

UNIVERSITÉ DU QUÉBEC À MONTRÉAL

NOUVEAUX ALGORITHMES POUR L'INFÉRENCE DE RÉSEAUX
PHYLOGÉNÉTIQUES

THÈSE
PRÉSENTÉE
COMME EXIGENCE PARTIELLE
DU DOCTORAT EN INFORMATIQUE

PAR
MATTHIEU WILLEMS

JUIN 2018

UNIVERSITÉ DU QUÉBEC À MONTRÉAL
Service des bibliothèques

Avertissement

La diffusion de cette thèse se fait dans le respect des droits de son auteur, qui a signé le formulaire *Autorisation de reproduire et de diffuser un travail de recherche de cycles supérieurs* (SDU-522 – Rév.07-2011). Cette autorisation stipule que «conformément à l'article 11 du Règlement no 8 des études de cycles supérieurs, [l'auteur] concède à l'Université du Québec à Montréal une licence non exclusive d'utilisation et de publication de la totalité ou d'une partie importante de [son] travail de recherche pour des fins pédagogiques et non commerciales. Plus précisément, [l'auteur] autorise l'Université du Québec à Montréal à reproduire, diffuser, prêter, distribuer ou vendre des copies de [son] travail de recherche à des fins non commerciales sur quelque support que ce soit, y compris l'Internet. Cette licence et cette autorisation n'entraînent pas une renonciation de [la] part [de l'auteur] à [ses] droits moraux ni à [ses] droits de propriété intellectuelle. Sauf entente contraire, [l'auteur] conserve la liberté de diffuser et de commercialiser ou non ce travail dont [il] possède un exemplaire.»

REMERCIEMENTS

Je tiens tout d'abord à remercier sincèrement mon directeur de recherches, Vladimir Makarenikov, pour tout le temps qu'il m'a consacré, pour ses encouragements à persévérer, pour ses conseils précieux, ainsi que pour ses relectures minutieuses de mes travaux. Il a su me guider dans un domaine qui ne m'était pas particulièrement familier, tout en me laissant beaucoup de liberté dans mes recherches.

Je remercie également mes collègues de bureau, notamment Nadia Tahiri et Étienne Lord, pour leurs fructueuses collaborations, Alix Boc pour son aide au début de mon parcours en bioinformatique, ainsi que tous les coauteurs des articles qui composent cette thèse.

Je remercie enfin le conseil de recherches en sciences naturelles et en génie du Canada ainsi que la Fondation de l'UQÀM pour leur soutien financier.

TABLE DES MATIÈRES

LISTE DES TABLEAUX	ix
LISTE DES FIGURES	xi
RÉSUMÉ	xvii
CHAPITRE I	
INTRODUCTION	1
1.1 Mise en contexte	1
1.2 Inférence phylogénétique	4
1.3 L'algorithme neighbor-joining	10
1.4 Le principe du maximum de vraisemblance	12
1.5 L'évolution réticulée	15
1.6 Réseaux phylogénétiques : définitions et méthodes d'inférence	20
1.7 Biolinguistique	27
1.7.1 Phylogénie et linguistique	27
1.7.2 Utilisation des réseaux phylogénétiques pour représenter l'évolution des langues	31
1.7.3 Description des données	32
1.8 Présentation des différentes parties de la thèse	36
1.9 Glossaire	38
CHAPITRE II	
UN NOUVEL ALGORITHME EFFICACE POUR L'INFÉRENCE DE RÉ- SEAUX D'HYBRIDATION EXPLICITES EN SUIVANT LE PRINCIPE DE NEIGHBOR-JOINING	41
RÉSUMÉ	43
2.1 Abstract	44
2.2 Introduction	44

2.3	Hybridization	48
2.4	Neighbor-joining for trees	50
2.5	Some properties of hybridization networks	53
2.5.1	Hybrids between neighbor parents	54
2.5.2	Hybrids between non-neighbor parents	54
2.5.3	Two important properties	56
2.6	Identification of hybrids	58
2.7	Algorithm for inferring hybridization networks	62
2.8	Results of simulations	65
2.8.1	Simulations with additive networks	65
2.8.2	Simulations with non-additive networks (i.e., with sequence-based networks)	69
2.8.3	Experiments with real data	73
2.9	Conclusion	77
CHAPITRE III		
UTILISATION DES RÉSEAUX D'HYBRIDATION POUR RETRACER L'ÉVOLUTION DES LANGUES INDO-EUROPÉENNES		79
RÉSUMÉ		81
3.1	Abstract	83
3.1.1	Background	83
3.1.2	Results	83
3.1.3	Conclusion	84
3.2	Background	84
3.3	Methods	89
3.3.1	Data description	89
3.3.2	Reconstruction of explicit linguistic hybridization networks	90
3.3.3	Reconstruction of split graph-based linguistic networks	102
3.3.4	Reconstruction of galled linguistic networks from word trees	104

3.4	Results and Discussion	111
3.4.1	Network relationships within the Germanic group	112
3.4.2	Network relationships within the Latin group	114
3.4.3	Network relationships within the Slavic group	115
3.4.4	Network relationships within the Persian and Sanskrit groups	116
3.4.5	Network relationships within the Celtic and French/Iberian groups	117
3.4.6	Network relationships within the West Germanic and French/Iberian groups	117
3.4.7	Network relationships between IE language groups	118
3.5	Conclusion	119
3.6	Declarations	120
3.6.1	Acknowledgements	120
3.6.2	Funding	120
3.6.3	Availability of data and materials	120
3.6.4	Authors' contributions	120

CHAPITRE IV

CONSTRUCTION DE RÉSEAUX D'HYBRIDATION EXPLICITES EN UTILISANT NEIGHBOR-JOINING ET LE PRINCIPE DU MAXIMUM DE VRAISEMBLANCE	123
---	-----

RÉSUMÉ	125
------------------	-----

4.1	Abstract	126
4.2	Introduction	126
4.3	Methods	128
4.4	Main algorithm	132
4.5	Results of simulations	134
4.6	Analysis of the mosquitoes dataset	136
4.7	Conclusion	140

CHAPITRE V

CONCLUSION	141
ANNEXE A	
CODE EN C++ DU PROGRAMME QUI IMPLÉMENTE L'ALGORITHME PRÉSENTÉ DANS LE CHAPITRE 2	145
ANNEXE B	
COMPARAISON ENTRE LA DISTANCE SCA ET LA DISTANCE DE LEVENSHTEIN	167

LISTE DES TABLEAUX

Tableau	Page
2.1 Identification of hybrids' parents for additive networks with one hybrid having two descendants.	69
2.2 True positive rates for $n = 32$ and one hybrid with hybridization level $\alpha = 0.5$ in additive networks.	69
2.3 Identification of hybrids' parents in non-additive networks with one hybrid and hybridization level $\alpha = 0.5$	71
2.4 Average number of iterations μ (and the corresponding standard deviation σ) after which hybrids were detected in networks with one hybrid and hybridization level $\alpha = 0.5$	73
2.5 Dataset of the restriction maps of the rDNA cistron of 16 species of mosquitoes constructed using 8 recognition restriction enzymes.	74
3.1 This table reports the results provided by the word borrowing event detection algorithm (Boc et al., 2010a) applied to the normalized Levenshtein (Levenshtein, 1966) and SCA (List, 2012a) distance matrices.	106

LISTE DES FIGURES

Figure	Page
1.1 Modèle de base d'un arbre phylogénétique.	5
1.2 Exemple d'une distance d'arbre sur un ensemble X de 6 taxons et l'arbre phylogénétique associé.	6
1.3 Scénario d'évolution le plus parcimonieux.	9
1.4 Buisson de taille 6.	11
1.5 Configuration où les nœuds i et j sont choisis comme voisins. . . .	12
1.6 Configuration de NJ utilisée avec un critère de maximum de vraisemblance.	14
1.7 Un réseau réticulé (c.-à-d., un réticulogramme) montrant une relation additionnelle entre les ancêtres des espèces b et d.	16
1.8 Le réseau réticulé représenterait mieux l'histoire de la vie qu'un arbre phylogénétique classique (Doolittle, 1999).	16
1.9 Un exemple de réseau d'hybridation.	18
1.10 Deux arbres possibles pour le phénomène d'hybridation présenté sur la figure 1.9.	19
1.11 Un exemple de split-graphe.	21
1.12 Un réseau de niveau 1 (a) et un réseau de niveau 2 (b).	25
1.13 L'arbre d'évolution des langues IE pour 14 des principaux groupes linguistiques. Les nombres sur les branches représentent leurs scores de bootstrap. Le nombre de langues appartenant à chaque groupe est indiqué entre parenthèses.	29
1.14 L'arbre complet d'évolution de 84 langues IE selon Gray & Atkinson (2003); Atkinson & Gray (2006). Les nombres sur les branches représentent leurs scores de bootstrap.	30

