Université de Montréal

Utilisation de réseaux en analyse phylogénétique : détection de taxons hybrides et combinaison d'arbres

par

Olivier Gauthier

Département de sciences biologiques

Faculté des arts et sciences

Thèse présentée à la Faculté des études supérieures

en vue de l'obtention du grade de Philosophiae doctor (Ph. D.)

en sciences biologiques

Avril 2006

Université de Montréal
Faculté des études supérieures

Cette thèse intitulée :

Utilisation de réseaux en analyse phylogénétique :
détection de taxons hybrides et combinaison d'arbres

présentée par :
Olivier Gauthier

a été évaluée par un jury composé des personnes suivantes :

M. Luc Brouillet, président rapporteur
M. François-Joseph Lapointe, directeur de recherche
M. Denis Barabé, membre du jury
M. David Bryant, examinateur externe
Mme. Nadia El-Mabrouk, représentante du doyen de la FES

Résumé

Une des préoccupations actuelles en analyse phylogénétique est notre capacité à détecter des événements d'évolution réticulée qui ne répondent pas au patron strictement divergent représenté par les arbres phylogénétiques. En dépit des algorithmes permettant l'estimation de réseaux plutôt que d'arbres, il n'existe pas encore d'évaluation formelle de la performance des différentes méthodes pour la détection de réels événements de réticulation. Cette thèse propose une évaluation comparée de deux algorithmes de reconstruction de réseau phylogénétique en présence d'hybrides connus et l'élaboration d'un nouveau test statistique de détection des hybrides basé sur l'analyse de quadruplets de taxons. La performance de ce test est évaluée par son application à des données réelles portant sur des hybrides connus, de même qu'à l'aide de simulations. La méthode proposée permet de détecter adéquatement des hybrides du genre *Aphelandra* (Acanthaceae) décrits à l'aide de caractères morphologiques. De même, l'hybridation ADN-ADN permet l'identification d'hybrides artificiels de kangourou du genre *Petrogale* (Marsupialia : Macropodidae). Par contre, des hybrides naturels entre ces mêmes espèces ne sont pas détectés par la méthode proposée. Des simulations sont présentées afin d'évaluer la performance statistique du critère de détection des hybrides. Ces résultats de simulations illustrent clairement que le test n'est pas sujet à une erreur de type I gonflée et tendent même à montrer que le test serait quelque peu conservateur.

Une autre préoccupation importante en analyse phylogénétique consiste en la combinaison d'hypothèses phylogénétiques multiples. Ces arbres multiples peuvent résulter de l'analyse séparée de plusieurs jeux de données, de l'existence de solutions également optimales pour un même jeu de données, ou encore de l'utilisation de méthodes d'analyse distinctes. Cette combinaison est effectuée à l'aide de méthodes de consensus. Il est démontré ici que des réseaux consensus permettent de conserver une plus grande partie de l'information phylogénétique contenue dans ces arbres multiples. Un cadre général pour construire et mesurer le contenu en information des réseaux

consensus est présenté.

Mots-clés : ADN, Consensus, Distance, Évolution, Information, Morphologie, Représentation par matrice, Réticulation, Simulation, Transfert latéral.

Abstract

A current preoccupation in phylogenetic analysis is our ability to detect events of reticulate evolution that do not meet the strictly divergent model of phylogenetic trees. In spite of the algorithms already available to infer networks rather than trees, no formal evaluation of the performance of these methods to detect hybrid taxa have been conducted yet. This thesis proposes a comparative analysis of two methods for inferring phylogenetic networks in the presence of known hybrids, as well as a new hybrid detection criterion based on the analysis of quartets of taxa. The performance of this test is assessed by its application to real data on known hybrids, and with computer simulations. The method is shown to adequately detect hybrids of the genus *Aphelandra* (Acanthaceae) scored for morphological characters. Furthermore, the test worked equally well with DNA-DNA hybridization data to detect artificial rock wallabies, *Petrogale* (Marsupialia: Macropodidae), hybrids. However, natural hybrids between members of the same genus eluded detection. Computer simulations are presented to illustrate the statistical performance of the method in a variety of conditions. The simulation results clearly show that the test does not exhibit inflated type I error, and that it is somewhat conservative.

Another important preoccupation in phylogenetic analysis is the combination of multiple phylogenetic hypotheses. These multiple trees can result from a separate analysis of many datasets, the existence of equally optimal solutions for a unique dataset, or the use of different analytical techniques. Consensus methods are used to combine these hypotheses. Here, it is shown that consensus networks can preserve a greater amount of the phylogenetic information embedded in these multiple trees. A general framework to compute consensus networks, as well as a way to measure their information content are presented.

Keywords: Consensus, Distances, Hybridization, Information, Matrix representation, Networks, Phylogenetic analysis, Reticulation, Simulations, Trees.

Table des matières

CHAPITRE 1 : INTRODUCTION 1
1.1. Introduction 2
1.1.1. Contexte 2
1.1.2. Concepts et définitions 5
1.1.2.1. Phylogénies 5
1.1.2.2. Arbres et réseaux : Définitions 5
1.1.2.3. Inférence phylogénétique 11
1.2. Le problème des réticulations en analyse phylogénétique 13
1.2.1. Contexte 13
1.2.2. Méthodes à l'étude 18
1.2.2.1. Décomposition des bipartitions 18
1.2.2.2. Réticulogrammes 19
1.3. Le problème du consensus en analyse phylogénétique 21
1.3.1. Contexte 21
1.3.2. Méthodes à l'étude 22
1.3.2.1. Les arbres consensus 22
1.3.2.2. Les réseaux consensus 24
1.4. Organisation de la thèse 27
Chapitre 2 : A COMPARISON OF ALTERNATIVE METHODS FOR DETECTING RETICULATION EVENTS IN PHYLOGENETIC ANALYSIS 28
2.1. Résumé 29
2.2. Abstract 30
2.3. Introduction 31
2.4. Distance based methods of reticulate analysis 31
2.5. Hybrid detection analysis 32
2.6. Hybrid detection through quartet analysis 33
2.7. Discussion 38
2.8. Acknowledgements 38
CHAPITRE 3 : HYBRIDS AND PHYLOGENETICS REVISITED. A STATISTICAL TEST OF HYBRIDIZATION USING QUARTETS 39
3.1. Résumé 40

3.2. Abstract 41
3.3. Introduction 42
3.4. Description of the test 44
3.4.1. Hypotheses 44
3.4.2. Statistical procedure 48
3.5. Application 49
3.5.1. Statistical analysis 49
3.5.2. Results 49
3.5.3. Discussion 50
3.6. Acknowledgements 55
Chapitre 4 : PHYLOGENY OF THE ROCK WALLABIES, *PETROGALE* (MARSUPIALIA: MACROPODIDAE), PART II: DETECTION OF HYBRIDISATION AMONG MACROPODINES 56
4.1. Résumé 57
4.2. Abstract 58
4.3. Introduction 59
4.4. Methods 62
4.5. Results 70
4.6. Discussion 72
4.7. Acknowledgements 77
Chapitre 5 : POWER AND ACCURACY OF THE HDC TEST 78
5.1. Résumé 79
5.2. Abstract 79
5.3. Introduction 80
5.4. Simulation procedure 82
5.5. Results and discussion 85
5.6. Acknowledgements 88
Chapitre 6 : GETTING MORE FROM YOUR TREES WITH CONSENSUS NETWORKS 89
6.1. Résumé 90
6.2. Abstract 90
6.3. Introduction 91
6.4. Computing a consensus network 93

6.5. The information content of consensus network 97
6.6. Application 102
6.7. Discussion 102
6.8. Acknowledgements 108
Chapitre 7 : CONCLUSION 109
BIBLIOGRAPHIE 115

Liste des tableaux

Table 2.1 The 17 hybrids used in the analysis and their Hybrid Detection Criterion (HDC) values. Values are given for each hybrid and each method. The maximum possible HDC value for this dataset is 10. High HDC values indicate possible hybridization.....34

Table 3.1 Values of the Hybrid Detection Criteria statistic value (HDC) and their associated probabilities (p) for quartet analyses using reticulograms and splitsgraphs ($*: p \leq 0.05$; $**: p \leq 0.01$). Hybrids where created by crosses between *Aphelandra campanensis* Durkee, *A. deppeana* Schltr. & Cham, *A. golfodulcensis* McDade, *A. gracilis* Leonard, *A. leonardii* McDade, *A. panamensis* McDade, *A. sinclairiana* Nees, *A. storkii* Leonard, and *A. terryae* Standley by McDade (1990). 51

Table 4.1 Registry numbers of parental *Petrogale* species and their natural hybrids. Artificial hybrids utilised extracts from specimens listed in our earlier paper. 63

Table 4.2 ΔT_ms among 19 *Petrogale* taxa, including two natural and two artificial hybrids; number of comparisons = 1051; average standard deviation (SD) = ±0.17; correlation of SDs with distance = 0.18. Columns are tracers, identified for the most part by the first four letters of the specific epithet. Names of hybrids and parental taxa are shown in ***bold italics***. First line of each cell lists average ΔT_m except for the homologues (**boldfaced**), where actual mean melting-temperature is given to permit comparison of tracer qualities; second line gives SD where applicable and number of replicates, separated by a solidus; na = not applicable. 64

Table 4.3 Four-taxon matrices (a-e) of ΔT_ms among parental *Petrogale* species and their natural and artificial hybrids. Conventions mostly as for Table 4.2, but in (e) four cells were reflected from their reciprocals (*italicised*) and one pair of reciprocals was estimated (underlined); homologous distances of the two hybrids are by definition zero. Negative values were set equal to zero. 67

Table 4.4 Scaled HDC s and their probabilities for the four hybrids included in Table 4.2 for both the reduced and full matrices and different threshold values (t=0.00, t=0.10, t=0.25, t=0.50, and t=0.75). Reduced matrices include all full species and only the hybrid under consideration (n=16) whereas the full matrix includes all species and all hybrids (n=19).....74

Table 5.1 Estimation of the power of the HDC test. Rejection of the null hypothesis for different number of taxa (n) and varying support for both partitions when sequences evolved along conflicting trees are combined to simulate reticulate evolution. Results are presented as percentage of rejection over 10 000 replicates for each combination of parameters ($\alpha = 0.05$). 86

Table 5.2 Estimation of the type I error of the HDC test. Rejection of the null hypothesis for different number of taxa (n) and varying support for both partitions when sequences evolved along topologically identical trees are combined to simulate evolution on the branches of a tree. Results are presented as percentage of rejection over 10 000 replicates for each combination of parameters ($\alpha = 0.05$). 87

Liste des figures

Figure 1.1 Le premier arbre phylogénétique (Darwin 1859). Darwin a utilisé l'unique figure de l'Origine des espèces afin d'illustrer l'apparition de nouvelles espèces par un processus de différentiation graduelle des populations. 3

Figure 1.2 Le *Monophyletischer Stammbaum der Organismen* ou arbre généalogique monophylétique des organismes de Haeckel (1866). La première d'une série de représentations stylisées de l'évolution des groupes d'organismes vivants réalisées par le biologiste allemand.....4

Figure 1.3 Deux visions réticulées de l'évolution. (a) Doolittle (1999) propose une vision selon laquelle les événements de réticulations jouent un rôle majeur dans la diversification et l'évolution des taxons. (b) Pour Rivera & Lake (2004), des fusions de génomes ont mené à l'apparition de grands groupes comme les eucaryotes, ce qui les mène à proposer le concept d'Anneau de la Vie (« *Ring of Life* ») pour expliquer la phylogénie des grands groupes. 6

Figure 1.4 Cet arbre (a) et ce réseau (b) sont des graphes phylogénétiques non enracinés. A, B, C, D, E, F, G et H sont des nœuds qui sont reliés par des branches identifiées par la paire de nœuds qu'elles joignent. A, B, C et D sont des nœuds terminaux ou feuilles qui correspondent aux taxons étudiés, les autres nœuds sont internes. Les traits pointillés illustrent un chemin reliant les feuilles B et D passant par les branches BE, EF et FD en (a) et un cycle, ou une réticulation, passant par les branches EF, FG, GH et HE en (b). Le cycle en (b) fait en sorte qu'il existe deux chemins différents entre les paires de nœuds terminaux dans le réseau, par exemple les chemins BF, FG, GH, HD et BF, FE, EH, HD relient les nœuds B et D. 8

Figure 1.5 Les 3 arbres binaires non enracinés présentés en (a), (b) et (c) sont les trois manières possibles de raffiner l'arbre (d) qui contient une polytomie. L'arbre (e) est un arbre étoile qui ne contient aucune information quant aux relations phylogénétiques entre les taxons. 9

Figure 1.6 Deux arbres non enracinés et leur matrice patristique. Les distances patristiques de l'arbre (a) sont additives, alors que celles de l'arbre (b) sont ultramétriques. La flèche en (b) montre la position de la racine de l'arbre ultramétrique. 12

Figure 1.7 Graphe de bipartitions (b), arbre (c), et réticulogramme (d) obtenus à partir de la même matrice de distances en (a). Le graphe de bipartitions (b) présente les bipartitions $AB|CD$ et $AC|BD$ étant donné que $d_{AB}+d_{CD}=6$, $d_{AC}+d_{BD}=8$ et $d_{AD}+d_{BC}=10$. De plus $S_{AB|CD}=2$ et $S_{AC|BD}=1$. Le graphe de bipartition représente parfaitement les distances dans la matrice initiale (a). Pour obtenir le réticulogramme (d), il faut tout d'abord obtenir un arbre (c) et sa matrice patristique (e), cette dernière n'est pas parfaitement ajustée à la matrice initiale (a). L'ajustement est amélioré en ajoutant une réticulation entre les nœuds A et C, telle qu'illustré par la matrice du réticulogramme (f). Les flèches unidirectionnelles représentent des unijections, les bidirectionnelles des bijections. 20

Figure 1.8 Trois arbres non-enracinés (a), (b), et (c), ainsi que leur consensus strict (d), leur consensus majoritaire (e), et leur consensus moyen (f). 23

Figure 1.9 Les trois bipartitions possibles avec quatre taxons (a) ne peuvent êtres représentées conjointement dans un plan à deux dimensions à moins d'avoir recours à l'arbre étoile (b). Il est possible de représenter deux à deux ces bipartitions à l'aide d'un graphe de bipartitions (c). Pour représenter les trois bipartitions dans un seul et même graphe, il faut utiliser une troisième dimension (d). 26

Figure 2.1 Illustration of the Hybrid Detection Criterion (HDC) for reticulogram quartets. (a) HDC is met if the hybrid (AB) is the sister taxa of either one of its parents (A or B) and has a reticulation to its other parent. HDC is not met if (b) the hybrid is the sister taxa of one of its parents, but does not have a reticulation to the other, or, (c) the hybrid is

the sister taxa of the other species (X). The positions of parents A and B are interchangeable. 36

Figure 2.2 Illustration of the Hybrid Detection Criterion (HDC) for splitsgraph quartets. (a) HDC is met if the hybrid (AB) forms a pair of weakly compatible splits with its parents (A and B). (b) HDC is not met if the hybrid forms a weakly compatible split with the other species (X). The positions of parents A and B are interchangeable. 37

Figure 3.1 Network topologies for which the Hybrid Detection Criterion (HDC) is met with (a) a reticulogram and, (b) a splitsgraph. (c) Reticulograms that contradict HDC because the putative hybrid (AB) is not a sister taxa to either of the parents or because the reticulation not added between the hybrid and one of its parents. (d) Split decomposition graph that contradicts HDC, the putative hybrid does not form a set of weakly compatible splits with both its parents. Parents A and B are interchangeable. The dashed lines in reticulograms are reticulations.....46

Figure 3.2 (a) A topology that offers no clear support for or against the Hybrid Detection Criterion (HDC), this case is arbitrarily given a weight of 0.5 in the calculations. (b) A topology for which the putative hybrid (AB) is not grouped with either of its parent; this contradicts HDC. (c) A star tree that provides no information on the relationships between the taxa, it neither supports nor contradicts HDC; such cases are ignored in the computation of the statistic. Parents A and B are interchangeable.....47

Figure 4.1 Distribution of *Petrogale* taxa in Australia. Modified from Eldridge & Close (1993).....60

Figure 4.2 Example of a splitsgraphs and trees and their interpretation under the HDC. (a) A splitsgraphs that groups the hybrid (H), its two parents (P1 and P2), and the outgroup (X) in a way that meets HDC (b) another that does not. (c) A tree for which the putative hybrid is not grouped with either of its parent; this also contradicts the criterion. (d) A tree that offers no clear support for or against the HDC, this case is arbitrarily given a weight of 0.5 in the calculations. (e) A star tree that provides no information on the relationships between the taxa, it neither

supports nor contradicts *HDC*; such cases are ignored in the computation of the statistic. 69

Figure 4.3 FITCH tree calculated from the data of Table 4.2, using the G, S, and Cavalli-Sforza & Edwards (P = 0) options; and randomising the input-order of taxa 50 times. All negative ΔT_ms were set equal to zero and 15 values in the last column (also negative and therefore set to zero) were reflected from their reciprocals, after taking into account row:column ratios according to the procedure of Springer & Kirsch (1991). Parental and hybrid taxa highlighted in boldface. 71

Figure 4.4 Splitsgraphs (a-e) calculated from the 4x4 matrices of Table 4.3a-e using SplitsTree4.0. The split highlighted with heavy branches in case corresponds to the tree that would be obtained by applying the FITCH algorithm to the same data. Notice that the artificial and natural hybrids are never paired together in the splitsgraphs, except for one case (e). Weights are given for internal branches.....73

Figure 5.1 Illustration of the Hybrid Detection Criterion (*HDC*). When considering splitsgraphs defined on quartets of taxa, a hybrid (H) occupies an intermediate phylogenetic position between its parents (A and B) with respect to any other taxa in the dataset (X). (a) Equal support for both underlying trees (1:1 ratio); and differential support for the underlying trees with (b) 1:3 ratio and (c) 1:7 ratio of the branch lengths supporting the weakly compatible splits. 81

Figure 5.2 Outline of the simulation procedure. (a) generate a random ultrametric tree; the position of the hybrid (taxa 16) is illustrated with dotted branches for clarity, it is not added at this point however; (b) duplicate tree and graft hybrid leave at the appropriate position in each tree; (c) go from ultrametric trees to additive trees; (d) evolve sequences along the trees; (e) compute distances; and, (f) submit to *HDC* test.....83

Figure 6.1 A matrix representation (MR) of a tree can take the form of a binary character matrix (left) or a distance matrix (right). Full lines indicate correspondence between trees and MR, as well as combination of MR from different trees in a unique MR. Dashed lines indicate equivalence

between two MR. Each split corresponds to a binary character for which the taxa on either side of the split are coded 1 and 0 respectively. Informative characters (partitioning the taxa in two groups of two) are numbered (1 to 3), while uninformative characters (partitioning the taxa in a group of three and a singleton) are only shown for the MR of single trees. Binary MR for two (or more) trees are obtained by combining the character matrices including the uninformative characters. Distance MR are obtained by summing the lengths of branches between pairs of taxa (here all equal 1), average distance MR for two (or more) trees contain the average path length distances between pairs of taxa.....95

Figure 6.2 (a) The three possible unrooted trees with four taxa; (b) the strict and majority rule consensus tree of any combination of the trees in (a); (c) the CN of any pair of the trees in (a); and, (d) the unconstrained median CN of all three trees in (a). 98

Figure 6.3 (a) the network with splits $AB|CDE$, $AE|BCD$, and $BC|ADE$ contains two maximally resolved trees (b and c). Tree (b) has three resolutions, while tree (c) is binary. The network thus allows four distinct fully resolved trees and as a $CIC_{rel} = 0.49$. 101

Figure 6.4 Phylogenetic trees for 11 mammalian species obtained from eight independent mitochondrial and nuclear genes (data from Springer *et al.* 1999). 103

Figure 6.5 Different consensus representations of the profile of trees in Fig. 6.4. (a) The majority rule consensus ($CIC_{rel} = 0.61$); (b) the median CN restricted to splits contained in at least two trees ($CIC_{rel} = 0.66$); (c) the split decomposition CN ($CIC_{rel} = 0.74$); (d) the median CN restricted to splits contained in at least three trees ($CIC_{rel} = 0.80$).....104

Liste des abréviations et symboles

HDC	Hybrid Detection Criterion
CDH	Critère de Détection des Hybrides
$\Delta T_m s$	Difference between median melting-temperature of hybridized sequences of homologous and heterologous hybrids
SD	Standard Deviation
t	Threshold
d_{ij}	Distance entre *i* et *j*
ij|kl	Bipartition partageant *i* et *j*, d'une part, et, *k* et *l*, d'autre part
$S_{ij\|kl}$	Indice d'isolation de la bipartition *ij|kl*
MR	Matrix Representation
MRD	Matrix Representation with Distances
MRP	Matrix Representation with Parsimony
CN	Consensus Network
CIC	Cladistic Information Content
CIC_{rel}	Relative Cladistic Information Content
ADN	Acide désoxyribonucléique
DNA	Deoxyribonucleic Acid
F_1	Première génération
K2P	Modèle de Kimura à deux paramètres

À Gaston

"Molecular phylogeneticists will have failed to find the "true tree", not because their methods are inadequate or because they have chosen the wrong genes, but because the history of life cannot properly be represented as a tree"

W. Ford Doolittle (1999)

Remerciements

Je voudrais tout d'abord remercier chaleureusement mon directeur de recherche. Je n'oublierais pas notre première rencontre au F-180, sombre local sans fenêtre, où tu avais répondu à mes questions de débutant en biostatistique au son de Motorhead. Dois-je te rappeler que ce jour-là j'étais convaincu que tu étais un étudiant. Il ne faut jamais se fier aux apparences, et sous tes cheveux hirsutes et ton vieux kangourou se cachaient un chercheur et un communicateur hors du commun, un créateur en effervescence, un être passionné partant sans cesse à la conquête de nouveaux horizons. Merci de m'avoir accompagné tout au long de cette aventure, et de m'avoir aidé à passer à travers les étapes les plus difficiles.

Comment pourrais-je passer sous silence la contribution inestimable de tous les membres passés et présents du Laboratoire d'Écologie Moléculaire Et Évolution (LEMEE) à cette thèse? Le LEMEE est plus qu'un milieu de recherche et de production scientifique, c'est une seconde famille. Merci pour les discussions et les commentaires scientifiques, mais aussi (surtout?) pour les longues heures de dîner à rire, les desserts fait maison et les sorties au Bilboquet ou chez Kilo, les midi-rolls! Pierre-Alexandre et Claudine, vous avez hardiment débroussaillé la voie que j'ai empruntée, il y a maintenant 5 ans, et sur laquelle Sarah, Sébastien et Véronique cheminent présentement. Nathalie, grande prêtresse du LEMEE, même si j'étais du « côté des ordinateurs », tu as su guider mon usage de la force et m'éviter de tomber du côté obscur. Anaïs, Anne-Marie, Antoine, Catherine, Émilie, Jeanne, Sophie et Yong, vous avez tous, d'une manière ou d'une autre, contribué à faire de notre laboratoire un endroit agréable et dynamique. Merci aussi à nos visiteurs occasionnels, Louise, Philippe et Simon (notre héros à tous), membres par adoption, dont les visites ont toujours été agréables, et le plus souvent accompagnées d'éclats de rires.

Merci à mes parents, Pierre et Viviane, et à ma famille, Pascale, Benoît et Dimitri, qui, malgré mon parcours sinueux et mes nombreux détours, ont toujours cru en mes capacités de mener à terme mes projets. Thank you Anna.

Merci tout particulièrement à Georges de m'avoir offert le support nécessaire pour retourner aux études après quelques années de décrochage afin d'obtenir un diplôme de premier cycle. Ce retour s'est avéré un tremplin qui a marqué le début d'une aventure tirant maintenant à sa fin.

Merci à mes amis, qui ont su m'encourager et croire en moi. Ils occupent une place importante dans ma vie même si je ne prends pas toujours le temps de leur faire savoir à quel point je les apprécie. Alexandre, Naïty, Maginnis, Jean-Philippe, Marième, Mathieu, Cynthia, Eve, Yohan, Johanna, Richard, Nicolas – vous êtes trop nombreux, je dois m'arrêter! - comme la vie serait monotone sans nos discussions qui ne mènent souvent nulle part mais qui sont toujours un parcours agréable à suivre et un bouillonnement d'idées incroyables.

Merci à tous les professeurs qui ont participé, par leurs commentaires ou leur participation à mes comités-conseil et jury, à la réalisation de ce projet. Merci donc à Bernard Angers, Denis Barabé, Luc Brouillet, Anne Bruneau, David Bryant et Pierre Legendre. Vos commentaires et encouragement ont été grandement appréciés.

Finalement, je voudrais remercier tous les organismes qui m'ont fourni les fonds nécessaires à la réalisation de mes études de cycles supérieurs. Merci donc au Conseil de Recherche en Sciences Naturelles et en Génie (CRSNG) du Canada, au Fonds Québécois de la Recherche sur la Nature et les Technologies (FQRNT) pour m'avoir octroyé des bourses de recherches, ainsi qu'au Fond de Bourses en Sciences Biologiques (FBSB) du Département de sciences biologiques de l'Université de Montréal pour m'avoir octroyé une bourse de rédaction.

Chapitre 1 :

INTRODUCTION

1.1. Introduction

1.1.1. Contexte

L'idée d'un Arbre du Vivant (« Tree of Life ») a été mise de l'avant par Darwin (1859) dans l'unique figure de l'Origine des espèces (Fig. 1.1) et développée par la suite par Haeckel (1866) dans ses célèbres représentations stylisées des relations d'ascendance et de descendance entre les groupes d'organismes vivants ou taxons (Fig. 1.2). Cette vision de relations hiérarchiques entre les taxons a donc imprégné les bases de l'analyse phylogénétique : dès sa naissance, l'objectif de la discipline a été de reconstruire cet arbre. En plus de correspondre à cette conception de l'évolution, les arbres sont des objets mathématiques qui possèdent plusieurs propriétés rendant leur construction et leur manipulation beaucoup plus aisée que celles d'autres graphes comme par exemple les réseaux (Bondy & Murty 1976; Felsenstein 2004). C'est pour ces deux raisons qu'ils occupent une si grande place en analyse phylogénétique. Avec l'avènement de la biologie moléculaire, la construction du véritable Arbre du Vivant devenait envisageable (Wolf *et al.* 2002). Le développement des nouvelles techniques moléculaires permettait en effet l'accès à de grandes quantités de données souvent jugées plus fiables que les caractères morphologiques (Hillis & Wiens 2000; Wolf *et al.* 2002). Les phylogénies basées sur l'ARN ribosomal (ARNr) ont d'ailleurs mené à la reconnaissance du domaine des Archaebactéries et à l'inférence d'une phylogénie des grands groupe qui a été désignée « modèle standard de l'évolution » (Wolf *et al.* 2002). Pourtant les contradictions fréquentes entre les phylogénies moléculaires et morphologiques de même qu'entre les phylogénies représentant plusieurs gènes rendent le projet difficile (Doolittle 1999; Daubin *et al.* 2002). En considérant l'importance de phénomènes biologiques qui violent la conception arborescente de l'évolution, tels que l'hybridation, le transfert latéral et l'introgression (Bullini 1994; Arnold 1997; Dowling & Secor 1997; Rieseberg 1997; Wolf *et al.* 2002), certains auteurs doutent même aujourd'hui de l'existence d'un arbre unique, même dans sa version la plus

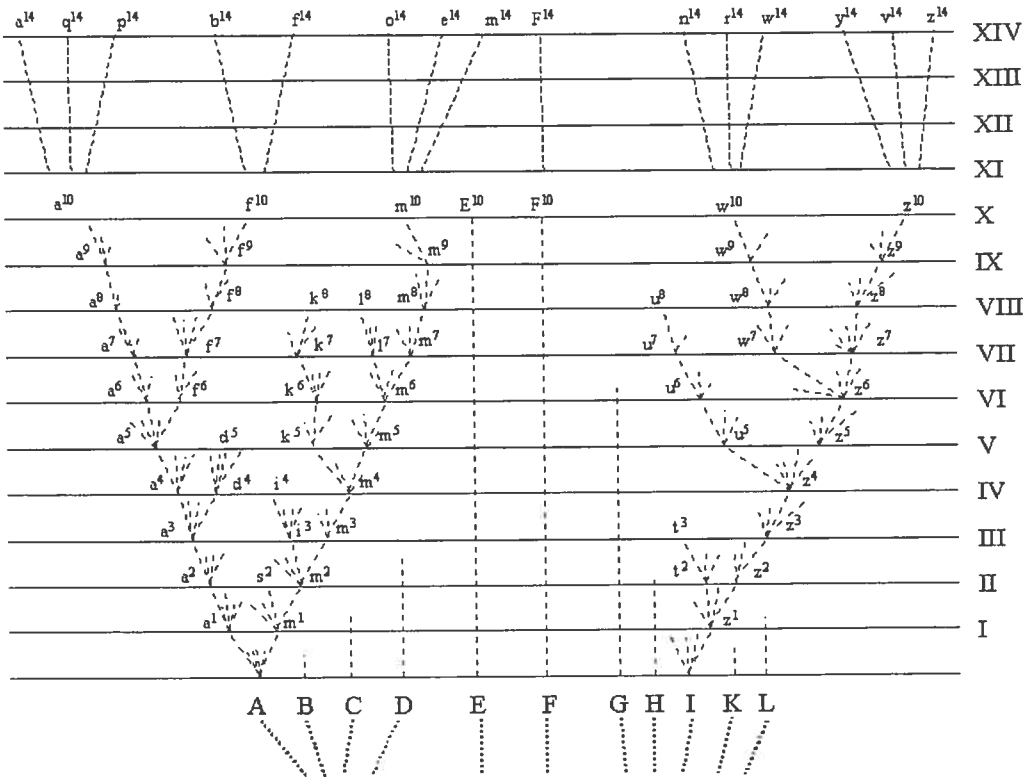


Figure 1.1 Le premier arbre phylogénétique (Darwin 1859). Darwin a utilisé l'unique figure de l'Origine des espèces afin d'illustrer l'apparition de nouvelles espèces par un processus de différentiation graduelle des populations.

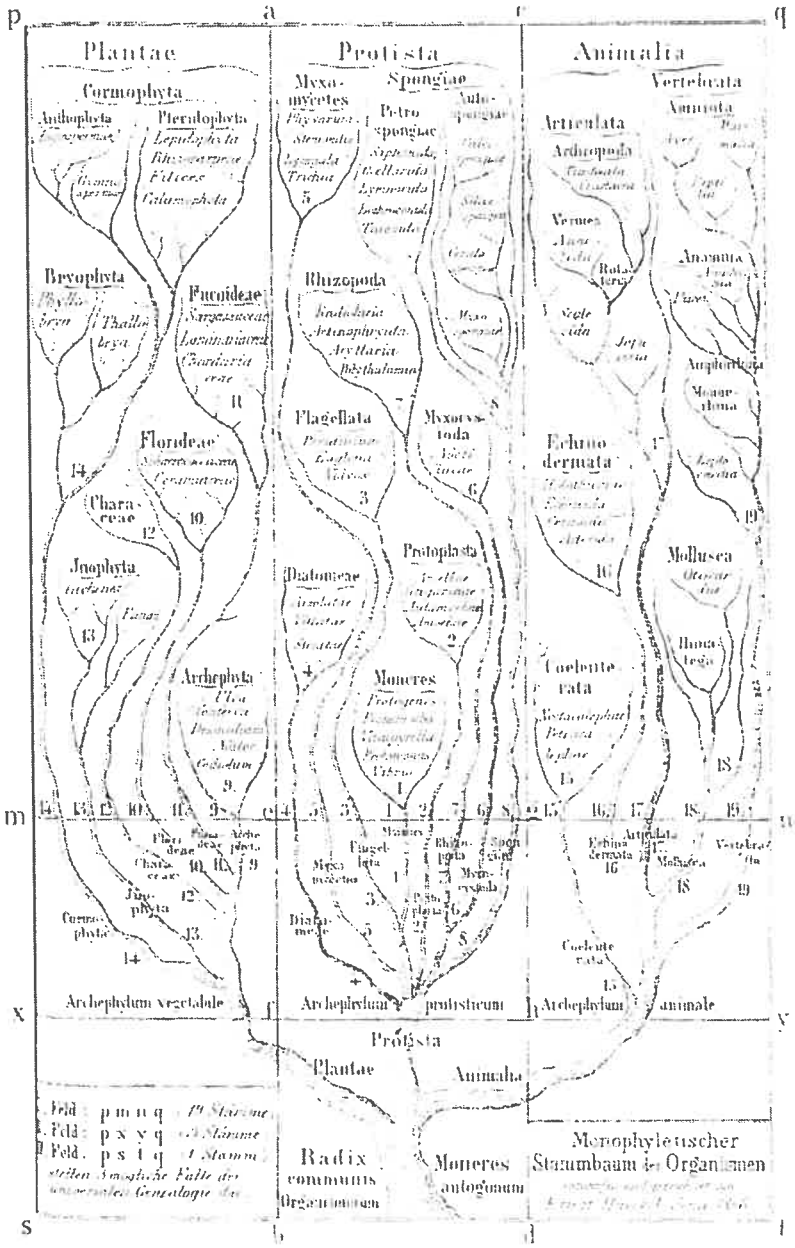


Figure 1.2 Le *Monophyletischer Stammbaum der Organismen* ou arbre généalogique monophylétique des organismes de Haeckel (1866). La première d'une série de représentations stylisées de l'évolution des groupes d'organismes vivants réalisées par le biologiste allemand.

simple où il représente l'évolution d'un noyau de gènes racontant la même histoire (Fig. 1.3; Doolittle 1999; Daubin *et al.* 2002; Rivera & Lake 2004). Si la notion d'Arbre du Vivant n'est plus compatible avec notre conception de l'évolution et qu'elle est constamment contredite par les résultats obtenus à ce jour, il est nécessaire de développer de nouvelles méthodes d'analyse pour retracer et représenter l'évolution des espèces. Cette recherche s'inscrit dans le contexte général de l'utilisation d'approches permettant la construction de réseaux, plutôt que d'arbres phylogénétiques. Elle aborde deux grands problèmes : 1) celui de la détection et du traitement des taxons issus de l'hybridation et de l'évolution réticulée; et, 2) celui de la combinaison de phylogénies définies sur un même ensemble de taxons en un réseau consensus.

