. In Chapter 6 we explore the suitability of AFLPs for detecting hybrid terminals, illustrated by an example of *Solanum*. While AFLPs are potentially strong markers to infer hybrids, most example studies include just several species and not many studies have been done that include a range of species. We conclude that in our example including 16 *Solanum* species, AFLPs do not reveal an incongruent or conflicting pattern. All separate analyses (Jackknife resampling, Bayesian and several network methods) do strongly support the same evolutionary relationships and do not indicate any underlying reticulate patterns.

Finally, in Chapter 7 we discuss whether it is possible to take the analysis of reticulate patterns one step further and be able to infer and represent organism-level evolutionary relationships. We consider several approaches to reconcile incongruent gene trees with as goal to reconstruct species trees. While this is currently not possible, promising approaches are on our way, including model-based approaches.

SAMENVATTING

Dit proefschrift beschrijft het fenomeen reticulate evolutie met de nadruk op de gevolgen voor fylogenie-reconstructie.

De term reticulatie is afgeleid van het Latijnse woord reticulum, een verkleinwood voor net. Het betekent dus letterlijk net-achtige evolutie, waarmee doorgaans evolutionaire processen worden beschreven waarbij verschillende afstammingslijnen genetisch materiaal uitwisselen, bijvoorbeeld bij het ontstaan van hybriden. In een evolutionaire stamboom worden dergelijke relaties weergeven in de vorm van evolutionaire lijnen (takken) die bij elkaar komen, zoals in een net of netwerk. Dit in tegenstelling tot een patroon van "vertakkende" evolutie, waarbij afstammingslijnen onafhankelijk van elkaar evolueren en de relaties in een stamboom worden weergegeven door splitsende takken.

Reticulate evolutie en de daaruit voortvloeiende patronen spelen een belangrijke rol binnen de plantensystematiek en daarom worden eerst in hoofdstuk 2 de evolutionaire gevolgen van reticulatie binnen de bloeiende planten behandeld. De nadruk ligt hierbij op hybridisatie en de mogelijke langetermijn gevolgen van hybride soorten. In het eerste deel van dit hoofdstuk wordt de geschiedenis gepresenteerd van studies gewijd aan hybriden, samen met een overzicht van terminologie betreffende hybridisatie, introgressie en andere verwante termen en processen. In het tweede deel wordt de evolutionaire significantie van hybridisatie besproken, en er wordt een overzicht gegeven van processen die mogelijk van invloed zijn op hybride soortvorming, alsmede een evaluatie van de invloed en frequentie van natuurlijke hybriden.

Hoofdstuk 3 beschrijft de algemene werkwijze waarop hybriden behandeld worden in fylogenetische analyses, binnen het onderzoeksgebied van de bloeiende planten. Er wordt een overzicht gepresenteerd van recente hybride-gerelateerde studies en daarnaast worden de consequenties van hybriden voor de fylogenie-reconstructie beschreven. Deze voorbeeldstudies laten opnieuw zien dat terminologie rondom hybriden vaak verwarrend is – mede veroorzaakt door de complexiteit aan mogelijke processen en factoren – en dit wordt geïllustreerd in een conceptueel kader. Daarnaast wordt een overzicht van netwerkreconstructie methoden gepresenteerd. Hoewel deze methoden (nog) niet vaak worden gebruikt op soortniveau, bespreken we hier de mogelijke bruikbaarheid van deze methoden voor het weergeven van reticulate patronen tussen soorten.

De mogelijke verstoring van hybriden op fylogenetische analyses wordt verder uitgewerkt in hoofdstuk 4, waar simulatiestudies zijn gebruikt om het effect te onderzoeken van hybride taxa (in de vorm van mozaïeksequenties) op de uiteindelijke geconstrueerde evolutionaire stambomen. Zowel Bayesiaanse als Jackknife methoden worden gebruikt om stambomen te maken waarbij de mogelijke verstoring van hybride taxa op uiteindelijke boomtopologie gemeten wordt met behulp van de "partition metric". Dit en eerder onderzoek laat zien dat de mozaïeksequenties geen drastisch effect hebben op de fylogenie-reconstructie, behalve in enkele uitzonderingsgevallen, zoals wanneer de oudersoorten van de hybride fylogenetisch ver van elkaar verwijderd zijn.