1.15 Split-graphes pour 9 langues IE.	33
1.16 Hybrides entre branches terminales.	35
1.17 Hybrides entre branches ancestrales.	35
2.1 An example of hybridization.	48
2.2 Two different trees for representing the same hybridization phenomenon of Fig. 2.1.	49
2.3 Configuration where nodes i and j are chosen as neighbors.	52
2.4 Hybrids between terminal branches.	53
2.5 Hybrids between ancestral branches.	53
2.6 Network configuration in which species h is a hybrid of two neighbor species i and j	55
2.7 Network configuration in which species h is a hybrid of two non-neighbor species i and j	56
2.8 Hybrid h whose parent i_1 has a direct neighbor i_2	58
2.9 Two networks corresponding to the same distance matrix of size 4.	64
2.10 Average simulation results for additive networks with hybridization level $\alpha = 0.3$ (Δ), $\alpha = 0.4$ ($\square$) and $\alpha = 0.5$ ($\diamond$), and with $DIFFMAX = 0.25$. Figure (a) (respectively (c)) shows the true (respectively false) positive rate as a function of the tree size. Figure (b) (respectively (d)) shows the true (respectively false) positive rate as a function of the number of hybrids.	67
2.11 Two network topologies used in our simulations with additive networks.	68
2.12 Average simulation results for non-additive networks with hybridization level $\alpha = 0.3$ (Δ), $\alpha = 0.4$ ($\square$) and $\alpha = 0.5$ ($\diamond$), and with $DIFFMAX = 0.25$. Figure (a) (respectively (c)) shows the true (respectively false) positive rate as a function of the tree size. Figure (b) (respectively (d)) shows the true (respectively false) positive rate as a function of the number of hybrids.	71

2.13	Average simulation results for non-additive networks with hybridization level $\alpha = 0.3$ (Δ), $\alpha = 0.4$ ($\square$) and $\alpha = 0.5$ ($\diamond$), and with $DIFFMAX = 0.35$. Figure (a) (respectively (c)) shows the true (respectively false) positive rate as a function of the tree size. Figure (b) (respectively (d)) shows the true (respectively false) positive rate as a function of the number of hybrids.	72
2.14	Split graph and galled network obtained for the rDNA cistron dataset in Tab. 2.5.	75
2.15	Hybridization network obtained with our new algorithm. The values of α are indicated on the reticulation branches (depicted by dashed lines).	76
3.1	(a) Configuration in which languages i and j are selected as neighbors by the NJ algorithm, and (b) configuration in which language h is identified as a recipient of lexical material from languages i and k by our algorithm for inferring explicit hybridization networks (here, the parameters α and $1 - \alpha$ represent the hybridization (i.e., reticulation) degree of donor languages i and k , respectively). . . .	91
3.2	This figure illustrates three possible network configurations (b-d), when our algorithm detects a hybrid, h , which is neighbor of one of its parents, $Nb(h)$, in the phylogenetic tree (a), e.g., in the IE language phylogeny inferred by Gray and Atkinson (see Fig. 3.1 in (Gray & Atkinson, 2003)). In configuration (b), language h receives the proportion, α , of its lexicon from its closest ancestor in the tree via direct inheritance and the remaining part of its lexicon, $(1 - \alpha)$, from a distant parent via word borrowing (e.g., see the case of Penn Dutch in Figs. 3.4 and 3.5(b)). In configuration (c), language h is a lexical hybrid of $Nb(h)$ and a distant parent (e.g., see the case of Sranan in Figs. 3.4 and 3.5(b)). In configuration (d), language h receives the proportion α (indicated, in this case, in parentheses) of its lexicon from both its closest ancestor via direct inheritance and from its neighbor $Nb(h)$ via word borrowing, and the remaining part, $(1 - \alpha)$, of its lexicon from a distant parent via word borrowing (e.g., see the case of Old Armenian in Fig. 3.4).	93
3.3	(a) Workflow chart of the new method for inferring explicit hybridization networks, and (b) an example of its application to a dataset consisting of 8 languages (including the hybrid language $L4$), 4 meanings and 16 cognate sets.	97

3.4	Explicit hybridization network given by our algorithm for the group of 84 IE languages originally considered by Dyen et al. (1992). Language groups are indicated on the left. The numbers at the arrows are the reticulation degrees corresponding to each of the donor languages and the numbers at the internal tree nodes are their age estimates.	99
3.5	(a) Split graph, (b) explicit hybridization network and (c) galled network, obtained for 8 languages of the West-Germanic group. .	100
3.6	(a) Split graph, (b) explicit hybridization network and (c) galled network, obtained for 7 languages of the North-Germanic group. .	101
3.7	(a) Split graph and (b) explicit hybridization network, obtained for 16 languages of the Latin group.	101
3.8	Split graph obtained for the entire set of 84 IE languages.	103
3.9	Partial galled network obtained for 12 IE languages. This is a maximum sub-network that includes reticulations of the complete galled network built for the entire set of 84 IE languages.	111
4.1	(a) A rooted phylogenetic tree; (b) A rooted phylogenetic (hybridization) network - here, Species 3 is a hybrid of Species 2 and 4; (c) An implicit phylogenetic network (split graph).	129
4.2	An intermediate NJ tree configuration used to compute the likelihood that Species 1 and 2 are neighbors. X and Y are the internal nodes of the presented intermediate NJ tree with n leaves and $n+1$ edges.	131
4.3	Configuration used to compute the likelihood that Species 3 is a hybrid of Species 1 and 2.	132
4.4	True positive and false positive rates with respect to the tree size (cases a and c) and the number of hybrids (cases b and d) obtained in simulations with 0 to 5 hybrids for trees with 8, 16 and 32 leaves and binary sequences of size 1000 using: (1) our previous distance-based method (Willems et al. (2014); $\diamond$), (2) our new method ($\square$), and (3) its refined version in which we corrected the likelihood with a Bayesian Information Criterion ($\triangle$). The averages over all parameter combinations except the fixed one (tree size or number of hybrids) are shown.	137

- 4.5 True positive (a) and false positive (b) rates shown with respect to the sequence size. These results were obtained from simulations with 0 to 5 hybrids and trees with 8 leaves. The sequences of sizes: 20, 50, 100, 200, 500 and 1000 were analyzed. The simulations were conducted using: (1) our new method ($\square$) and (2) its refined version in which we corrected the likelihood with a Bayesian Information Criterion ($\triangle$). The averages over all parameter combinations except the fixed one (sequence size) are shown. 138
- 4.6 Network obtained for the restriction map of the 16-species mosquitoes dataset (Kumar et al., 1998) using our new method. Five hybrids, linked to the rest of the network by dashed lines, were identified. The numbers on dashed network edges represent the respective contributions of parents to hybrids. 139

RÉSUMÉ

La théorie de l'évolution de Darwin, puis la découverte de l'ADN (acide désoxyribonucléique) dans les années 1950 ont donné naissance à la phylogénie moléculaire, dont le but principal est de construire des arbres d'espèces à partir de données moléculaires (essentiellement de l'ADN ou des protéines). De nombreux algorithmes ont été développés dans ce contexte. Les méthodes de distances sont les plus rapides, tandis que les méthodes basées sur les caractères donnent généralement de meilleurs résultats mais nécessitent des temps de calcul beaucoup plus importants.

Par ailleurs, plusieurs phénomènes évolutifs fondamentaux, comme l'hybridation, ne peuvent pas être représentés par un arbre phylogénétique. Il faut alors considérer des réseaux phylogénétiques. Plusieurs méthodes d'inférence de tels réseaux ont été introduites depuis une vingtaine d'années. La plupart d'entre elles produisent des réseaux implicites, qui peuvent s'avérer très difficiles à interpréter. Notre premier projet a ainsi consisté à développer un algorithme pour inférer un réseau d'hybridation explicite à partir d'une matrice de distances entre un certain nombre d'espèces. Des simulations et des tests sur des données réelles nous ont permis de mettre en lumière l'efficacité de notre nouvel algorithme. Notre programme est disponible en ligne pour l'ensemble de la communauté scientifique.

Dans un deuxième temps, nous avons appliqué notre nouvelle méthodologie dans le cadre de la biolinguistique. L'évolution des langues peut en effet être représentée sous la forme d'un arbre ou d'un réseau, et plusieurs études ont démontré la pertinence d'appliquer des méthodes d'inférence phylogénétique à des données linguistiques. Nous avons ainsi reconstruit un réseau d'hybridation explicite représentant l'histoire de 84 langues indo-européennes, et nous avons comparé ce réseau à ceux obtenus à partir d'autres méthodes.

Enfin, dans notre troisième projet, nous avons développé une méthode d'inférence de réseaux phylogénétiques basée sur les caractères. Notre algorithme prend en entrée une séquence binaire (correspondant, par exemple, à la présence ou l'absence de certains gènes) pour chaque espèce considérée, et reconstruit un réseau phylogénétique explicite, dont les feuilles sont en correspondance avec les espèces considérées. Plusieurs simulations ont montré que cet algorithme donne

de meilleurs résultats que notre méthode de distances, même si les temps de calculs sont généralement plus longs.