1.1.2. Concepts et définitions

1.1.2.1. Phylogénies

Une phylogénie « raconte » l'histoire des relations de parenté entre des groupes d'organismes vivants, habituellement des espèces, et par extension consiste en la représentation graphique de cette histoire. Les phylogénies ont tout d'abord été construites en fonction des connaissances et de l'intuition des biologistes qui se basaient majoritairement sur des données morphologiques. L'analyse phylogénétique permet aujourd'hui d'estimer les relations entre les espèces à l'aide de critères et d'algorithmes objectifs, et ce à partir, plus souvent qu'autrement, de données moléculaires.

1.1.2.2. Arbres et réseaux : Définitions

En théorie des graphes, un *graphe* est une paire $G = (V, E)$ composée d'un ensemble de *nœuds* (V) et de *branches* (ou *arêtes*; E) reliant chacune deux nœuds entre eux (Fig. 1.4). Un *chemin* est un graphe $P = (V, E)$ reliant deux nœuds distincts en ne passant jamais plus d'une fois par la même branche. On dit qu'un graphe est *connexe* s'il existe un

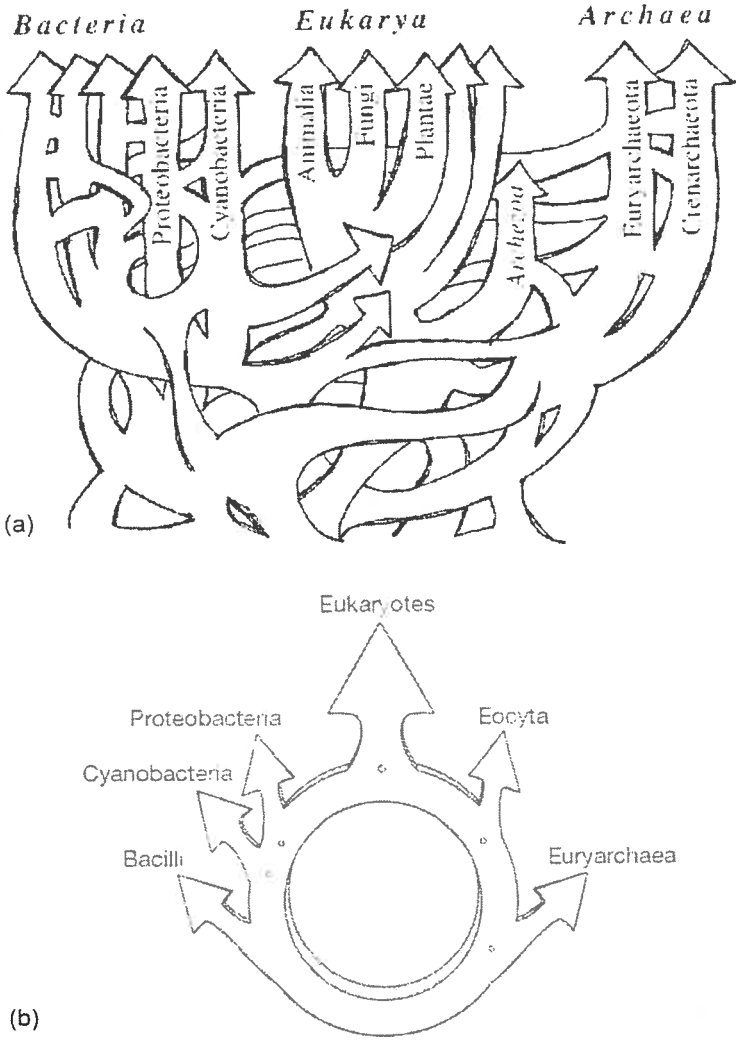
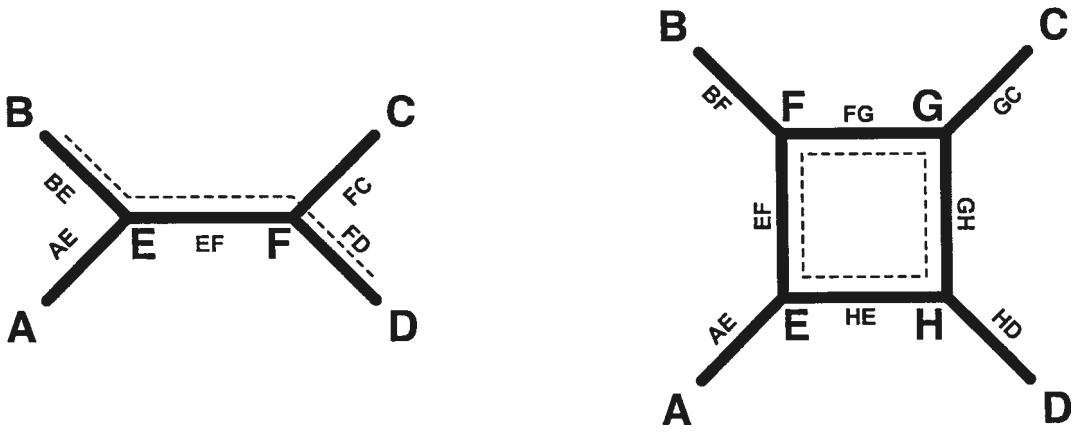


Figure 1.3 Deux visions réticulées de l'évolution. (a) Doolittle (1999) propose une vision selon laquelle les événements de réticulations jouent un rôle majeur dans la diversification et l'évolution des taxons. (b) Pour Rivera & Lake (2004), des fusions de génomes ont mené à l'apparition de grands groupes comme les eucaryotes, ce qui les mène à proposer le concept d'Anneau de la Vie (*« Ring of Life »*) pour expliquer la phylogénie des grands groupes.

chemin entre toutes les paires de nœuds. Un *cycle* est un chemin reliant un nœud à lui-même. Pour le biologiste, un cycle représente une réticulation. Un *arbre*, qu'il soit phylogénétique ou non, est un graphe connexe acyclique (Fig. 1.4a). Par extension, un *réseau phylogénétique* est un graphe connexe pouvant contenir un ou plusieurs cycles (Fig. 1.4b). Le terme graphe phylogénétique fait globalement référence aux arbres et aux réseaux phylogénétiques. Bien que l'on utilise parfois dans la littérature la terminologie « d'arbre réticulé » il serait plus juste d'utiliser le terme de réseau phylogénétique en présence de réticulations, car un arbre n'en est plus un au sens mathématique du terme à partir du moment ou des réticulations y sont autorisées. Un graphe peut être *enraciné* ou non. Un graphe enraciné possède une *direction* de sa base, ou *racine*, vers ses feuilles, représentant ainsi une relation de descendance des nœuds *parents* vers les nœuds *enfants*. Le *degré* (ou la *valence*) d'un nœud est égal au nombre de branches y étant rattaché. Les nœuds de degré 1 sont dits *terminaux* (ou *feuilles*). Dans le cadre de l'analyse phylogénétique il s'agit des taxons à l'étude. On dira alors que les graphes phylogénétiques portent des *étiquettes* uniquement sur les nœuds terminaux. Un graphe dont tous les nœuds portent une étiquette est dit complètement étiqueté. Les nœuds de degré supérieur à 1 sont dits *internes*. Un graphe phylogénétique ne contient jamais de nœuds de degré 2. Si tous les nœuds internes d'un arbre sont de degré 3, l'arbre est dit *binaire* (Fig. 1.5a, b et c). Un arbre binaire est complètement *résolu* car il est impossible de le *raffiner*, c'est-à-dire d'y ajouter de nouveaux nœuds de degré 3. Un arbre contenant un ou des nœuds de degré plus élevé que 3, mais au moins un nœud de degré 3, n'est que partiellement résolu (Fig. 1.5d). Le nombre d'arbres binaires enracinés (T_R) différents croît de manière exponentielle en fonction du nombre (n) de feuilles (Cavalli-Sforza & Edwards 1967; Felsenstein 1978) :

$$T_R(n) = \prod_{i=2}^{n} (2i-3) \qquad (1.1)$$



(a) (b)

Figure 1.4 Cet arbre (a) et ce réseau (b) sont des graphes phylogénétiques non enracinés. A, B, C, D, E, F, G et H sont des nœuds qui sont reliés par des branches identifiées par la paire de nœuds qu'elles joignent. A, B, C et D sont des nœuds terminaux ou feuilles qui correspondent aux taxons étudiés, les autres nœuds sont internes. Les traits pointillés illustrent un chemin reliant les feuilles B et D passant par les branches BE, EF et FD en (a) et un cycle, ou une réticulation, passant par les branches EF, FG, GH et HE en (b). Le cycle en (b) fait en sorte qu'il existe deux chemins différents entre les paires de nœuds terminaux dans le réseau, par exemple les chemins BF, FG, GH, HD et BF, FE, EH, HD relient les nœuds B et D.

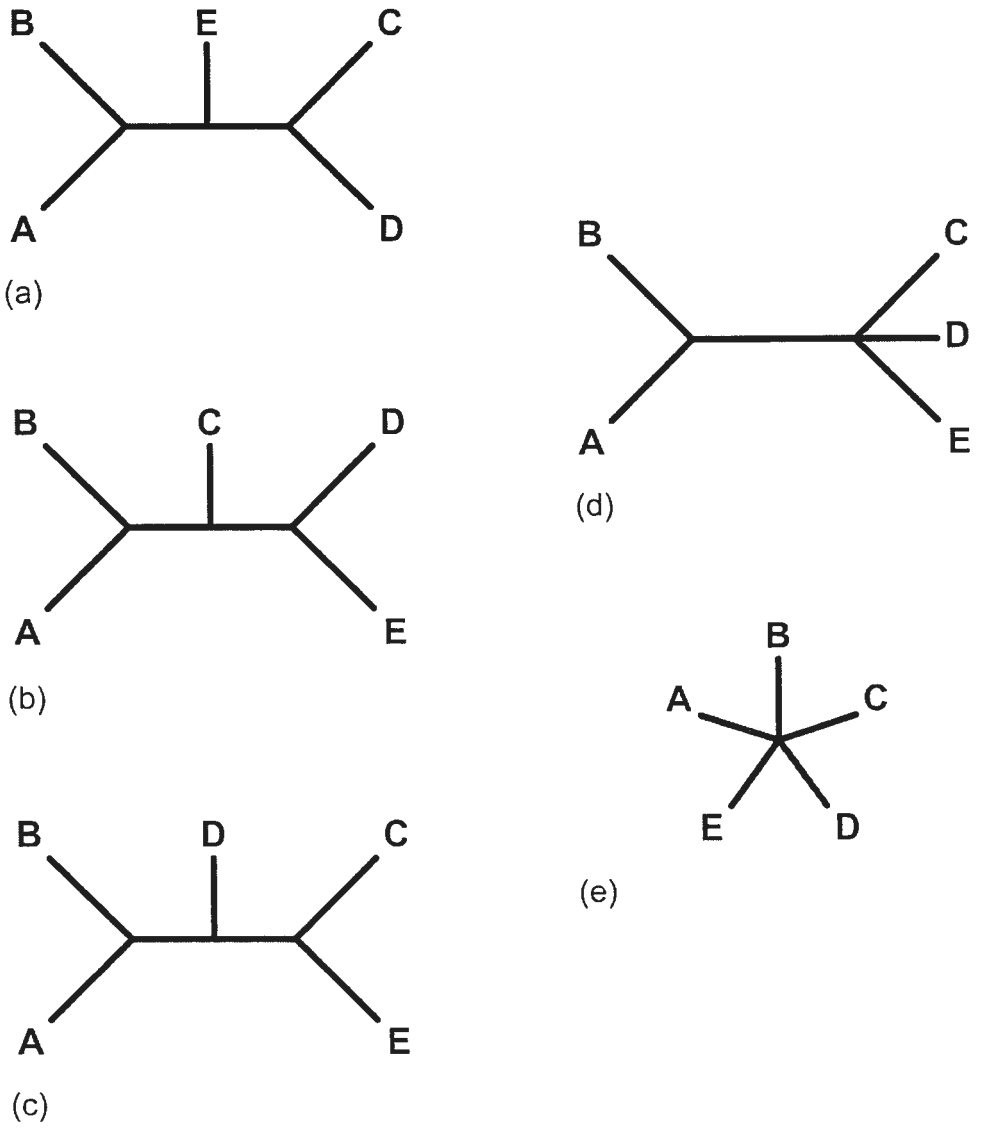


Figure 1.5 Les 3 arbres binaires non enracinés présentés en (a), (b) et (c) sont les trois manières possibles de raffiner l'arbre (d) qui contient une polytomie. L'arbre (e) est un arbre étoile qui ne contient aucune information quant aux relations phylogénétiques entre les taxons.

Similairement, le nombre d'arbres binaires non enracinés (T_U) croît de manière exponentielle en fonction du nombre (n) de feuilles (Edwards & Cavalli-Sforza 1964; Felsenstein 1978) :

$$T_U(n) = \prod_{i=3}^{n} (2i-5) \qquad (1.2)$$

Un arbre ne contenant qu'un nœud interne de degré *n* est dit *arbre étoile*, celui-ci n'est pas du tout résolu (Fig. 1.5e). Chaque nœud de degré supérieur à 3 correspond à une *polytomie*, qui peut être résolue de plusieurs façons différentes (Rohlf 1982; Mickevich & Platnick 1989).

Chaque branche d'un arbre définit une *partition* des taxons en deux groupes (ou *bipartition*) disjoints; c'est-à-dire que si une branche est retirée d'un arbre il en résulte deux arbres disjoints. Par exemple, la branche EF de l'arbre de la figure 1.4a définit les bipartitions $\{A, B\}$ et $\{C, D\}$, on pourra également parler de la bipartition $AB|CD$. Dans un réseau phylogénétique, une branche n'est pas toujours suffisante pour définir une bipartition : seules les branches ne faisant pas partie d'un cycle définissent une bipartition. Il faut retirer deux ou plus de ces branches pour faire apparaître une bipartition, par exemple dans le réseau de la figure 1.4b les branches FG et HE définissent la bipartition $AB|CD$, alors que les branches EF et GH définissent la bipartition $AD|BC$.

Peu importe l'approche utilisée pour l'obtenir, chaque arbre peut être représenté par une matrice de distances unique. Cette matrice *patristique* est parfaitement équivalente à l'arbre qu'elle représente : il existe une bijection qui permet de passer d'un arbre à sa matrice patristique et vice-versa (Buneman 1971). À chaque branche d'un arbre correspond une valeur ou *longueur*. Ces valeurs peuvent être toutes égales ou varier d'une branche à l'autre. La distance entre deux feuilles dans la matrice patristique est égale à la somme des longueurs des

branches se trouvant sur le chemin les reliant (Fig. 1.6). Une matrice de distances d'arbre est toujours *additive*, mais peut également être *ultramétrique*. Une matrice de distance est dite additive si elle satisfait la condition des quatre points (Buneman 1971; Fig. 1.6a) :

$$d_{ij} + d_{kl} \leq \max\left(d_{ik} + d_{jl}; d_{il} + d_{jk}\right) \text{ pour tous les } i,\ j,\ k \text{ et } l \in V \tag{1.3}$$

La condition ultramétrique, ou des trois points parce que définie sur des triplets, est plus restrictive. Un arbre ultramétrique est enraciné et toutes ses feuilles sont équidistantes de la racine (Fig. 1.6b), ainsi :

$$d_{ik} \leq \max\left(d_{ij}, d_{jk}\right) \text{ pour tous les } i,\ j \text{ et } k \in V \tag{1.4}$$

1.1.2.3. Inférence phylogénétique

Il existe de nombreuses façons d'estimer des arbres phylogénétiques. Felsenstein (2004) fait état de plus de 3000 articles portant sur ces méthodes, un nombre impressionnant pour une jeune discipline ayant à peine 40 ans. Bien que le but ne soit pas ici d'effectuer une revue exhaustive des différentes méthodes disponibles, il est nécessaire de dresser un portait sommaire des principales approches.

Tous les algorithmes de reconstruction phylogénétique reposent sur un principe commun : trouver la phylogénie optimale en fonction des données observées et d'un critère d'optimisation défini *a priori*. Les trois principales familles de méthodes utilisées aujourd'hui par les biologistes diffèrent quant à leur critère d'optimisation. Les *méthodes de parcimonie* préfèrent la phylogénie qui minimise le nombre de pas évolutifs; c'est-à-dire le nombre de changements d'états de caractères sur l'arbre (Hennig 1979). Les *méthodes de maximum de vraisemblance* sélectionnent plutôt l'arbre qui, lorsque jumelé à un modèle d'évolution des caractères, maximise la probabilité d'avoir engendré les données observées aux feuilles de l'arbre (Felsenstein 1981). Finalement, les *méthodes de distances* considèrent les distances mesurées entre les espèces comme

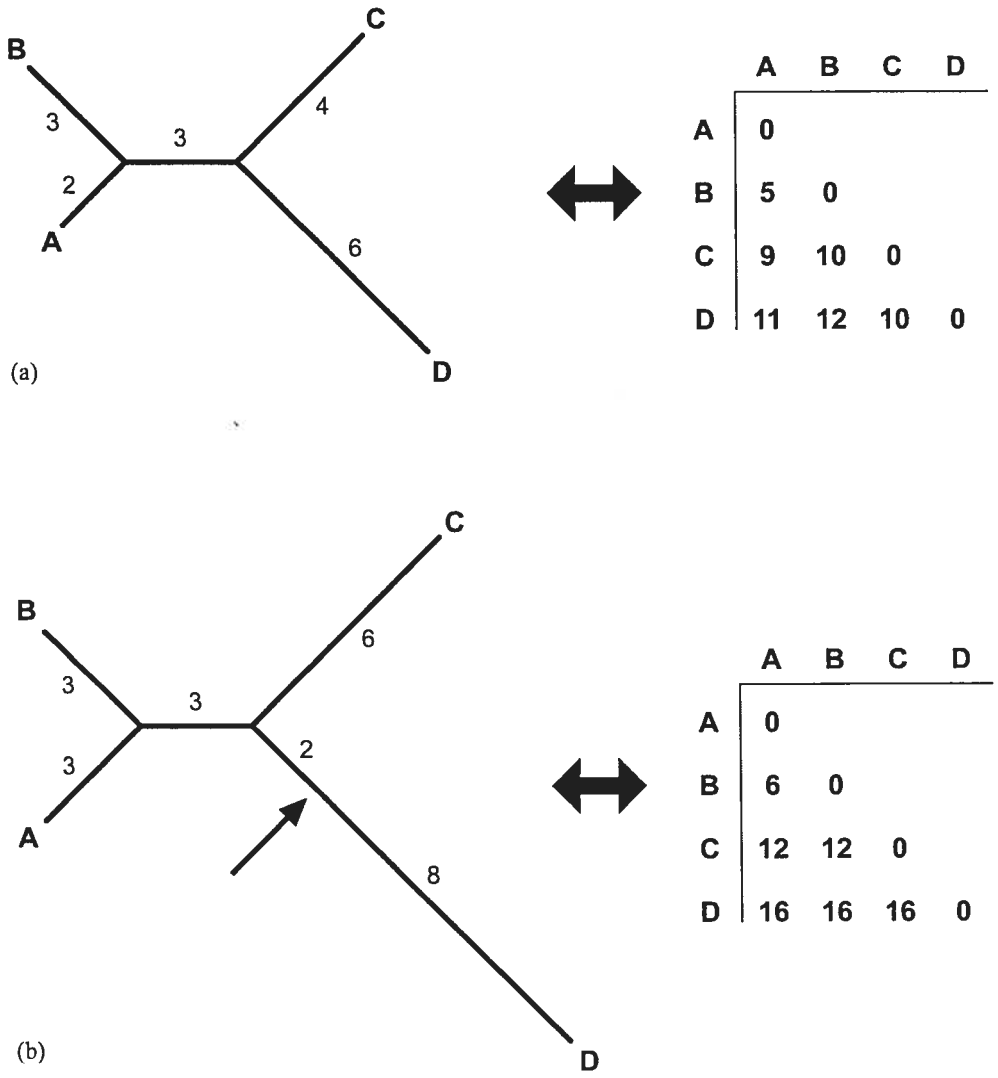


Figure 1.6 Deux arbres non enracinés et leur matrice patristique. Les distances patristiques de l'arbre (a) sont additives, alors que celles de l'arbre (b) sont ultramétriques. La flèche en (b) montre la position de la racine de l'arbre ultramétrique.

des estimations des vraies distance évolutives (c'est-à-dire la somme des longueurs des branches se trouvant sur le chemin les rejoignant dans la vraie phylogénie) et recherchent l'arbre dont la matrice patristique s'ajuste le mieux à la matrice de distances observées (Cavalli-Sforza & Edwards 1967; Fitch & Margoliash 1967). Le critère particulier utilisé lors de l'inférence phylogénétique peut varier d'une méthode à l'autre pour une même famille d'algorithmes.

Seules des méthodes de distances seront abordées dans le cadre de cette thèse. Ces méthodes présentent l'avantage de pouvoir êtres utilisées avec toutes les données, peu importe leur type (qualitatif, semi quantitatif, quantitatif) et leur provenance (éthologique, morphologique, génétique). Elles sont également les seules à permettre l'analyse de données qui ne se présentent que sous la forme de distances, comme les matrices d'hybridation ADN-ADN et de sérologie comparée. Ces algorithmes incluent les méthodes de groupement (Sokal & Sneath 1963), comme le neighbor-joining (Saitou & Nei 1987), qui groupent successivement les taxons les plus similaires jusqu'à l'obtention d'une phylogénie complète, ainsi que les algorithmes de moindres carrés qui recherchent parmi toutes les solutions possibles l'arbre minimisant la somme de carrés des écarts entre les distances observées et les distances d'arbre (Cavalli-Sforza & Edwards 1967; Fitch & Margoliash 1967).

1.2. Le problème des réticulations en analyse phylogénétique

1.2.1. Contexte

Le problème de l'évolution réticulée est un sujet des plus actuels en analyse phylogénétique et plusieurs auteurs démontrent un intérêt pour la phylogénie de groupes ayant une histoire réticulée (Comes & Abbot 1999; Doolittle 1999; Lapointe 2000; Legendre 2000a, 2000b; Rohlf 2000; Sneath 2000; Xu 2000; Comes & Abbot. 2001; Gandolfi *et al.* 2003). Les taxons

hybrides sont définis comme étant issus de l'hybridation, de l'introgression ou du transfert latéral entre deux lignées indépendantes (Dowling & Secor 1997) et diffèrent donc des taxons dits « normaux », qui sont eux issus de la divergence des lignées (Wagner 1969). L'*hybridation* se définit comme la reproduction entre des individus provenant de populations, ou groupes de populations, distinguables sur la base d'un, ou de plusieurs, caractères héréditaires (Harrison 1990, 1993). Cette définition inclut les croisements entre membres d'espèces différentes. Étant donné que les hybrides peuvent êtres plus ou moins aptes que leur parents (Arnold & Hodges 1995; Emms & Arnold 1997; Johnston *et al.* 2001), l'hybridation peut mener à l'effondrement ou au renforcement des barrières à la reproduction entre les deux espèces (Barton & Hewitt 1985) et même donner lieu à la création de nouvelles espèces (Arnold 1992, 1997). Lorsqu'une partie du matériel génétique d'une espèce pénètre le génome de l'autre sans apparition de nouvelles espèces, on parlera d'*introgression* (Barton et Hewitt 1985; Arnold 1997). Le *transfert latéral*, ou *horizontal*, réfère généralement à l'échange de matériel génétique, parfois sous la forme de plasmides, entre procaryotes (Sonea & Matthieu 2000). Il est également utilisé, de manière plus générale et en opposition au *transfert vertical*, pour signifier tout échanges génétiques entre taxons qui ne suivent pas un modèle arborescent.

Traditionnellement les chercheurs ont favorisé différentes approches pour aborder le problème de la réticulation lors de l'inférence phylogénétique. Certains l'ont ignoré en supposant qu'il n'était pas important dans le groupe étudié, ou simplement, parce qu'ils ne disposaient pas de méthodes appropriées pour son étude (Skala & Zrzavy 1994). D'autres, jugeant la présence d'hybrides comme une nuisance, ont préconisé leur retrait de l'analyse phylogénétique pour les replacer entre les parents supposés dans la phylogénie obtenue (Wagner 1969, 1983). Les derniers ont investi leurs efforts dans le développement de méthodes permettant l'analyse conjointe des taxons hybrides et non hybrides (Rieseberg & Morefield 1995; Makarenkov & Legendre 2000). Dans le cadre de cette thèse ces deux

dernières approches sont explorées.

En effet, l'hybridation étant aujourd'hui reconnue comme un processus important, tant chez les plantes que chez les animaux, il n'est plus possible d'ignorer ce phénomène lors de l'inférence phylogénétique. La première formulation de l'hypothèse de la spéciation par hybridation chez les plantes est attribuée à Linnée (Rieseberg 1997), mais son acceptation par la communauté scientifique attendra les travaux d'Anderson (1936) et remonte aux années 1950 (Stebbins 1950; Anderson & Stebbins 1954). Dernièrement, Ellstrand et ses collaborateurs (1996) ont estimé, qu'en moyenne 11% des espèces végétales pouvaient être issues de l'hybridation, soit un total de 27 500 hybrides parmi les 250 000 espèces de plantes décrites à ce jour. D'autre part, Bullini (1994) souligne que le rôle de l'hybridation a longtemps été sous-estimé dans le règne animal, bien que l'on ait identifié un nombre croissant d'exemples. Dowling & Secor (1997) énumèrent à cet égard plusieurs taxons animaux pour lesquels l'hypothèse de l'hybridation a été émise (38 espèces 16 genres et 3 familles). Cette liste comprend aussi bien des invertébrés que des vertébrés. Holliday (2003) suggère même que l'hybridation entre lignées pré-humaines ait pu jouer un rôle important dans l'évolution des hominidés. La compréhension des phénomènes et processus de l'hybridation interspécifique et de son impact sur la biodiversité est également préoccupante (Raybould & Gray 1994; Rhymer & Simberloff 1996; Allendorf *et al.* 2001; Grosholz 2002; Perry *et al.* 2002).

Il est clair que les techniques de reconstruction phylogénétique classiques qui représentent l'évolution à l'aide de diagrammes arborescents sont incapables de représenter adéquatement l'évolution des taxons issus de l'hybridation. Pour résoudre ce problème, plusieurs méthodes et logiciels permettant la construction de phylogénies réticulées ont été proposées (Sneath 1975; Humphries 1983; Nelson 1983; Wagner 1983; Wanntorp 1983; Bandelt & Dress 1992; Bandelt *et al.* 1995, 2000; Bandelt 1994; Rieseberg et Morefield 1995; Dress *et al.* 1996; Jakobsen & Easteal 1996; Jakobsen *et al.* 1997; Fitch 1997; Sosef 1997; Makarenkov & Legendre

2000; Sang & Zhong 2000; Strimmer & Moulton 2000; Xu 2000; Baccam *et al.* 2001; Holder *et al.* 2001; Strimmer *et al.* 2001, 2003; Bryant & Moulton 2002, 2004; Holland *et al.* 2002; Legendre & Makarenkov 2002; von Haeseler & Churchill 1993; Nakhleh *et al.* 2003, 2004; Moret *et al.* 2004). Lapointe (2000) ainsi que Posada & Crandall (2001b) ont passé en revue plusieurs de ces méthodes. Ces approches sont beaucoup moins restrictives que les méthodes arborescentes et sont donc très utiles pour illustrer les incompatibilités d'un jeu de données à l'aide de réticulations. Cependant il apparaît nécessaire d'identifier les qualités et les défauts de ces différentes méthodes afin de pouvoir en faire une utilisation adéquate. Dans le cadre de cette thèse une évaluation de méthodes phylogénétiques de détection de taxons hybrides est proposée. À cet effet, deux stratégies de comparaison ont été adoptées. La première consiste à appliquer ces méthodes à des données réelles comprenant des espèces non hybrides et des hybrides synthétisés expérimentalement. La seconde se base sur l'estimation de la performance relative des méthodes, notamment par une approche de simulations.

McDade (1990, 1992, 1997) a effectué une série d'expériences portant sur les patrons de caractères morphologiques chez des hybrides de première génération (F_1) et leur comportement en analyse phylogénétique en ayant recours à des croisements contrôlés entre des espèces du genre *Aphelandra* (Acanthaceae; McDade 1984). Cette série d'article est d'une importance capitale car elle constitue la première évaluation formelle de l'impact de la présence d'hybrides dans un jeu de données sur les méthodes classiques de reconstruction d'arbres phylogénétiques. Ses conclusions principales sont (1) que les méthodes de parcimonie et de distances sont incapables de distinguer un hybride d'un taxon non hybride, (2) qu'un taxon hybride aura tendance à se placer à la base d'un groupe contenant un de ses parents, et (3) que l'inclusion de taxons hybrides n'entraîne pas de perturbations majeures des phylogénies, à moins d'être très nombreux ou que les parents soient très éloignés.

Le jeu de données de McDade (1990) offre donc la possibilité de mettre à

l'épreuve des méthodes d'analyse phylogénétique permettant la détection de taxons hybrides tout en œuvrant dans la continuité de ces travaux. Dans le cadre du présent projet, deux méthodes spécifiques ont été comparées en les appliquant aux données de McDade : la décomposition des bipartitions (Bandelt & Dress 1992) et les réticulogrammes (Makarenkov & Legendre 2000). D'un point de vue purement pratique, ces deux méthodes ont été sélectionnées parce que leurs logiciels permettaient le traitement automatisé d'analyses multiples, qualité nécessaire pour leur évaluation à l'aide de simulations. De plus, la décomposition des bipartitions est la méthode de reconstruction de réseaux la plus largement connue et utilisée en analyse phylogénétique et en génétique des populations. De son côté, la construction de réticulogrammes constituait une nouvelle approche prometteuse pour l'indentification de taxons hybrides.

Mentionnons tout de suite que l'application directe de ces méthodes à la détection de taxons hybrides a donné des résultats décevants (Chapitre 2). Ayant illustré cette incapacité des méthodes à l'étude il a été nécessaire de développer un nouvel indice basé sur un critère de détection des hybrides (Chapitre 2) ainsi que son test de signification (Chapitre 3). Ce test permet la détection des taxons hybrides, leur retrait de l'analyse phylogénétique, puis leur ajout à la phylogénie contenant les espèces parentales. Les données de McDade (1990) ont également été utilisées afin de mettre cette nouvelle approche à l'épreuve (Chapitre 2 et 3).

Afin d'étoffer l'évaluation de ce nouvel outils statistique, un second jeu de données a également été analysé. Les relations phylogénétiques entre les kangourous du genre *Petrogale* (Marsupialia : Macropodidae) sont reconnues comme étant difficiles à résoudre (Campeau-Péloquin *et al.* 2001). Ces difficultés ont été attribuées à des cas d'introgression et à l'existence d'espèces issues de l'hybridation (e.g. Eldridge & Close 1993; Eldridge 1997; Eldridge & Pearson 1997; Campeau-Péloquin *et al.* 2001). Le jeu de données d'hybridation ADN-ADN de Campeau-Péloquin *et al.* (2001), augmenté de données sur des hybrides connus a donc été utilisé afin d'évaluer la performance du test de détection des hybrides développé

dans cette thèse (Chapitre 4). Finalement, la puissance et l'erreur de type I du test ont été évaluées à l'aide de simulations (Chapitre 5). Une description sommaire de la décomposition des bipartitions et de la méthode de construction de réticulogrammes complète cette section, la description du nouvel indice et de son test de signification se retrouvant dans le corps de la thèse.

1.2.2. Méthodes à l'étude

1.2.2.1. Décomposition des bipartitions

La décomposition des bipartitions (« split decomposition ») a été développée afin de représenter graphiquement le signal phylogénétique d'un jeu de données présenté sous forme de distances (Bandelt & Dress 1992). Un des objectifs de la méthode est d'illustrer l'ajustement des distances observées à un modèle arborescent. Pratiquement, la méthode représente les groupements les plus forts, qui seraient les seuls retenus par une méthode de reconstruction d'arbre, mais également des groupements alternatifs et contradictoires, qui sont définis par les données.

Cette approche repose sur l'analyse indépendante de tous les ensembles de quatre taxons (quadruplets) suivi de l'assemblage d'un graphe généralement *planaire*; c'est-à-dire représenté dans un plan à deux dimensions, et défini sur l'ensemble des taxons. Le résultat est un *graphe de bipartitions* (ou « splitsgraph »). Chaque quadruplet est tour à tour soumis à la condition des quatre points. Une méthode arborescente choisira parmi les trois bipartitions possibles pour quatre taxons $\{i,j,k,l\}$ la bipartition $ij|kl$ pour laquelle la somme $d_{ij}+d_{kl}$ est minimale (Fig. 1.7). Les deux autres bipartitions sont alors rejetées. La décomposition des bipartitions procède plutôt en rejetant la bipartition $ij|kl$ présentant la somme $d_{ij}+d_{kl}$ maximale. Un indice d'isolation est ensuite calculé à partir des deux bipartitions restantes :

$$S_{ij|kl} = \left[\left(d_{il} + d_{jk} \right) - \left(d_{ij} + d_{kl} \right) \right] / 2 \qquad (1.5)$$

Cet indice représente une mesure du « support » de chacun des deux arbres définis sur le quadruplet $\{i, j, k, l\}$. Il est utilisé pour représenter le conflit entre ces *bipartitions faiblement compatibles*; c'est-à-dire qui peuvent êtres représentées conjointement par un réseau planaire, mais pas par un arbre (Fig. 1.7). Si les distances entre tous les taxons sont parfaitement additives, une seule bipartition sera retenue pour chaque quadruplet et le graphe complet est un arbre. Dress *et al.* (1996) présentent en détails l'algorithme permettant d'établir la liste des bipartitions faiblement compatibles pour chacun des quadruplets et de construire la représentation planaire sur l'ensemble des taxons. La méthode a été implémentée dans le logiciel SplitsTree (Huson 1998; Huson & Bryant 2006).