In hoofdstuk 5 is een selectie gemaakt van mogelijk geschikte netwerkreconstructie methoden om te testen in hoeverre deze methoden in staat zijn reticulate patronen te visualiseren. We hebben hiervoor de mozaïeksequenties uit de simulaties van hoofdstuk 4 gebruikt en getest waar deze in de geconstrueerde netwerken terecht komen. Daarnaast is een aantal bestaande (gepubliceerde) DNA-sequentie data sets getest, waarin vermoedelijke hybride taxa voorkomen. In sommige gevallen wordt het hybride taxon in de verwachte positie geplaatste, namelijk met verbindingen naar beide ouderlijke soorten. Maar geen enkele methode doet dit voor alle voorbeeld data sets. De meeste methoden plaatsen te veel extra verbindingen ("reticulaties") tussen de betrokken soorten, waarschijnlijk doordat de data sets sterke variatie tussen de soorten vertonen. De meeste methoden zijn namelijk oorspronkelijk ontworpen voor het representeren van relaties op populatie niveau, waar doorgaans minder variatie optreedt.

Naast een incongruent patroon kan reticulatie ook resulteren in een additief patroon, zoals bijvoorbeeld verwacht kan worden in hybride taxa bij het gebruik van AFLP-markers. In hoofdstuk 6 onderzoeken we de bruikbaarheid van AFLPs voor het detecteren van hybride taxa, met als voorbeeld een groep verwante soorten binnen het geslacht *Solanum*. AFLPs zijn potentieel zeer geschikte markers om hybriden mee te kunnen aantonen, maar de meeste voorbeeldstudies gebruiken slechts een klein aantal soorten en de bruikbaarheid voor grotere data sets met meerdere (minder verwante) soorten is moeilijker vast te stellen. In onze studie met 16 *Solanum* soorten laten alle afzonderlijke analyses (Bayesiaans, Jackknife en networkmethoden) dezelfde evolutionaire relaties zien en geven geen indicatie voor onderliggende reticulate patronen. Wij concluderen daarom dat AFLPs in dit geval geen duidelijk incongruent of conflicterend patroon blootleggen.

Tenslotte behandelen we in hoofdstuk 7 de mogelijkheid om de reticulate patronen tussen de op gen-gebaseerde stambomen te interpreteren in termen van

evolutionaire relaties op organismeniveau, in plaats van genniveau. We bespreken de verschillende manieren om incongruente stambomen van verschillende genen samen te voegen in een gezamenlijke representatie van de soortrelaties. Hoewel dit momenteel niet mogelijk is, zijn er veelbelovende methoden in ontwikkeling, vooral gebaseerd op het gebruik van modellen.