Mots clés : Réseaux phylogénétiques, hybridation, neighbor-joining, maximum de vraisemblance, biolinguistique.

CHAPITRE I

INTRODUCTION

1.1 Mise en contexte

Le terme phylogenèse (du grec *phulon*, signifiant « race, tribu ») a été défini par Haeckel en 1860 comme « l'histoire du développement paléontologique des organismes par analogie avec l'ontogénie ou histoire du développement individuel ». Un arbre phylogénétique (ou phylogénie ou *X*-arbre (Barthélemy & Guénoche, 1991)) est une représentation graphique de la phylogenèse d'un groupe d'espèces (ou de taxons). Une telle représentation est basée sur le concept de « descendance des espèces avec modification de leurs caractères », c.-à-d., sur le principe de la transmission des caractères d'une génération impliquant d'éventuelles modifications comme les mutations de certains nucléotides qui composent la molécule d'ADN. Les arbres phylogénétiques ne peuvent cependant pas être utilisés pour représenter des phénomènes comme l'hybridation ou le transfert horizontal de gènes. Il faut alors ajouter des réticulations dans l'arbre, et on obtient ainsi des réseaux phylogénétiques, qui sont des structures beaucoup plus complexes.

Depuis une cinquantaine d'années, de nombreux algorithmes ont été développés pour inférer des arbres phylogénétiques à partir de données moléculaires (généralement des séquences d'ADN ou de protéines). Les méthodes de distances transforment l'ensemble des séquences initiales en une unique matrice de distances

entre les espèces étudiées. Ces méthodes sont très rapides mais ne prennent pas en compte toute l'information disponible. Les méthodes basées sur les caractères (maximum de vraisemblance, maximum de parcimonie et méthodes bayésiennes) utilisent directement les séquences moléculaires. Elles donnent souvent de meilleurs résultats mais demandent beaucoup plus de temps de calcul, malgré des implémentations de plus en plus performantes.

Les méthodes d'inférence de réseaux phylogénétiques sont quant à elles beaucoup plus récentes. Elles prennent généralement en entrée un ensemble d'arbres phylogénétiques, ou plus généralement un ensemble de regroupements d'espèces, qui présentent certaines contradictions. Les réseaux phylogénétiques ainsi obtenus peuvent être abstraits ou explicites. Un réseau abstrait met en relief les incompatibilités dans les données initiales, sans donner d'explications précises sur les phénomènes qui justifient ces incompatibilités. Un réseau explicite est beaucoup plus utile pour les biologistes, puisqu'il précise les réticulations (hybridation ou transfert de gènes par exemple), qui expliquent les contradictions identifiées.

Dans notre premier projet présenté dans le chapitre 2, nous avons développé un algorithme qui prend en entrée une matrice de distances entre n espèces, et qui construit un réseau d'hybridation explicite, dont les nœuds terminaux correspondent aux n espèces initiales. Le réseau ainsi construit explicite les hybrides et leurs parents. Il s'agit de la première implémentation d'une méthode de distances qui infère des réseaux explicites. Chan et al. (2006), Bordewich & Tokac (2016) et Bordewich et al. (2018) ont déterminé certains principes de reconstruction de réseaux phylogénétiques explicites à partir de distances. Cependant, dans la plupart des cas, il faut disposer de plusieurs matrices de distances. De plus, les réseaux ainsi obtenus vérifient certaines restrictions topologiques assez importantes. Par ailleurs, notre algorithme détermine le degré d'hybridation correspondant à chacun des deux parents de l'hybride (pourcentage de données reçues par chaque

noeud hybride de la part de ses deux parents). Or, il n'existe que peu de modèles (Willson, 2013; Francis & Steel, 2015) qui prennent en compte ces pourcentages.

Depuis une vingtaine d'années, l'inférence phylogénétique est utilisée en linguistique pour construire des arbres, qui représentent l'histoire de l'évolution d'un certain ensemble de langues. Dans notre deuxième projet présenté dans le chapitre 3, nous avons démontré que le concept de réseau phylogénétique pouvait être utilisé de manière pertinente dans ce contexte. De plus, nous avons utilisé notre méthode de distances pour inférer un réseau d'hybridation de 84 langues indo-européennes. Ce réseau a ensuite été justifié par des arguments linguistiques, géographiques et historiques.

Dans notre troisième projet présenté dans le chapitre 4, nous avons adapté notre algorithme initial pour le transformer en une méthode basée sur les caractères, dans le cadre de données binaires (p. ex., présence/absence de certains gènes ou de certains sites de restriction). Notre algorithme, basé sur le principe du maximum de vraisemblance, prend ainsi en entrée une séquence binaire pour chaque espèce étudiée, et construit un réseau d'hybridation explicite. Même si cette méthode est plus lente que notre méthode de distances, les résultats obtenus sur des données simulées sont sensiblement meilleurs, surtout dans le cas d'un petit nombre d'espèces. De plus, le principe de reconstruction du réseau nous permet d'inférer les séquences ancestrales (c.-à-d., les séquences des ancêtres hypothétiques des espèces actuelles), ce qui est un enjeu majeur en bioinformatique.

Notons que nos algorithmes sont basés sur des fondements mathématiques rigoureux, mais aussi sur de nombreuses expérimentations. Nous avons ainsi parfois retenu certains critères d'optimisation, qu'on ne pouvait pas justifier mathématiquement, mais qui donnaient de meilleurs résultats.

Le plan de cette introduction est le suivant. Dans la section 1.2, nous présen-

tons en détails les principes de l'inférence d'arbres phylogénétiques. La section 1.3 décrit l'algorithme neighbor-joining (Saitou & Nei, 1987), qui est la méthode de distances sur laquelle est basée l'essentiel de nos travaux. Dans la section 1.4, nous expliquons le principe du maximum de vraisemblance utilisé dans le chapitre 4. Les sections 1.5 et 1.6 présentent l'évolution réticulée, les réseaux phylogénétiques et les principales méthodes d'inférence de tels réseaux. Dans la section 1.7, nous expliquons comment utiliser les principes de l'inférence phylogénétique en linguistique. La section 1.8 décrit les différentes parties de cette thèse. Un glossaire des principaux termes techniques utilisés se trouve à la section 1.9.

1.2 Inférence phylogénétique

Nous présentons tout d'abord les principales caractéristiques des arbres phylogénétiques. Ces arbres sont composés de feuilles ou nœuds externes qui sont en correspondance avec les espèces les plus récentes (généralement contemporaines) pour lesquelles on dispose de données biologiques, de nœuds internes qui représentent des ancêtres virtuels, et de branches (ou arêtes) qui définissent les relations entre les taxons. Si l'arbre est enraciné, ces branches permettent de définir une relation *ancêtre - descendant* entre deux nœuds successifs. Dans ce cas, la racine représente l'ancêtre commun de toutes les espèces étudiées. Un arbre non enraciné ne prend pas en compte le sens de l'évolution, et on ne peut plus définir des relations de descendance au niveau des nœuds internes. Il permet de classer différents groupes d'espèces sans considérer de notion temporelle. La figure 1.1 présente un exemple d'arbre enraciné.

Le degré d'un nœud est le nombre d'arêtes adjacentes à ce nœud. Si ce degré est strictement supérieur à trois, ce nœud est dit non résolu, ce qui peut signifier soit la divergence simultanée de plusieurs espèces, soit l'incapacité de déterminer l'ordre de divergence de ces espèces. Si l'arbre est enraciné, on peut distinguer les

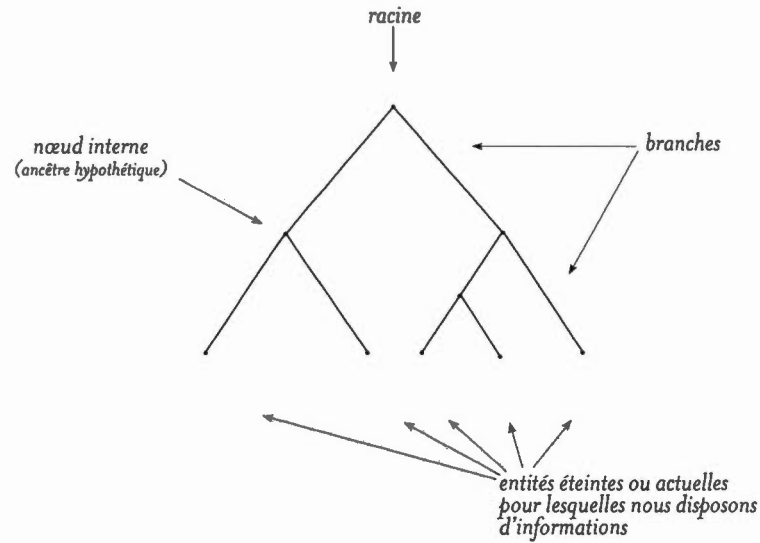


Figure 1.1 Modèle de base d'un arbre phylogénétique.

arêtes rentrantes et les arêtes sortantes.