1.2.2.2. Réticulogrammes

Cherchant à élucider des histoires évolutives réticulées, Makarenkov & Legendre (2000) et Legendre & Makarenkov (2002) ont développé une méthode permettant la reconstruction d'un autre type de réseau phylogénétique qu'ils ont baptisé *réticulogramme*. Partant d'un arbre additif, l'algorithme procède par l'ajout de branches supplémentaires afin d'augmenter l'ajustement de la représentation à la matrice de distances observées (Fig. 1.7). Pour les taxons entre lesquels il n'existe qu'un chemin, la distance de réticulogramme est égale à la distance calculée sur l'arbre initial. Pour les autres taxons, on utilise la distance minimale entre ceux-ci dans le réticulogramme. L'algorithme procède de manière itérative : à chaque itération il ajoute parmi toutes les réticulations possibles celle qui minimise la somme des carrés des écarts entre les distances observées et les distances du réticulogramme. Le nombre de réticulations ainsi ajoutées peut être défini *a priori* par le biologiste ou être déterminé à l'aide d'un critère objectif. Makarenkov & Legendre (2000) proposent deux critères alternatifs qui permettent d'arrêter l'ajout

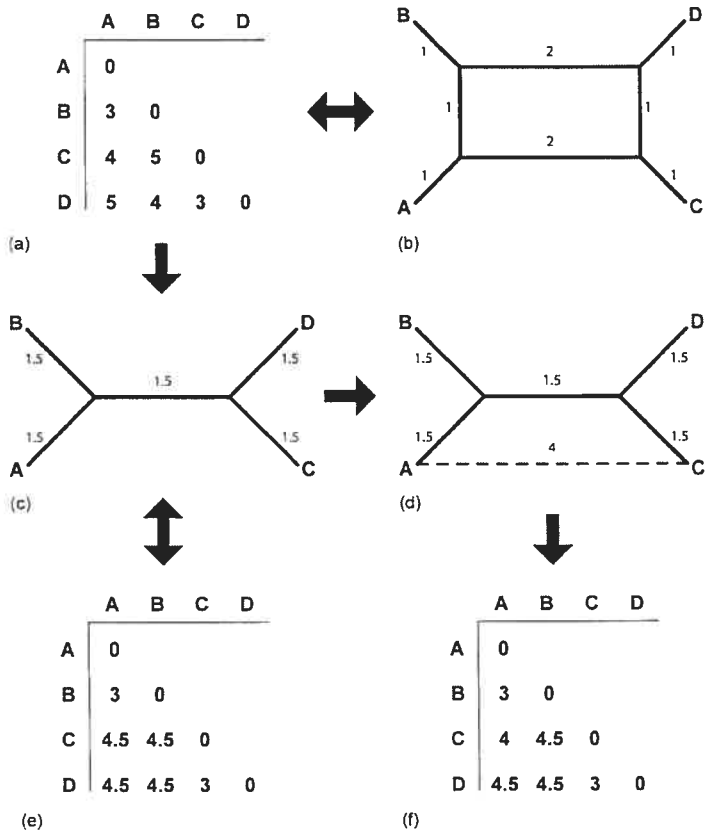


Figure 1.7 Graphe de bipartitions (b), arbre (c), et réticulogramme (d) obtenus à partir de la même matrice de distances en (a). Le graphe de bipartitions (b) présente les bipartitions $AB|CD$ et $AC|BD$ étant donné que $d_{AB}+d_{CD}=6$, $d_{AC}+d_{BD}=8$ et $d_{AD}+d_{BC}=10$. De plus $S_{AB|CD}=2$ et $S_{AC|BD}=1$. Le graphe de bipartition représente parfaitement les distances dans la matrice initiale (a). Pour obtenir le réticulogramme (d), il faut tout d'abord obtenir un arbre (c) et sa matrice patristique (e), cette dernière n'est pas parfaitement ajustée à la matrice initiale (a). L'ajustement est amélioré en ajoutant une réticulation entre les nœuds A et C, telle qu'illustré par la matrice du réticulogramme (f). Les flèches unidirectionnelles représentent des unijections, les bidirectionnelles des bijections.

de réticulations supplémentaires lorsque l'amélioration de l'ajustement ne justifie plus la complexification du modèle. Aucune réticulation n'est ajoutée sur une matrice additive étant donné qu'elle est parfaitement ajustée à un arbre. Legendre & Makarenkov (2002) proposent diverses applications de leur méthode, notamment en analyse phylogénétique et en biogéographie. Le logiciel TRex (Makarenkov 2001) permet, entre autres, la construction des réticulogrammes à partir de matrices de distances.

1.3. Le problème du consensus en analyse phylogénétique

1.3.1. Contexte

Toutes les méthodes de reconstruction phylogénétique sont susceptibles de générer des arbres multiples pour un même jeu de données. Ces arbres représentent des solutions équivalentes, en fonction d'un critère d'optimisation choisi. De même, l'analyse séparée de jeux de données indépendants, par exemple pour différents gènes (Brower *et al.* 1996; Maddison 1997) ou encore pour des gènes et des données morphologiques, peut produire des arbres phylogénétiques distincts qui pourront représenter des hypothèses différentes. Dans les deux cas, il est d'intérêt pour le biologiste de combiner ces arbres afin d'illustrer les zones d'accord et de désaccord entre eux.

Le second objectif de la recherche présentée dans cette thèse concerne la combinaison de ces arbres multiples. Les méthodes de consensus ont été développées spécifiquement pour identifier les conflits entre plusieurs arbres, ou encore pour illustrer les groupes phylogénétiques qui sont le mieux soutenus par ces arbres distincts (Bryant 2003). Formellement, une méthode de consensus est généralement définie comme une fonction opérant sur un ensemble d'arbres portant sur un ensemble de taxons donnés, et fournissant en retour un arbre unique, illustrant les relations entre ces mêmes taxons (Steel *et al.* 2000; Bryant

2003; Lapointe & Cucumel 2003). Bien que ces méthodes diffèrent selon le type d'information considéré et la manière dont elles traitent le conflit entre les arbres, les algorithmes de consensus les plus couramment utilisés répondent tous à la même propriété : la solution est toujours un arbre.

1.3.2. Méthodes à l'étude

1.3.2.1. Les arbres consensus

Les méthodes de consensus dites « classiques » sont très nombreuses et Bryant (2003) en présente une classification selon les propriétés intrinsèques des algorithmes. Dans le cadre de cette thèse, seulement trois de ces méthodes seront abordées : le consensus strict (Rohlf 1982), le consensus majoritaire (Margush & McMorris 1981) et le consensus moyen (Lapointe & Cucumel 1997). Les deux premières approches, qui sont aussi les plus populaires, ne se basent que sur l'information topologique des arbres, c'est-à-dire contenue dans les bipartitions, sans tenir compte des longueurs de branches, alors que la troisième approche est l'une des rares à tenir compte des longueurs de branches lors de la construction du consensus. Bien que le consensus strict et le consensus majoritaire soient généralement appliqués à des arbres enracinés, ils sont également applicables à des arbres non-enracinés. C'est ce dernier type d'arbres qui sera considéré pour le reste de la thèse.

En bref, le consensus strict d'un ensemble d'arbres contient uniquement les bipartitions présentes dans tous les arbres initiaux (Fig. 1.8d). Cette méthode très sensible sera donc affectée par la position variable des taxons dans les arbres combinés (Wilkinson & Thorley 2001). En conséquence, la solution manque souvent de résolution (Fig. 1.8d). Le consensus majoritaire est moins conservateur que le strict car il comprend les bipartitions qui se retrouvent dans une majorité d'arbres. Il aura donc tendance à produire des arbres plus résolus que le consensus strict (Fig. 1.8e). Le consensus moyen, quant à lui, prend en compte

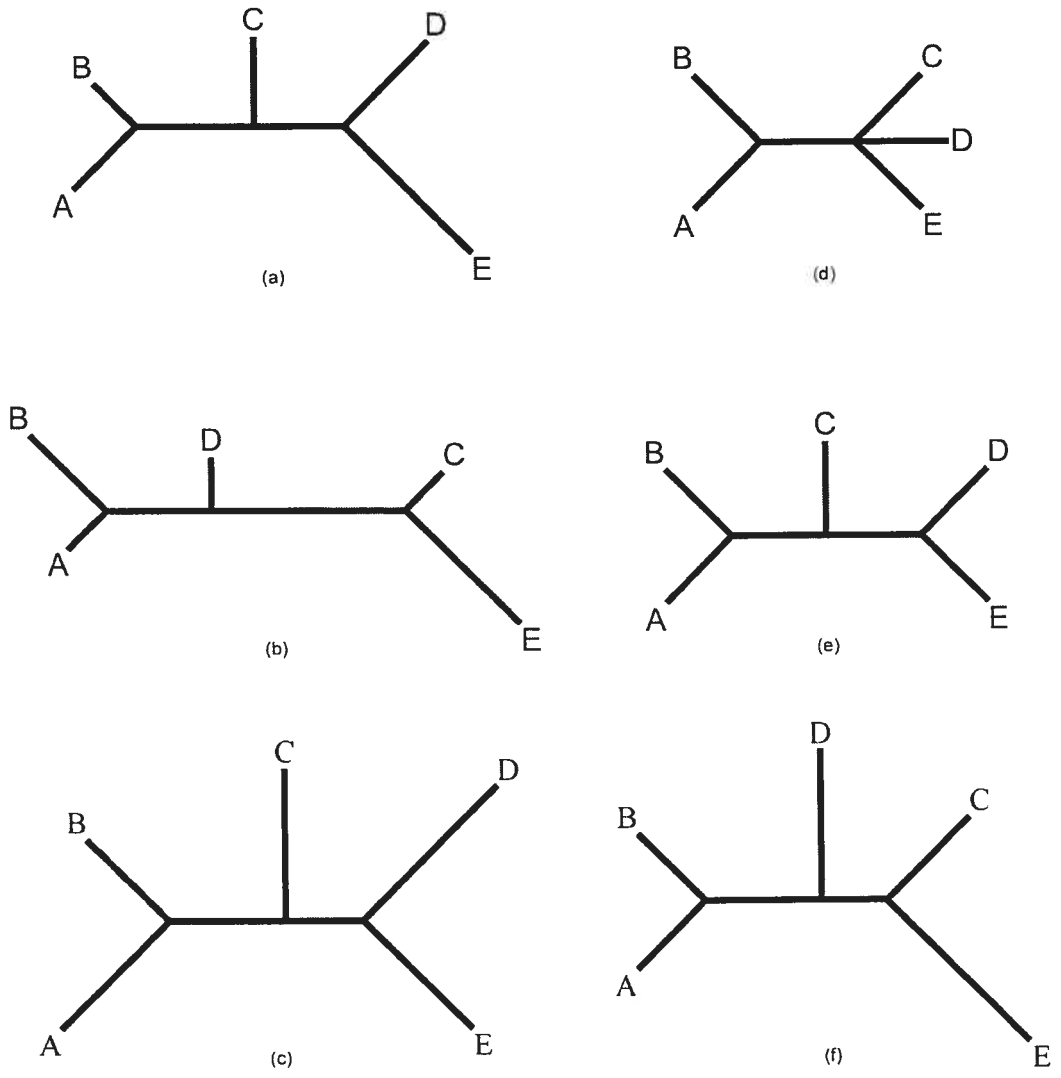


Figure 1.8 Trois arbres non-enracinés (a), (b), et (c), ainsi que leur consensus strict (d), leur consensus majoritaire (e), et leur consensus moyen (f).

l'information supplémentaire contenue dans les longueurs des branches des arbres considérés (Fig. 1.8f). Cette information est directement codée dans la matrice de distances patristique de chacun des arbres. Le consensus moyen est l'arbre qui minimise la somme des carrés des écarts entre sa matrice de distances patristique et les matrices de distances des arbres du profil. Lapointe & Cucumel (1997) ont démontré que cet arbre peut être estimé en minimisant la somme des carrés des écarts entre le consensus et une matrice de distances moyenne calculée à partir du profil d'arbres. Pratiquement, le consensus moyen est donc obtenu en appliquant un algorithme des moindres carrés à la matrice moyenne.

1.3.2.2. Les réseaux consensus

Si les arbres multiples doivent êtres considérés comme autant d'hypothèses phylogénétiques valables, et, si la nature de l'évolution n'est pas toujours arborescente, il semble nécessaire de proposer des méthodes de consensus permettant de résumer sous forme de réseau l'information conflictuelle contenue dans un ensemble d'arbres.

Des méthodes de reconstruction phylogénétique ont déjà été proposées dans le but de représenter sous forme de réseau tous les arbres les plus parcimonieux soutenus par un jeu de données. Les *réseaux médians* (« median networks ») (Bandelt *et al.* 1995) sont apparentés à la décomposition des bipartitions (Bandelt & Dress 1992). Tout comme cette dernière, leur reconstruction procède par l'analyse successive des quadruplets et l'assemblage d'un graphe complet. Toutefois, toutes les bipartitions retrouvées dans le jeu de données sont retenues et le graphe final est souvent multidimensionnel. En effet, la représentation conjointe des trois bipartitions possibles sur quatre taxons $\{A,B,C,D\}$, soient $AB|CD$, $AC|BD$ et $AD|BC$, ne peut être réalisée en deux dimensions (Fig. 1.9). Holland & Moulton (2003) ont d'ailleurs proposé de construire des réseaux consensus en utilisant des réseaux médians pour

représenter toutes les bipartitions contenues dans l'ensemble d'arbres initial.

Fitch (1997) propose le « netting », un autre algorithme permettant de construire directement un réseau contenant les arbres les plus parcimonieux. Bien que l'algorithme lui-même ne soit pas plus efficace, il propose des règles permettant d'obtenir le profil des arbres les plus parcimonieux à partir du réseau. De telles règles sont nécessaires si l'on veut éviter que la décomposition du réseau ne fasse apparaître de nouveaux arbres.

Les réseaux médians et le « netting » sont d'une application limitée étant donné leur restriction au seul critère de maximum de parcimonie. Ils ne peuvent d'ailleurs être considérés comme de réelles méthodes de consensus étant donné qu'ils permettent de passer directement des données au réseau sans passer par l'inférence des arbres eux-mêmes. Bandelt (1995) a été le premier à suggérer l'utilisation des réseaux pour visualiser l'information contenue dans un ensemble d'arbres. Dans le cadre de cette thèse, je proposerai un cadre général ainsi qu'une approche particulière pour la construction de réseaux consensus. Ce cadre flexible permet la combinaison d'arbres dans un réseau consensus selon différents critères d'optimalité. J'explorerai les propriétés, les avantages et les inconvénients de cette approche en la comparant aux arbres consensus décrits ci-dessus, ainsi qu'aux réseaux consensus d'Holland & Moulton (2003). J'insisterai principalement sur la plus grande efficacité des réseaux consensus par rapport aux méthodes de consensus traditionnelles; c'est-à-dire leur plus grande capacité à conserver l'information phylogénétique contenue dans les arbres de départ (Thorley *et al.* 1998; Wilkinson & Thorley 2001).

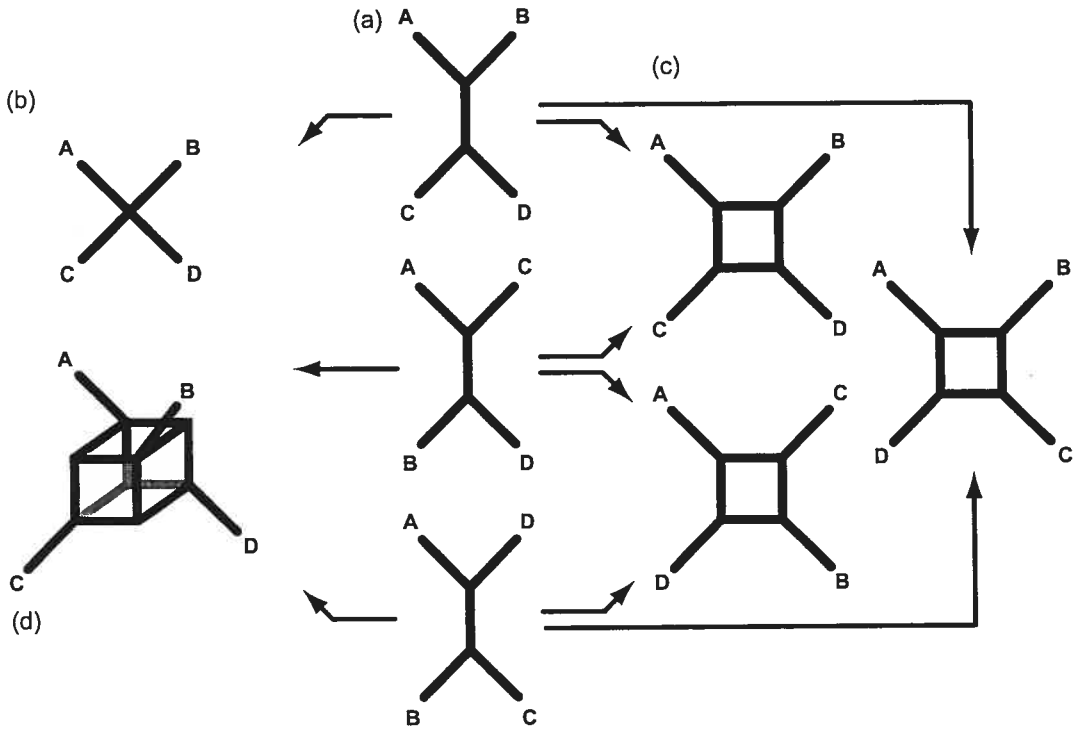


Figure 1.9 Les trois bipartitions possibles avec quatre taxons (a) ne peuvent êtres représentées conjointement dans un plan à deux dimensions à moins d'avoir recours à l'arbre étoile (b). Il est possible de représenter deux à deux ces bipartitions à l'aide d'un graphe de bipartitions (c). Pour représenter les trois bipartitions dans un seul et même graphe, il faut utiliser une troisième dimension (d).

1.4. Organisation de la thèse

La thèse est organisée en 7 chapitres. Le présent chapitre a permis d'introduire les bases théoriques et la problématique du projet. Le Chapitre 2 présente ensuite les résultats d'une étude comparée de la performance de deux méthodes alternatives de détection d'hybrides : les graphes de bipartitions et les réticulogrammes. Les deux méthodes sont appliquées directement aux données de McDade (1990, 1992, 1997). Les résultats obtenus ayant menés à la définition d'un critère de détection des hybrides permettant le calcul d'un indice de détection, les deux méthodes y sont également évaluées en fonction de ce critère. Le Chapitre 3 fait suite aux conclusions du Chapitre 2 en proposant un test statistique de signification de l'indice de détection des hybrides. La performance de ce test y est également confirmée en l'appliquant au jeu de données de McDade. Le Chapitre 4 est une autre application du test de détection des hybrides aux données d'hybridation ADN-ADN de Campeau-Péloquin *et al.* (2001) portant sur des kangourous du genre *Petrogale* (Marsupialia : Macropodidae) dont certains sont des hybrides connus. L'erreur de type I ainsi que la puissance du test sont évaluées à l'aide de simulations dans le Chapitre 5. Le Chapitre 6 traite de l'application des réseaux au problème du consensus et des avantages de cette approche pour la combinaison d'arbres multiples incompatibles. Le dernier chapitre présente une conclusion critique des résultats de la thèse. Le format de la présentation est sous forme d'articles scientifiques et les Chapitres 2 à 6 sont rédigés en anglais alors que les Chapitres 1 et 7 sont en français.

Chapitre 2 :

A COMPARISON OF ALTERNATIVE METHODS FOR DETECTING RETICULATION EVENTS IN PHYLOGENETIC ANALYSIS

Une version précédente de cet article a été publiée :

Gauthier, O. & Lapointe, F.-J. 2002. A comparison of alternative methods for detecting reticulation events in phylogenetic analysis, pp. 341-347 dans *Classification, Clustering, and Data Analysis: Recent Advances and Applications* édité par K. Jajuga, A. Sokolowski, et H.-H. Bock, Springer-Verlag, Berlin.

2.1. Résumé

Une préoccupation actuelle en analyse phylogénétique concerne notre capacité à détecter des événements d'évolution réticulée (ex : hybridation) qui ne répondent pas au patron strictement divergent représenté par les arbres phylogénétiques. En dépit de la disponibilité d'algorithmes permettant l'estimation de réseaux plutôt que d'arbres, il n'existe pas à ce jour d'évaluation formelle de leur habileté à détecter de réels événements de réticulation. Dans cet article, nous évaluons la performance des réticulogrammes et de la décomposition des bipartitions pour l'identification d'événements d'hybridation connus dans une phylogénie. Nos résultats montrent qu'aucune de ces deux techniques ne permet d'identifier sans ambiguïté des hybrides. Nous proposons donc une approche basée sur l'analyse des quadruplets de taxons, en combinaison avec ces deux méthodes. Cette nouvelle approche mène à une identification quasi parfaite des hybrides à l'étude. Nous proposons également des pistes pouvant mener à l'amélioration de l'algorithme de construction des réticulogrammes.

2.2. Abstract

A growing concern in phylogenetic analysis is our ability to detect events of reticulate evolution (e.g. hybridization) that deviate from the strictly branching pattern depicted by phylogenetic trees. Although algorithms for estimating networks rather than trees are available, no formal evaluation of their ability to detect actual reticulations has been performed. In this paper, we evaluate the performance of reticulograms and split decomposition graphs (or splitsgraphs) for the identification of known hybridization event. Our results show that neither technique permits unambiguous identification of hybrids. We thus introduce a quartet-based approach used in combination with these two methods and show that quartet analysis of splitsgraphs lead to a near perfect identification of hybrids. We also suggest ways in which the reticulogram reconstruction algorithm could be improved.

2.3. Introduction

The problem of reticulate evolution represents a growing concern in phylogenetic analysis (see Legendre 2000). Lateral gene transfer, introgression, and hybridization, among other phenomena, bring upon reticulate events of evolution. They do not result in a strictly bifurcating branching pattern such as those depicted by phylogenetic trees, but rather in a graph with multiple paths between some of the nodes, which cannot be elucidated by classical phylogenetic reconstruction methods. McDade (1990, 1992, 1997) has studied the impact of hybrids in phylogenetic analysis and compared the behavior of parsimony and distance-based tree reconstruction methods. Her results illustrated the poor performance of these techniques and showed that new methods were badly needed to solve the so-called 'hybrid problem'. Although some algorithms of network estimation are currently available (for reviews see Lapointe 2000; Posada & Crandall 2001b) their relative performance has never been determined for detecting hybridization events in a phylogeny. Using McDade's data we compared the behavior of two distance-based methods of reticulate analysis. Based on our results, a new quartet-based approach for hybrid detection is introduced.

2.4. Distance based methods of reticulate analysis

The so-called reticulogram method of Makarenkov & Legendre (2000) was designed specifically to detect events of reticulate evolution. It starts from an additive tree and adds additional edges, or reticulations, until a goodness of fit criterion is minimized, or a fixed number of reticulations specified by the user are added. Two criteria have been proposed by Makarenkov & Legendre (2000) as different stopping rules:

$$Q1 = \frac{\sqrt{Q(N)}}{\left(n(n-1)/2\right) - N} \qquad (2.1)$$

and

$$Q2 = \frac{Q(N)}{\left(n(n-1)/2\right) - N} \qquad (2.2)$$

where $Q(N) = \sum\sum\left(d_{ij} - \delta_{ij}\right)^2$, d_{ij} and δ_{ij} are the original dissimilarities and the reticulogram (or tree) distances respectively, n is the number of objects, and N the number of edges in the reticulogram (or tree). The result is presented in the form of a tree with extra edges superimposed onto it to depict possible reticulation events. This method will return a tree if the input data satisfies the four-point condition. Reticulogram reconstruction is implemented in the TRex program (Makarenkov 2001) available from the WWWeb at < http://www.info.uqam.ca/~makarenv/trex.html>.

Bandelt & Dress (1992) developed split decomposition in a totally different perspective. Their method aims at representing the conflicting signals in a phylogenetic data set. It uses the four-point condition on quartets to reject the least fitting tree among the three distinguishable topologies involving four objects. If the two remaining trees are both supported by the data, a pair of weakly compatible splits is shown to represent the conflict. The full representation on all quartets of the set of weakly compatible splits is a usually planar graphed called a splitsgraph. If no conflict is present in the data set, the split decomposition method will output a tree. Splitsgraphs were computed with the SplitsTree 3.2 program (Huson 1998), available from the WWWeb at <www.splitstree.org>. A newer version, SplitsTree 4 or JSplits (Huson & Bryant 2006), that adds numerous features to the original is now available at this address.

2.5. Hybrid detection analysis

To assess the relative performance of the reticulograms and splitsgraphs to detect actual hybridization events, we applied both techniques to a data set containing known hybrids. The morphological data collected by McDade (1984 1990) included 12 species of the plant genus *Aphelandra* and 17 hybrids produced in the lab by crossing parents representing 9 of the 12 species (see

Table 2.1). A species of this genus, the zebra plant (*Aphelandra squarrosa*), is a common houseplant. For simplicity, our analyses were conducted on 17 different data sets, each containing the 12 species and one single hybrid. Given the important proportion of empty cells, similarities between pairs of taxa were computed used Gower's similarity coefficient (Gower 1971) that accounts for missing data, and then converted to distances. These 17 distance matrices were submitted to TRex and SplitsTree and the position of the hybrids with respect to their parents were noted to determine the hybrid detection rate of the competing approaches. An hybrid was detected in a reticulogram if it was the sister taxa to one of its parents and had a reticulation to the other. In splitsgraphs, it had to form a pair of weakly compatible splits with both of its parents.

Our results indicated that a direct application of reticulate methods did not enable hybrid detection. In reticulograms, an average of 8.4 reticulations were added to the tree by the algorithm (7.0 and 9.7 for $Q1$ and $Q2$ respectively), making the interpretation of the resulting graph difficult, if at all possible. Hybrids were unambiguously detected in one single case with $Q1$ and three times with $Q2$ (these reticulations were always among the last ones added to minimize the criteria). Direct application of split decomposition did not provide better results, with only two hybrids detected in total. These were hybrids between closely related parents that grouped together in all of the analyses. It should also be noted that the dataset contains some homoplasy or even possible hybrid species (McDade 1984, 1990). Moreover, splitsgraphs and reticulograms computed on the 12 species only contain numerous weakly compatible splits and reticulations (not shown).

2.6. Hybrid detection through quartet analysis

Given the poor performance of the reticulate analysis methods in our comparisons, we have developed a different approach for detecting hybridization events in a data set. Based on morphological character patterns, it has long been known that hybrids should be placed between their parents in

Table 2.1 The 17 hybrids used in the analysis and their Hybrid Detection Criterion (HDC) values. Values are given for each hybrid and each method. The maximum possible HDC value for this dataset is 10. High HDC values indicate possible hybridization.

Parents		Method	
Ovulate	Staminate	Reticulogram	Split Decomposition
A. deppeana	*A. panamensis*	5	10
	A. sinclairiana	5	10
	A. storkii	4	9
A. golfodulcensis	*A. deppeana*	3	10
	A. leonardii	3	9
	A. sinclairiana	7	10
A. leonardii	*A. campanensis*	4	9
	A. golfodulcensis	1	7
	A. sinclairiana	6	10
A. panamensis	*A. deppeana*	6	10
	A. golfodulcensis	6	9
	A. leonardii	4	9
	A. sinclairiana	4	6
A. sinclairiana	*A. deppeana*	5	9
	A. golfodulcensis	2	10
	A. gracilis	1	10
	A. terryae	1	10

a phylogenetic tree (Wagner 1969), but the presence of additional species may obscure these relationships (McDade 1990, 1992, 1997). This led us to turn towards quartet analysis using a Hybrid Detection Criterion (HDC), a technique that can be applied in combination with any method of phylogenetic analysis. HDC is defined in the following way: in quartets made up of one hybrid (AB), its two parents (A and B) and any other species (X), over all quartets the hybrid should never group with X but should be positioned next to A half of the time and next to B the other half of the time, or, if the method allows for hybridization events, with both A and B in every quartet. For each of the 17 data sets, we looked at all possible quartets {A, B, AB, X}, for a total of ten quartets per analysis. For each data set and each method, the number of quartets satisfying HDC was used as a measure of hybrid detection, yielding values ranging from zero (no quartet meet HDC) to ten (all quartets meet HDC). In order to detect hybrids through quartet analysis using the reticulogram algorithm, we fixed the number of reticulations to one; for a given quartet, HDC was met if the hybrid grouped with one parent and had a reticulation to the other (Figure 2.1). For split decomposition quartets, HDC was satisfied if the hybrid was placed between its two parents (Figure 2.2). The criterion was also not satisfied when split decomposition resulted in a tree. With both methods, star trees were treated as uninformative and did not satisfy HDC.

Overall, quartet analysis increased the hybrid detection rate compared to the direct application of reticulate methods (Table 2.1). While split decomposition permitted unambiguous identification of the majority of hybrids, reticulograms produced spurious results. Closer examination of individual reticulograms indicated that the single reticulation was added between one of the parents and the more distant species (X) in the vast majority of quartets; this situation violates HDC.

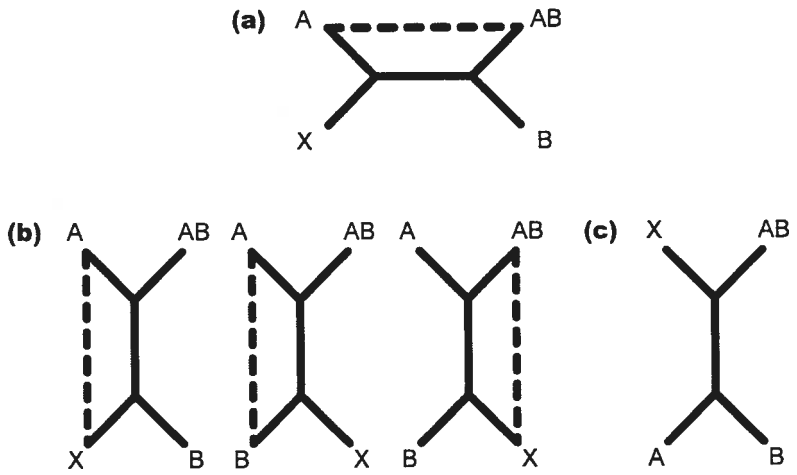


Figure 2.1 Illustration of the Hybrid Detection Criterion (HDC) for reticulogram quartets. (a) HDC is met if the hybrid (AB) is the sister taxa of either one of its parents (A or B) and has a reticulation to its other parent. HDC is not met if (b) the hybrid is the sister taxa of one of its parents, but does not have a reticulation to the other, or, (c) the hybrid is the sister taxa of the other species (X). The positions of parents A and B are interchangeable.

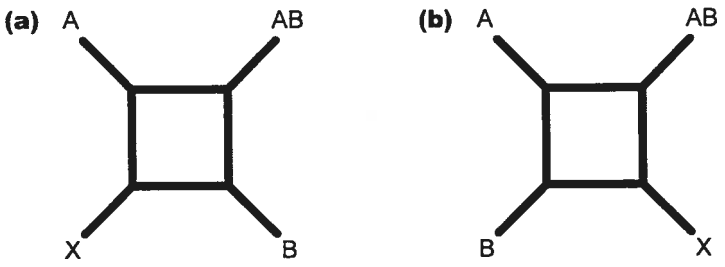


Figure 2.2 Illustration of the Hybrid Detection Criterion (HDC) for splitsgraph quartets. (a) HDC is met if the hybrid (AB) forms a pair of weakly compatible splits with its parents (A and B). (b) HDC is not met if the hybrid forms a weakly compatible split with the other species (X). The positions of parents A and B are interchangeable.

2.7. Discussion

Our results showed that neither reticulograms nor splitsgraphs allow for efficient hybrid detection when these methods are directly applied. Whereas split decomposition only identified hybrids between closely related parents, the main problem with the reticulogram approach appears to be caused by the goodness of fit criterion ($Q1$ or $Q2$) for adding reticulations. In trying to maximize the fit of a reticulogram to a distance matrix, it follows that the first reticulations usually connect the most distant pairs of nodes in the tree. Therefore, closely related parents and their hybrids will rarely be detected by this method. However, we believe that this could be corrected by adding some constraints to the algorithm, such that reticulations are never added over a certain distance threshold, or by prohibiting reticulations to or from internal nodes.

On the other hand, quartet analysis provided very good hybrid detection rates, depending on the algorithm selected. Reticulograms performed rather poorly and we believe, here again, that being able to define a distance threshold for adding a reticulation could correct this problem. Interestingly, quartet-based split decomposition proved to be very efficient to detect hybrids. By producing a pair of weakly compatible splits, this technique clearly shows whether a putative hybrid could be identified simply by its position between the two parent species. In the light of those results, we recommend to use quartet analysis of split decomposition graphs to accurately detect hybridization events in phylogenetic data sets. Future work will focus on developing a statistical test of the significance of HDC, and its application to other datasets as well as simulation work to assess its performance under different conditions.