NAWOORD

In mijn eerste weken zei een collega-aio tegen me: "O, wat fijn voor je dat je nog maar net bent begonnen: dan heb je nog helemaal geen fouten gemaakt!" Ik heb hier later nog vaak aan teruggedacht: Waarom was ik er niet eerder achter gekomen dat methode X eigenlijk een "primitive trash approach" is (Hans Bandelt – thanks for helping me sorting out what directions *not* to take..!), in plaats van maanden achter crashende computers te zitten en er niet te vergeten een student mee heb opgezadeld (Benno – bedankt voor het door blijven worstelen ondanks ook je eigen scepsis!)
En had ik niet beter meteen iemand met gevoel voor planten de aardappels kunnen laten kruisen, in plaats van het tegen beter weten in eerst zelf te gaan proberen? (Theo – dank voor het maken van onze enige levensvatbare hybride!)
Achteraf lijkt het vaak alsof alles wel in 4 maanden had gekund in plaats van 4 jaar! Maar ik geloof dat het uiteindelijk om het proces gaat en niet (alleen) om het resultaat. En juist momenten waarop je denkt "Waarom wilde ik dit ook alweer? En waarom heb ik geen normale baan? " leveren vaak de beste herinneringen op. (In een zware sneeuwstorm een veel te oude computer vervoeren om er onze eerste linux-versie op te zetten (wurlug-groep – bedankt!) of tot 3 uur 's nachts in de kroeg nog een titel voor een congrespraatje in elkaar knutselen (om er vervolgens achter te komen dat het 5-regelige epistel toch echt niet werd geaccepteerd als titel).

En uiteraard heb ik alles zeker niet alleen gedaan. Dank voor allen die op wat voor manier dan ook hebben bijgedragen aan het onderzoek, ideeën hebben geleverd of gewoon de dagen vrolijker hebben gemaakt –sorry dat ik niet iedereen kan noemen hier!
Een aantal mensen ben ik zeer erkentelijk voor de hulp met mijn geworstel met Linux, scripts, computerclusters of alles tegelijk, enkele daarvan: Mark Fiers, Jack Leunissen en Niels de Keijzer. Daarnaast dank aan Michiel Kunst voor de stimulans om überhaupt met het clustergebeuren te beginnen!
Natuurlijk mijn begeleiders! Ronald voor de vaak relativerende en luchtige kijk op het geheel en Marc voor alle hulp en het uitstralen van rust op stressmomenten.
Freek heeft meer tijd aan dit project besteed dan enig ander en ik denk dat we ondanks ietwat botsende karakters een erg goede samenwerking hebben gehad. Dank voor je geduld, het me door de laatste weken heen slepen (het perfecte moment om toch te besluiten alles uit het raam te gooien), maar vooral voor je enthousiasme ("het is eigenlijk heel leuk!"): hartelijk dank voor alles!

Als computer-vastgekluisterde aio is het erg fijn om af ent toe wat ideeën te kunnen uitwisselen en mensen te zien. Gelukkig waren er de Taskforce, Tracks of Evolution en andere formele of informele bijeenkomsten om anderen te ontmoeten. Daarnaast was het erg fijn dat we door beperkte kantinefaciliteiten genoodzaakt waren dagelijks ook de echte planten te bekijken en samen te lunchen. De samenstelling was vaak wisselend, maar in ieder geval Thomas, Lars, Freek, Miguel, Timo, Jurriaan, Robin en Marleen en alle anderen: bedankt voor alle gezelligheid!
Ook het "gewone" vakgroepsleven met koffie, werkbesprekingen, af en toe uitjes, etc. was erg prettig. Ondanks dat ik me realiseer dat ik zelf niet altijd de meest sociale collega was heb ik de dagelijkse contacten met iedereen altijd erg gewaardeerd. Marleen en Ties wil ik hier nog speciaal noemen voor het immer tonen van extra belangstelling, zelfs en vooral in mijn laatste stressmaanden: veel dank!

Gelukkig is er ook nog een leven naast de universiteit. Marije & Wouter: dank voor alle relativering en de mooie illusie dat het onvolwassen leven kan blijven voortbestaan! Miranda, Servan, Willem, Joris, Annette, Deby en anderen: bedankt voor de afleiding en het proberen te begrijpen waar ik mee bezig was.
Birgit, ik ben erg blij met onze vriendschap! Bedankt dat je me de afgelopen jaren altijd het gevoel hebt gegeven er voor me te zijn. Je bent een van de weinige goede redenen om *wel* in Wageningen te blijven!
And of course – my new life and love in New York; Jack. You are my complete opposite, but thank you for believing in us!