Nous commençons avec quelques définitions de base concernant les arbres phylogénétiques et les métriques d'arbre, en suivant la terminologie de Barthélemy & Guénoche (1991). La distance $d(x, y)$ entre deux sommets x et y dans un arbre valué T est définie comme la somme de toutes les longueurs des arêtes de l'unique chemin reliant x et y dans T . Une feuille est un sommet de degré 1. La figure 1.2 donne un exemple du calcul d'une telle distance.

Définition 1.1. Soit X un ensemble fini de n taxons. Une dissimilarité d sur X est une fonction positive ou nulle sur $X \times X$ telle que pour tout x, y appartenant à X :

(1) $d(x, y) = d(y, x)$, et

(2) $d(x, y) = d(y, x) \geq d(x, x) = 0$.

Définition 1.2. Une dissimilarité d sur X satisfait la condition des quatre points

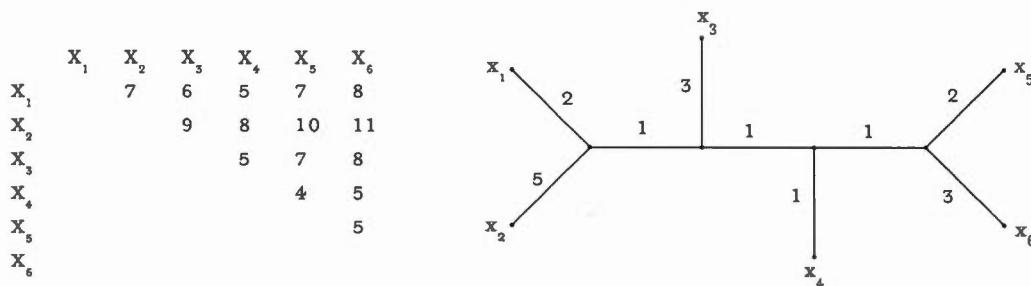


Figure 1.2 Exemple d'une distance d'arbre sur un ensemble X de 6 taxons et l'arbre phylogénétique associé.

si pour tout x, y, z , et w de X :

$$d(x, y) + d(z, w) \leq \text{Max}\{d(x, z) + d(y, w); d(x, w) + d(y, z)\}.$$

Définition 1.3. Pour un ensemble fini X non vide, un X -arbre est un couple (T, ϕ) consistant en un arbre T , avec un ensemble de sommets V et une relation $\phi : X \rightarrow V$, ayant la propriété que, pour tout $v \in V$ de degré au plus 2, $v \in \phi(X)$. Un X -arbre est un arbre phylogénétique si ϕ est une bijection de X dans l'ensemble des feuilles de T . Il est dit binaire si chaque sommet interne a un degré égal à 3.

Le théorème principal reliant la condition des quatre points et la représentabilité d'une dissimilarité par un arbre phylogénétique (c.-à-d., une phylogénie) est le suivant (Buneman, 1971, 1974; Dobson, 1974; Patrinos & Hakimi, 1972; Zaretskii, 1965) :

Théorème 1.1. Toute dissimilarité satisfaisant la condition des quatre points peut être représentée par un arbre phylogénétique tel que pour tout x, y appartenant à X , $d(x, y)$ est égale à la longueur du chemin reliant les feuilles x et y dans T .

Cette dissimilarité est appelée une distance d'arbre, une distance additive ou une métrique d'arbre. Cet arbre est unique.

La reconstruction d'un arbre phylogénétique commence par l'analyse des données biologiques associées aux espèces étudiées. Il peut s'agir de séquences nucléotidiques (ADN), d'acides aminés (protéines) ou de données binaires. Dans cette section, nous nous restreindrons au cas des séquences d'ADN. Une séquence nucléotidique représente l'ADN et est un assemblage linéaire de quatre types de base : les cytosines (C) et thymine (T) (famille des pyrimidines), et les adénine (A) et guanine (G) (famille des purines). Il y a ainsi 4^N séquences d'ADN différentes de longueur N . Certaines séquences d'ADN représentent des gènes qui seront exprimés en protéines. Ces séquences doivent préalablement être alignées pour pouvoir être comparées.

Trois approches principales ont été développées pour construire des arbres phylogénétiques : la phénétique, la cladistique et la probabiliste.

L'approche phénétique ne tient pas compte du processus de l'évolution. Le meilleur arbre possible est reconstruit à partir des distances entre les espèces en utilisant une stratégie de regroupement hiérarchique. Ces distances sont calculées en fonction du nombre de nucléotides différents dans un ensemble de gènes dont les séquences d'ADN ont été alignées. La somme des longueurs des branches de l'unique chemin entre deux feuilles est censée être la plus proche possible de la distance réelle entre les deux espèces représentées par ces feuilles. Ce n'est pas le cas si le taux d'évolution n'est pas constant dans tout l'arbre ou si l'hypothèse de l'horloge moléculaire n'est pas vérifiée (voir le glossaire pour la définition de cette hypothèse). On peut alors corriger les distances par différentes transformations (Jukes & Cantor, 1969; Kimura, 1980). Les deux principales méthodes de distances sont neighbor-joining (Saitou & Nei, 1987) (que nous expliquons en détails dans la

section suivante) et UPGMA (Sneath & Sokal, 1973) qui est l'acronyme de « Un-weighted Pair Group Method with Arithmetic mean ». Comme on le verra dans le cas de neighbor-joining (NJ), la complexité de ces algorithmes est polynomiale en fonction du nombre d'espèces étudiées (c.-à-d., le nombre de feuilles de l'arbre reconstruit). Ce sont les méthodes les plus rapides.

L'approche cladistique cherche, quant à elle, à établir des relations de parenté en s'intéressant directement aux nucléotides partagés par les taxons. On prend alors en entrée une séquence d'ADN pour chaque espèce étudiée, on considère tous les scénarios d'évolution en inférant les caractères des ancêtres potentiels à chaque noeud interne, et on détermine l'arbre qui correspond au meilleur scénario d'évolution selon un critère préalablement choisi. Ces méthodes sont principalement basées sur le critère du maximum de parcimonie (Fitch, 1971). Le scénario le plus parcimonieux est celui qui nécessite le moins de modifications des séquences au cours de l'évolution. La figure 1.3 montre le scénario d'évolution le plus parcimonieux pour les séquences CAAG, CCAG, GCAT et GCTT. Pour une topologie d'arbre donnée, l'algorithme de Fitch (Fitch, 1971) permet de retrouver un des scénarios les plus parcimonieux en $O(n * L * k)$, où n est le nombre de feuilles de l'arbre, L la longueur des séquences d'ADN et $k = 4$ le nombre d'états possibles. Cet algorithme est basé sur les principes de la programmation dynamique. Pour chaque position des séquences considérées, on part des états des feuilles de l'arbre pour remonter progressivement : pour chaque noeud, on construit l'ensemble de ses états possibles en fonction des ensembles des états de ses descendants. Le problème de retrouver la topologie qui porte le scénario le plus parcimonieux est cependant NP-difficile, ce qui restreint le nombre d'espèces qu'on peut considérer. Notons que les longueurs de branches ne sont pas prises en compte dans ce contexte.

L'approche probabiliste (ou maximum de vraisemblance), quant à elle, tient com-

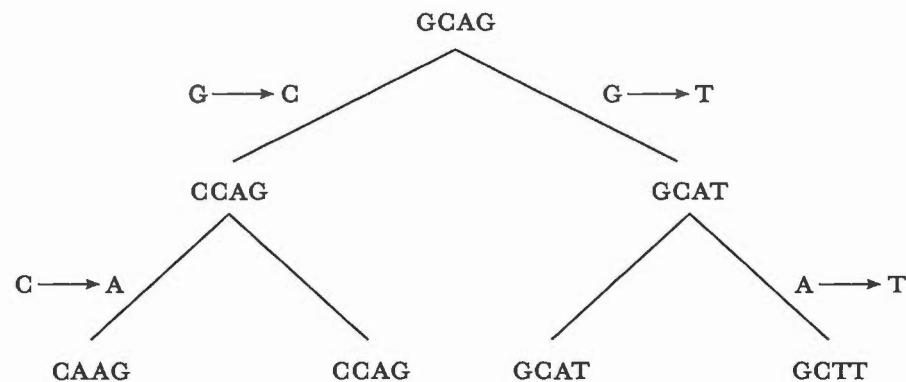


Figure 1.3 Scénario d'évolution le plus parcimonieux.

pte des longueurs de branches. Elle évalue en termes de probabilités l'ordre des branchements et la longueur des arêtes d'un arbre. On doit ainsi disposer d'un modèle d'évolution, c.-à-d., on doit définir la probabilité d'une mutation d'un nucléotide en un autre le long d'une branche d'un arbre en fonction de la longueur de cette branche et des deux nucléotides. Pour un arbre phylogénétique donné, on peut alors calculer la vraisemblance de cet arbre, c.-à-d., la somme des probabilités de tous les scénarios d'évolution le long de cet arbre (voir la section 1.4). Une des difficultés majeures est l'optimisation des longueurs de branches pour une topologie d'arbre fixée. Une fois ces longueurs optimisées, la vraisemblance d'un arbre est calculée en utilisant un principe de programmation dynamique (Felsenstein, 1981). Le problème de retrouver l'arbre le plus vraisemblable est NP-difficile.