2.8. Acknowledgements

The authors are grateful to L. A. McDade for providing the data set used in the present study. This work was made possible by a NSERC scholarship to O. Gauthier and by NSERC grant no. OGP0155251 to F.-J. Lapointe.

Chapitre 3 :

HYBRIDS AND PHYLOGENETICS REVISITED. A STATISTICAL TEST OF HYBRIDIZATION USING QUARTETS

Cet article a été accepté pour publication dans Systematic Botany :

Gauthier, O. & Lapointe, F.-J. 2006. Hybrids and phylogenetics revisited. A statistical test of hybridization using quartets.

3.1. Résumé

Un test statistique par permutations est présenté afin de tester la signification de l'indice de détection des hybrides (HDC) proposé par Gauthier & Lapointe (2002). Ce test permet d'évaluer une hypothèse d'hybridation donnée en considérant la probabilité d'observer une valeur aussi ou plus élevée de l' HDC en absence d'hybridation. Le test a été appliqué aux données de McDade (1990) sur des hybrides artificiels entre des espèces du genre *Aphelandra* (Acanthaceae). Les résultats montrent que, lorsque utilisé conjointement avec la décomposition des bipartitions, le test de l' HDC fournit des résultats fiables dans la plupart des cas. Par contre, l'utilisation des réticulogrammes entraîne une perte de puissance et peut mener à des résultats erronés. La procédure proposée ici peut également être utilisée dans le cadre d'analyses préliminaires de jeux de données portant sur des taxons pour lesquels l'hybridation est suspectée, mais pour lesquels aucune hypothèse formelle ne peut être formulée.

3.2. Abstract

The occurrence of reticulations in the evolutionary history of species poses serious challenges for all modern practitioners of phylogenetic analysis. Such events, including hybridization, introgression and lateral gene transfer, lead to evolutionary histories that cannot be adequately represented in the form of phylogenetic trees. Although numerous methods that allow for the reconstruction of phylogenetic networks have been proposed in recent years, the detection of reticulations still remains problematic. In this paper we present a Hybrid Detection Criterion (HDC) along with a statistical procedure that allows for the identification of hybrid taxa. The test assesses whether a putative hybrid is systematically intermediate between its postulated parents, with respect to the other taxa. The performance of the statistical method is evaluated using known hybrids of the genus *Aphelandra* (Acanthaceae) using two network methods: reticulograms and split decomposition graphs. Our results indicate that the HDC test is reliable when used jointly with split decomposition. On the other hand the test lacks power and gives misleading results when using reticulograms. We then show how the procedure can be used as a tool to identify putative hybrids.

3.3. Introduction

The problem of reticulate evolution is of great interest in phylogenetic analysis, and several authors have proposed ways of resolving and representing the phylogeny of taxa with a reticulate history (e.g. Sneath 1975; Nelson 1983; Wanntorp 1983; Jakobsen & Easteal 1996; Lapointe 2000; Makarenkov & Legendre 2000; Sang & Zhong 2000; Xu 2000; Holder *et al.* 2001; Gauthier & Lapointe 2002; Nakhleh *et al.* 2004). Although traditionally recognized as an important phenomenon among plants and bacteria (Rieseberg 1997; Doolittle 1999), reticulate evolution is also occurring within the animal kingdom (Bullini 1994; Dowling & Secor 1997). It has even been postulated to exist among pre-human lineages (Holliday 2003). Reticulate taxa are defined as the product of hybridization, introgression or lateral gene transfer between two independent lineages (Dowling & Secor 1997). Thus, they differ from so-called "normal" taxa that are the result of bifurcating events (Wagner 1969). Hybridization is defined as the reproduction between individuals stemming from populations, or groups of populations, that are distinguishable on the basis of one, or many, hereditary characters (Harrison 1990, 1993). This broad definition also includes crosses between members of different species. Hybridization can give rise to new species (Arnold 1992, 1997), or lead to a reinforcement or a breakdown of reproductive barriers among populations (Barton & Hewitt 1985). Introgression takes place when part of the genetic material from one species permeates into the genome of another, without the creation of a new species. Lateral, or horizontal, gene transfer generally refers to the exchange of genetic material, in the form of plasmids, between prokaryotes (Sonea & Matthieu 2000), but is also used in contrast to vertical gene transfer, which refers to standard genetic exchange along the branches of a phylogenetic tree.

One of three attitudes towards reticulate evolution has traditionally been adopted in phylogenetic analysis. Some researchers chose to ignore it, either because it was not considered important in the group under study, or simply because they lacked appropriate methods to take it into account (Skala & Zrzavy 1994). Others considered the presence of hybrids in phylogenetic data

sets as a nuisance and a possible source of spurious results; they suggested to remove the offending taxa prior to the analysis and then to re-graft them on the inferred tree (Wagner 1969, 1983; Posada & Crandall 2002). This approach, however, leaves open the problem of identifying a hybrid taxon and its parents. Finally, others have developed different strategies for the joint analysis and representation of "normal" and reticulate taxa with phylogenetic networks (e.g. Sneath 1975; Nelson 1983; Wanntorp 1983; Rieseberg & Morefield 1995; Jakobsen & Easteal 1996; Makarenkov & Legendre 2000; Sang & Zhong 2000; Holder *et al.* 2001; Bryant & Moulton 2002; Legendre & Makarenkov 2002; Nakhleh *et al.* 2004).

So far, no single method of network reconstruction has been shown to identify hybrid taxa unambiguously. However, it was suggested that reticulations could be better detected by methods that did not build networks or trees (Wiuf *et al.* 2001; Posada & Crandall 2001a; Posada 2002). Because phylogenetic networks contain multiple trees (Bandelt & Dress 1992; Nakhleh *et al.* 2004), a reticulate phylogeny can, at best, be partially recovered by traditional phylogenetic inference of evolutionary trees (Posada & Crandall 2002). In a previous paper (Gauthier & Lapointe 2002) we evaluated the performance of two network methods for the detection of known hybrids: reticulograms (Makarenkov & Legendre 2000; Legendre & Makarenkov 2002) and split decomposition (Bandelt & Dress 1992). We showed that neither method allowed for the unambiguous identification of hybrid taxa, and thus proposed a Hybrid Detection Criterion (HDC; Gauthier & Lapointe 2002) based on quartets analysis. The HDC states that a putative hybrid should systematically occupy an intermediate position between its postulated parents in a network, with respect to all other taxa in the data set. The number of quartets that satisfy the HDC is taken as a measure of support for the hybridization hypothesis. This index takes high values when the hypothesis is true, and low values otherwise. The application of HDC to first generation hybrids produced high values when used in conjunction with split decomposition, whereas lower values were obtained with reticulograms (Gauthier & Lapointe 2002). In order to make sure that such results were not

obtained by chance alone it is necessary to evaluate the significance of HDC values with a statistical test. Precisely, we must determine whether a given taxon is the hybrid of two parent taxa, the null hypothesis being that it is not. The statistical decision is made on the basis of the probability of observing a value of HDC that is greater than or equal to the observed value, when the null hypothesis is true. If this probability is smaller than a predetermined significance level (e.g. $\alpha = 0.05$), the test rejects the null hypothesis, and the alternative hypothesis of hybridization is accepted. The distribution of HDC values being unknown, a permutation procedure is used to compute the probability (Edginton 1995; Manly 1997). In the present paper, we describe the statistical test in details, and then apply it to McDade's (1990, 1992, 1997) data on *Aphelandra* (Acanthaceae) hybrids. Our objective is to address the following questions: (1) does the HDC significance test allows for the detection of first generation (F_1) hybrids when used jointly with split decomposition?; (2) does the HDC significance test allows for the detection of F_1 hybrids when used jointly with reticulograms?; (3) which one of these competing approaches performs better than the other?

3.4. Description of the test

3.4.1. Hypotheses

Consider a data set S with n taxa. If AB is a putative hybrid and A and B are its putative parents, $HDC_{A,B,AB}$ is computed by sequentially analyzing the $n-3$ quartets composed of these tree taxa $\{A, B, AB\}$ and any other taxon $\{X\}$ from the dataset. The value of $HDC_{A,B,AB}$ is exactly the number of quartets that agree with the hybridization hypothesis. It depends on the particular network reconstruction that is used. Although many network methods are available, we only focus on the same two that were evaluated by Gauthier & Lapointe (2002): reticulograms (Makarenkov & Legendre 2000; Legendre & Makarenkov 2002) and split decomposition graphs (or splitsgraph; Bandelt & Dress 1992). For a reticulogram with one reticulation,

HDC is satisfied if AB is the sister taxa of one of its putative parents and is linked to the other with a reticulation (Fig. 3.1a). A splitsgraph agrees with HDC only if AB forms a pair of weakly compatible splits with both its putative parents, not with X (Fig. 3.1b). All other reticulogram (Fig. 3.1c) and splitsgraph (Fig. 3.1d) topologies disagree with HDC. When the result is a tree partial support (Fig. 3.2a) or no support (Fig. 3.2c and 3.2d) is given for HDC.

The null hypothesis (H_0) states that AB is not the result of hybridization between A and B; it is associated with small values of $HDC_{A,B,AB}$. On the other hand, the alternative hypothesis (H_1) states that AB is a hybrid taxon of A and B; it is associated with high values of HDC_{AB}. Formally, these statistical hypotheses can be defined as:

$$H_0 : HDC_{A,B,AB} \leq HDC_{i,j,ij} \text{ for all } i,j \in S; i \neq j; \{i,j\} \neq \{A,B\} \qquad (3.1)$$

and

$$H_1 : HDC_{A,B,AB} > HDC_{i,j,ij} \text{ for all } i,j \in S; i \neq j; \{i,j\} \neq \{A,B\} \qquad (3.2)$$

where $HDC_{i,j,ij}$ is the value of the criterion for all other hybridization hypotheses; i.e. all hypotheses that do not postulate AB to be an hybrid of A and B. The test is one-tailed by definition. If the probability of observing a $HDC_{i,j,ij}$ value that is greater than or equal to $HDC_{A,B,AB}$ under H_0 is smaller than or equal to a given significance level (α), the null hypothesis is rejected; it is very likely that AB is the result of hybridization between taxa A and B. When this probability is larger than the nominal significance level α, the test fails to reject the null hypothesis; AB is not likely to be a hybrid of A and B.

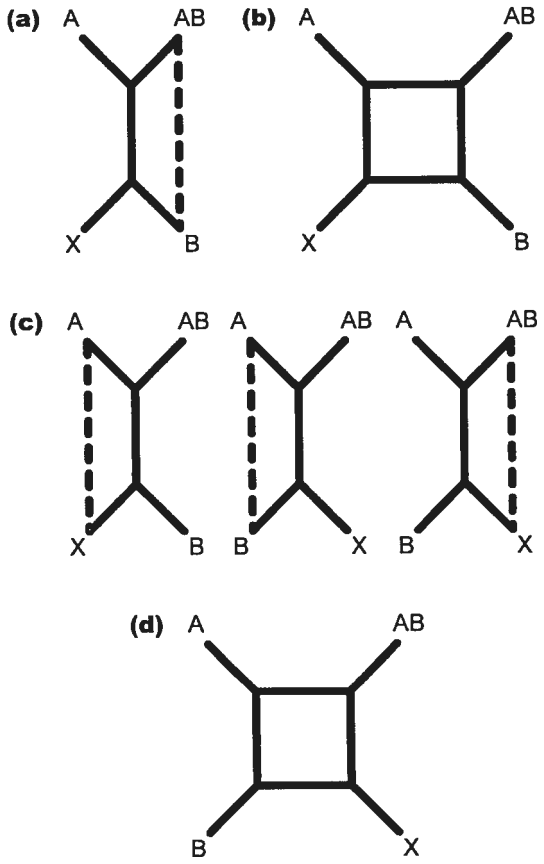


Figure 3.1 Network topologies for which the Hybrid Detection Criterion (HDC) is met with (a) a reticulogram and, (b) a splitsgraph. (c) Reticulograms that contradict HDC because the putative hybrid (AB) is not a sister taxa to either of the parents or because the reticulation not added between the hybrid and one of its parents. (d) Split decomposition graph that contradicts HDC, the putative hybrid does not form a set of weakly compatible splits with both its parents. Parents A and B are interchangeable. The dashed lines in reticulograms are reticulations.

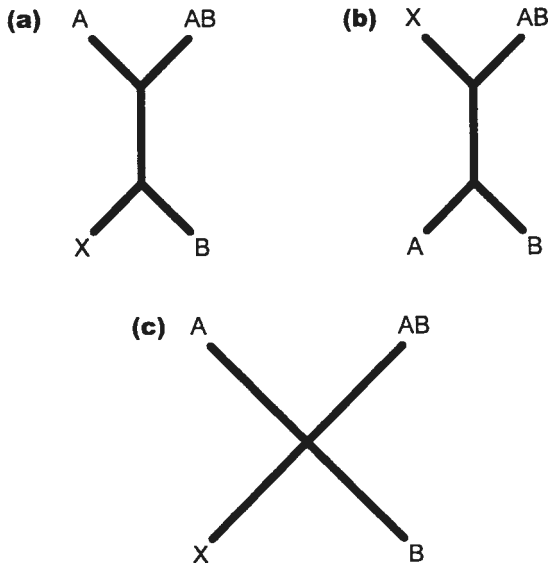


Figure 3.2 (a) A topology that offers no clear support for or against the Hybrid Detection Criterion (HDC), this case is arbitrarily given a weight of 0.5 in the calculations. (b) A topology for which the putative hybrid (AB) is not grouped with either of its parent; this contradicts HDC . (c) A star tree that provides no information on the relationships between the taxa, it neither supports nor contradicts HDC ; such cases are ignored in the computation of the statistic. Parents A and B are interchangeable.

3.4.2. Statistical procedure

The probability (p) of $HDC_{A,B,AB}$ under H_0 is obtained through a standard permutation procedure (Edginton 1995; Manly 1997). A permutation ($P(n,r)$) is an ordered set without repetition of a certain number (r), of objects taken from a set of n objects with $r \leq n$; in this case $r = 3$. There are $P(n,3) = \frac{n!}{(n-3)!}$ permutations of three objects drawn amongst n, but only $P(n,3)/2$ different hybridization hypotheses to be evaluated because the order of the parents, A and B, is not important. The test proceeds as follows:

(i) Compute the reference statistic $HDC_{A,B,AB}$

(ii) Pick three taxa at random

(iii) Assign the roles of putative parents (A and B) and putative hybrid (AB) at random among these three taxa

(iv) Compute $HDC^*_{A,B,AB}$, the permuted value of the statistic, on these three taxa and compare this value with $HDC_{A,B,AB}$; note if $HDC^*_{A,B,AB} \geq HDC_{A,B,AB}$ or not

(v) Repeat steps (ii) to (iv) a large number of times (k); for small datasets all possible cases can be enumerated

(vi) The probability of HDC_{AB} under H_0 is computed as the number of times that $HDC^*_{A,B,AB} \geq HDC_{A,B,AB}$ divided by the number of permutations (k+1):

$$p = \frac{number\ of\ times\ that\left[HDC^*_{A,B,AB} \geq HDC_{A,B,AB}\right]}{k+1} \qquad (3.3)$$

(vii) H_0 is rejected if $p < \alpha$ where α is the significance level as determined *a priori*.

Here, $HDC_{A,B,AB}$ and $HDC^{*}_{A,B,AB}$ are computed using splitsgraphs or reticulograms, but other suitable network methods may be used instead to detect hybridization events (see Lapointe 2000; Posada & Crandall 2001b). Every hypothesis requires the analysis of $n-3$ quartets (Gauthier & Lapointe 2002). A program to carry out these calculations is available upon request from the authors.

3.5. Application

3.5.1. Statistical analysis

The different hybrid taxa included in McDade's (1990, 1992, 1997) studies were submitted to the HDC test (Table 3.1). As in Gauthier & Lapointe (2002) 17 datasets, each composed of 12 parent taxa and one artificial hybrid, were tested independently using split decomposition and reticulograms. These hybridization hypotheses thus correspond to the crosses generated by McDade (1990). For each test, all possible permutations were enumerated ($P(13,3)/2 = (13!/(13-3)!)/2 = 858$ cases).

The original dataset included 50 morphological characters, coded as binary, semi-quantitative and multi-class variables (McDade 1984, 1990). These data were then recoded as 95 additive binary characters (Sneath & Sokal 1973) by McDade (1997). Overall, $6.39\,\%$ of the cells in the global matrix were scored as missing, while the percentage of missing cells ranged from $0.00\,\%$ to $36.84\,\%$ ($\bar{x} = 10.90\,\%$) for the 17 individual datasets. Given the important proportion of empty cells, similarities between pairs of taxa were computed used Gower's similarity coefficient (Gower 1971) that accounts for missing data, and then converted to distances.

3.5.2. Results

The HDC test, used in conjunction with reticulograms, was able to reject the null hypothesis in 7 cases out of 17, for a success rate of $41\,\%$ (Table

3.1). On the other hand, the use of split decomposition provided the correct result in 14 cases out of 17, for a success rate of $82\,\%$ (Table 3.1). Moreover, the probability of not rejecting the null hypothesis when it is wrong (type II error) is more variable and twice as large for reticulograms ($\bar{x} = 0.097$; $0.002 - 0.361$) then for split decomposition ($\bar{x} = 0.049$; $0.016 - 0.232$): the test exhibits reduced power when using reticulograms. Of the 17 hybrids tested: six were correctly identified by both techniques; eight were identified by split decomposition only; one by reticulograms only; and two by neither method (Table 3.1).

3.5.3. Discussion

The proposed significance test of the Hybrid Detection Criterion allows for the identification of F_1 hybrids. This statistical procedure was shown to be twice as effective when split decornposition is used. However, the test did not produce significant results for all hybrids; and some hybrids that are detected with reticulograms elude detection by split decomposition. In order to understand the conditions affecting the result of the test and its behavior, it is necessary to perform a character analysis of the hybrids relative to their parents.

McDade (1990) defined several patterns of character states shared by parent taxa and their hybrids, and the frequency of these patterns influences the performance of the two methods under comparison. The three character patterns observed here are the following: (1) the hybrid has the same character state as both parents; (2) the hybrid has the same character state as one parent only; and (3) the hybrid has a different character state then both parents. HDC is based on the intermediacy of the hybrid with respect to its parents. Thus, we expect that a large number of character patterns of type 3 will affect the HDC test negatively. The largest proportion of type 3 pattern was observed in hybrids that were not identified by either methods ($5.79\,\%$) or by reticulograms only ($6.32\,\%$). However, such characters were practically absent from hybrids that were

Table 3.1 Values of the Hybrid Detection Criteria statistic value (HDC) and their associated probabilities (p) for quartet analyses using reticulograms and splitsgraphs ($*: p \leq 0.05$; $**: p \leq 0.01$). Hybrids where created by crosses between *Aphelandra campanensis* Durkee, *A. deppeana* Schltr. & Cham, *A. golfodulcensis* McDade, *A. gracilis* Leonard, *A. leonardii* McDade, *A. panamensis* McDade, *A. sinclairiana* Nees, *A. storkii* Leonard, and *A. terryae* Standley by McDade (1990).

Parents		Reticulograms		Splitsgraphs	
Ovulate	Staminate	*HDC*	*p*	*HDC*	*p*
	A. panamensis	5	0.019*	10	0.016*
A. deppeana	*A. sinclairiana*	5	0.024*	10	0.016*
	A. storkii	4	0.066	9	0.050*
	A. deppeana	3	0.057	10	0.016*
A. golfodulcensi	*A. leonardii*	3	0.059	9	0.049*
	A. sinclairiana	7	0.063	10	0.016*
	A. campanensis	4	0.361	9	0.050*
A. leonardii	*A. golfodulcensi*	1	0.145	7	0.156
	A. sinclairiana	6	0.002**	10	0.017*
	A. deppeana	6	0.026*	10	0.016*
A. panamensis	*A. golfodulcensi*	6	0.333	9	0.050*
	A. leonardii	4	0.029*	9	0.052
	A. sinclairiana	4	0.143	6	0.232
	A. deppeana	5	0.138	9	0.049*
A. sinclairiana	*A. golfodulcensi*	2	0.141	10	0.017*
	A. gracilis	1	0.009**	10	0.020*
	A. terryae	1	0.029*	10	0.016*

identified by both methods ($0.88\,\%$) or by split decomposition alone ($2.24\,\%$). There is a monotone positive correlation (Kendall 1938) between the sum of distances between the hybrid and each of its parents, that is proportional to the number of type 3 characters, and the probability of $HDC_{A,B,AB}$ under H_0 when splitsgraphs are used ($\tau = 0.497$, $p < 0.01$), but not when reticulograms are used ($\tau = 0.082$, $p > 0.05$).

The results presented here show that reticulograms cannot detect the intermediate position of hybrids between their parents. Moreover, they seem to pick out signals that contradict intermediacy. For example, the hybrid between *Aphelandra panamensis* and *A. leonardii* presents the greatest proportion of type 3 characters ($6.32\,\%$), but it is still identified as a hybrid when the test is conducted using reticulograms (Table 3.1). As suggested by Gauthier & Lapointe (2002), it appears that the first reticulations are always added between the most dissimilar nodes in the tree. The bad performance of reticulograms could also be explained by a lack of robustness to missing data. As a matter of fact, the eight hybrids that were only identified using splitsgraphs show higher proportions of missing data ($13.03\,\%$) than those identified by both methods ($7.54\,\%$).

The need for a statistical test, rather than simply looking at HDC values, is illustrated by the hybrids between *A. sinclairiana* and *A. gracilis*, and between *A. sinclairiana* and *A. terryae*. In both of these cases, the HDC computed with reticulograms equals one, a very small value, but nonetheless a significant one. An identical HDC value, but clearly non significant, was obtained for the hybrid between *A. leonardii* and *A. golfodulcensis*. On the contrary, large values (e.g. 7 for the hybrid between *A. golfodulcensis* and *A. sinclairiana*) were not significant. As proposed by Gauthier & Lapointe (2002), HDC is a measure of the support for a given hybridization hypothesis. When used in conjunction with split decomposition, the statistical test presented here enables to further assess the probability of this hypothesis.

In the last of her papers on the effect of hybrids on phylogenetic analysis, McDade (1997) concluded that neither parsimony nor distance-based methods would be of any help to identify hybrid taxa. Such classical approaches were inherently incapable of inferring reticulate relationships and the inclusion of hybrids in the analysis did not reveal any discernable pattern or affect goodness of fit measures such as the retention index. Her results also indicate that the best approach might be to look for taxa with intermediate distances to its two parents, but she offers no way of doing this other than looking at distances or analyzing ordination plots. The HDC statistic can not only detect hybrids as intermediates between their parents in the overall dataset, but also, by breaking the problem down to quartets of taxa, searches for systematic intermediacy of the taxa with regard to each of the other taxa individually.

Besides allowing for testing the significance of HDC, and assessing hybridization hypothesis, the procedure presented here can be used to perform preliminary and exploratory analysis on a dataset for which hybridization is hypothesized. The enumeration of all possible hybridization hypotheses provides a global picture of the relationships among the taxa at hand. This approach can thus be useful to analyze taxa for which hybridization is suspected to have played a historical role; for example, with taxa that are shown to hybridize in the wild or in the laboratory, but for which it is impossible to formulate specific hybridization hypotheses. By enumerating all possible scenarios, without *a priori* knowledge, it is possible to identify probable hybridization events. Using such a procedure with the nine hybrids that scored an HDC of 10 with split decomposition leads to the retention of 15 to 18 different most probable hybridization hypotheses. The true hypotheses are all found among these.

The HDC test also allows assessing that any given taxon is a hybrid, without specifying its parents: the test then allows for the identification of the most probable parents. With the exception of the hybrid between *Aphelandra sinclairiana* and *A. gracilis*, the parents of all hybrids that scored an HDC of 10 could have been identified unambiguously in this manner.

Using the same rationale, it is also possible to perform the test by specifying the putative hybrid and only one of its putative parents. This leads to the proper identification of the other parent in 24 out of 34 cases (71%). In three other cases, it was statistically impossible to choose between the proper parent and another putative parent. Finally, given two parents, the test also allows the identification of the most probable hybrid. In the present case, the correct hybrid is selected in 16 cases out of 17 (94%), the only exception being the hybrid between *A. panamensis* and *A. golfodulcensis*. In 14 of these 17 cases (82%) the identification is unambiguous.

Proper care must be taken when using the procedure in any of the exploratory ways described here. It would be unsound to consider finding the hybridization hypothesis that maximizes HDC as sufficient evidence to invoke actual hybridization because this would, undoubtedly, lead to many erroneous conclusions. In these conditions, it is necessary that hypothesis identified with this procedure be further tested using the HDC test on a new dataset.

The hybrid detection criterion computed with split decomposition, along with its associated statistical test, represents a promising approach for the study of reticulate patterns of evolution such as those resulting from hybridization events. The procedure presented here can be used to test specific hypotheses, or to conduct preliminary and exploratory analyses. While direct reconstruction of phylogenetic networks is not yet efficient and accurate, identification of hybrid taxa can be used to circumvent the problem. Such taxa can be removed prior to phylogenetic analysis and grafted in their proper position on the inferred tree. Future work will focus on a further validation of the procedure with different datasets and using simulations.

3.6. Acknowledgements

The authors are grateful to L. A. McDade for providing the data set used in the present study and to V. Makarenkov for providing the source code for his reticulogram reconstruction software. The authors also wish to thank members of the Laboratoire d'Écologie Moléculaire Et Évolution (LEMEE) for their constructive comments on a previous version of this article. This work was made possible by NSERC and FQRNT scholarships to O. Gauthier and by NSERC grant no. OGP0155251 and FQRNT grant no. PR88559 to F.-J. Lapointe.

Chapitre 4 :

PHYLOGENY OF THE ROCK WALLABIES, *PETROGALE* (MARSUPIALIA: MACROPODIDAE), PART II: DETECTION OF HYBRIDISATION AMONG MACROPODINES

Cet article sera soumis prochainement :

Kirsch, J. A. W., O. Gauthier, A. Campeau-Péloquin, M. D. B. Eldridge, & F.-J. Lapointe. Phylogeny of the rock wallabies, *Petrogale* (Marsupialia: Macropodidae), part II: Detection of hybridisation among macropodines. *Sera soumis à Australian Journal of Zoology.*

4.1. Résumé

Les relations phylogénétiques entre les wallabies des rochers, *Petrogale* (Marsupialia: Macropodidae), sont difficiles à résoudre. En raison des nombreux cas documentés d'hybridation interspécifique en nature et de la facilité avec laquelle des croisements peuvent êtres effectués en laboratoire, l'introgression et la spéciation suite à l'hybridation ont été invoquées comme des causes possibles de ces difficultés. Dans le cadre de cette étude, une approche phylogénétique est utilisée pour identifier des hybrides de *Petrogale* d'origine connue. Le test du critère de détection des hybrides (HDC) est appliqué à des données d'hybridation ADN-ADN pour 15 espèces, deux hybrides naturels résultants de croisements en captivité, et deux hybrides artificiels issus des mêmes paires d'espèces parentales. Si les hybrides naturels ne sont pas détectés par l' HDC , les hybrides artificiels, qui sont des mélanges équimolaires d'extraits parentaux, sont aisément identifiés. De plus, des graphes de décomposition des bipartitions construits à partir de cinq paires d'hybrides naturels et artificiels, incluant ceux évalués à l'aide de l' HDC , ainsi que de leurs parents, montrent, à une exception près, que ces deux types d'hybrides ne se regroupent pas ensemble. Étant donné que l' HDC assume que l'hybride est en position intermédiaire entre ses parents potentiels, il est probable que la recombinaison génétique inégale, ou un autre type de recombinaison, affecte les résultats du test. Ces conclusions nous permettent de douter de la possibilité de détecter des hybrides de *Petrogale* par une approche phylogénétique.

4.2. Abstract

Phylogenetic relationships among rock-wallabies, *Petrogale* (Marsupialia: Macropodidae), have proven difficult to resolve. Given the documented inter-specific hybridisation in the wild and the ease with which hybrids can be bred in captivity, introgression and hybrid speciation are likely explanations for these difficulties. In this paper, an attempt is made at using a phylogenetic approach to identify *Petrogale* hybrids of known origin. The Hybrid Detection Criterion (HDC) test is applied to DNA-DNA hybridisation data for 15 full species, two natural yard-bred hybrids, and two artificial hybrids from the same pairs of parental species. While the yard-bred hybrids elude detection with this technique, the artificial hybrids, consisting of equimolar mixture of parental extracts, are easily identified. Moreover, splitsgraphs constructed from five pairs of natural and artificial hybrids, including those evaluated with HDC, and their parents show that, in all cases but one, these two kinds of hybrids do not groups together. Because the HDC assumes an intermediate phylogenetic position of the hybrid between its postulated parents, it is likely that unequal crossing-over, or another recombination event, affects the results of the test. These conclusions cast some doubt on the possibility of accurately detecting *Petrogale* hybrids with a phylogenetic approach.

4.3. Introduction

The difficulty of resolving relationships among *Petrogale* species has been attributed, at least in part, to introgression over contact zones and possibly even to the hybrid origin of some taxa (Eldridge & Close 1993; Campeau-Péloquin *et al.* 2001). Given the geographic distribution of *Petrogale* taxa (Fig. 4.1), where several species or 'races' form an almost continuous linear (parapatric) series along the eastern coast, one might expect - and indeed does find - that the majority of natural hybrids or cases of introgression are found among members of this group, whose mutual relationships have proved difficult to resolve (Eldridge & Close 1993; Eldridge 1997; Campeau-Péloquin *et al.* 2001). At the same time, hybrids have also been bred among allopatric but karyotypically more primitive Western Shield taxa (Eldridge & Pearson 1997). The ease with which interspecific hybrids can be bred in captivity enhances the inference of such natural occurrences in the wild (Close & Bell 1997). Thus, in cases where unusual or complex chromosomal rearrangements are found to have taken place more than once (Eldridge *et al.* 1990), hybridisation may be a more likely explanation than repeated and identical transformations.

The *representation* of hybridisation in a phylogeny is relatively straightforward, by the placement of hybrid taxa at the nodal point joining the two parental taxa; or by anastomoses of lineages, creating a network (Nelson 1983; Lapointe 2000; Posada & Crandall 2001b). More problematic, and probably the main reason for the neglect of anastomosis, is the *detection* of hybridisation to begin with. One reason is that phylogenetic algorithms in common use, such as those using parsimony, maximum likelihood or minimum-evolution criteria, produce trees, and thus, do not permit reticulations. Wholesale transfer or combination of large portions of the genome between hybridising species, as in the origin of eucaryotes, should be easily detectable (Katz 1999). However, closely related species differ in only a small percentage of their genomes or characters, and in these instances the chief difficulty remains that the shared presence of a few 'foreign' features in two taxa might be confounded with

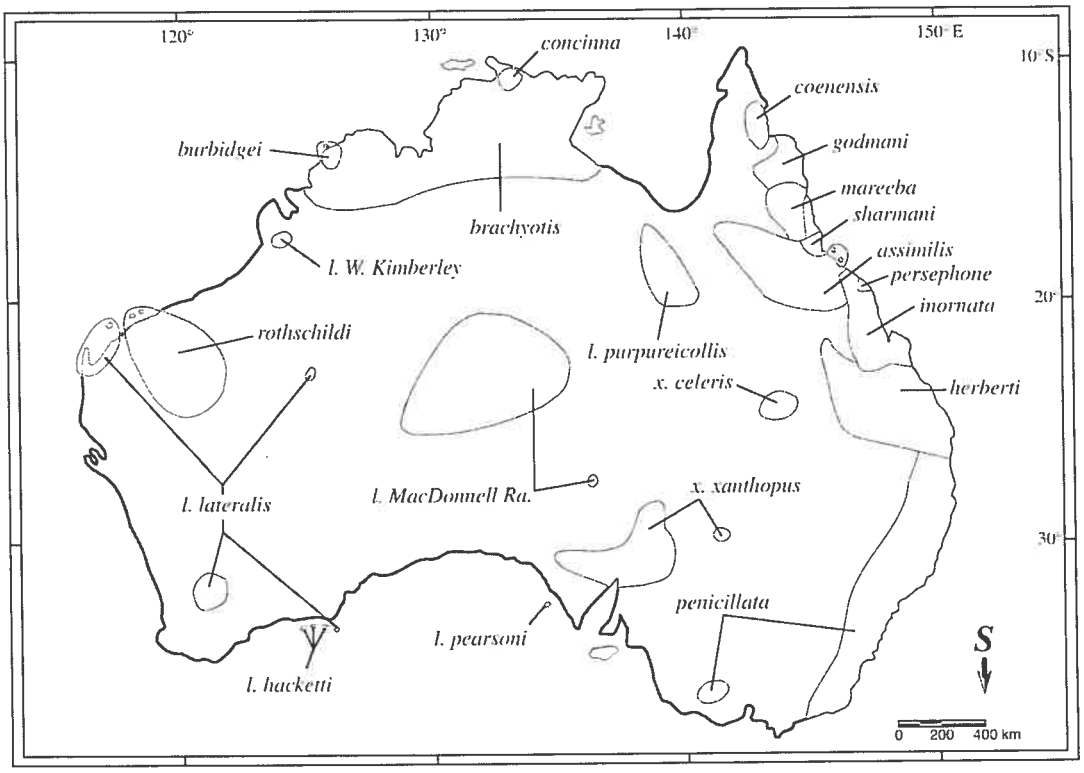


Figure 4.1 Distribution of *Petrogale* taxa in Australia. Modified from Eldridge & Close (1993).

homoplasy,

Several methods specifically designed for the detection of hybrids or the representation of conflicting signal have been devised (Lapointe 2000; Posada & Crandall 2001b; Nakhleh *et al.* 2004), but rarely tested for their performance against known and suspected instances of introgression or hybridisation (Gauthier & Lapointe 2002; Legendre & Makarenkov 2002). Gauthier & Lapointe (2002) compared two distinct approaches, using a set of 17 known hybrids among 12 species of the plant genus *Aphelandra*. Whereas a reticulogram is obtained by adding one or more edges to an additive tree-graph to improve the fit to the data (Makarenkov & Legendre 2000), split decomposition is used to detect conflicts in the data, which are then represented by weakly compatible splits in a splitsgraph (Bandelt & Dress 1992). Gauthier & Lapointe (2002) found that neither method was very effective when applied to large data sets. However, they defined an Hybrid Detection Criterion (HDC) based on the analysis of quartets of taxa, and were able to correctly identify 14 of the 17 hybrids when using this approach jointly with split decomposition. In a second paper, Gauthier & Lapointe (2006) designed a permutation test to assess the probabilities of their results.