Tot slot; waar zou je zijn zonder familie? Willem en Frans: het gemis aan regelmatig contact wordt ruimschoots gecompenseerd door de vertrouwdheid en relaxedheid waarmee jullie bijvoorbeeld tot diep in de nacht uren kunnen blijven doorpraten, los van hoe druk, hoe moe of hoelang geleden; geweldig om zulke broers te hebben!
En uiteraard boven alles mijn ouders voor het geduld en ondersteuning in wat voor carrière- of huizenswitch ik me elke keer weer op meende te moeten storten. Bedankt voor jullie onvoorwaardelijke en onbegrensde steun!

CURRICULUM VITAE

Bastienne Vriesendorp werd op 27 december 1974 geboren in Ottoland (Zuid-Holland) and groeide op in Ottoland en Bilthoven. Na het doorlopen van het Christelijk Gymnasium in Utrecht begon zij in 1993 met de studie biologie aan de Landbouwuniversiteit Wageningen. Haar eerste afstudeervak Diertaxonomie had als onderwerp akoestische signalen bij spoorcicaden en voor een tweede afstudeervak Dierecologie (gericht op vegetatiekartering en olifantenschade) verbleef ze 6 maanden in Maputo in Mozambique. Als stage heeft ze bij Ouwehands Dierenpark in Rhenen gewerkt binnen de afdeling educatie. Ze ontving haar doctoraaldiploma in maart 1999.

Na haar studie heeft ze een aantal jaren een mix van activiteiten gedaan buiten de academische wereld. Na ruim een halfjaar de postdoctorale lerarenopleiding aan de Vrije Universiteit in Amsterdam te hebben gevolgd, heeft ze dit onderbroken en is vervolgens naar Sherkin Island Marine Station aan de zuidwest kust van Ierland gegaan, om fytoplankton te monitoren en de schoonheid ende troost van de algenwereld te ontdekken. Na terugkomst is ze gaan werken bij Noldus Information Technology in Wageningen, ontwikkelaar van software voor gedragsonderzoek. Hier was zij sales engineer zoölogie en biologie, maar concludeerde uiteindelijk dat ze toch graag weer de wetenschap in wilde.

Zij is in juli 2003 begonnen als AIO bij de vakgroep Biosystematiek aan de Wageningen Universiteit binnen het Nationaal Herbarium Nederland om te werken aan reticulate evolutie. De resultaten van dit onderzoek zijn in dit proefschrift beschreven.

Training and education within the Graduate School Biodiversity

Name PhD student: Bastienne Vriesendorp
Institute: National Herbarium of the Netherlands – Wageningen branch, Biosystematics Group, Wageningen University

	Credit hours
1. PhD Courses	
Molecular phylogenies (Nov 2003)	32
Scientific writing course (Oct 2005)	60
Advanced topics in phylogenetic reconstruction (Feb 2005)	100
Supervision and organisation MSc projects (Feb 2005)	16
2. Annual PhD meetings	108
3. Essay and seminar on the background and framework of the project	152
4. Literature study resulting in written report	76
5. Presentation of results at international conferences	
Symposium on Phylogenetic Combinatorics, Uppsala, 2004	32
Work visit and presentation, Zurich, Dec 2004	16
International Botanical Congress, Vienna, 2005	38
Young Systematists' Forum, London 2005	8
Evolution meeting Stony Brook University, New York, 2006	32
6. Activities within the NHN Task Force Molecular Systematics Phylogenetics and Biogeography	76
7. Facultative elements	
Activities within the Scientific Discussion Group of the Biosystematics Group, Wageningen University	100
Training in various laboratory techniques	38
Work visit Prof. Bandelt, Hamburg, Mar 2005	24
Organizing PhD Day 2005	8
Total credit hours	916

Printed by: Ponsen & Looijen, Wageningen, the Netherlands

Cover design: Heleen Vriesendorp (illustration) & Frans Vriesendorp (design & lay-out)