La plupart des méthodes précédemment citées sont implémentées dans les logiciels PAUP (Swafford, 2002), Phylip (Felsenstein, 2005) et T-Rex (Boc et al., 2012).

Notons que deux méthodes de maximum de vraisemblance sont particulièrement efficaces : la méthode PhyML (Guindon & Gascuel, 2003), implémentée dans le

logiciel PhyML 3 (Guindon et al., 2010), et la méthode RAxML (Stamatakis et al., 2008) implémentée dans le logiciel RAxML-Light (Stamatakis et al., 2012).

Deux interfaces récentes, BEAGLE (Ayres et al., 2012) et PLL (Flouri et al., 2015), permettent de faire des calculs à grande échelle dans des temps de plus en plus raisonnables.

Des approches bayésiennes (Rannala & Yang, 1996) ont également permis d'utiliser le maximum de vraisemblance pour des données plus importantes. Dans ce contexte, l'hypothèse optimale est celle qui maximise la probabilité *a posteriori*. Cette probabilité *a posteriori* est proportionnelle à la vraisemblance multipliée par la probabilité *a priori* de l'hypothèse. On peut ainsi développer des algorithmes plus rapides qui peuvent incorporer des modèles d'évolution plus complexes. Dans ce cadre, les logiciels MrBayes (Ronquist et al., 2012), BEAST (Drummond et al., 2012) et RevBayes (Höhna et al., 2016) utilisent les MCMC (chaînes de Markov Monte-Carlo) pour parcourir l'espace de tous les arbres possibles en vue d'obtenir l'arbre le plus vraisemblable.

Certaines approches plus récentes proposent de mélanger des méthodes phylogénétiques classiques avec des algorithmes de regroupement. Ruan et al. (2014) construisent par exemple des phylogrammes sphériques à partir d'un ensemble de séquences de nucléotides.

1.3 L'algorithme neighbor-joining

L'algorithme neighbor-Joining (Saitou & Nei, 1987) (NJ) est la méthode de distances la plus utilisée pour inférer des arbres phylogénétiques. Atteson (1999) a démontré que cet algorithme trouve la bonne phylogénie si les distances utilisées sont suffisamment proches des vraies distances d'évolution.

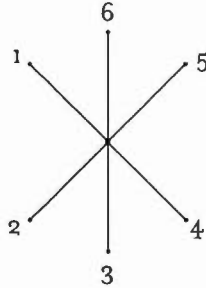


Figure 1.4 Buisson de taille 6.

On prend en entrée une matrice de distances quelconque $D = \{D[i][j]\}_{1 \leq i \leq n, 1 \leq j \leq n}$ sur un ensemble de n espèces. On a $D[i][i] = 0$ pour tout $1 \leq i \leq n$, et $D[i][j] = D[j][i]$ pour tout $1 \leq i \leq n$ et $1 \leq j \leq n$.

NJ est un algorithme de regroupement qui commence avec un buisson de n feuilles et n branches, où n est le nombre d'espèces étudiées (voir la figure 1.4 dans le cas $n = 6$). Cet arbre est graduellement transformé en un arbre phylogénétique non enraciné avec les mêmes n feuilles et avec $2n - 3$ branches. La i -ème étape consiste à choisir deux voisins parmi $n - i + 1$ candidats. On considère ainsi toutes les $\frac{(n-i+1)(n-i)}{2}$ configurations similaires à celle représentée dans la figure 1.5. Pour chacune de ces configurations, on calcule les longueurs de branches qui minimisent un critère des moindres carrés, où on compare les dissimilarités données en entrée avec une distance additive.

Saitou & Nei (1987) ont montré que la somme des longueurs de branches de la topologie de la figure 1.5 est égale à :

$$S_{i,j} = \frac{1}{2}D[i][j] + \frac{\sum_{1 \leq k \leq n; k \neq i,j} [D[i][k] + D[j][k]]}{2(n-2)} + \frac{\sum_{1 \leq k < l \leq n; k,l \neq i,j} D[k][l]}{n-2}. \quad (1.1)$$

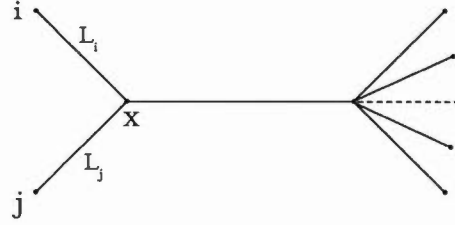


Figure 1.5 Configuration où les nœuds i et j sont choisis comme voisins.

On joint les nœuds i et j qui minimisent l'évolution totale, c.-à-d., la somme des longueurs de toutes les branches $S_{i,j}$. On remplace les nœuds choisis i et j par le nœud X (leur ancêtre commun direct) et on obtient une matrice de distances de taille $n - 1$. On calcule les nouvelles distances de X aux feuilles restantes de l'arbre en utilisant la formule suivante :

$$d(X, k) = \frac{1}{2}(D[i][k] + D[j][k]), \quad k \neq i, j. \quad (1.2)$$

Après $n - 3$ étapes, on obtient un arbre phylogénétique non enraciné dont les longueurs de branches sont calculées à chaque étape en utilisant les équations suivantes :

$$L_i = \frac{1}{2}D[i][j] + \frac{1}{2(n-2)}(P - Q), \quad L_j = \frac{1}{2}D[i][j] - \frac{1}{2(n-2)}(P - Q), \quad (1.3)$$

où

$$P = \sum_{1 \leq k \leq n, k \neq i, j} D[i][k], \text{ et } Q = \sum_{1 \leq k \leq n, k \neq i, j} D[j][k].$$

Cet algorithme a une complexité algorithmique de $O(n^3)$, où n est le nombre d'espèces étudiées.

1.4 Le principe du maximum de vraisemblance

Nous présentons ici en détails le principe du maximum de vraisemblance dans le cas de données binaires, puisque c'est dans ce contexte que nous l'utilisons dans le

chapitre 4. On considère n séquences binaires de taille L , et un modèle d'évolution $\text{Pr}(t, N_1, N_2)$, où $\text{Pr}(t, N_1, N_2)$ est la probabilité que le caractère N_1 (0 or 1) évolue vers le caractère N_2 (0 or 1) durant le temps t . Dans le cas des données binaires, on utilise le modèle F81 (Felsenstein, 1981) : soit π_0 (respectivement π_1) la proportion de 0's (respectivement, 1's) dans les données d'entrée, on pose $\beta = \frac{1}{1-\pi_0^2-\pi_1^2}$. Les probabilités de transition sont données par les formules suivantes :

$$\text{Pr}(t, i, j) = \begin{cases} e^{-\beta t} + \pi_j (1 - e^{-\beta t}) & , \text{ si } i = j, \\ \pi_j (1 - e^{-\beta t}) & , \text{ si } i \neq j. \end{cases} \quad (1.4)$$

La vraisemblance d'un arbre se calcule avec la formule suivante :

$$\mathcal{L}(T) = \prod_{l=1}^L \mathcal{L}_l(T), \quad (1.5)$$

où $\mathcal{L}_l(T)$ est la somme des probabilités de tous les scénarios d'évolution possibles à la position l . Un scénario d'évolution consiste à choisir un caractère pour chaque nœud interne. Il y a donc 2^{n-2} scénarios pour chaque position d'un arbre binaire de taille n . La vraisemblance d'un arbre se calcule malgré tout en temps polynomial en utilisant une méthode de programmation dynamique (Felsenstein, 1981). La détermination de l'arbre le plus vraisemblable reste cependant un problème NP-difficile à cause du nombre de topologies à considérer. De nombreuses heuristiques ont ainsi été développées dans ce contexte. De plus, pour chaque topologie, il est nécessaire d'optimiser les longueurs de branches, comme nous l'expliquons dans le cas particulier des configurations que nous utiliserons dans le chapitre 4 (voir la figure 1.6).