In a previous article (Campeau-Peloquin *et al.* 2001), we presented a phylogeny of *Petrogale* based on single-copy DNA/DNA hybridisation, remarking that this effort was part of a wider study investigating the detection and possible role of hybrids in the evolution of species within this genus. Here we elaborate on further experiments with rock wallabies not reported in the earlier paper, involving such hybrids. We apply split decomposition and the Hybrid Detection Criterion (HDC) to a series of hybrids of the rock wallabies examined by DNA hybridisation to assess the performance of the method. One special feature of DNA hybridisation is that it permits fabrication of artificial hybrids from a combination of parental extracts, and a direct comparison of the behaviour of these constructs with that of natural hybrids. Our *expectation* was that the two types of hybrids would behave similarly; our *hope* was that by their characterisation vis-à-vis each other and their parental taxa we might be in a position to detect other, suspected hybrids among the more inclusive group of

Petrogale species, or uncover hints of hybridisation in the more remote past, or specify a general means of hybrid-detection in any taxonomic group.

4.4. Methods

Tissue samples and their extracts were identical to those listed in our earlier paper on rock wallabies (Campeau-Peloquin *et al.* 2001), with the addition of the several hybrids noted in Table 4.1. All protocols for extraction, labelling, hybridisation, and data analysis were similarly as earlier specified, again except for the exceptional treatment of information pertaining to the natural and artificial hybrids. Natural hybrids were yard-bred individuals of known parentage, while artificial hybrids consisted of equimolar mixtures of parental extracts. The hybridisation data were recorded as ΔT_ms and are shown in Tables 4.2 and 4.3a-e. Tree-construction was carried out using FITCH (Felsenstein 2005) with the global branch-swapping (G), subreplicate (S), and Cavalli-Sforza & Edwards (1967) or P = 0 options enabled. Splitsgraphs were calculated with SplitsTree4.0 (Huson & Bryant 2006) from the series of 4x4 matrices, and an original program was used for calculating HDC values. Statistical distributions were generated from the data themselves by permutation to provide estimates of probabilities for the resulting values (Gauthier & Lapointe, 2006)

Split decomposition proceeds by considering each four-taxon case derivative from among the *n* taxa in a dataset. In such an instance, there are three possible resolved topologies representing the different ways of splitting the taxa. Because the optimisation criterion for the tree is minimum-length, the topology implied by the shortest split is considered least supported, and the other two as giving conflicting support for the remaining weakly compatible splits. Such conflict is represented as a central 'box' joining the four taxa, and which may be divided in two ways. A splitsgraph displays the weakly compatible splits for all taxa in a data set, with boxes placed at the appropriate junctures indicating possible instances of hybridisation.

Table 4.1 Registry numbers of parental *Petrogale* species and their natural hybrids. Artificial hybrids utilised extracts from specimens listed in our earlier paper.

Taxon	Field Numbers
P. persephone	S-691
P. xanthopus	AZ-1
Natural hybrid	S-870
P. assimilis	S-442
P. penicillata	S-672
Natural hybrids	S-956; S-1003
P. inornata	S-462
P. purpureicollis	S-888
Natural hybrid	S-995
P. godmani	X-3
P. purpureicollis	X-1
Natural hybrid	X-9
P. herberti	S-889 (of S-975); S-1087 (of S-1098 & S-1247)
P. penicillata	S-775 (of S-975); S-1028 (of S-1098 & S-1247)
Natural hybrids	S-975; S-1098; S-1247

Table 4.2 ΔT_ms among 19 *Petrogale* taxa, including two natural and two artificial hybrids; number of comparisons = 1051; average standard deviation (SD) = ±0.17; correlation of SDs with distance = 0.18. Columns are tracers, identified for the most part by the first four letters of the specific epithet. Names of hybrids and parental taxa are shown in ***bold italics***. First line of each cell lists average ΔT_m except for the homologues (**boldfaced**), where actual mean melting-temperature is given to permit comparison of tracer qualities; second line gives SD where applicable and number of replicates, separated by a solidus; na = not applicable.

(continued next page)

Table 4.2 (continued)

	Brac	*Burb*	*Herb*	*Late*	*Mare*	*Shar*	***Xant***	***Pers***	*Godm*	*Inor*
P. brachyotis	**84.44**	0.92	1.18	1.16	1.08	1.21	1.18	1.12	1.28	1.08
	0.23/6	0.15/3	0.17/3	0.04/2	0.06/3	0.33/3	0.12/3	0.07/3	0.25/3	0.03/2
P. burbidgei	0.89	**84.45**	1.11	1.27	1.15	1.20	1.22	1.18	1.17	1.01
	0.19/3	0.15/6	0.00/2	0.14/3	0.08/3	0.18/3	0.08/3	0.17/3	0.12/2	0.10/3
P. herberti	1.57	1.52	**83.86**	0.89	0.41	0.51	1.33	1.19	0.65	0.26
	0.13/3	0.13/3	0.08/6	0.07/3	0.04/3	0.06/3	0.22/3	0.07/3	0.11/3	0.14/3
P. lateralis	1.32	1.34	0.59	**84.17**	0.52	0.72	1.04	0.90	0.65	0.46
	0.10/3	0.10/2	0.09/3	0.08/6	0.04/3	0.11/2	0.06/3	0.13/3	0.30/3	0.26/3
P. mareeba	1.32	1.46	0.18	0.75	**83.29**	0.20	1.03	0.92	0.20	-0.01
	0.21/2	0.25/3	0.09/3	0.16/3	0.07/6	0.06/3	0.11/3	0.08/3	0.04/3	0.12/3
P. sharmani	1.62	1.44	0.19	0.86	0.09	**84.44**	1.25	1.14	0.44	-0.05
	0.25/3	0.30/3	0.02/3	0.12/3	0.14/3	0.26/6	0.31/2	0.32/3	0.33/3	0.17/3
P. xanthopus	1.72	1.45	1.05	1.14	1.16	1.19	**84.10**	1.06	1.13	1.19
	0.11/3	0.07/3	0.08/3	0.04/3	0.02/3	0.12/3	0.10/5	0.08/2	0.09/3	0.02/3
P. persephone	1.69	1.76	1.35	1.50	1.53	1.49	1.55	**83.64**	1.42	1.20
	0.07/2	0.18/3	0.13/3	0.25/2	0.03/3	0.21/3	0.30/3	0.11/6	0.11/3	0.13/3
P. godmani	1.58	1.31	0.33	0.93	0.30	0.38	1.20	1.13	**84.09**	0.10
	0.07/3	0.21/3	0.13/3	0.07/3	0.09/3	0.03/3	0.08/3	0.15/3	0.09/12	0.10/3
P. inornata	1.55	1.47	0.22	0.88	0.28	0.31	1.15	1.03	0.50	**83.80**
	0.22/3	0.18/3	0.02/3	0.12/3	0.06/3	0.03/3	0.09/3	0.05/3	0.19/3	0.35/12
P. purpurei	1.57	1.27	0.36	0.68	0.51	0.40	1.01	1.13	0.45	0.53
	0.22/3	0.27/3	0.08/3	0.09/3	0.06/3	0.06/3	0.08/3	0.27/3	0.12/8	0.28/9
P. penicillata	1.35	1.36	-0.10	0.74	0.34	0.44	1.12	0.96	0.57	-0.01
	0.16/3	0.23/3	0.27/3	0.12/3	0.22/3	0.25/3	0.19/3	0.16/3	0.27/3	0.27/3
assi.* X *peni.	1.60	1.32	0.28	0.97	0.45	0.32	1.11	1.09	0.50	0.25
(natural)	na/1	na/1	na/1	na/1	na/1	na/1	na/1	na/1	na/1	na/1
P. assimilis	2.27	2.08	1.05	1.65	0.97	1.14	2.00	1.93	1.19	1.10
	0.24/3	0.14/3	0.14/3	0.11/3	0.09/3	0.24/3	0.14/3	0.18/3	0.11/3	0.17/3
P. rothschildi	2.51	2.36	1.43	1.60	1.80	1.92	2.36	1.98	1.59	1.72
	0.01/2	0.23/2	0.01/2	0.19/2	0.08/2	0.04/2	0.06/2	0.08/2	0.02/2	0.18/2
P. coenensis	1.77	1.47	0.48	1.05	0.45	0.55	1.37	1.28	0.89	0.37
	0.13/3	0.44/3	0.00/2	0.13/3	0.07/3	0.06/3	0.09/3	0.03/3	0.59/2	0.19/3
persXxant.	2.99	1.82	2.26	2.28	2.24	1.95	2.60	1.29	1.92	1.85
(natural)	1.04/2	0.49/2	0.85/2	0.56/2	na/1	na/1	1.93/2	0.04/2	0.10/2	0.18/2
***assi.*Xpeni.**	1.63	1.46	0.19	0.89	0.42	0.42	1.24	1.25	0.55	0.03
(artificial)	0.10/3	0.22/3	0.06/3	0.07/3	0.21/3	0.10/3	0.04/3	0.09/3	0.13/3	0.13/3
pers.* X *xant.	1.58	1.46	1.17	1.49	1.34	1.22	0.75	0.49	1.22	1.12
(artificial)	0.10/3	0.27/3	0.08/3	0.37/3	0.02/3	0.02/3	0.20/2	0.40/3	0.17/3	0.20/3

Table 4.2 (continued)

	Purp	***Peni***	***AsPeN***	***Assi***	*Roth*	*Coen*	***PeXaN***	***AsPeA***	***PeXaA***
P. brachyotis	1.31	2.28	1.30	0.93	0.16	0.39	-0.36	0.65	na
	0.19/3	0.18/2	0.33/3	0.13/3	0.10/3	0.09/3	0.10/2	0.08/3	
P. burbidgei	1.37	1.53	1.18	0.85	0.18	0.60	-0.32	0.95	na
	0.12/3	0.20/3	0.29/3	0.21/3	0.10/3	0.14/3	0.09/3	0.19/3	
P. herberti	0.51	0.52	0.33	-0.38	-0.20	0.04	-0.27	-0.15	na
	0.03/3	0.27/3	0.17/3	0.38/3	0.09/3	0.12/3	0.24/3	0.23/3	
P. lateralis	0.84	0.76	0.64	-0.09	-0.49	0.17	-0.41	0.09	na
	0.21/3	0.04/2	0.08/3	0.41/3	0.13/3	0.22/3	0.17/3	0.16/3	
P. mareeba	0.34	0.42	0.10	-0.48	-0.32	-0.26	-0.39	-0.19	na
	0.08/3	0.08/3	0.05/3	0.41/3	0.10/3	0.22/3	0.34/3	0.21/3	
P. sharmani	0.47	0.54	0.22	-0.90	-0.36	-0.33	-0.42	-0.30	na
	0.09/3	0.16/3	0.18/3	0.35/3	0.07/3	0.13/3	0.03/3	0.08/3	
P. xanthopus	1.15	1.53	0.99	0.76	0.07	0.56	-0.86	0.66	0.33
	0.04/3	0.31/3	0.13/3	0.32/2	0.06/3	0.48/3	0.10/3	0.28/3	0.13/5
P. persephone	1.47	1.60	1.36	1.08	0.29	0.98	-0.91	0.81	0.07
	0.15/3	0.29/3	0.09/3	0.21/3	0.14/3	0.23/3	0.10/3	0.48/3	0.27/6
P. godmani	0.56	0.75	0.22	-0.53	-0.22	-0.35	-0.35	-0.14	na
	0.21/9	0.26/3	0.06/3	0.30/3	0.13/3	0.23/3	0.18/3	0.28/3	
P. inornata	0.81	0.59	0.28	-0.06	-0.30	-0.11	-0.37	-0.32	na
	0.12/9	0.18/3	0.17/3	0.47/3	0.06/3	0.16/3	0.13/3	0.05/3	
P. purpurei	**83.99**	0.70	0.41	0.17	-0.29	-0.21	-0.48	0.09	na
	0.16/15	0.12/3	0.21/3	0.10/3	0.08/3	0.11/3	0.14/3	0.26/3	
P. penicillata	0.55	**84.52**	0.11	-0.10	-0.37	-0.20	-0.52	0.01	na
	0.11/3	0.30/5	0.16/3	0.23/3	0.20/3	0.11/2	0.16/3	0.16/3	
assi.* X *peni.	0.92	0.51	**84.06**	-0.35	-0.40	-0.42	-0.04	0.46	na
(natural)	na/1	na/1	0.22/4	na/1	na/1	na/1	na/1	na/1	
P. assimilis	1.28	1.60	1.06	**82.49**	0.48	0.59	0.43	0.74	na
	0.10/3	0.15/3	0.32/3	0.53/6	0.07/3	0.08/3	0.21/3	0.14/3	
P. rothschildi	1.70	2.21	1.60	1.08	**82.51**	0.89	0.53	1.32	na
	0.04/2	0.08/2	0.08/2	0.08/2	0.13/5	0.06/2	0.05/2	0.37/2	
P. coenensis	0.89	0.85	0.44	0.11	0.14	**83.07**	-0.23	0.32	na
	0.16/3	0.21/2	0.23/3	0.19/3	0.30/3	0.33/6	0.07/3	0.11/3	
persXxant.	2.02	2.23	2.35	1.32	2.35	1.27	**81.57**	1.72	0.10
(natural)	0.14/2	0.05/2	0.32/2	0.02/2	0.89/2	0.39/2	0.25/5	0.47/2	0.14/7
***assi.*Xpeni.**	0.59	0.53	0.22	-0.16	-0.15	0.15	-0.36	**80.71**	na
(artificial)	0.09/3	0.24/3	0.24/3	0.04/3	0.13/3	0.17/3	0.12/3	0.27/6	
pers.* X *xant.	1.16	1.54	1.28	0.76	0.16	0.42	-1.03	0.89	**79.40**
(artificial)	0.09/3	0.03/3	0.17/3	0.62/3	0.25/3	0.20/3	0.15/2	0.10/3	0.21/6

Table 4.3 Four-taxon matrices (a-e) of ΔT_ms among parental *Petrogale* species and their natural and artificial hybrids. Conventions mostly as for Table 4.2, but in (e) four cells were reflected from their reciprocals (*italicised*) and one pair of reciprocals was estimated (<u>underlined</u>); homologous distances of the two hybrids are by definition zero. Negative values were set equal to zero.

(a)	*P. xanthopus*	*P. persephone*	Natural hybrid	Artificial hybrid
P. xanthopus	0.00/5	1.06/2	0/3	0/5
P. persephone	1.55/3	0.00/6	0/3	0/6
Natural hybrid	2.60/2	1.29/2	0.00/5	0.10/7
Artificial hybrid	0.75/2	0.49/3	0/2	0.00/6

(b)	*P. penicillata*	*P. assimilis*	Natural hybrid	Artificial hybrid
P. penicillata	0.00/5	0/3	0.11/3	0.01/3
P. assimilis	1.60/3	0.00/6	1.06/3	0.74/3
Natural hybrid	0.51/1	0/1	0.00/4	0.46/1
Artificial hybrid	0.53/3	0/3	0.22/3	0.00/6

(c)	*P. inornata*	*P. purpureicollis*	Natural hybrid	Artificial hybrid
P. inornata	0.00/6	0.92/6	0/6	0.16/6
P. purpureicollis	0.36/6	0.00/6	0/6	0/6
Natural hybrid	0.72/6	1.16/6	0.00/6	0.82/6
Artificial hybrid	0/6	0.52/6	0/6	0.00/6

(d)	*P. godmani*	*P. purpureicollis*	Natural hybrid	Artificial hybrid
P. godmani	0.00/6	0.55/6	0/6	0.07/6
P. purpureicollis	0.58/6	0.00/6	0/6	0/6
Natural hybrid	0.80/6	0.86/6	0.00/6	0.59/6
Artificial hybrid	0.15/6	0.24/6	0/6	0.00/6

(e)	*P. herberti*	*P. penicillata*	Natural hybrid	Artificial hybrid
P. herberti	0.00/6	0.52/3	*0/1*	*1.04/1*
P. penicillata	0/3	0.00/5	*0/1*	*0/1*
Natural hybrid	0.34/3	0.44/3	0.00/1	<u>0.21/1</u>
Artificial hybrid	0/3	0.19/2	<u>0.21/1</u>	0.00/1

To calculate the value of the HDC, all possible quartets among a set of *n* taxa are formed that include the parental taxa (P1 and P2), a putative hybrid (H), and a fourth or 'outgroup' taxon (X), one each of the remaining *n*-3 taxa being added in turn. We consider the corresponding splitsgraphs as evidence for hybridisation when the two weakly compatible splits associate the putative hybrid (H) with either one of its parents (P1 or P2), but not with the outgroup taxon (X; Fig 4.2a). The HDC score is then computed as the number of times such quartets are consistent with or support the hypothesis of hybridisation (H_1), and can be scaled by dividing by the total number of relevant quartets. More precisely, two positive splits, each pairing H with one or the other parent (P1 or P2), meet HDC and provide full support (1.0) for hybridisation. However, a split supporting the grouping of H and X would be considered as evidence against hybridisation, and scored zero (0.0) in the numerator (Fig. 4.2b and 4.2c). On the other hand, a situation where one of the two splits is of zero length could be considered as providing partial support (0.5) for H_1 as long as the other split does not group H with X (Fig. 4.2d). Finally, when both splits are null, resulting in a star tree, no evidence against or for H_1 is provided, and the denominator of the scaled HDC is reduced by one (Fig. 4.2e). DNA-DNA hybridisation experiments can lead to negative distance values and these were treated as null. Recognising that some random experimental error exists in most data, the question immediately arises as to what degree of conflict between positive signals should be judged as full support for hybridisation (scored as 1.0). How similar need two positive splits be? We employed threshold values ranging from 0.10 to 0.75, figures related to the ratios between the weakly compatible splits, where the more equal splits imply greater conflict - a more squared box - and therefore stronger support for H_1.

For the data of Table 4.2, separate HDC scores were calculated for each of the four hybrids (two natural and two artificial). However, we also wished to assess the significance of these scores. To do so, all possible hybridisation hypotheses - not just those involving one of the four hybrids – were generated

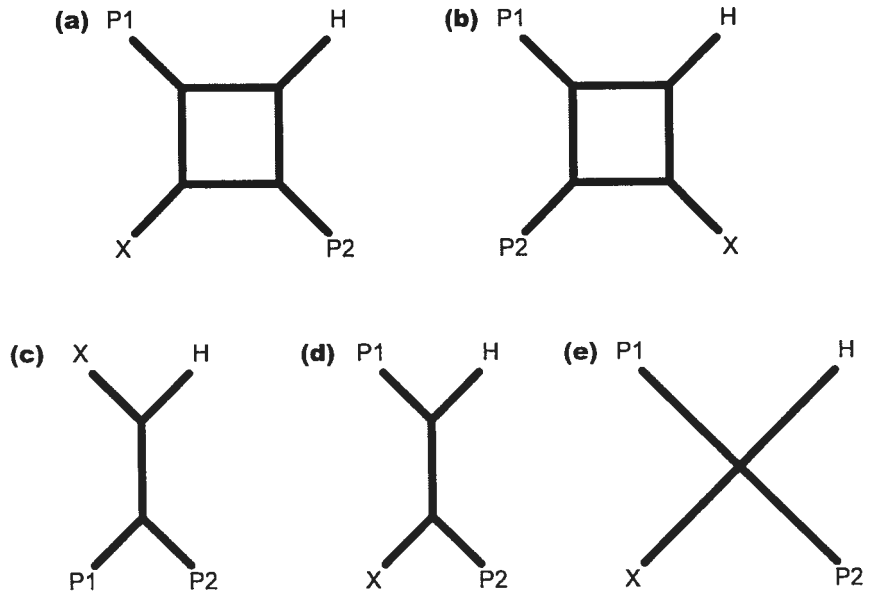


Figure 4.2 Example of a splitsgraphs and trees and their interpretation under the *HDC*. (a) A splitsgraphs that groups the hybrid (H), its two parents (P1 and P2), and the outgroup (X) in a way that meets *HDC* (b) another that does not. (c) A tree for which the putative hybrid is not grouped with either of its parent; this also contradicts the criterion. (d) A tree that offers no clear support for or against the *HDC*, this case is arbitrarily given a weight of 0.5 in the calculations. (e) A star tree that provides no information on the relationships between the taxa, it neither supports nor contradicts *HDC*; such cases are ignored in the computation of the statistic.

and the HDC score of each of these hypotheses were compared with the HDC score for quartets involving a hybrid. The number of times a score for a randomly-generated hypothesis is as great or greater than the actual HDC score, divided by the total number of permutations, gives the probability of the actual HDC score under the null hypothesis (H_0) of no hybridisation. Because the presence of hybrid taxa in the full matrix may affect the significance of the test, reduced matrices were also analysed by considering only one artificial and one natural hybrid between the same parental species. The probability distribution of the HDC scores for the full matrix was based on the complete enumeration of all possible 2907 permutations. For the reduced matrices, all 1680 permutations were analysed.

4.5. Results

The ΔT_m data of Table 4.2 are equivalent to those presented in our previous paper except for the addition of hybrid data and for tracers of *Petrogale assimilis, P. coenensis,* and *P. rothschildi*, all three of which were treated as unlabeled in the earlier analysis because repeated labelling gave mostly negative distances. We included them here because we wished to provide the broadest possible context for the consideration of the hybrids. Negative distances generally have no evolutionary meaning; they probably represent experimental error, especially when the label is compressed and therefore relatively undiscriminating. Thus, the negative distances were set equal to zero for all analyses. Otherwise, no manipulations were performed on these data except for reflecting 15 missing entries into the last column, while taking row:column ratios into account, according to the procedure of Springer & Kirsch (1991).

Figure 4.3 is a tree constructed from these data using FITCH. This tree is compatible with the figure 3b tree presented in our earlier *Petrogale* paper (Campeau-Peloquin *et al.* 2001), and has the same 'backbone' whether or not the artificial or natural hybrids, or both, are excluded from the computations. Importantly, the topology is identical if actual negative values (rather than

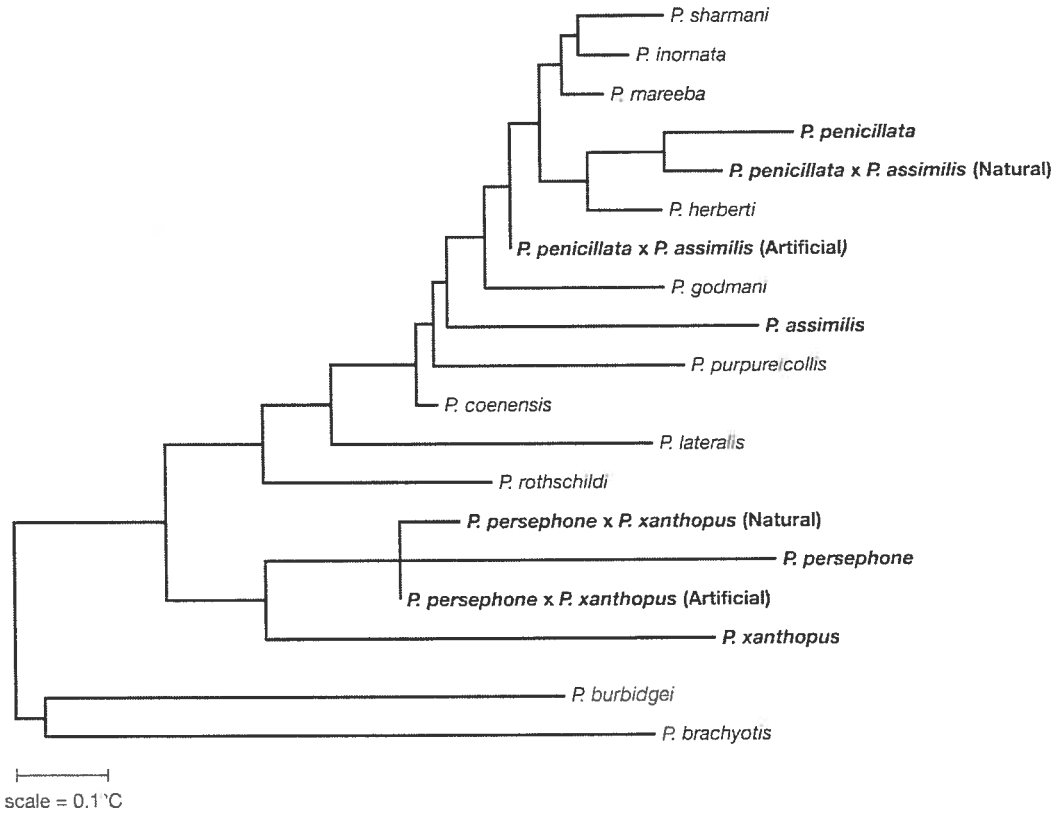


Figure 4.3 FITCH tree calculated from the data of Table 4.2, using the G, S, and Cavalli-Sforza & Edwards (P = 0) options; and randomising the input-order of taxa 50 times. All negative ΔT_ms were set equal to zero and 15 values in the last column (also negative and therefore set to zero) were reflected from their reciprocals, after taking into account row:column ratios according to the procedure of Springer & Kirsch (1991). Parental and hybrid taxa highlighted in boldface.

zeros) are used.

Table 4.3a-e presents the data for all five 4x4 experiments, two of which are abstracted from the Table 4.2 data. The *P. herberti/P. penicillata* hybrids were also originally included in the larger table as drivers, but labelling of these taxa failed, and so Table 4.3e represents the single case for which we had to estimate (as opposed to reflect) values, namely the reciprocal comparisons of the two hybrids, which we did using the additive procedure of Landry *et al.* (1996).

Figure 4.4a-e shows splitsgraphs calculated from the data in Table 4.3. In each splitsgraph, the larger split corresponds to the topology of the tree that would be obtained by applying the FITCH algorithm to the same data. These results show that the artificial and natural hybrids would never pair together in any of the corresponding trees, whereas they would be supported by weakly compatible splits in one case only (Fig. 4.4e). Of course, without inclusion of a true outgroup, this means that either the artificial or natural non-parental taxon is equally likely to represent a hybrid.

Table 4.4 summarises the results of the HDC calculations and probabilities for the examples of four hybrids (two natural and two artificial) included in the 19-taxon matrix, under several threshold, and for both reduced and full matrices. Most scaled HDC scores for the artificial hybrids provided significant support for H_1 at either = 0.01 or better (for *P. persephone* plus *P. xanthopus*) or = 0.10 or better (for *P. assimilis* plus *P. penicillata*), but did not provide significant evidence of hybridisation for the natural hybrids.

4.6. Discussion

Our initial intent was to construct a series of 4x4 matrices comparing pairs of parental species with their natural hybrid offspring and with equivalent artificial hybrids fabricated by combining equimolar amounts of tissue-extracts from the two parents, such constructs being uniquely permitted by the technique of DNA hybridisation. The expectation was that the artificial and natural hybrids would behave similarly, and would then provide models for detecting cryptic hybrids

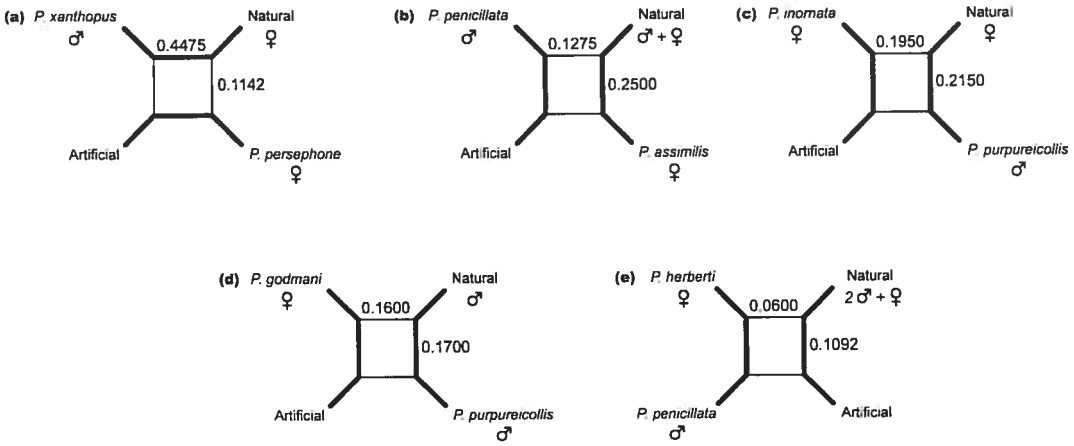


Figure 4.4 Splitsgraphs (a-e) calculated from the 4x4 matrices of Table 4.3a-e using SplitsTree4.0. The split highlighted with heavy branches in case corresponds to the tree that would be obtained by applying the FITCH algorithm to the same data. Notice that the artificial and natural hybrids are never paired together in the splitsgraphs, except for one case (e). Weights are given for internal branches.

Table 4.4 Scaled *HDC* s and their probabilities for the four hybrids included in Table 4.2 for both the reduced and full matrices and different threshold values (t=0.00, t=0.10, t=0.25, t=0.50, and t=0.75). Reduced matrices include all full species and only the hybrid under consideration (n=16) whereas the full matrix includes all species and all hybrids (n=19).

				Reduced matrix		Full matrix	
t=0.00				*HDC*	*p*	*HDC*	*P*
P. persephone	x	*P. xanthopus*	Natural	0.2308	0.6042	0.3125	0.4840
P. persephone	x	*P. xanthopus*	Artificial	1.0000	0.0173	1.0000	0.0145
P. assimilis	x	*P. penicillata*	Natural	0.2308	0.6054	0.3125	0.4840
P. assimilis	x	*P. penicillata*	Artificial	0.7692	0.1119	0.7500	0.1142
t=0.10				*HDC*	*p*	*HDC*	*p*
P. persephone	x	*P. xanthopus*	Natural	0.2308	0.5589	0.3125	0.4303
P. persephone	x	*P. xanthopus*	Artificial	1.0000	0.0077	1.0000	0.0089
P. assimilis	x	*P. penicillata*	Natural	0.1923	0.5958	0.2500	0.5184
P. assimilis	x	*P. penicillata*	Artificial	0.7692	0.0786	0.7188	0.0949
t=0.25				*HDC*	*p*	*HDC*	*p*
P. persephone	x	*P. xanthopus*	Natural	0.1923	0.5542	0.2813	0.4142
P. persephone	x	*P. xanthopus*	Artificial	1.0000	0.0036	1.0000	0.0052
P. assimilis	x	*P. penicillata*	Natural	0.1154	0.6929	0.1875	0.5690
P. assimilis	x	*P. penicillata*	Artificial	0.6923	0.0667	0.6563	0.0843
t=0.50				*HDC*	*p*	*HDC*	*p*
P. persephone	x	*P. xanthopus*	Natural	0.1923	0.4994	0.2188	0.4472
P. persephone	x	*P. xanthopus*	Artificial	1.0000	0.0006	0.9688	0.0010
P. assimilis	x	*P. penicillata*	Natural	0.1154	0.6524	0.1563	0.5703
P. assimilis	x	*P. penicillata*	Artificial	0.6154	0.0458	0.5938	0.0588
t=0.75				*HDC*	*p*	*HDC*	*P*
P. persephone	x	*P. xanthopus*	Natural	0.1154	0.6315	0.1563	0.5273
P. persephone	x	*P. xanthopus*	Artificial	0.9231	0.0006	0.8750	0.0007
P. assimilis	x	*P. penicillata*	Natural	0.1154	0.6292	0.1563	0.5273
P. assimilis	x	*P. penicillata*	Artificial	0.5000	0.0631	0.4688	0.0826

in larger trees. However, the proposition that artificial and natural hybrids would act equivalently bears some modification based on the special assumptions and properties of distance analyses.

Single-copy DNA hybridisation and other distance-generating techniques such as comparative serology assume that the entities being compared are equally complex, i.e., contain the same number of distinct genetic sequences (or their protein products). If so, then the measured reciprocal distances will - within experimental error - be equal, satisfying one important requirement of (mathematical) distances. In practice, reciprocal distances are *not* always equal, due to systematic error (e.g., compression of labels or antisera) as well as random experimental error; but the former can be corrected for (Springer & Kirsch 1991), and studies of absolute genome-sizes and reassociation-kinetics usually indicate no major complexity differences among at least moderately closely-related taxa (Kirsch *et al.* 1990).

However, in the special case of the two types of hybrids examined here, the assumption of equal complexity clearly does not hold. Considering only those sequences in which two taxa *differ*, the *natural* hybrid should present on average half of the sequences characteristic of each parent - as with any F_1 hybrid. So, tracers prepared with DNA of either parental taxon tested against the hybrid will be 50% of the distance separating the two parental taxa, and similarly for a tracer made with an extract of that natural hybrid with respect to either parent. Thus we can say that the hybrid is *intermediate* between the parental taxa and presumably has no special features of its own.