On veut calculer la vraisemblance des configurations semblables à la configuration de la figure 1.6. Les caractères évoluant indépendamment les uns des autres, nous avons quatre scénarios d'évolution possibles pour chaque position, et la vraisem-

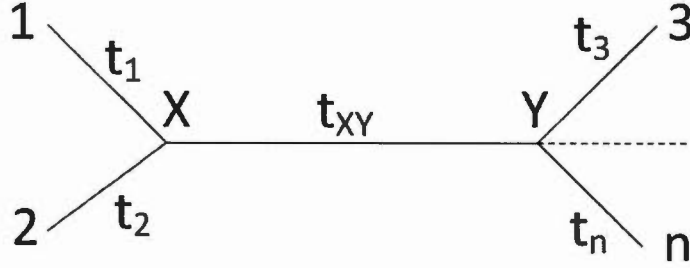


Figure 1.6 Configuration de NJ utilisée avec un critère de maximum de vraisemblance.

blance de la configuration de la figure 1.6 est égale à :

$$f(t_1, t_2, \dots, t_n, t_{XY}) = \prod_{1 \leq l \leq L} \left(\sum_{(\epsilon_X, \epsilon_Y) \in \{0;1\}^2} \left(\Pr(t_1, N_l^1, \epsilon_X) \Pr(t_2, N_l^2, \epsilon_X) \right. \right. \\ \left. \left. \Pr(t_{XY}, \epsilon_X, \epsilon_Y) \prod_{3 \leq k \leq n} \Pr(t_k, \epsilon_Y, N_l^k) \right) \right), \quad (1.6)$$

où N_l^k est le caractère en position l de la séquence k . Nous devons ainsi déterminer les longueurs de branches $t_1, t_2, \dots, t_n, t_{XY}$ qui maximisent cette fonction pour chacune des $\frac{n(n-1)}{2}$ configurations de NJ. Pour cela nous appliquons la méthode de Newton-Raphson explicitée par Kishino et al. (1990) dans le cadre du principe du maximum de vraisemblance, et utilisée dans la plupart des implémentations de ce principe en phylogénie. Il s'agit d'une méthode itérative qui permet de déterminer un maximum local d'une fonction à plusieurs variables. Rappelons le principe de cette méthode dans le cas d'une fonction $g(x)$ à une variable. On cherche à déterminer une valeur de x telle que $g'(x) = 0$. On doit choisir une valeur initiale x_0 , et on définit alors récursivement une suite x_i par la formule $x_{i+1} = x_i - \frac{g'(x_i)}{g''(x_i)}$. Notons que la convergence de cette suite n'est pas garantie et que sa limite, si elle existe, ne correspond pas forcément à un extremum global de la fonction $g(x)$. Le choix de la valeur initiale est crucial pour obtenir une

convergence rapide vers le maximum recherché. On trouvera plus de détails sur cette méthode, notamment dans le cas de plusieurs variables, dans Press et al. (2007). La procédure itérative suivante est alors utilisée (Kishino et al., 1990) :

- On détermine les valeurs initiales $t_1^0, t_2^0, \dots, t_n^0, t_{XY}^0$ en utilisant les longueurs de branches obtenues par l'algorithme NJ à partir de la matrice de distances calculée avec les séquences binaires initiales et le modèle d'évolution F81.
- On met à jour t_1 et t_2 par la méthode de Newton-Raphson.
- On met à jour $t_3, t_4, \dots, t_n$ par la même méthode.
- On met à jour t_{XY} par la même méthode.
- On itère les trois étapes précédentes jusqu'à ce que les variations de toutes les longueurs de branches soient inférieures à 1%.

1.5 L'évolution réticulée

L'évolution réticulée correspond aux processus d'évolution des espèces qui ne peuvent pas être représentés parfaitement par le modèle arborescent de bifurcation utilisé classiquement en analyse phylogénétique (Doolittle, 1999; Legendre, 2000). La figure 1.7 montre un réseau réticulé (plus exactement un réticulogramme dans ce cas). Le trait situé entre les arêtes 1 et 2 représente une arête de réticulation ajoutée à l'arbre original.

La figure 1.8 présentée par Doolittle (1999) montre que l'évolution des espèces se produit selon un modèle en réseau plutôt que selon un modèle en arbre. Pour Doolittle, l'impossibilité de trouver le vrai arbre de la vie n'est pas due aux méthodologies ou aux données utilisées, mais au fait que l'histoire de la vie ne peut être représentée correctement par un arbre. Nous allons détailler les différents processus biologiques conduisant à des réticulations en insistant tout particulièrement sur l'hybridation, qui est le phénomène pris en compte dans notre thèse.

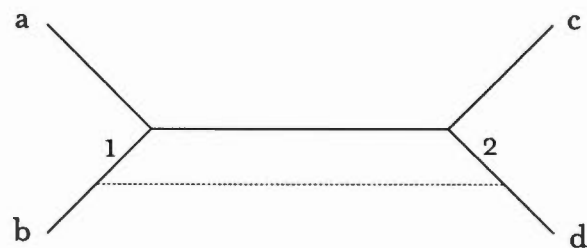


Figure 1.7 Un réseau réticulé (c.-à-d., un réticulogramme) montrant une relation additionnelle entre les ancêtres des espèces b et d.

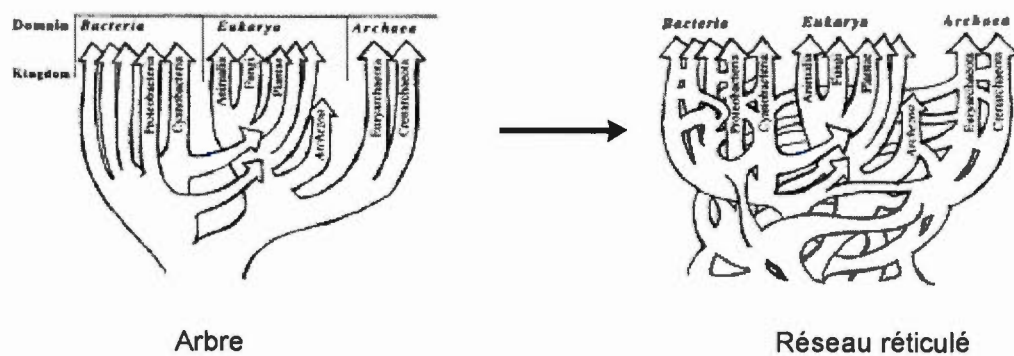


Figure 1.8 Le réseau réticulé représenterait mieux l'histoire de la vie qu'un arbre phylogénétique classique (Doolittle, 1999).

Le transfert horizontal (ou latéral) de gènes (Koonin, 2003; Doolittle et al., 2003) est un transfert direct de matériel génétique d'une lignée à une autre. Ces transferts sont fréquents chez les bactéries. Les Bacteria et les Archaea ont ainsi développé la capacité de s'adapter à de nouveaux environnements en acquérant des nouveaux gènes par transfert plutôt qu'en modifiant leur patrimoine génétique par une série de mutations (Gogarten et al., 2002; Zhaxybayeva et al., 2004; Boc et al., 2010b), ce qui leur permet notamment d'augmenter leur caractère pathogène (Gyles & Boerlin, 2014). Plus récemment, de nombreux transferts horizontaux ont été identifiés entre des bactéries et des animaux (Hotopp, 2011; Robinson et al., 2013).

L'homoplasie est le développement au sein de différentes espèces, qui n'ont pas d'ancêtres communs, d'organes qui se ressemblent et qui ont les mêmes fonctions (McGhee, 2011). Il s'agit d'évolutions convergentes qui compliquent les inférences phylogénétiques (Smouse, 2000).

Les recombinaisons génétiques, quant à elles, sont des modifications du matériel génétique à l'intérieur d'une même lignée, comme le réassortiment des gènes parentaux lors de la formation des gamètes (crossing-over), ou l'échange de matériel génétique entre chromosomes homologues.

L'hybridation est un des exemples les plus importants d'évolution réticulée (Arnold, 1997). Dans la figure 1.9, deux lignées (Racine-Espèce 2 et Racine-Espèce 3) se recombinent pour créer une nouvelle espèce (Espèce 4). La nouvelle espèce possède soit le même nombre de chromosomes que les espèces parents (hybridation *diploïde*), soit la somme du nombre de chromosomes de ses parents (hybridation *polyploïde*). Il existe trois principaux mécanismes d'hybridation (Makarenskov et al., 2006b) :

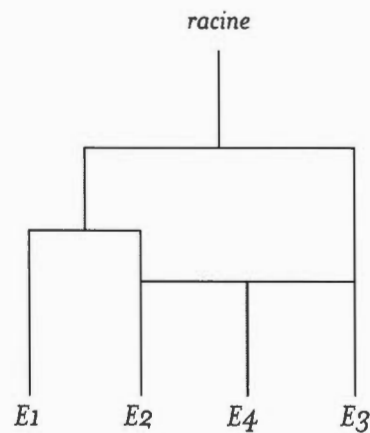


Figure 1.9 Un exemple de réseau d'hybridation.