In contrast, an *artificial* hybrid, being an equimolar mixture of parental extracts will, unlike the natural hybrid, encompass *all* differences between the parental species. The genome of an artificial hybrid thus represents a doubling in complexity over either parental taxon or their natural hybrid. The effect of the difference in complexity will be a profound asymmetry in the reciprocal distances between the artificial hybrid and any of the other three taxa. What will be the consequence of this asymmetry for the representations of the relationship among all four?

Initially we imagined that artificial and natural hybrids would behave similarly; that is, that the two hybrids should group together, with parental species forming a distinct group. Our results clearly show that this is not the case. Indeed, in all instances but one, the splitsgraphs (Fig. 4.4a-e) highlight the differences between both types of hybrids, and a treelike representation of the data would never pair the natural and artificial hybrids together. Interestingly, the only exception to the general pattern is the matrix for which half the data were estimated (Table 4.3e).

That the pattern of association among parental and the two types of hybrid taxa is consistent also receives statistical support from the HDC analyses. Comparison of all possible quartets including both parents, either hybrid (artificial or natural), and a fourth taxon (Table 4.4) show that the probabilities that the experimental constructs represent hybrids are almost always significant, while the probabilities for the natural hybrids are no better than those for non-hybrids (in fact, near the middle of the distributions). Moreover, for the artificial hybrids, support for H_1 increases as the threshold requirements for satisfying the criterion are made more stringent, at least up to 0.50, while that for the natural hybrids remains about the same. Although, with only two observed cases, this conclusion is somewhat tenuous, we note once again that the consistency of splitsgraphs derived from the 4x4 matrices is strong circumstantial evidence that addition of the remaining three (or more) cases to the full matrix would give similar results.

Nonetheless, the failure of natural hybrids to behave in an expected manner raises the interesting question of why that might be so. A non-exhaustive list of possible reasons for our results might include: (1) a maternal effect; (2) a sex-chromosome-related difference; (3) a replication phenomenon; or (4) some combination of these or other factors. It is particularly tempting to adduce a maternal effect, given the nearly uniform matrilineal inheritance of mitochondrial DNA (which is certainly present in our extracts). Indeed, as indicated in Figure 4.4a-e, where the sexes of the parents and natural hybrids are noted, in four out of the five cases the natural hybrid does pair with the

female parent (considering the largest split only), irrespective of the sex of the hybrid. Of course, many more cases would be needed to conclude that this pattern is not due to chance. Similarly, we might imagine that, the male being the heterogametic sex in therian mammals, the natural hybrid might tend to be associated with the female parent, because so many genes are lacking on the Y-chromosome. However, the sex chromosomes are relatively small in marsupials, and could only account for at most an amount proportional to the total number or length of the chromosomal set, assuming a relatively uniform distribution of unique sequences throughout the karyotype. So, even in combination with a mitochondrial effect, a sex-chromosome difference could not account for the marked asymmetry. More likely is the third possibility, that unequal crossing-over or some other replication phenomenon selectively eliminates a large part of one or the other parental genome, but here again we lack sufficient data bearing on this explanation. Unfortunately, the overall conclusion from this study must therefore be that the prospect for detecting natural hybridisation among rock wallabies is not encouraging: such hybrids will show a marked affinity with one parent or another, not intermediacy and certainly not a tendency to impose anastomosing on the phylogram.

4.7. Acknowledgements

The authors are grateful to all members of the LEMEE for constructive comments on an earlier version of this manuscript. This work was supported by NSERC and FQRNT scholarships to OG, a Faculté des Études supérieures de l'Université de Montréal scholarship to ACP, NSERC grant no. OGP0155251 and FQRNT grant no. PR88559 to FJL, funds from the Australian Research Council and From Macquarie University to MDBE, and by funds for the hybridization experiments provided by JAWK.

Chapitre 5 :

POWER AND ACCURACY OF THE HDC TEST

Cet article sera soumis prochainement :

Gauthier, O., E. L. Bui, &, F.-J. Lapointe. Power and accuracy of the *HDC* test.

5.1. Résumé

La validité du test statistique de la signification de l'indice de détection des hybrides (HDC) proposé par Gauthier & Lapointe (2002, 2006) est vérifiée à l'aide de simulations par ordinateur. Des séquences d'ADN ont été simulées le long des branches de phylogénies réticulées et non-réticulées afin d'estimer la puissance et l'erreur de type I du test. Les résultats montrent que le test est en mesure de détecter le caractère intermédiaire des hybrides dans de nombreuses conditions, et qu'il ne présente pas une erreur de type I plus grande qu'attendue. Ces simulations tendent également à montrer que le test serait quelque peu conservateur.

5.2. Abstract

The validity of the statistical test of the Hybrid Detection Criterion (HDC ; Gauthier & Lapointe 2002, 2006) is assessed with computer simulations. DNA sequences are simulated along the branches of reticulate and non-reticulate phylogenies in order to estimate both the statistical power and the type I error of the test. Results indicate that the HDC test is able to detect the intermediate phylogenetic position of the hybrids under varying conditions and that it is does not exhibit larger than expected type I error. The test also appears to be rather conservative.

5.3. Introduction

The detection of reticulate evolution and hybrid taxa is a major challenge for current phylogenetic analysis. Indeed, following the realization that hybridization could play a role in the creation of new species (Stebbins 1950; Anderson & Stebbins 1954; Arnold 1992, 1997) provisions where made to illustrate such events with reticulate phylogenies (Sneath 1975; Nelson 1983; Wagner 1983; Wanntorp 1983). Despite new developments in the field (*e.g.* Makarenkov & Legendre 2000; Xu 2000; Nakhleh *et al.* 2003, 2004) no phylogenetic inference method constitutes an adequate way of identifying hybrid taxa. While we cannot ignore the problem completely, we must find a way to identify hybrid taxa and their parental lineages in order to reconstruct the treelike part of the phylogeny on which the hybrid taxa can be later grafted in the appropriate location (Wagner 1969, 1983). Gauthier & Lapointe (2002, 2006) have proposed a Hybrid Detection Criterion (HDC) statistic based on quartets analysis along with a test to assess its significance. The HDC states that, with respect to all other taxa in the data set, a putative hybrid should systematically occupy an intermediate position between its postulated parents in a network (Fig. 5.1). The number of quartets that satisfy the HDC is then taken as a measure of support for the hybridization hypothesis. This index takes high values when the hypothesis is true, and low values otherwise. In previous papers (Gauthier & Lapointe 2002, 2006), the performance of this statistic has been evaluated with morphological characters scored for known hybrids of the genus *Aphelandra* (Acanthaceae; McDade 1984, 1990) and the results showed that the test performed rather well in such cases. Furthermore, the test was used to identify Rock Wallaby (*Petrogale*) hybrids from DNA-DNA hybridization data. Both yard-bred and artificial – constructed from the combination of parental DNA extracts – hybrids were tested. While the test performed extremely well to detect the intermediacy of the artificial hybrids, the yard-bred hybrids did not exhibit this intermediacy and were thus not detected (Kirsch *et al.* in prep.).

In the present paper, the usefulness of the HDC test is further assessed using

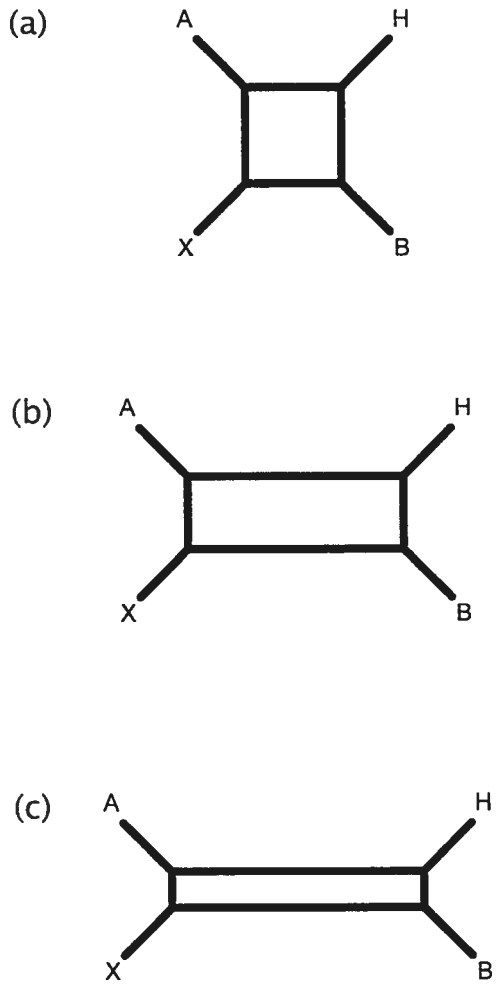


Figure 5.1 Illustration of the Hybrid Detection Criterion (HDC). When considering splitsgraphs defined on quartets of taxa, a hybrid (H) occupies an intermediate phylogenetic position between its parents (A and B) with respect to any other taxa in the dataset (X). (a) Equal support for both underlying trees (1:1 ratio); and differential support for the underlying trees with (b) 1:3 ratio and (c) 1:7 ratio of the branch lengths supporting the weakly compatible splits.

computer simulations. By evolving DNA sequences along the branches of reticulate (network) and non-reticulate (tree) phylogenies, we estimate both the power and the type I error of the statistical procedure. The test was evaluated for different sample sizes, as well as varying levels of noise in the reticulate phylogenetic signal.

5.4. Simulation procedure

In order to estimate the power of the HDC test, we simulated DNA sequences along the branches of phylogenetic networks containing only one reticulation. As the HDC only allows for the formulation of hypothesis concerning terminal, taxa the reticulation was always located at the tip of the phylogeny. The protocol used in the simulations takes advantage of the fact that a network displays multiple trees (Bandelt & Dress 1992; Nakhleh et al. 2004) and that two trees can always be uniquely represented by a network (Bandelt & Dress 1992; Bandelt 1995). Sequence evolution on a network was thus simulated as on two incongruent trees that, when combined, yield the appropriate network. This procedure can be viewed as analogous to a situation where two different sequences, that each originate from a distinct parental species, are sampled from the hybrid. However this need not be the case because each site is treated as independent in the analysis. What is simulated here is not the actual process of hybridization, only the resulting data that it can lead to (Posada & Crandall 2002).

Each run starts with the generation of a random binary ultrametric tree with $(n-1)$ leaves (Fig. 5.2a). This tree is duplicated and two leaves are then randomly chosen as parental lineages (Fig. 5.2b). An additional leaf is grafted on each of the corresponding branches and labeled as the hybrid taxon. It is grafted halfway along the shortest of the two branches, and at the same distance from the tip on the longer branch (Fig. 5.2b). Branch lengths are then independently multiplied by a factor drawn from the uniform [0.5,1.5] in order to depart from the ultrametric condition (Fig. 5.2c). DNA sequences are then evolved along the branches of both trees under the Kimura-2-parameter model

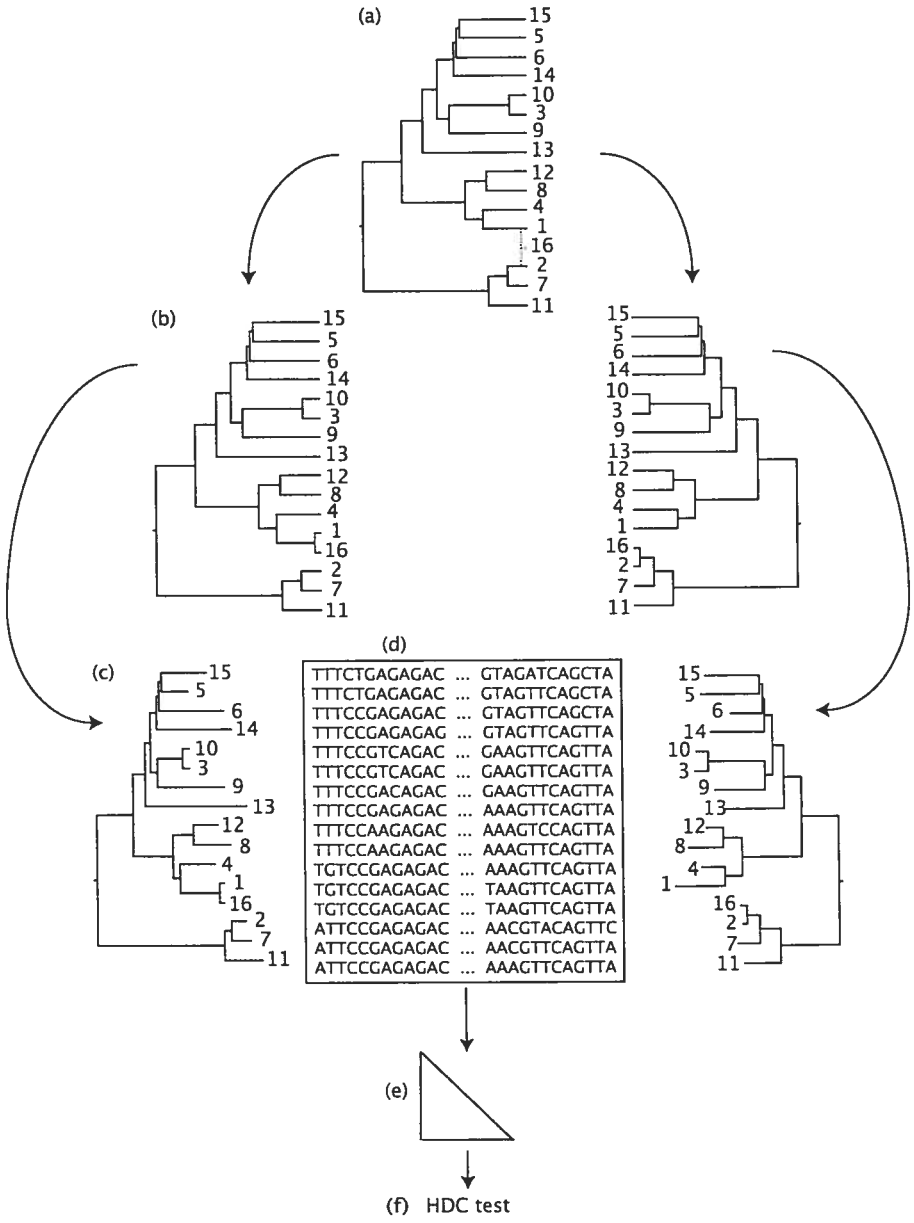


Figure 5.2 Outline of the simulation procedure. (a) generate a random ultrametric tree; the position of the hybrid (taxa 16) is illustrated with dotted branches for clarity, it is not added at this point however; (b) duplicate tree and graft hybrid leave at the appropriate position in each tree; (c) go from ultrametric trees to additive trees; (d) evolve sequences along the trees; (e) compute distances; and, (f) submit to *HDC* test.

(K2P; Kimura 1980) with the transition/transversion ratio set to 2, using SEQ-GEN v 1.3.2 (Rambaut & Grassly 1997; Fig. 5.2d). The two trees are treated as two partitions so that SEQ-GEN outputs concatenated sequences. K2P corrected distances are then computed using the DNADIST program in PHYLIP (Felsenstein 2005; Fig. 5.2e). This distance matrix is then submitted to the HDC test and the result is the probability to observe this HDC value under H_0 (Fig. 5.2f). All HDC tests were conducted with split decomposition (Bandelt & Dress 1992), since earlier work showed that it gave accurate results with known hybrids (Gauthier & Lapointe 2002, 2006).

A similar procedure was used to estimate the type I error rate of the test. The initial tree is generated with n leaves and this single tree is duplicated before applying random branch length variations independently to both trees. Thus, no hybrid taxon is added to the tree, and both trees have the same topology. To generate a false hybridization hypothesis, two taxa are randomly designated as parents, and a third one as their hybrid. The rest of the simulations are carried out as before (see Fig. 5.2).

Since the HDC test is based on the intermediate phylogenetic position of the hybrid taxon with regards to its parental taxa in a reticulate phylogeny, an additional parameter was used in order to assess the effect of varying degrees of intermediacy on the results. The reticulate signal is strongest when both the underlying trees are equally supported by the data; whereas it is absent when only one of the underlying trees is supported by the data. These situations are easily simulated by varying the number of base pairs evolved along the branches of each underlying trees. The power of the HDC test was evaluated for situations where the ratio of base pairs supporting each tree was 1:1, 1:3, and 1:7 (Fig. 5.1). While the evaluation of type I error is effectively a situation where this ratio is 0:1 when considering only the topology, the same ratios of base pairs were used in these simulations for the sake of comparison. Moreover, the simulations were run for different number of taxa: 16, 32, and 48. For each combination of parameters, 10 000 replicates were generated and the simulated DNA sequences were always 4 000 bp. All tests where

conducted using all possible permutations, totaling 1 680, 14 880, and 51 888 permutations for 16, 32 and 48 taxa respectively (Gauthier & Lapointe 2006). Results are presented as the number of times the null hypothesis of no hybridization is rejected.

5.5. Results and discussion

We addressed specific questions concerning the reliability of the HDC test. Our first simulations addressed the power of the test, and the results reveal that the null hypothesis is appropriately rejected in 92.8% to 99.3% of cases (Table 5.1). The test also behaves as expected with respect to the strength of the reticulate signal, leading to a rejection of the null hypothesis more often when the branch length ratio was 1:1. Moreover, reticulations were more often detected for larger data sets (i.e. more numerous taxa).

Our second objective was to evaluate the type I error of the test. Our results indicate that the test is valid since the error rate varies from 3.35% to 4.27% and is always inferior to the nominal significance level ($\alpha = 0.05$; Table 5.2). However, we conjecture that in some cases the HDC test might be too conservative. No consistent effect of the support ratio for both trees was observed, which implies that branch length heterogeneity is not affecting the results of the test. However, inaccurate rejection of the null hypothesis was more frequent with more taxa. This is consistent with the results obtained in the first simulations.

As for all permutation tests the power of the HDC test increases with the number of permutations under consideration. All possible permutations where enumerated here, and, since their number is dependent on the number of taxa, this can explain part of the greater power of the test with larger datasets. However, another important issue that must be raised here is the proportion of reticulate relationships compared to that of tree-like relationships within the datasets. In our simulations only one hybridization event was simulated regardless of the number of taxa, thus the proportion of reticulate relationships decreases in larger datasets. This should lead to a larger proportion of

Table 5.1 Estimation of the power of the HDC test. Rejection of the null hypothesis for different number of taxa (n) and varying support for both partitions when sequences evolved along conflicting trees are combined to simulate reticulate evolution. Results are presented as percentage of rejection over 10 000 replicates for each combination of parameters ($\alpha = 0.05$).

n	Base pairs per tree (tree1:tree2)		
	500:3500	1000:3000	2000:2000
16	96.5 %	95.1 %	92.8 %
32	98.7 %	98.2 %	97.0 %
48	99.3 %	99.0 %	98.0 %

Table 5.2 Estimation of the type I error of the HDC test. Rejection of the null hypothesis for different number of taxa (n) and varying support for both partitions when sequences evolved along topologically identical trees are combined to simulate evolution on the branches of a tree. Results are presented as percentage of rejection over 10 000 replicates for each combination of parameters ($\alpha = 0.05$).

n	Base pairs per tree (tree1:tree2)		
	500:3500	1000:3000	2000:2000
16	3.82 %	3.35 %	3.47 %
32	3.87 %	3.94 %	4.23 %
48	4.12 %	4.11 %	4.27 %

permutations having a smaller HDC score, and an increased probability of detecting the actual hybrid. Furthermore, this raises the question of the reliability of the test in the presence of multiple hybrids. While this has not been investigated here, it can be conjectured that this would make the detection of hybrids more difficult because it would lead to larger proportions of reticulate relationships and permutations with high HDC scores. This needs to be confirmed with new simulations.

These results, as well as our previous evaluations using real morphological and molecular data, all lead to the conclusion that the HDC allows for accurate and consistent detection of intermediate hybrid taxa. Furthermore, the test does not exhibit inflated type I error. However, as expected, departure from the intermediacy criterion can lead to spurious results and undetectable hybrids, but this discrepancy needs to be rather large to really affect the statistical conclusion. These new results thus confirm the utility of HDC test as an analytical and exploratory tool in the search for hybrid taxa and reticulate evolution.

5.6. Acknowledgements

The authors are grateful to all members of the LEMEE for constructive comments on an earlier version of this manuscript. This work was supported by NSERC and FQRNT scholarships to OG and NSERC grant no. OGP0155251 and FQRNT grant no. PR88559 to FJL.

Chapitre 6 :

GETTING MORE FROM YOUR TREES WITH CONSENSUS NETWORKS

Cet article sera soumis prochainement :

Gauthier, O. & Lapointe, F.-J. 2006. Getting more from your trees with consensus networks. Sera soumis à *Systematic Biology*.

6.1. Résumé

L'analyse phylogénétique mène souvent à l'inférence d'arbres multiples définis sur un même ensemble de taxons. Les méthodes de consensus sont alors utilisées pour identifier les zones d'accord et de désaccord entre ces hypothèses concurrentes, ou encore pour ne retenir que les relations phylétiques soutenues unanimement ou majoritairement. Le choix d'une méthode de consensus se base sur différents critères. D'un point de vue axiomatique, les méthodes répondant aux propriétés Pareto et co-Pareto devraient êtres sélectionnées. D'un autre côté, on pourra préférer des méthodes menant à des solutions contenant une plus grande quantité d'information, par exemple, des solutions mieux résolues. Dans cet article, nous présentons un cadre général pour la construction de réseaux consensus contenant plus d'information que les traditionnels arbres consensus. La mesure de Contenu en Information Cladistique (CIC) est adaptée aux réseaux consensus.

6.2. Abstract

Phylogenetic inference often results in the production of multiple trees on a given set of leaves. Consensus methods are commonly used to identify areas of conflict and agreement among trees or only retain the relationships that are supported either unanimously or by a majority of trees while discarding other, less supported relationships. The choice of a given consensus method can be based on different criteria. From an axiomatic perspective, methods that are Pareto and co-Pareto should be selected. On the other hand, one may prefer methods that produce consensus solutions containing more information, such as consensus trees that are better resolved. In this paper, we discuss different ways to produce consensus containing more phylogenetic information in the form of Consensus Networks. We extend the notion of Cladistic Information Content (CIC) to measure the information content of consensus networks.

6.3. Introduction

Phylogenetic inference often leads to solutions made up of multiple trees on a given set of leaves or taxa. These competing hypotheses might be the equally optimal trees obtained from the analysis of a single matrix using maximum parsimony (Hennig 1979) or maximum likelihood (Felsenstein 1981), or the set of most probable trees produced by Bayesian analysis (Rannala & Yang 1996; Larget & Simon 1999; Mau *et al.* 1999). They might also have been inferred independently from different data sets, or even be the result of re-sampling methods such as the bootstrap or the jackknife (Felsenstein 1985; Penny & Hendy 1985, 1986). Consensus methods are commonly used to identify areas of conflict and agreement among such multiple trees; they can represent the relationships that are supported either unanimously or by a majority of trees while discarding other, less supported relationships (Bryant 2003). Thus, a consensus method is usually defined as a function that takes as input a set, or profile, of trees on the same set of taxa and returns a single tree on the same set of taxa (Leclerc & Cucumel 1987; Steel *et al.* 2000; Bryant 2003). This function can take different forms, and many consensus methods have been proposed (see Swofford 1991; and Bryant 2003 for reviews), amongst which the strict consensus (Rohlf 1982) and majority-rule consensus (Margush & McMorris 1981) are perhaps the most widely used and understood. These various functions differ in two major respects: (1) the kind of information they preserve, and (2) the way they deal with conflict among input trees. Indeed, the information contained in a tree can be considered in terms of nesting, three or four taxa statements, components, and branch lengths, while conflict can be left unresolved or dealt with using different criteria (Bryant 2003). Notwithstanding these differences, most methods abide to the prevailing phylogenetic model: that of a tree embedding a hierarchy of descent. This model puts forward a fully resolved, or binary, tree as the ideal representation of evolution. In practice consensus trees are seldom binary and they embed – i.e. are compatible with – multiple binary trees. Indeed, since it is often impossible to distinguish between hard and soft polytomies (Nelson & Platnick

1980; Maddison 1989) and since the latter can be resolved in a number of different ways, an unresolved tree can be refined by a set of binary trees (Rohlf 1982; Mickevich & Platnick 1989; Steel *et al.* 2000).

To circumvent this practical problem, a consensus method could be defined as a function that takes as input a profile of trees on the same set of taxa, and returns one or multiple binary trees on the same set of taxa. Interestingly, on top of accounting for the fact that consensus trees are often unresolved, this definition also allows for the formulation of consensus methods that produce multiple trees (Wilkinson 1994; Lapointe & Cucumel 2002) and networks (Bandelt 1995; Holland & Moulton 2003; Lapointe & Cucumel 2003). This is desirable because incongruence among phylogenetic trees often leads to poorly resolved consensus solutions, and using multiple trees or networks can improve the representation of shared information among the fundamentals. Also, the treeness of phylogenies has been questioned and the so-called Tree of Life most probably ranges in complexity from that of a 'simple' tree, to that of an entangled network, or even an inscrutable web (Doolitle 1999; Daubin *et al.* 2002; Rivera & Lake 2004).

The diversity of consensus methods already available (Bryant 2003) makes it necessary to choose among them, or among the solutions they produce. This decision can be based on axiomatic properties of the competing approaches. This is the position championed by Page (1992) who bases the choice of a given consensus method solely on what is judged to be an important "aspect of tree structure" and on the "level of agreement between trees" that should be represented in the consensus. From this standpoint, it can be argued that methods that are Pareto and co-Pareto are preferable over those that are not. Indeed, a method that is Pareto will only produce solutions that include statements made by all input trees. Likewise, a method that is co-Pareto will only display relationships that are present in at least one input tree. Both these properties are desirable when a consensus is perceived as a mean to summarize the statements made by the input trees. However, the choice of a consensus solution could also be based on the amount of phylogenetic information it conveys. How many statements does it make about the taxa

under consideration? Are these statements unambiguous? How many taxa are found in the consensus? How many statements that could be made about the taxa are disallowed by the consensus? How many binary trees, out of all the possible trees with the same number of leaves as the consensus, are compatible with the consensus? All these questions reflect what biologists consider relevant in a phylogeny: grouping. They have to be answered in order to measure the phylogenetic information content of a consensus, and, for this to be practical, an objective measure of information such as the Cladistic Information Content (CIC; see section 6.5) is needed (Thorley *et al.* 1998; Thorley 2000). Evidently, the CIC of solutions produced by a given consensus method depends both on the input trees and the procedure itself, and ultimately, we might seek a balance between these two criteria.

In this paper, we discuss different ways to produce solutions containing more phylogenetic information in the form of Consensus Networks (CN). Although we are not the first to propose a relaxation of the treeness criterion for consensus (Bandelt 1995; Holland & Moulton 2003; Lapointe & Cucumel 2003), we present a general framework to obtain a CN from any input profile of weighted or unweighted trees. Furthermore, the CIC, a measure of phylogenetic information content originally devised for strict consensus trees (Thorley *et al.* 1998), is extended here to measure the information content of consensus networks. Finally, this measure and some properties of CN are illustrated with an application to the phylogeny of mammalian orders.

6.4. Computing a consensus network

In the spirit of the many flavors of matrix representation (MR) in the consensus setting, a general approach for constructing a CN of a set of k trees defined on the same set of leaves S involves two steps (see Lapointe *et al.* 2003; Wilkinson *et al.* 2005): (1) obtain a MR from the input profile, and (2) compute a network from this MR. Both these steps can be achieved in a number of ways (e.g. Holland & Moulton 2003; Lapointe *et al.* 2003; Holland *et al.* 2004, 2005). We will only consider weighted unrooted trees, but the approach presented here can be used with rooted and unweighted trees as well. At least

two different $n \times n$ distance matrices d can represent a tree T defined on the set $S = \{1,...,n\}$ (Fig. 6.1). Discarding branch lengths and thus focusing only on topological information, the branch distance between two taxa is obtained by counting the number of branches separating them (Zaretskii 1965). Accounting for branch lengths, the path-length distance between two leaves is the sum of the weight of the branches between them (Buneman 1971). Computing branch or path length distance matrices for every tree, the input profile of trees is recoded in a series of matrix representations with distances (MRD) that must be combined in a single matrix prior to network reconstruction.

Alternatively, the k input trees could be coded as binary characters in a pseudo- character matrix as for Matrix Representation with Parsimony (MRP; Baum 1992; Ragan 1992; Purvis 1995). This is the approach proposed by Holland & Moulton (2003) for the construction of CN. Since we are dealing with unrooted trees we focus on bipartitions, or splits, of S rather than components. A split defines two non empty subsets S' and S'' of S such that $S' \cup S'' = S$ and $S' \cap S'' = \varnothing$. A tree is composed of a set of compatible splits and each split in the input trees gives rises to a binary character for which taxa in S' and S'' are assigned the values 0 and 1 respectively.

We propose to use the average path-length distance matrix in which the distance between leaves A and B is the arithmetic mean of the distance between A and B computed over the k matrix representations as the MR of the input profile of trees. It should be noted that the median or another parameter could be used as well (Lapointe & Cucumel 1997). Also, computing a Hamming distance (or single character difference) on the MR of splits and dividing by k would yield the mean branch distance matrix (Lapointe *et al.* 2003). At this point it should be clear that not only a collection of trees, but also a collection of networks, or simply splits, whether they are compatible or not, could be combined in this way.

The second step towards the construction of a CN, inferring the network from the MR, necessitates the choice of an adequate method. Recent years have seen the appearance of numerous techniques and softwares to infer

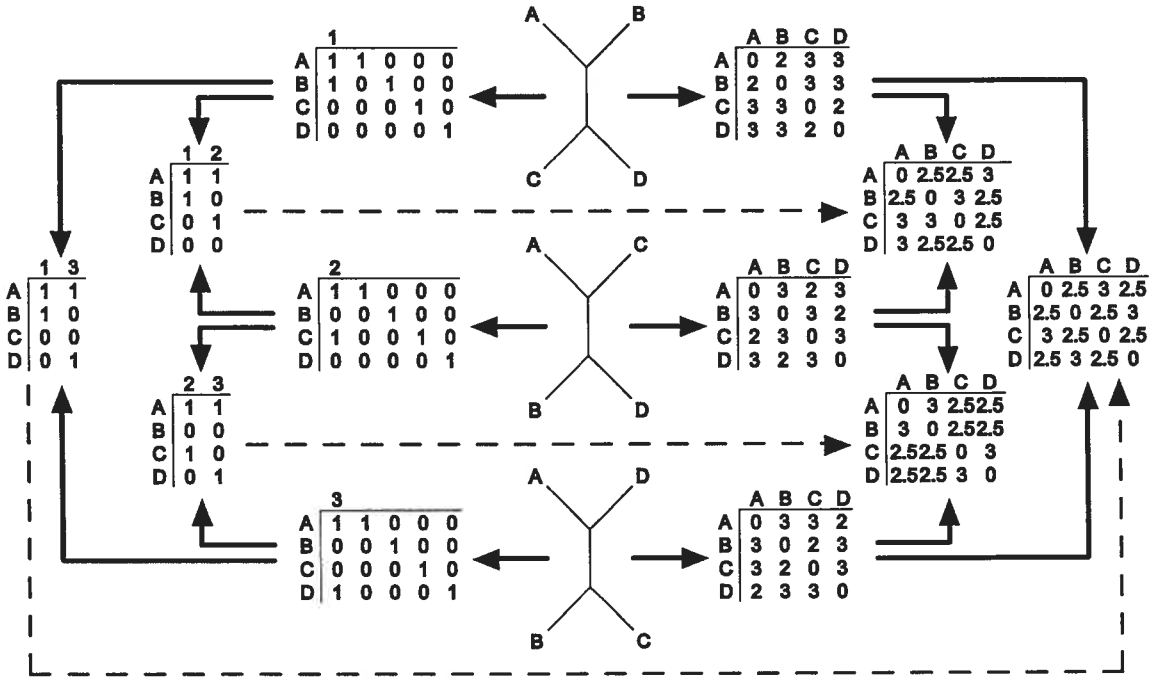


Figure 6.1 A matrix representation (MR) of a tree can take the form of a binary character matrix (left) or a distance matrix (right). Full lines indicate correspondence between trees and MR, as well as combination of MR from different trees in a unique MR. Dashed lines indicate equivalence between two MR. Each split corresponds to a binary character for which the taxa on either side of the split are coded 1 and 0 respectively. Informative characters (partitioning the taxa in two groups of two) are numbered (1 to 3), while uninformative characters (partitioning the taxa in a group of three and a singleton) are only shown for the MR of single trees. Binary MR for two (or more) trees are obtained by combining the character matrices including the uninformative characters. Distance MR are obtained by summing the lengths of branches between pairs of taxa (here all equal 1), average distance MR for two (or more) trees contain the average path length distances between pairs of taxa.

phylogenetic networks and detect reticulate taxa (e.g. Fitch 1997; Lapointe 2000; Strimmer & Moulton 2000; Xu 2000; Posada & Crandall 2001b; Strimmer *et al.* 2001; Nakhleh *et al.* 2004). The motivation for these developments was either to graphically illustrate conflicting signals in a given dataset (e.g. Bandelt & Dress 1992), to elucidate reticulate relationships between taxa brought about by recombination, lateral gene transfer, or hybridization events (e.g. Xu 2000; Legendre & Makarenkov 2002), or to produce a single graphical representation of the set of MP trees directly from the dataset (Bandelt *et al.* 1995; Fitch 1997).

We propose to use the split decomposition method of Bandelt & Dress (1992) on the average MRD of the input trees (but see Holland & Moulton 2003 for a similar application using bipartitions and median networks). Split decomposition is a distance-based method that was developed to identify and represent conflicting signals in any phylogenetic data set (Bandelt & Dress 1992). It considers in turn all the quartets of taxa in the dataset with regards to the four-point condition and the three distinguishable unrooted trees involving four objects ($AB|CD$, $AC|BD$ and $AD|BC$). It then rejects the most-violating split, and, unless the data is perfectly tree-like, keeps the remaining two which are termed weakly compatible because they can be represented by a planar circular split system, but not by a tree (Bandelt & Dress 1992). Differential support, in the form of an isolation index, is attributed to both bipartitions (see also Winkworth *et al.* 2005). The full representation on all quartets of the set of weakly and fully (i.e. tree-like) compatible splits is called a splitsgraph. When given a set of fully compatible splits the splitsgraph is a tree. Computing a splitsgraph on the MRD of an input profile of trees gives a consensus splitsgraph (or more generally a consensus network). Interestingly, a consensus splitsgraph of any two input trees will always display both these input trees (Bandelt & Dress 1992). However, the consensus splitsgraph of three or more trees does not always satisfy this property.

Consider the three unrooted phylogenies for four taxa depicted in Fig. 6.2a. The strict consensus of this profile contains all and only those bipartitions that

are present in all trees and is thus completely unresolved (Fig. 6.2b). The majority rule consensus would also be a star tree. On the other hand, a consensus splitsgraph of any pair of these trees combines both weakly compatible splits in a single representation (Fig. 6.2c). Thus, in a consensus splitsgraph it is possible to retain two weakly compatible splits rather than throwing away valuable information. For example, considering the two topmost trees in Figure 6.2a, both splits $AB|CD$ and $AC|BD$ are supported by the data, split $AD|BC$ is not however. The consensus splitsgraph embeds these weakly compatible signals (Fig 6.2c), whereas a strict consensus concludes to an absence of shared information (Fig 6.2b). A median CN of the three trees in Fig. 6.1a will always display all tree splits at once (Fig. 6.2d). Just like the star tree (Fig. 6.2b), it is compatible with all of them.

6.5. The information content of consensus network

Many recent researches focusing on the use of networks in phylogenetics have pointed out that network methods were desirable because they often convey more phylogenetic information than trees (Wilkinson *et al.* 2003; Holland *et al.* 2005; Winkworth *et al.* 2005). They all fail, however, to provide an objective and efficient way to quantify the information content of networks and to compare it to that of trees. Although the notion of how much information a tree contains can seem rather intuitive, formal definitions are still needed; numerous information indices have been proposed in the past, but most are flawed (Page 1992; Thorley 2000). While the information contained in branch lengths is vital to phylogenetic reconstruction, the amount of information conveyed by a phylogeny is usually understood in terms of grouping because the phyletic relationships among species are defined only by cladogenesis. An appropriate information measure should tell us something about the uncertainty placed on the existence and composition of such groups. Namely, a star tree is compatible with all possible groupings of the leaves, and thus maximally uncertain; it contains no cladistic information. A fully resolved tree, on the other hand, leaves no uncertainty as to the way the taxa are grouped; it is thus maximally informative. Similarly, an unresolved tree, or a network, will

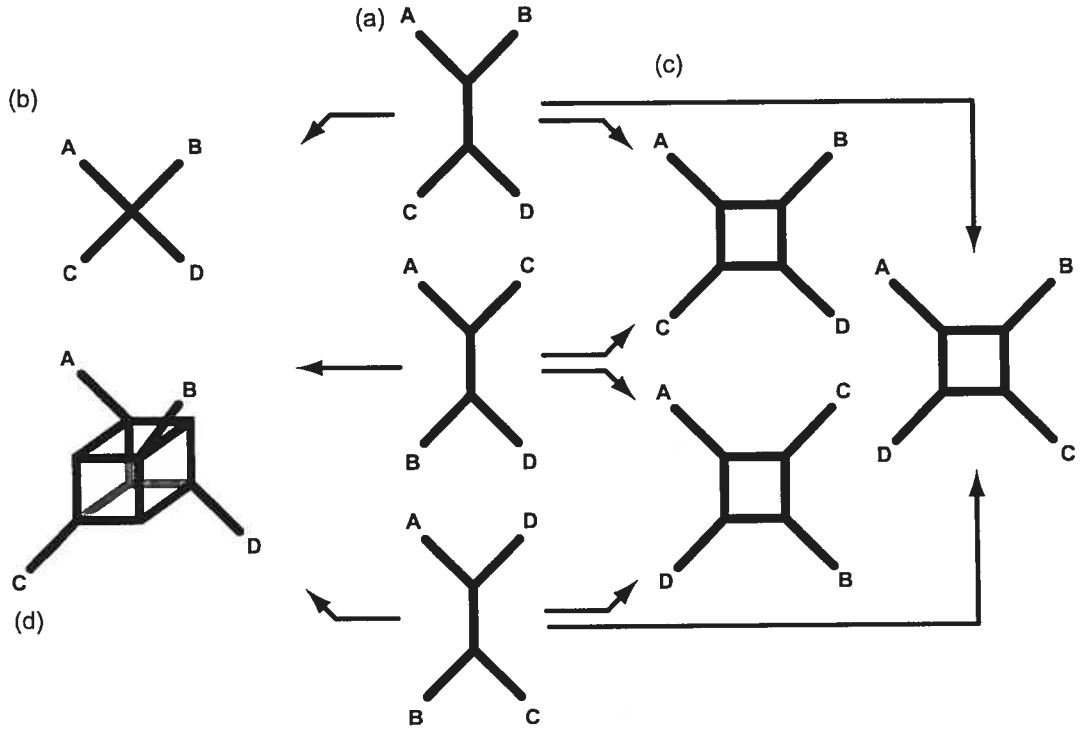


Figure 6.2 (a) The three possible unrooted trees with four taxa; (b) the strict and majority rule consensus tree of any combination of the trees in (a); (c) the CN of any pair of the trees in (a); and, (d) the unconstrained median CN of all three trees in (a).

exhibit varying degrees of uncertainty between these two extremes, and should be attributed intermediate information values. Moreover, while parameters such as tree shape and labeling should not influence a measure of the information content of phylogenies, tree size should be taken into account. Thus, a fully resolved phylogeny on 50 taxa should come out as more informative than a fully resolved tree defined on 20 taxa. Thorley *et al.* (1998) have proposed a measure of the information content of phylogenetic trees that fits these requirements and takes its roots in information theory, the general form of which is their Cladistic Information Content (CIC).

The information conveyed by an observation (I) depends both on the number of equally probable possibilities before the observation (P_0) and the number of equally probable possibilities after the observation (P_1; Brillouin 1962):

$$I = -\log P_0 / P_1 \tag{6.1}$$

The CIC of a tree T on a set of leaves $S = \{1,2,...,n\}$ is thus defined as being inversely proportional to the ratio of the number of binary trees permitted by T (n_R) to the number of possible binary trees for n taxa (n_T; Thorley *et al.* 1998):

$$CIC_T = -\log \frac{n_R}{n_T} \tag{6.2}$$

where, in the case of unrooted trees (Edwards & Cavalli-Sforza 1964):

$$n_R = \prod_{i \in V(T)} T_U(d_i) \tag{6.3}$$

where d_i is the degree of node i in the set of nodes $V(T)$ from tree T, and:

$$T_U(n) = \prod_{i=3}^{n} (2i-5) \tag{6.4}$$

In order to compute the CIC of a network one has to enumerate the multiple trees embedded in the network. Just like each of the networks in Fig. 6.2c displays two of the three trees in Fig. 6.2a, a network always displays two or

more trees. Information is said to be redundant when more bits are used to transmit a message then the number of actual bits contained in the message itself (Brillouin 1962). In the present case redundancy would follow from counting the same tree more then once in a given network, for example by counting an unresolved tree and one or more trees that refine it. In order to eliminate any redundancy, the trees contained in the network are obtained by extracting all the largest sets of compatible splits from the set of weakly compatible, or incompatible, splits that make up the network with a brute force approach; this yields only the maximally resolved (although not necessarily binary) trees embedded in the network. For example, the network with splits $AB|CDE$, $AE|BCD$, and $BC|ADE$ (Fig. 6.3a) contains two maximally resolved trees. One of these (Fig. 6.3b) has three resolutions (Fig. 6.3d), the other is binary (Fig. 6.3c). The CIC of the network is then computed as that of a tree, by summing the n_R over all maximally resolved trees contained in the network. This is equivalent to calculating the combined CIC of multiple trees that do not contain redundant information (Thorley 2000). The CIC of the network (N) is then:

$$CIC_N = -\log \frac{\sum n_R \; of \; trees \; in \; N}{n_T} \qquad (6.5)$$

Thus, the CIC of a consensus splitsgraph is the sum of the CIC of all themaximally resolved trees embedded in this graph. Note that if N is a tree, then equation 6.4 is equivalent to equation 6.2 and we get back to the original CIC (Thorley *et al.* 1998). The maximum CIC value, $CIC_{\max} = -\log 1/n_T$, depends only on n. To render the comparison of different CIC values easier, the relative CIC is computed as:

$$CIC_{rel} = CIC/CIC_{\max} \qquad (6.6)$$

For example, the network in Fig. 6.3a allows four fully resolved trees out of the 25 possible unrooted binary trees for five taxa and has a $CIC_{rel} = 0.49$. Although the relative measure has no units, information is usually expressed either in bits or in nats, depending on the base of the logarithm, 2 or e

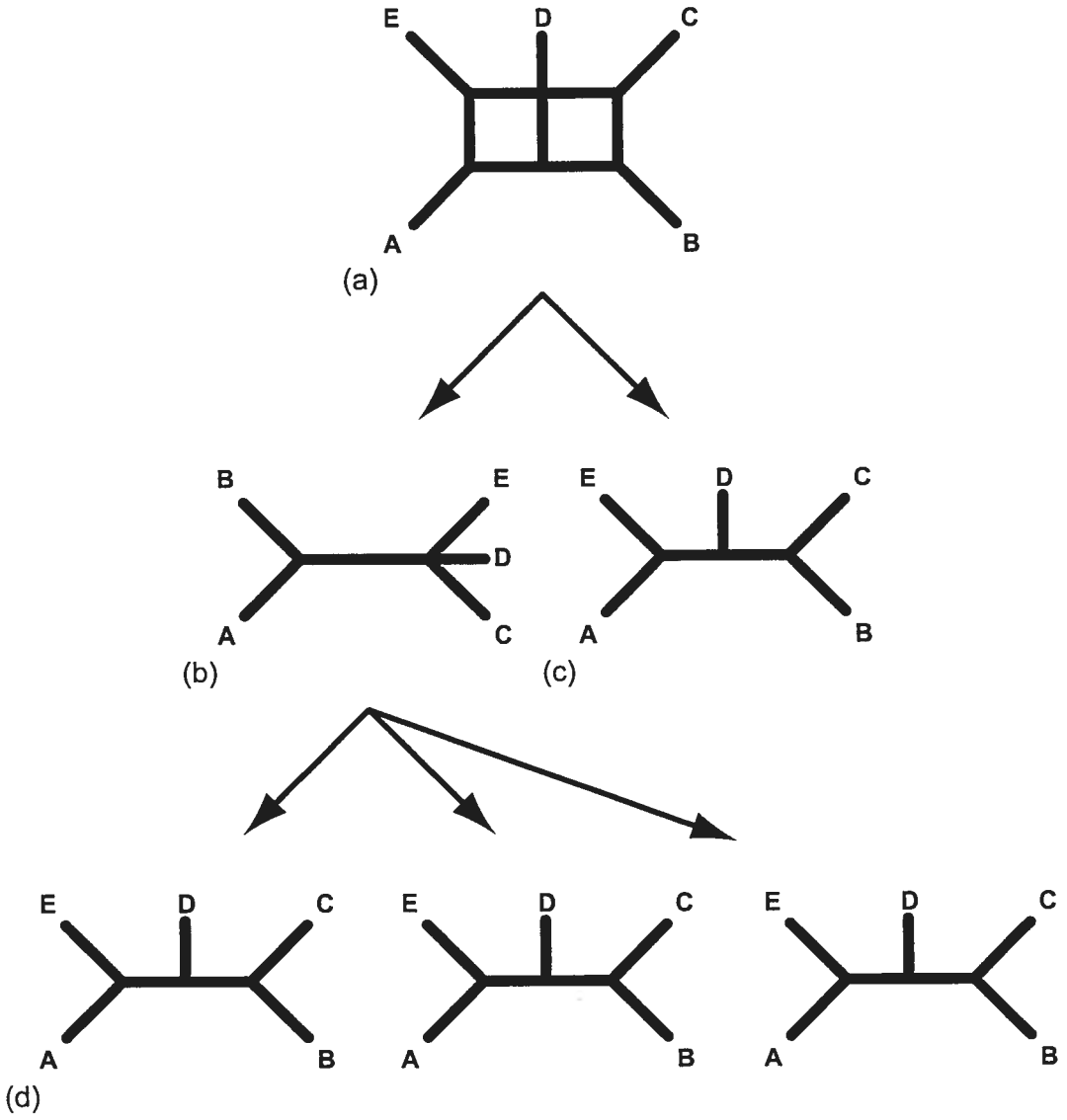


Figure 6.3 (a) the network with splits $AB|CDE$, $AE|BCD$, and $BC|ADE$ contains two maximally resolved trees (b and c). Tree (b) has three resolutions, while tree (c) is binary. The network thus allows four distinct fully resolved trees and as a $CIC_{rel} = 0.49$.

respectively. If T is a binary tree, then n_R equals one; otherwise n_R equals the number of binary trees that are compatible with T; i.e. the number of binary trees that refine T. The CIC and the resolution of individual trees can be computed within RadCon (Thorley & Page 2000).

6.6. Application

In order to illustrate the use of CN and CIC we constructed different consensus of the eight mammalian gene trees presented in Fig. 6.4 (data from Springer *et al.* 1999). The strict consensus (not shown) is completely unresolved, and has a corresponding CIC of 0. The majority rule consensus offers some resolution (Fig. 6.5a); it can be refined by 945 different unrooted binary trees, and has a CIC_{rel} of 0.61. The CIC of the consensus is improved by using a CN. The median CN containing the splits included in at least two trees (Fig. 6.5b) has a CIC_{rel} of 0.66 (396 trees), whereas the split decomposition CN (Fig. 6.5c) has a CIC_{rel} of 0.74 (90 trees). If only the splits included in at least three trees are accounted for, the CIC_{rel} of the median CN is 0.80 (30 trees; Fig. 6.5d). The unrestricted median CN (including all the splits in the input trees) contained little more information than a star tree; and exhibited the highly incompatible signals in the input trees.

6.7. Discussion

Many information measures have been proposed for consensus trees (e.g. Mickevich & Farris 1981; Mickevich & Platnick 1989) but, as reviewed by Thorley (2000), they have many shortcomings. More importantly, most are affected by tree shape, such that pectinate trees are considered more informative than balanced trees. The CIC, on the other hand, is shapeless and takes into account the relative position and size of the components (Thorley 2000). It is an appropriate measure of the information content of phylogenies. This measure can thus help researchers choose among competing consensus methods and solutions: the most informative ones should be preferred as long as the information they convey is relevant.

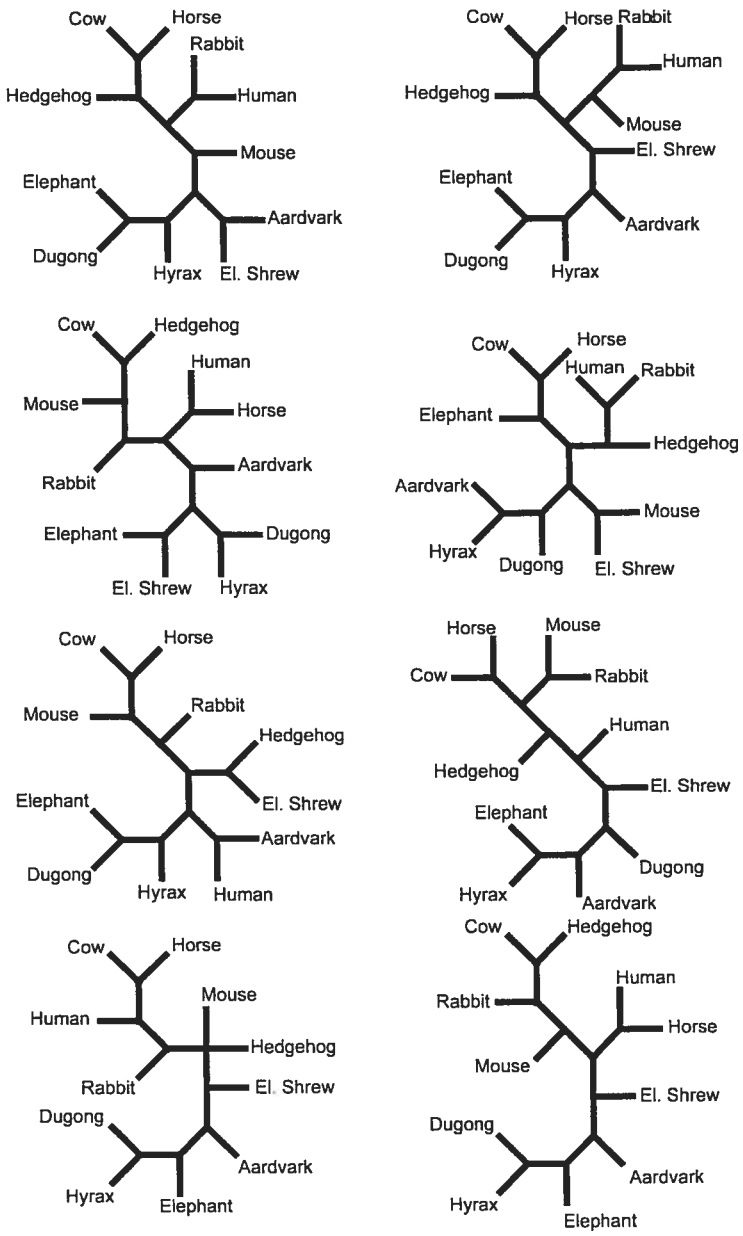


Figure 6.4 Phylogenetic trees for 11 mammalian species obtained from eight independent mitochondrial and nuclear genes (data from Springer *et al.* 1999).

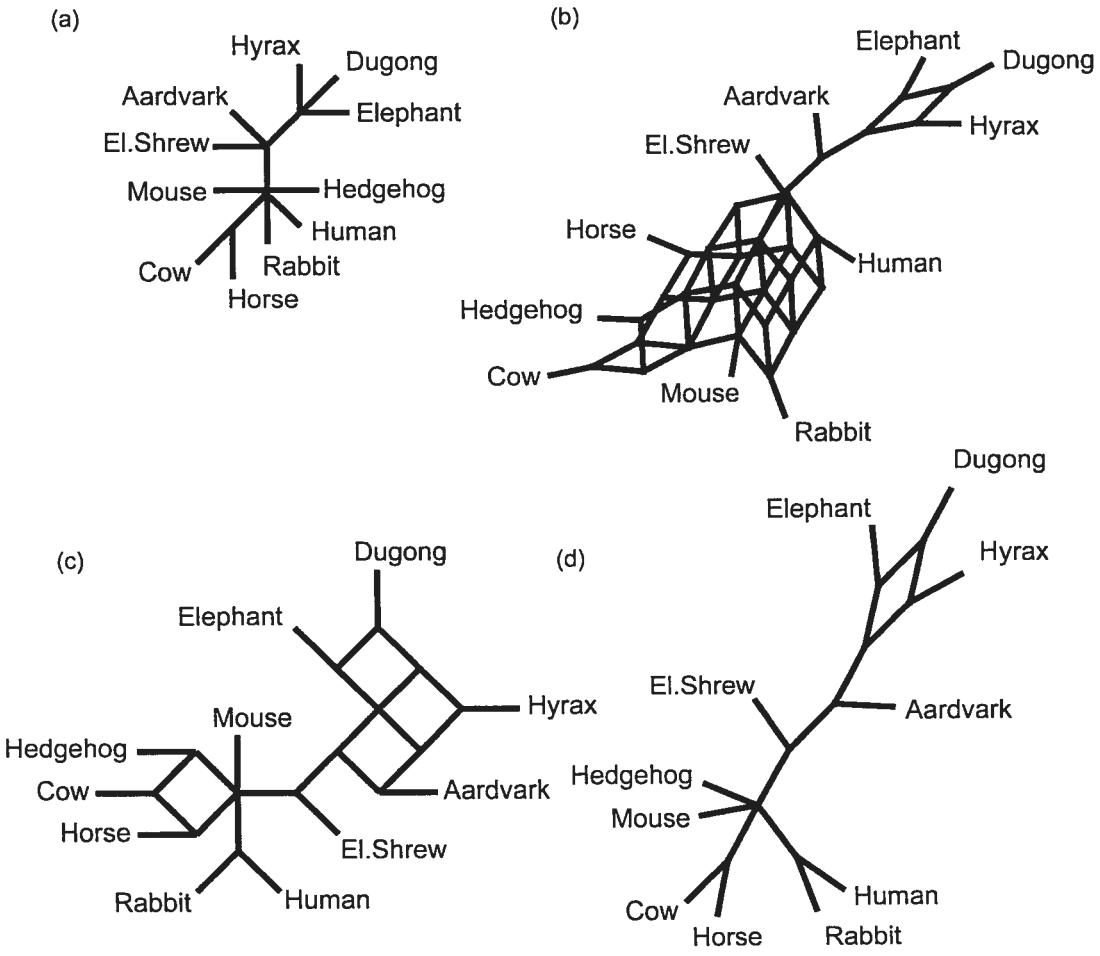


Figure 6.5 Different consensus representations of the profile of trees in Fig. 6.4. (a) The majority rule consensus ($CIC_{rel} = 0.61$); (b) the median CN restricted to splits contained in at least two trees ($CIC_{rel} = 0.66$); (c) the split decomposition CN ($CIC_{rel} = 0.74$); (d) the median CN restricted to splits contained in at least three trees ($CIC_{rel} = 0.80$).

Relevancy is dependent on the axiomatic properties of the methods, such that quantity of information should not be equated with its quality. Indeed, a consensus method that returns a random binary tree would score maximally with a measure such as the CIC: it apparently provides a very informative solution. Nonetheless, it contains no relevant phylogenetic information because it preserves only the names of the taxa in the input trees. The information contained in a split decomposition CN is relevant because it consists only of the primary and secondary signals in the input profile of trees.

When T is a consensus tree, Thorley *et al.* (1998) state that the CIC can only be used if T is strict (*sensu* Wilkinson 1994); i.e. if T contains only relationships that are present in all the input trees. For example, the strict consensus tree is strict for components because it contains only components present in all input trees, while the Adams consensus is strict on nesting because it contains only nestings present in all input trees (Adams 1972). Hence they consider a consensus as being only a summary or representation of the source trees and equate the trees counted by n_R not with the binary trees that are compatible with T, but with the trees that could have been represented among the input trees. Following this definition, majority rule consensus trees, all MR consensus trees, as well as the consensus networks presented here, contain no phylogenetic information because it is impossible "to deduce from the consensus tree" *or network* "alone which of the possible trees could not have been represented among the fundamentals" (Thorley *et al.* 1998). The problem we have with this assertion is that it concludes that most of the phylogenetic supertrees published so far do not contain any phylogenetic information. Indeed, supertree methods are generalizations of consensus methods for input trees defined on partially overlapping sets of taxa; the vast majority of supertree methods are not strict (but see the strict component supertree of Bryant 2002).

We see no reason not to consider a consensus method as a meta-analytical tool. Consequently it is not required to be strict in order for the above measure of phylogenetic information to be applicable. This allows considering as

phylogenetically informative the solutions of consensus methods that preserve only the most supported relationships (e.g. majority-rule consensus), strip the least supported ones (e.g. consensus splitsgraphs), or optimize a given criterion (e.g. average consensus of Lapointe & Cucumel 1997). Moreover, it allows for the measure of the information content of the numerous supertrees published so far.

While Thorley *et al.* (1998) did not consider branch lengths in their CIC, this important element could be taken into account in a weighted version of the information measure of CN. This might prove difficult, since we must not only deal with the relative order of internal nodes as in the Dendritic Information Content (Thorley 2000), but also with the relative support of competing hypotheses. Although such developments are beyond the scope of this paper, it should be pointed out that, in this context, a multidimensional CN (using the method of Holland & Moulton 2003 for example) might be more informative than a planar CN. However, a high dimensionality of the solution will still leave it overwhelmingly difficult to interpret and most probably contain much noise.

However, when considering only the topology of the consensus solution to calculate its CIC, a CN using split decomposition will usually contain more information than one using median networks. Indeed, as Cassens *et al.* (2005) rightfully noted, "a graph with maximum compatibility (i.e., [...] a complete graph [...] where each node is connected to all others) has a 100% compatibility [...] yet it is of little value since it conveys no genealogical information". Thus, unless it is constrained to low dimensionality, a median CN can carry little information while being hard to read. However, the CN based on split decomposition will return a star tree only when the three topologies are equally supported (Wilkinson *et al.* 2003). When the trees are differentially supported, the least supported topology is eliminated, thus increasing the phylogenetic information content of the CN.

One way to further maximize the information content of consensus solutions without any more relaxation of the treeness criterion would be to use reduced CN profiles. Reduced consensus methods are more sensible than standard

consensus to information shared among competing trees (Wilkinson 1994, 1995, 1996). Although the formal definition of a reduced CN is beyond the scope of this paper, a surrogate to reduced CN would be to sequentially remove problematic taxa from the matrix representation before computing the solution. To convey more information than a conventional consensus tree, a CN should not contain complex multidimensional sets of incompatible splits because these cannot be more informative than polytomies. CN can thus be used to increase the CIC of consensus solutions, not with arbitrary conflict resolution or by displaying all the input trees, but by weeding out less supported relationships.

A current concern in phylogenetics is our ability to construct large phylogenies using supermatrix or supertree approaches (Sanderson *et al.* 1998; Bininda-Emonds *et al.* 2002; Wilkinson *et al.* 2005). Although we are not concerned here with supermatrices or the debate on whether we should combine data or trees, we would like to point out that the framework we have outlined here for inferring CN is easily extendable to the construction of super networks, at least in theory. However, combining the MR of trees (or networks) defined on partially overlapping sets of taxa leads to a number of new problems. In the first place, it generates missing data in the MR of the input trees; pseudo-characters are coded as missing for all the taxa not present in a given tree, while no distance exist for pairs of taxa that are never present in the same tree. Missing distances can be estimated using tested and accurate methods (Levasseur *et al.* 2003; Makarenkov & Lapointe 2004), and most network construction algorithms can accommodate missing values in a character matrix; for example the distances required for split decomposition can be computed on a data matrix with missing cells. However, the impact of these new elements, missing data and the recourse to estimated distances, have not been assessed in the context of phylogenetic network inference and should be addressed before we can confidently recommend the construction of such supernetworks, and measure their information content.

6.8. Acknowledgements

The authors are grateful to all members of the LEMEE for constructive comments on an earlier version of this manuscript. This work was supported by NSERC and FQRNT scholarships to OG, NSERC grant no. OGP0155251 and FQRNT grant no. PR88559 to FJL. Thanks are extended to M. Springer for making available the data used in the application.

Chapitre 7 :

CONCLUSION

Dans le cadre de cette thèse, je me suis penché sur la pertinence et l'utilisation des méthodes de réseau en analyse phylogénétique. Je me suis intéressé à deux grandes applications de ces méthodes : (1) la détection de taxons étant issus d'un événement de réticulation, notamment ceux qui sont issus de l'hybridation interspécifique; et (2) la combinaison d'arbres multiples à l'aide de réseaux consensus. Il est d'intérêt historique de rappeler que les réseaux ont d'abord été introduits en analyse phylogénétique dans le but de représenter, puis éventuellement de détecter, l'évolution réticulée (Sneath 1975; Nelson 1983; Jakobsen & Easteal 1996; Fitch 1997; Xu 2000; Legendre & Makarenkov 2002; Moret *et al.* 2004). Toutefois, si la représentation de réticulations connues sur une phylogénie constitue une opération simple, la détection de tels événements à l'aide d'une méthode de reconstruction phylogénétique demeure encore aujourd'hui, et malgré les résultats que j'ai obtenus, une opération difficile. Par contre, la recherche de nouvelles applications pour les réseaux phylogénétiques a mené à des développements fort intéressants. En effet, l'application des réseaux au problème du consensus m'apparaît aujourd'hui comme étant, à court terme, l'avenue la plus profitable pour ces méthodes. Si les réseaux étaient pratiquement absents de la littérature phylogénétique, il y a cinq ans, alors que j'entamais la recherche dont les résultats sont présentés dans cette thèse, ils se retrouvent aujourd'hui dans de nombreuses publications (e.g. : Bryant & Moulton 2002; Cassens *et al.* 2003, 2005; Holland *et al.* 2004, 2005; Nakhleh *et al.* 2003, 2004). Toutefois, la plupart des articles publiés qui les abordent, tout comme cette thèse, traitent encore et toujours des propriétés de ces méthodes, en présentent de nouvelles, et argumentent pour une plus grande utilisation de ces dernières. De nombreuses questions restent encore sans réponse et beaucoup de travail reste à faire.

La première conclusion de ma thèse porte sur la complexité de l'inférence d'une phylogénie réticulée. En effet, tel que présenté au Chapitre 2, aucune des deux méthodes mise à l'épreuve n'a été en mesure d'identifier correctement des hybrides d'origine connue. Si la décomposition des bipartitions n'a pas été initialement développée pour ce type d'analyse, les

réticulogrammes eux l'ont été, et il apparaît donc qu'ils ne remplissent pas leur fonction première. Bien que cette méthode présente un intérêt certain pour des applications biogéographiques, par exemple, retracer la conquête du territoire après la fonte des glaciers par des routes de colonisation primaires et secondaires (Legendre & Makarenkov 2002), sa validité dans le contexte de la détection des taxons hybrides est fortement mise en doute par mes résultats. La performance de la décomposition des bipartitions dans ce contexte n'a pas été plus encourageante, cette approche ne permettant, dans les cas étudiés, que l'identification de certains hybrides entre lignées sœurs. Ceci découle directement de la manière dont procède la méthode : en présence d'un taxon résultant de l'hybridation entre lignées éloignées le signal réticulé aura des répercussions sur l'ensemble des quadruplets contenant ces trois taxons, et donc, sur l'ensemble du graphe qui prendra la forme d'une toile complexe. Le problème que constitue la détection de taxons hybrides n'a donc pu être résolu à l'aide de l'application directe de méthodes d'analyse phylogénétique.

En effet, il a été nécessaire de réduire le problème à des quadruplets de taxons afin de pouvoir mettre en évidence la position phylogénétique intermédiaire des hybrides (Chapitre 2). À l'aide du test statistique élaboré au Chapitre 3 il est effectivement possible d'identifier avec justesse des taxons hybrides ainsi que les lignées leur ayant donné naissance. Les résultats obtenus tant avec les hybrides connus d'*Aphelandra* (Chapitre 3), les hybrides artificiels de *Petrogale* (Chapitre 4), ainsi qu'avec la simulation de séquences d'ADN le long des branches de phylogénies réticulées (Chapitre 5) confirment la validité de la méthode. Par contre, les résultats négatifs obtenus avec les hybrides naturels de kangourous *Petrogale* mettent en évidence la portée limitée de l'approche à certains types de données (Chapitre 4). En effet, si les hybrides peuvent présenter en moyenne des états de caractères morphologiques intermédiaires entre leurs parents, cela ne va pas nécessairement de soi avec les données moléculaires pour lesquelles une des deux lignées parentales peut être exprimée majoritairement (McDade 1995, Rieseberg 1998). Bien que les données utilisées et les résultats obtenus ne permettent pas de tirer de conclusion définitive à ce sujet, il est probable que

des portions du génome de l'un ou l'autre des parents sont éliminées par un phénomène de réplication comme la recombinaison inégale. Une telle éventualité expliquerait le comportement différent des hybrides artificiels et naturels, et l'incapacité de la méthode proposée ici à détecter ces derniers à l'aide de telles données. Par contre, dans le cas des simulations (Chapitre 5) ainsi que des hybrides dits artificiels de *Petrogale* (Chapitre 4), la construction expérimentale des taxons résulte en une position phylogénétique intermédiaire des hybrides par rapport à leurs parents. Nous sommes donc en mesure de nous demander si ce caractère intermédiaire des hybrides est purement artificiel et ne reflète en rien la réalité. Il apparaît évident que cela est probablement le cas en ce qui concerne les hybrides artificiels de *Petrogale*, considérant les résultats obtenus avec les hybrides naturels. Partant de là, je pourrais également critiquer la méthode de simulations, étant donné que, dans les faits, combiner des extraits cellulaires équimolaires des deux parents et concaténer des séquences d'ADN de mêmes longueurs provenant également des deux parents, sont des procédés qui semblent très similaires. Si cette méthode n'est pas parfaite, je crois toutefois que la combinaison de séquences de longueurs différentes, jusqu'à un ratio de 1:7, simule adéquatement une situation ou l'hybride n'est pas parfaitement intermédiaire entre ses deux lignées parentales. Il faut également garder à l'esprit que c'est avant tout le résultat, et non le processus, qui est simulé ici. Les résultats positifs obtenus dans ces conditions me portent à croire que la méthode est appropriée avec des données de séquence, mais qu'elle est probablement inadéquate pour des données d'hybridation ADN-ADN. Le test de HDC permet donc la détection de lignées hybrides situées à la cime d'une phylogénie. Son utilisation permet donc également d'inférer une phylogénie sous la forme d'un arbre à laquelle ces hybrides peuvent êtres greffés *a posteriori*. Toutefois, cette approche devra être plus amplement étudiée si elle doit être appliquée à d'autres groupes et différents types de données. Je pense qu'il est également d'intérêt de rappeler que la procédure présentée au Chapitre 3 peut-être utilisée à des fins exploratoires, sans avoir à formuler une hypothèse d'hybridation complète. Ainsi, on pourra rechercher dans un jeu de données quels sont les hybrides

potentiels en ne proposant qu'un, ou voire même aucun, parent. De même, il est possible de rechercher le taxon qui est l'hybride le plus probable entre deux parents donnés, par exemple, entre deux espèces qui sont reconnues pour s'hybrider en nature. Il est toutefois nécessaire de tester les hypothèses ainsi identifiées à l'aide d'un nouveau jeu de données.