1. L'autopolyploïdisation est un évènement de spéciation impliquant le doublement des chromosomes au sein d'une même espèce. Elle peut être représentée par une bifurcation dans l'arbre phylogénétique. On parle alors d'hybridation intraspécifique.
2. L'allopolyplôidisation est une hybridation où la nouvelle espèce acquiert l'ensemble des compléments des chromosomes *diploïdes* des deux parents. Les parents n'ont pas forcément le même nombre de chromosomes. La spéciation est instantanée car tout croisement de la nouvelle espèce avec ses parents risque de produire une espèce *triploïde* stérile.
3. La spéciation par l'hybridation *diploïde* est un évènement sexuel normal entre deux parents d'espèces distinctes mais assez proches et ayant en général le même nombre de chromosomes. La spéciation n'est pas immédiate et nécessite que la nouvelle espèce soit séparée de ses parents.

Nous nous concentrons dans nos travaux sur les phénomènes de spéciation par

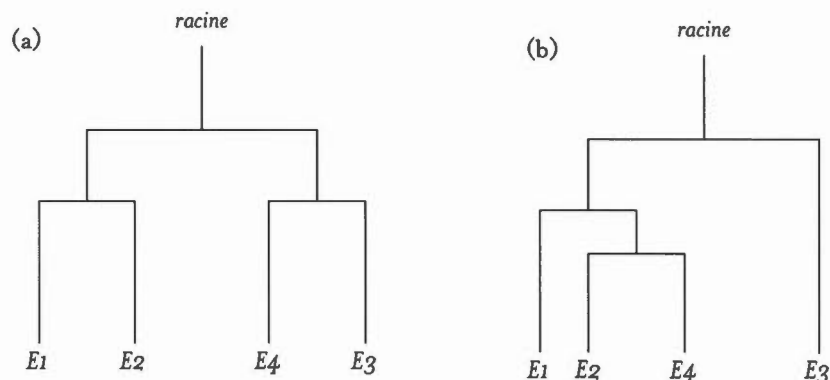


Figure 1.10 Deux arbres possibles pour le phénomène d'hybridation présenté sur la figure 1.9.

hybridation *diploïde*. Au cours de ce processus, l'hybride hérite d'un des deux chromosomes homologues de chaque paire de chromosomes de chacun de ses deux parents. L'évolution des gènes hérités de chacun des deux parents peut ainsi être représentée dans deux arbres différents. L'analyse phylogénétique classique des quatre espèces associées aux nœuds terminaux du réseau d'hybridation de la figure 1.9 donnera un des deux arbres de la figure 1.10, selon les gènes utilisés pour l'inférence de l'arbre.

L'hybridation est très fréquente chez les plantes parmi lesquelles il y aurait plus de 70000 hybrides naturels (Stace, 1991), ainsi que des hybrides créés par l'homme pour introduire certaines caractéristiques chez des espèces cultivées (Judd, 2008). De nombreuses espèces de poissons, d'amphibiens et de reptiles sont également le fruit de phénomènes d'hybridation, suite à la gynogenèse, un mode de reproduction permettant à des hybrides femelles unisexués de se reproduire en utilisant le sperme d'une espèce bisexuée proche pour stimuler le développement des oeufs (Dawley, 1989). Les hybridations sont très rares chez la majorité des autres

groupes, notamment les oiseaux, les mammifères, et la plupart des arthropodes.

1.6 Réseaux phylogénétiques : définitions et méthodes d'inférence

Toutes les définitions de cette section sont tirées de Huson et al. (2010). La définition la plus générale d'un réseau phylogénétique est la suivante.

Définition 1.4. *On appelle réseau phylogénétique tout graphe utilisé pour représenter des relations d'évolution entre un ensemble de taxons qui sont associées à certains des nœuds du graphe (généralement les feuilles).*

Le réseau est dit explicite s'il s'agit d'un arbre auquel on rajoute des réticulations qui représentent explicitement certains des phénomènes évolutifs définis dans la section précédente, en identifiant les espèces hybrides et leurs parents, par exemple.

Il est dit abstrait (ou implicite) s'il ne permet que de visualiser certaines incompatibilités dans les données sans expliciter des phénomènes biologiques particuliers.

Les réseaux abstraits les plus utilisés sont les réseaux de bipartition. Soit X un ensemble de taxons. Une bipartition (ou « split » en anglais) consiste en deux sous-ensembles non vides de X : A et B , tels que $A \cap B = \emptyset$ et $A \cup B = X$. Soit $\mathcal{S}$ un ensemble de bipartitions sur X , un réseau de bipartitions (Bandelt & Dress, 1992a,b) qui représente $\mathcal{S}$ est un réseau phylogénétique tel que chaque bipartition de $\mathcal{S}$ est représentée par un ensemble d'arêtes parallèles (voir la figure 1.11 pour un exemple). Ces arêtes ne représentent pas des phénomènes biologiques mais peuvent être vues comme des hypothèses d'évolution alternatives ou des similarités entre espèces. Plus formellement, on définit la notion de split-graphe et on associe un réseau de bipartitions à tout split-graphe (voir Huson et al. (2010) pour plus de détails). On confondra par la suite ces deux notions. Bryant & Moulton (2004), ainsi que Bandelt & Dress (1992b), ont développé des méthodes d'inférence de split-graphes à partir de matrices de distances. Neighbor-Net (Bryant &

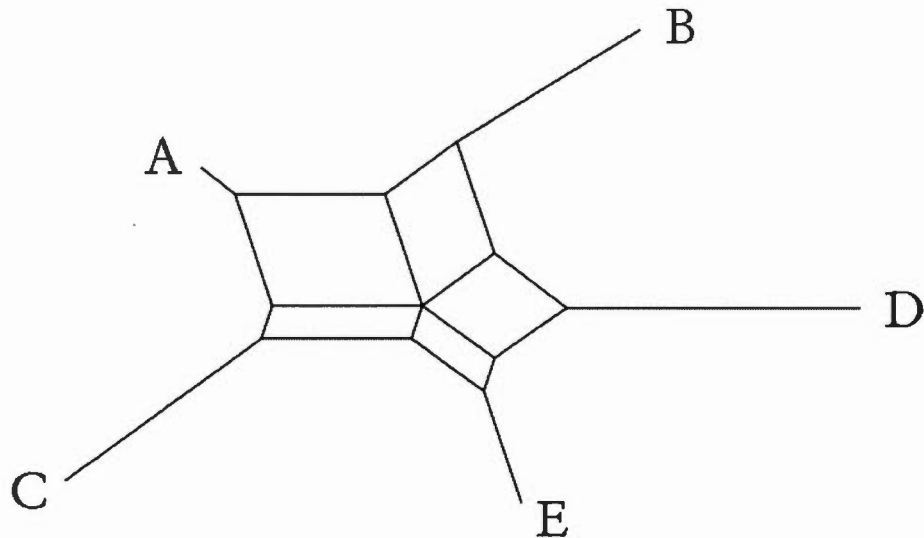


Figure 1.11 Un exemple de split-graphe.

Moulton, 2004; Huson & Bryant, 2006) est ainsi la méthode la plus utilisée pour la reconstruction de split-graphes planaires. Holland & Moulton (2003) ont introduit la notion de super réseau de bipartitions obtenu à partir d'un ensemble d'arbres contradictoires. Des réseaux médians sont inférés par Bandelt et al. (1995) à partir d'un ensemble de séquences. Ce sont des split-graphes dont les arêtes parallèles sont étiquetées par les séquences qui justifient leur présence dans le réseau. Tous ces réseaux sont souvent difficiles à interpréter. Cependant, Gambette & Huson (2008) en ont amélioré la visualisation, et le logiciel SplitsTree (Huson & Bryant, 2006) est l'outil le plus utilisé actuellement.

Legendre et Makarenkov (Legendre & Makarenkov, 2002; Makarenkov & Legendre, 2004) ont proposé d'utiliser les réticulogrammes pour détecter les réticulations dans des données évolutives. Ils ont développé une méthode basée sur les distances qui infère des phylogénies réticulées. Cette méthode utilise la topologie

d'un arbre phylogénétique comme une structure de base sur laquelle on ajoute, au fur et à mesure et suivant un critère d'optimisation, des arêtes de réticulation pour construire un réticulogramme (voir la figure 1.7 pour un exemple de réticulogramme à une arête de réticulation). Un enjeu majeur et délicat est alors de déterminer à quel moment l'algorithme doit cesser d'ajouter des branches.