Finalement, une des conclusions les plus importantes de ma thèse concerne l'application des réseaux au problème de la combinaison d'arbres en analyse phylogénétique (Chapitre 6). J'ai tout d'abord défini un cadre méthodologique général pour la construction de réseaux consensus. Par la suite, en généralisant la mesure du contenu en information phylogénétique de Thorley *et al.* (1998) au domaine des réseaux, j'ai pu démontrer l'intérêt de ces méthodes à l'aide d'exemples théoriques et pratiques. Il est donc profitable pour le praticien d'avoir recours aux réseaux phylogénétiques non seulement pour illustrer, et éventuellement détecter l'évolution réticulée, mais également pour représenter l'incertitude, le flou, qui accompagne habituellement toute hypothèse phylogénétique. En effet, de nombreuses situations mènent à la définition d'arbres multiples sur un même ensemble de taxons. Qu'il s'agisse de solutions également optimales, d'arbres de gènes, ou du résultat de la validation interne d'un arbre, il est avantageux d'utiliser un réseau consensus, plutôt qu'un arbre consensus, afin de maximiser le contenu en information dudit consensus. Ainsi, les réseaux consensus permettent de raffiner les arbres consensus qui ne sont que rarement binaires. De cette manière, il est possible de visualiser conjointement des hypothèses concurrentes réunies en une seule représentation graphique, plutôt que de les ignorer et aboutir avec des polytomies multiples. La généralisation de cette approche à la combinaison d'arbres définis sur des ensembles de taxons ne se chevauchant que partiellement, le domaine des super-arbres et des super-réseaux, mérite certainement d'être étudiée, mais dépasse largement le cadre de cette thèse.

En conclusion, cette thèse n'a pas pour but de répondre à toutes les questions que nous nous posons, ou que nous devrions nous poser, sur l'utilisation de réseaux en analyse phylogénétique. Par contre, elle apporte un

certain nombre de réponses qu'il était nécessaire d'obtenir avant de pouvoir utiliser adéquatement ces méthodes. Dans ce contexte, je crois, et j'espère, que les analyses et les développements qui sont présentés dans cet ouvrage permettront, d'une part, de mieux utiliser ces méthodes et, d'autre part, de guider ceux qui chercheront à développer de nouvelles méthodes d'inférence de réseau et de détection d'événements de réticulation. Une partie du travail à venir se rapporte à la promotion de ces méthodes qui permettent de s'écarter du modèle arboré, même dans des situations où l'évolution réticulée n'est pas mise en cause. Ceci pourra être difficile alors que les chercheurs font déjà face à une multitude de méthodes d'analyse et de manières de procéder en compétition les unes avec les autres, et s'en remettent le plus souvent à celles qui sont disponibles dans les logiciels d'analyse phylogénétique les plus communs et les plus conviviaux. À ce titre le logiciel SplitsTree4.0 (Huson & Bryant 2006) qui implémente de nombreuses méthodes d'inférence de réseau et permet, entre autres, la construction de réseaux consensus, à l'aide d'une interface conviviale, aura un rôle essentiel à jouer. Une plus grande utilisation des réseaux est souhaitable en analyse phylogénétique, particulièrement dans le contexte actuel ou les données moléculaires sont abondamment disponibles. Non seulement pourront-ils nous permettre de reconnaitre de réelles réticulations, mais ils nous aideront aussi à concilier les hypothèses phylogénétiques parfois contradictoires que nous rencontrons dans la recherche de l'Arbre du Vivant, arbre qui semble être affublé d'un bon nombre de lianes se projetant entre ses branches.

BIBLIOGRAPHIE

Adams, E. N., III. 1972. Consensus techniques and the comparison of taxonomic trees. *Systematic Zoology* **21**: 390–397.

Allendorf, F. W., R. F. Leary, P. Spruell, & J. K. Wenburg. 2001. The problem with hybrids: setting conservation guidelines. *Trends in Ecology and Evolution* **16**:613–622.

Anderson, E. 1936. Hybridization in American *Tradescantias*. *Annals of the Missouri Botanical Garden* **23**:511–525.

Anderson, E. & G. L. Stebbins Jr. 1954. Hybridization as an evolutionary stimulus. *Evolution* **8**:378–388.

Arnold, M. L. 1992. Natural Hybridization as an evolutionary process. *Annual Reviews of Ecology and Systematics* **23**:237–261.

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

Arnold, M. L. & S. A. Hodges. 1995. Are natural hybrids fit or unfit relative to their parents? *Trends in Ecology and Evolution* **10**:67–71.

Baccam, P., R. J. Thompson, O. Fedrigo, S. Carpenter, & J. L. Cornette. 2001. PAQ: Partition analysis of quasispecies. *Bioinformatics* **17**:16–22.

Bandelt, H.-J. 1994. Phylogenetic networks. *Verhandl Naturwiss Vereins Hamburg* **34**:51–71.

Bandelt, H.-J. 1995. Combination of data in phylogenetic analysis. *Plant Systematics and Evolution Supplement* **9**:355–361.

Bandelt, H.-J. & A. W. M. Dress. 1992. Split decomposition: A new and useful approach to phylognetic analysis of distance data. *Molecular phylogenetics and Evolution* **1**:242–252.

Bandelt, H.-J., P. Forster, B. C. Sykes, & M. B. Richards. 1995. Mitochondrial portraits of human populations using median networks. *Genetics* **141**:743–753.

Bandelt, H.-J., V. Macaulay, & M. Richards. 2000. Median networks: Speedy construction and greedy reduction, one simulation, and two case studies from human mtDNA. *Molecular phylogenetic and Evolution* **16**:8–28.

Barton, N. H. & G. M. Hewitt. 1985. Analysis of hybrid zones. *Annual reviews of ecology and systematics* **16**:113–148.

Baum, B. R. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon* **41**:3–10.

Bininda-Emonds, O. R. P., J. L. Gittleman, & M. A. Steel. 2002. The (super)tree of life: Procedures, problems, and prospects. *Annual Review of Ecology and Systematics* **33**:265–289.

Bondy, J. A. & U. S. R. Murty. 1976. Graph theory with applications. Macmillan Press, London.

Brillouin, L. 1962. *Science and information theory*. 2nd English edition. Academic Press, New York.

Brower, A. V. Z., R. DeSalle, & A. Vogler. 1996. Gene trees, species trees, and systematics: A cladistic perspective. *Annual reviews of ecology and systematics* **27**:423–450.

Bryant, D. 2002. *Strict consensus supertrees*. Technical report, School of Computer Science, McGill university, Montréal, Canada.

Bryant, D. 2003. A classification of consensus methods for phylogenetics. pp. 163–184 dans *Bioconsensus* édité par M. Janowitz, F.-J. Lapointe, F. McMorris, B. Mirkin, & F. B. Roberts.

Bryant, D. & V. Moulton. 2002. NeighborNet: an agglomerative algorithm for the construction of phylogenetic networks, pp. 375–391 dans *Workshops on Algorithms in Bioinformatics (WABI) 2002* édité par R. Guigo & D. Gusfield, Springer-Verlag, Berlin.

Bryant, D. & V. Moulton. 2004. NeighborNet: an agglomerative algorithm for the construction of phylogenetic networks. *Molecular Biology and Evolution* **21**:255–265.

Bullini, L. 1994. Origin and evolution of animal hybrid species. *Trends in Ecology and Evolution* **9**:422–426.

Buneman, P. 1971. The recovery of trees from measures of dissimilarity, pp. 387–395 dans *Mathematics in the Archaeological and Historical Sciences* édité par F. R. Hodson, D. G. Kendall, & P. Tautu, Edinburgh University Press, Edinburgh.

Campeau-Peloquin, A., J. A. W. Kirsch, M. D. B. Eldridge, & F.-J. Lapointe. 2001. Phylogeny of the rock-wallabies, Petrogale (Marsupialia : Macropodidae) based on DNA/DNA hybridisation. *Australian Journal of Zoology* **49**:463–486.

Cassens, I., P. Mardulyn, & M. C. Milinkovitch. 2005. Evaluating Intraspecific "Network" Construction Methods Using Simulated Sequence Data: Do Existing Algorithms Outperform the Global Maximum Parsimony Approach? *Systematic Biology* **54**:363–372.

Cassens, I., K. van Waerebeeek, P. B. Best, E. A. Crespo, J. Reyes, & M. C. Milinkovitch. 2003. The phylogeography of dusky dolphins (*Lagenorhyncus obscurus*): A critical examination of network methods and rooting procedures. *Molecular Ecology* **12**:1781–1792.

Cavalli-Sforza, L. L. & A. W. F. Edwards. 1967. Phylogenetic analysis. Models and estimation procedures. *American Journal of Human Genetics* **21**:550–570.

Close, R. L., & J. N. Bell. 1997. Fertile hybrids in two genera of wallabies: Petrogale and Thylogale. Journal of Heredity 88:393–397.

Comes, H. P. & R. J. Abbot. 1999. Reticulate evolution in the Mediterranean species complex of *Senecio* sect. *Senecio*: Uniting phylogenetics and population level approaches, pp. 171–198 dans *Molecular Systematics and Plant Evolution* édité par P. M. Hollingsworth, R. M. Bateman, & R. J. Gornall, Taylor and Francis, Londre.

Comes, H. P. & R. J. Abbot. 2001. Molecular phylogeography, reticulation, and lineage sorting in mediterranean *Senecio* sect. *Senecio* (Asteraceae). *Evolution* **55**:1943–1962.

Darwin, C. 1859. *On the origin of species by means of natural selection*, 1ère édition anglaise, John Murray, Londre.

Daubin, V., M. Gouy, & G. Perriere. 2002. A phylogenomic approach to bacterial phylogeny: Evidence of a core of genes sharing a common history. *Genome Research* **12**:1080–1090.

Doolittle, W. F. 1999. Phylogenetic classification and the universal tree. *Science* **284**:2124–2128.

Dowling, T. E. & C. L. Secor. 1997. The role of hybridization and introgression in the diversification of animals. *Annual Reviews of Ecology and Systematics* **28**:593–619.

Dress, A., D. Husson, & V. Moulton. 1996. Analyzing and visualizing sequence and distance data using Splitstree. *Discrete Applied Mathematics* **71**:95–109.

Edginton, E. S. 1995. *Randomization tests*, 3ième édition, Marcel Dekker Inc., New York, 341 pp.

Edwards, A. W. F. & L. L. Cavalli-Sforza. 1964. Reconstruction of evolutionary trees, pp. 67–76 dans *Phenetic and Phylogenetic Classification* édité par W. H. Heywood & J. McNeill, Systematics Association Publication No. 6, London.

Eldridge, M. D. B. 1997. Taxonomy of rock-wallabies, Petrogale (Marsupialia: Macropodidae). II. An historical review. Autralian Mammalogy 19:113–122.

Eldridge, M. D. B., & R. L. Close. 1993. Radiation of chromosome shuffles. Current Opinionn Genetics and Developments 3:915–922.

Eldridge, M. D. B., R. L. Close, & P. G. Johnston. 1990. Chromosomal rearrangements in rock wallabies, Petrogale (Marsupialia: Macropododae). III. G-banding analysis of Petrogale inornata and Petrogale penicillata. Genome 33:798–802.

Eldridge, M. D. B. & D. J. Pearson. 1997. Chromosomal rearrangements In rock wallabies, Petrogale (Marsupialia, Macropodidae). 9. Further G-branding studies of the Petrogale lateralis complex – P. lateralis pearsoni, the west Kimberley race, and a population heterozygous for a centris fusion. Genome 40:84–90.

Ellstrand, N. C., R. Whotkus, & R. Rieseberg. 1996. Distribution of spontaneous plant hybrids. *Proceedings of the National Academy of Sciences of the United States of America* **93**:5090–5093.

Emms, S. K. & M. L. Arnold. 1997. The effect of habitat on parental and hybrid fitness: transplant experiments with Louisiana Irises. *Evolution* **51**:1112–1119.

Felsenstein, J. 1978. The number of evolutionary trees. *Systematic Zoology* **17**:27–33.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution* **16**:368–376.

Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39**:783–791.

Felsenstein, J. 2004. *Inferring phylogenies.* Sinauer, Sunderland.

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

Fitch, W. M. 1997. Networks and viral evolution. *Journal of Molecular Evolution* **44**:S65–S75.

Fitch, W. M. & E. Margoliash. 1967. Construction of phylogenetic trees. *Science* **155**:279–284.

Gandolfi, A., I. R. Sanders, V. Rossi., & P. Menozzi. 2003. Evidence for recombination in putative ancient hybrid asexuals. *Molecular Biology and Evolution* **20**:754–761.

Gauthier, O. & F.-J. Lapointe. 2002. A comparison of alternative methods for detecting reticulation events in phylogenetic analysis, pp. 341–347 dans *Classification, Clustering, and Data Analysis: Recent Advances and Applications* édité par K. Jajuga, A. Sokolowski, & H.-H. Bock, Springer-Verlag, Berlin.

Gauthier, O. & F.-J. Lapointe. 2006. Hybrids and phylogenetics revisited. A statistical test of hybridization using quartets. Accepté pour publication dans *Systematic Botany*

Gower, J. C. 1971. A general coefficient of similarity and some of its properties. *Biometrics* **17**:857–871.

Grosholz, E. 2002. Ecological and evolutionary consequences of coastal invasions. *Trends in Ecology and Evolution* **17**:22–27.

Haeckel, E. 1866. *Generelle Morphologie der Organismen*, Berlin.

Harrison, R. G. 1990. Hybrid zones: Windows on evolutionnary process. *Oxford Surveys in Evolutionnary Biology* **7**:69–128.

Harrison, R. G. 1993. Hybrids and hybrid zones: Historical perspective, pp. 3–12 dans *Hybrid Zones and the evolutionary process* édité par R. G. Harrison, Oxford University Press, Oxford.

Hennig, W. 1979. *Phylogenetic systematics.* University of Illinois Press, Urbana, Illinois, USA.

Hillis, D. M. & J. J. Wiens. 2000. Molecular versus morphological systematics: conflicts, artifacts, and misconceptions pp. 1–19 dans *Phylogenetic analysis of morphological data* édité par J. J. Wiens, Smithsonian Institution Press, Washington, D.C.

Holder, M. T., J. A. Anderson, & A. K. Holloway. 2001. Difficulties in detecting hybridization. *Systematic Biology* **50**:978–982.

Holland, B. R., F. Delsuc, & V. Moulton. 2005. Visualizing conflicting evolutionary hypotheses in large collections of trees: Using consensus networks to stuffy the origins of placentals and hexapods. *Systematic Biology* **54**:66–76.

Holland, B. R., K. T. Huber, A. Dress, & V. Moulton. 2002. δ plots: A tool for analyzing phylogenetic distance data. *Molecular Biology and Evolution* **19**:2051–2059.

Holland, B. R., K. T. Huber, V. Moulton, &, P. J. Lockhart. 2004. Using consensus networks to visualize contradictory evidence for species phylogeny. *Molecular Biology & Evolution* **21**:1459–1461.

Holland, B. & V. Moulton. 2003. Consensus networks: a method for visualising incompatibilities in a collection of trees pp. 165–176 dans *Third International Workshop in Algorithms in Bioinformatics (WABI)*, Springer.

Holliday, T. W. 2003. Species concepts, reticulations, and human evolution. *Current anthropology* **44**:653–673.

Humphries, C. J. 1983. Primary data in hybrid analysis. pp. 89–103 dans *Advances in Cladistics: Proceedings of the Second meeting of the Willi Hennig Society* édité par N. I. Platnick & V. A. Funk, Columbia University Press, New York.

Huson, D. H. 1998. SplitsTree: A program for analyzing and visualizing evolutionary data. *Bioinformatics* **14**:68–73.

Huson, D. H. & D. Bryant. 2006 Application of Phylogenetic Networks in Evolutionary Studies. *Molecular Biology and Evolution* **23**:254–267.

Jakobsen, I. B. & S. Easteal. 1996. A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. *Computer Applications in the BIOSciences* **12**:291–295.

Jakobsen, I. B., S. R. Wilson, & S. Easteal. 1997. The partition matrix: Exploring variable phylogenetic signals along nucleotide sequence alignments. *Molecular Biology and Evolution* **14**:474–484.

Johnston, J. A., R. A. Wesselingh, A. C. Bouck, L. A. Donovan & M. L. Arnold. 2001. Intimately linked or hardly speaking? The relationship between genotype and environmental gradients in a Louisiana Iris hybrid population. *Molecular Ecology* **10**:673–681.

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

Kendall, M. G. 1938. A new measure of rank correlation. *Biometrika* **30**:81–93.

Kirsch, J. W. A., C. Krajewski, M. S. Springer, & M. Archer. 1990. DNA-DNA hybridization studies of carnivorous marsupials. II. Relationships among dasyurids (Marsupialia: Dasyuridae). Australian Journal of Zoology 38:673-696.

Landry, P.-A., F.-J. Lapointe, & J. A. W. Kirsch. 1996. Estimating phylogenies from lacunose distance matrices: additive is superior to ultrametric estimation. *Molecular Biology and Evolution* **13**:818–823.

Lapointe, F.-J. 2000. How to account for reticulation events in phylogenetic analysis: A comparison of distance based methods. *Journal of classification* **17**:175–184.

Lapointe, F.-J. & G. Cucumel. 1997. The average consensus procedure: Combination of weighted trees containing identical or overlapping sets of objects. *Systematic Biology* **46**:306–312.

Lapointe, F.-J. & G. Cucumel. 2002. Multiple Consensus Trees. pp. 359–364 in Classification, Clustering, and Data Analysis, K. Jajuga, A. Sokolowski, H.-H. Bock [eds.], Springer, Berlin.

Lapointe, F.-J. & G. Cucumel. 2003. How good can a consensus get? Assessing the reliability of consensus trees in phylogenetic studies. pp. 205–219 dans *Bioconsensus* édité par M. Janowitz, F. J. Lapointe, F. McMorris, B. Mirkin, & F. B. Roberts, DIMACS series in discrete mathematics and theoretical computer science, American Mathematical Society, Providence, Rhode Island.

Lapointe, F.-J., M. Wilkinson, &, D. Bryant. 2003. Matrix representations with parsimony or with distances: Two sides of the same coin? *Systematic Biology* **52**:865–868.

Larget, B. & D. L. Simon 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Molecular Biology and Evolution 16:750–759.

Leclerc, B. & Cucumel, G. 1987. Consensus en classification : une revue bibliographique. Mathématique en Sciences Humaines 100 :109-128.

Legendre, P. 2000. Special Section on Reticulate Evolution. *Journal of Classification* **17**:153–195 comprend des articles de F.-J. Lapointe, P. Legendre, F. J. Rohlf, P. E. Smouse, & P. H. A. Sneath.

Legendre, P. 2000a. Biological applications of reticulation analysis. *Journal of Classification* **17**:191–195.

Legendre, P. 2000b. Reticulate evolution: From bacteria to philosopher. *Journal of Classification* **17**:153–157.

Legendre, P. & V. Makarenkov. 2002. Reconstruction of biogeographic and evolutionary networks using reticulograms. Systematic Biology 51:199–216.

Levasseur, C., P. A. Landry, V. Makarenkov, J. A. W. Kirsch, & F. J. Lapointe. 2003. Incomplete distance matrices, supertrees and bat phylogeny. *Molecular Phylogenetics & Evolution* **27**:239–246.

Maddison, W. 1989. Reconstructing character evolution on polytomous cladograms. *Cladistics* **5**:365-377.

Maddison, W. P. 1997 Gene trees in species trees. *Systematic Biology* **46**:523–536.

Makarenkov, V. 2001. TRex: Reconstructing and Visualizing Phylogenetic Trees and Reticulation Networks, *Bioinformatics* **17**:664–668.

Makarenkov, V. & F.-J. Lapointe. 2004. A weighted least-squares approach for inferring phylogenies from incomplete distance matrices. *Bioinformatics* **20**:2113-2121–2004.

Makarenkov, V. & P. Legendre. 2000. Improving the additive tree representation of a given dissimilarity matrix using reticulations. pp. 35–46 dans *Data Analysis, Classification, and Related Methods* édité par H. A. L. Kiers, J. P. Rasson, P. J. F. Groenen, & M. Schade, Springer, Berlin

Manly, B. J. F. 1997. *Randomization, bootstrap and Monte Carlo methods in biology*, 2ième edition, Chapman and Hall, Londre, 399 pp.

Margush, T. & F. R. McMorris. 1981. Consensus *n*-trees. *Bulletin of Mathematical Biology* **43**:239–244.

Mau, B., M. A. Newton, & B. Larget. 1999. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics* **55**:1–12

McDade, L. A. 1984. Systematics and reproductive biology of the central american species of the *Aphelandra pulcherrima* complex (Acanthaceae). *Annals of the Missouri Botanical Garden* **71**:104–165.

McDade, L. A. 1990. Hybrids and phylogenetic systematics I. Patterns of character expression in hybrids and their implication for cladistic analysis. *Evolution* **44**:1685–1700.

McDade, L. A. 1992. Hybrids and phylogenetic systematics II. The impact of hybrids on cladistic analysis. *Evolution* **46**:1329–1346.

McDade, L. A. 1995. Hybridization and phylogenetics. pp. 305–331 dans *Experimental and molecular approaches to plant biosystematics Proceedings of the Fifth International Symposium of the International Organization of Plant Biosystematists (IOPB)* édité par P. C. Hoch & A. G. Stephenson, Monographs in Systematic Botany from the Missouri Botanical Garden 53.

McDade, L. A. 1997. Hybrids and phylogenetic systematics III. Comparison with distance methods. *Systematic Botany* **22**:669–683.

Mickevich, M. F. & J. S. Farris. 1981. The implications of congruence in Menidia. *Systematic Zoology* **30**:351–370.

Mickevich, M. F. & N. I. Platnick. 1989. On the information content of classifications. *Cladistics* **5**:33–47.

Moret, B.M.E., L. Nakhleh, T. Warnow, C.R. Linder, A. Tholse, A. Padolina, J. Sun, & R. Timme. 2004. Phylogenetic networks: Modeling, reconstructibility, and accuracy." *Transactions on Computational Biology and Bioinformatics* **1**:13–23.

Nakhleh, L., J. Sun, T. Warnow, C.R. Linder, B.M.E. Moret, & A. Tholse. 2003. Towards the Development of Computational Tools for Evaluating Phylogenetic Network Reconstruction Methods dans *Proceedings of the Eighth Pacific Symposium on Biocomputing (PSB 03)*.

Nakhleh, L., T. Warnow, & C. R. Linder. 2004. Reconstructing Reticulate Evolution in Species - Theory and Practice. pp. 337–346 dans *Proceedings of the Eighth Annual International Conference on Research in Computational Molecular Biology (RECOMB 2004)* édité par P. E. Bourne & D. Gusfield, ACM Press, New York, U.S.A.

Nelson, G. J. 1983. Reticulaton in cladograms. pp. 105–111 dans *Advances in Cladistics: Proceedings of the Second meeting of the Willi Hennig Society* édité par N. I. Platnick & V. A. Funk, Columbia University Press, New York.

Nelson, G. J. & N. I. Platnick. 1980. Multiple branching in cladograms: two interpretations. *Systematic Zoology* **29**:86–91.

Page, R. D. M. 1992. Comments on the information content of classifications. *Cladistics* **8**:87–95.

Penny, D. & M. D. Hendy. 1985. Testing methods of evolutionary tree construction. *Cladistics* **1**:266–278.

Penny, D. & M. D. Hendy. 1986. Estimating the reliability of evolutionary trees. *Molecular Biology and Evolution* **3**:403-417.

Perry, W. L., D. M. Lodge, & J. L. Feder. 2002. Importance of hybridization between indigenous and nonindigenous freshwater species: An overlooked threat to north American biodiversity. *Systematic Biology* **51**:255–275.

Posada, D. 2002. Evaluation of methods for detecting recombination from DNA sequences: Empirical data. *Molecular Biology and Evolution* **19**:708–717.

Posada, D. & K. A. Crandall. 2001a. Evaluation of methods for detecting recombination from DNA sequences. Computer simulations. *Proceedings of the National Academy of Sciences* **98**:13757–13762.

Posada, D. & K. A. Crandall. 2001b. Intraspecific gene genealogies: Trees grafting into networks. *Trends in Ecology and Evolution* **16**:37–45.

Posada, D. & K. A. Crandall. 2002. The effect of recombination on the accuracy of phylogeny estimation. *Journal of Molecular Evolution* **54**:396–402.

Purvis, A. 1995. A Modification to Baum and Ragans method for combining phylogenetic trees. *Systematic Biology* **44**:251–255.

Ragan, M. A. 1992. Phylogenetic inference based on matrix representation of trees. *Molecular Phylogenetics and Evolution* **1**:53–58.

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

Rannala, B. & Z. Yang. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *Journal of Molecular Evolution* **43**:304–311.

Raybould, A. F. & A. J. Gray. 1994. Will hybrids of genetically modified crops invade natural communities? *Trends in Ecology and Evolution* **9**:85–89.

Rhymer, J. M. & D. Simberloff. 1996. Extinction by hybridization and introgression. *Annual Reviews of Ecology and Systematics* **27**:83–109.

Rieseberg, L. H. 1997. Hybrid origins of plant species. *Annual Reviews of Ecology and Systematics* **28**:359–389.

Rieseberg, L. H. 1998. Molecular ecology of hybridization, pp. 243–265 dans *Advances in Molecular Ecology* édité par G. R. Carvalho, Volume 306, NATO Science Series: Life Sciences, IOS Press, Amsterdam, Pays-Bas.

Rieseberg, L. H. & J. D. Morefield. 1995. Character expression, phylogenetic reconstruction, and the detection of reticulate evolution. 333–354 dans *Experimental and molecular approaches to plant biosystematics Proceedings of the Fifth International Symposium of the International Organization of Plant Biosystematists (IOPB)* édité par P. C. Hoch & A. G. Stephenson, Monographs in Systematic Botany from the Missouri Botanical Garden 53.

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

Rohlf, F. J. 1982. Consensus indices for comparing classications. *Mathematical Biosciences* **59**:131–144.

Rohlf, F. J. 2000. Phylogenetic models and reticulations. *Journal of Classification* **17**:185–189.

Saitou, N, & M. Nei. 1987. The neighbor-joining method: A new method for reconstructing phylogentic trees. *Molecular Biology and Evolution* **4**:406–425.

Sanderson, M. J., A. Purvis, & C. Henze. 1998. Phylogenetic supertrees: Assembling the trees of life. *Trends in Ecology & Evolution* **13**:105–109.

Sang, T. & Y. Zhong. 2000. Testing hybridization hypotheses based on incongruent gene trees. *Systematic Biology* **49**:422–434.

Skala, Z. & J. Zrzavy, J. 1994. Phylogenetic reticulations and cladistics – Discussion of methodological concepts. *Cladistics* **10**:305–313.

Sneath, P. H. A. 1975. Cladistic representation of reticulate evolution. *Systematic Zoology* **24**:360–368.

Sneath, P. H. A. 2000. Reticulate evolution in bacteria and other organisms: How can we study it? *Journal of Classification*. **17**:159–163.

Sneath, P. H. A. & R. R. Sokal. 1973. Principles of numerical taxonomy. Freeman, San Francisco, USA

Sokal, R. R. & P. H. A. Sneath. 1963. *Principles of Numerical taxonomy*. W. H. Freeman, San Francisco.

Sonea, S. & L. G. Mathieu. 2000. *Prokaryotology: A coherent view*, Presses de l'Université de Montréal, Montréal.

Sosef, M. S. M. 1997. Hierarchical models, reticulate evolution and the inevitability of paraphyletic supraspecific taxa. *Taxon* **46**:75–85.

Springer, M. S., H. M. Amrine, A. Burk, & M. J. Stanhope. 1999. Additional support for Afrotheria and Paenungulata, the performance of mitochondrial versus nuclear genes, and the impact of data partitions with heterogeneous base composition. *Systematic Biology* **48**:65–75.

Springer, M. S. & J. A. W. Kirsch. 1991. DNA hybridization, the compression effect, and the radiation of diprotodontian marsupials. Systematic Zoology 40:131–151.

Stebbins, G. L. 1950. *Variation and Evolution in Plants*. Columbia University Press, New York.

Steel, M, A. W. M. Dress, & S. Bocker. 2000. Simple but fundamental limitations on supertree and consensus methods. *Systematic Biology* **49**:363–368.

Strimmer, K., K. Forslund, B. Holland, & V. Moulton. 2003. A novel exploratory method for visual recombination detection. *Genome Biology* **4**:R33.

Strimmer, K. & V. Moulton. 2000. Likelihood analysis of phylogenetic networks using directed graphical models. *Molecular biology and evolution* **17**:875–881.

Strimmer, K., C. Wiuf, & V. Moutlon. 2001. Recombination analysis using directed graphical models. *Molecular Biology and Evolution* **18**:97–99.

Swofford, D. L. 1991. When are phylogeny estimates from molecular and morphological data incongruent?, pp. 295–333 dans *Phylogenetic analysis of DNA sequences*, édité par M. M. Miyamoto & J. Cracraft, Oxford University Press, New York, USA.

Thorley, J. L. 2000. Cladistic information, leaf stability and supertree construction, Thêse de Doctorat, University of Bristol.

Thorley, J. L. & R. D. M. Page. 2000. RadCon: Phylogenetic tree comparison and consensus. *Bioinformatics* **16**:486–487.

Thorley, J. L., M. Wilkinson, & M. A. Charleston. 1998. The information content of consensus trees. pp. 91–98 dans *Advances in data science and classification* édité par A. Rizzi, M. Vichi, & H.-H. Bock, Springer, Berlin.

von Haeseler, A. & G. A. Churchill. 1993. Network models for sequence evolution. *Journal of Molecular Evolution* **37**:77–85.

Wagner, W. H. Jr. 1969. The Role and Taxonomic Treatment of Hybrids. *Bioscience* **19**:785–789.

Wagner, W. H. Jr. 1983. Reticulistics: The recognition of hybrids and their role in cladistics and classification. pp. 63–79 dans *Advances in Cladistics: Proceedings of the Second meeting of the Willi Hennig Society* édité par N. I. Platnick & V. A. Funk, Columbia University Press, New York.

Wanntorp, H.-E. 1983. Reticulated cladograms and the identification of hybrid taxa. pp. 81–88 dans *Advances in Cladistics: Proceedings of the Second meeting of the Willi Hennig Society* édité par N. I. Platnick & V. A. Funk, Columbia University Press, New York.

Wilkinson, M. 1994. Common cladistic information and its consensus representation: reduced Adams and reduced cladistic consensus trees and profiles. *Systematic Biology* **43**:343–368.

Wilkinson, M. 1995. More on reduced consensus methods. *Systematic Biology* **44**:435–439.

Wilkinson, M. 1996. Majority-rule reduced consensus trees and their use in bootstrapping. *Molecular Biology & Evolution* **13**:437–444.

Wilkinson, M., J. A. Cotton, C. Creevey, O. Eulenstein, S. R. Harris, F. J. Lapointe, C. Levasseur, J. O. McInerney, D. Pisani, &, J. L. Thorley. 2005. The shape of supertrees to come: tree shape related properties of fourteen supertree methods. *Systematic Biology* **54**:419–31.

Wilkinson, M., F.-J. Lapointe, & D. J. Gower. 2003. Branch lengths and support. *Systematic Biology* **52**:127–130.

Wilkinson, M. & J. L. Thorley. 2001. Efficiency of strict consensus trees. *Systematic Biology* **50**:610–613.

Winkworth, R. C., D. Bryant, P. J. Lockhart, D. Havell, & V. Moulton. 2005. Biogeographic interpretation of splits graphs: Least squares optimization of branch lengths. *Systematic Biology* **54**:56–65.

Wiuf, C., T. Christensen, & J. Hein. 2001. A simulation study of the reliability of recombination detection methods. *Molecular Biology and Evolution* **18**:1929–1939.

Wolf, Y. I., I. B. Rogozin, N. V. Grishin, & E. V. Koonin. 2002. Genome trees and the Tree of Life. *Trends in Genetics* **18**:472–479.

Xu, S. 2000. Phylogenetic analysis under reticulate evolution. *Molecular Biology and Evolution* **17**:897–907.

Zaretskii, K. 1965. Constructing a tree on the basis of a set of distances between the hanging vertices. *Uspekhi Matematicheskikh Nauk* **20**:90–92 (en russe).

« Richard sighed.
"Well." He said, "it's to do with the project which first made the software incarnation of the company profitable. It was called *Reason*, and in its own way it was sensational."
"What was it?"
"Well, it was a kind of back-to-front program. It's funny how many of the best ideas are just old idea back-to-front. You see there have already been several programs written that help you to arrive at decisions by properly ordering and analysing all the relevant facts so that they then point naturally towards the right decision. The drawback with these is that the decision which all the properly ordered and analysed facts point to is not necessarily the one you want."
"Yeeeess ..." said Reg's voice from the kitchen.
"Well, Gordon's great insight was to design a program which allowed you to specify in advance what decision you wished it to reach, and only then to give it all the facts. The program's task, which it was able to accomplish with consummate ease, was simply to construct a plausible series of logical-sounding steps to connect the premises with the conclusion."
"And I have to say that it worked brilliantly. Gordon was able to by himself a Porsche almost immediately despite being completely broke and a hopeless driver. Even his bank manager was unable to find fault with his reasoning. Even when Gordon wrote it off three weeks later." »

Douglas Adams, Dirk Gently's Holistic Detective Agency