Un réseau d'hybridation est un arbre phylogénétique auquel on rajoute des réticulations qui correspondent à des événements d'hybridation comme dans la figure 1.9. La première méthode générale d'inférence de tels réseaux a été développée par Albrecht et al. (2012) qui ont proposé un algorithme parallèle pour trouver un réseau d'hybridation minimum à partir de deux arbres contradictoires. Cet algorithme est cependant très lent même s'il est exécuté sur un ordinateur à plusieurs processeurs (Chen et al., 2012). Wu (2010), ainsi que Chen & Wang (2012), ont présenté des algorithmes qui permettent de prendre en entrée plus que deux arbres contradictoires. Rieseberg & Morefield (1995) ont développé un programme, RETICLAD, qui permet d'identifier les hybrides en se basant sur le principe qu'ils combinent les caractères de leurs parents. Cependant, ce programme ne permet de trouver des réticulations qu'entre les arêtes terminales d'un arbre. Rieseberg & Ellstrand (1993) ont explicité des exemples pour lesquels le programme semble bien fonctionner. Plus récemment, Mirzaei & Wu (2016) ont développé une méthode rapide pour inférer un réseau d'hybridation parcimonieux à partir d'un grand nombre d'arbres de gènes. Olave et al. (2017) ont proposé, quant à eux, un nouvel algorithme rapide pour détecter une hybridation dans le cas des tris de lignées incomplets.

Un réseau de recombinaisons est un arbre phylogénétique auquel on rajoute des réticulations qui correspondent à des événements de recombinaison. De plus, chaque nœud (respectivement chaque branche de l'arbre) doit être étiqueté(e) par une séquence (respectivement par les positions des séquences où ont lieu les mutations

le long de cette branche). Hein (1993) a développé une extension de la méthode de parcimonie pour prendre en compte les recombinaisons. Plus récemment, pour pouvoir considérer un plus grand nombre d'espèces, Gusfield et al. (2003) ont restreint leur attention aux réseaux de niveau 1 (« galled trees » en anglais), tandis que Huson & Klöpper (2007) ont élaboré deux méthodes pour détecter des événements de recombinaison à partir de séquences binaires en utilisant des réseaux de niveau 1 et de niveaux supérieurs. Ces réseaux sont définis en détails ci-dessous.

Tout d'abord, Huson et al. (2010) proposent une définition plus précise des réseaux phylogénétiques dans le cas des réseaux enracinés. Soit G un graphe orienté, on notera V l'ensemble de ses sommets et E l'ensemble de ses arêtes. De plus, pour tout nœud v dans un graphe orienté, on peut définir une notion de degré entrant et de degré sortant de v , en distinguant les arêtes orientées vers v des arêtes orientées à partir de v . On a alors la définition suivante.

Définition 1.5. *Soit X un ensemble de taxons. Un réseau phylogénétique enraciné $N = (V, E, \lambda)$ sur X consiste en un graphe orienté et acyclique $G = (V, E)$, et en un étiquetage $\lambda : X \rightarrow V$ qui établit une bijection entre X et les feuilles de V (c.-à-d., les nœuds de degré sortant 0). De plus, on suppose qu'il existe exactement un nœud de degré entrant 0. Ce nœud est la racine du réseau.*

Notons que l'existence de cette racine implique la connexité du graphe.

On définit différents types de nœuds et d'arêtes de la façon suivante.

Définition 1.6.

- Un nœud dont le degré entrant est égal au plus à 1 est appelé un nœud d'arbre.
- Un nœud dont le degré entrant est égal au moins à 2 est appelé un nœud réticulé.

- Un nœud réticulé de degré entrant égal à 2 est appelé une bicomposition.
- Une arête menant à un nœud d'arbre est appelée une arête d'arbre.
- Une arête menant à un nœud réticulé est appelée une arête réticulée (ou réticulation).

Un réseau phylogénétique enraciné dont tous les nœuds réticulés sont des bicompositions est dit bicombinant. Cela signifie que tous les nœuds réticulés ont un degré entrant au plus égal à 2.

Rappelons qu'en théorie des graphes, un point d'articulation (respectivement un isthme) dans un graphe est un nœud (respectivement une arête) dont le retrait augmente le nombre de composantes connexes du graphe. Une composante biconnexe d'un graphe est un sous-graphe maximal induit par un ensemble d'arêtes qui ne contient aucun point d'articulation. On peut alors définir la notion de réseau de niveau k de la manière suivante (Choy et al., 2005).

Définition 1.7. Soit N un réseau phylogénétique enraciné bicombinant sur X . Le réseau N est dit de niveau k si le nombre maximum de réticulations contenues entièrement dans une composante biconnexe de N est égal à k .

La figure 1.12 représente un réseau de niveau 1 et un réseau de niveau 2. En effet, il y a deux réticulations dans la composante biconnexe du réseau (b) qui contient les nœuds h_3 et h_4 . Toutes les autres composantes biconnexes des réseaux (a) et (b) contiennent une seule réticulation. Dans un réseau de niveau 1, les réticulations sont indépendantes les unes des autres. Ce n'est plus le cas dans les réseaux de niveaux supérieurs. Van Iersel & Kelk (2011) ont élaboré un algorithme polynomial pour construire des réseaux de niveaux 1 et 2 en prenant en entrée un ensemble de triplets. Un triplet est un arbre phylogénétique à 3 feuilles. Van Iersel et al. (2010) ont développé une méthode générale pour inférer des réseaux de niveau k

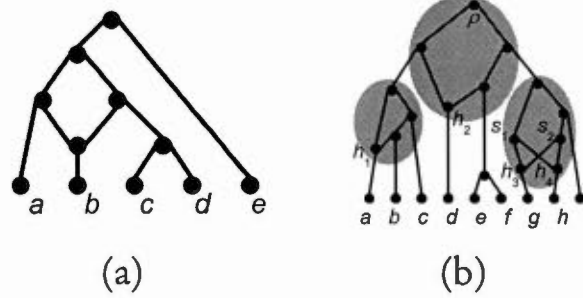


Figure 1.12 Un réseau de niveau 1 (a) et un réseau de niveau 2 (b).

à partir de différentes sortes de données. Huber et al. (2017) ont, quant à eux, trouvé un algorithme polynomial pour inférer un réseau de niveau 1 à partir de certains ensembles de réseaux à 2 et 3 feuilles.

D'autres réseaux phylogénétiques enracinés jouent un rôle important : les réseaux de clusters (Huson & Rupp, 2008). Pour tout nœud v dans un réseau phylogénétique enraciné, on note $L(v)$ l'ensemble des feuilles qui sont des descendants de v . On a alors la définition suivante.

Définition 1.8. *Un réseau de clusters est un réseau phylogénétique enraciné sur un ensemble de taxons X qui vérifie les propriétés suivantes :*

- $L(v) \subseteq L(w)$ si et seulement si v est un descendant de w .
- $L(v) = L(w)$ si et seulement si $v = w$, ou v (respectivement w) est un nœud réticulé et est parent de w (respectivement v).
- Si v est un enfant de w , alors il n'existe aucun nœud u tel que $L(v) \subsetneq L(u) \subsetneq L(w)$.
- Tout nœud réticulé a exactement une arête sortante et cette dernière est une arête d'arbre.

Pour toute arête d'arbre e d'un sommet v vers un sommet w , on définit le cluster $\gamma(e)$ de e comme l'ensemble des feuilles qui sont des descendants de w .

De plus, on définit un cluster sur X comme n'importe quel sous-ensemble strict de X . On peut voir X comme un ensemble d'espèces qui sont exactement les descendants d'une espèce ancestrale. Huson & Rupp (2008) ont élaboré un algorithme qui prend en entrée un ensemble de clusters sur X et qui construit un réseau de clusters ayant exactement cet ensemble de clusters au sens de la définition 1.8. Plusieurs méthodes d'inférence de réseaux de clusters ont été implémentées dans le logiciel Dendroscope (Huson & Scornavacca, 2012).

D'autres types de réseaux ont été définis, et d'autres méthodes d'inférence ont été développées. Par exemple, Makarenkov et al. (2006a), Boc et al. (2010b), ainsi que Boc & Makarenkov (2011) ont proposé des algorithmes pour identifier et valider statistiquement des transferts horizontaux de gènes à partir d'arbres contradictoires pour un même ensemble d'espèces. Ces méthodes sont implémentées dans le logiciel T-REX (Boc et al., 2012). Dans le cas d'un ensemble d'individus d'une même population, Bandelt et al. (1999) utilisent un critère de parcimonie pour obtenir un réseau en combinant plusieurs arbres de couverture minimale. Doyon et al. (2010) proposent, quant à eux, une méthode de parcimonie pour réconcilier un arbre d'espèces et des arbres de gènes, en prenant en compte les transferts horizontaux, les pertes de gènes et les duplications. Solís-Lemus & Ané (2016) ont présenté une nouvelle méthode, basée sur le principe de la pseudo-vraisemblance, pour inférer des réseaux phylogénétiques à partir de données génétiques. Wen et al. (2016) sont, quant à eux, les premiers à avoir utilisé des méthodes bayésiennes dans le cadre des réseaux phylogénétiques. Pour une vue d'ensemble sur la question, on se référera au livre de Huson et al. (2010).

Certaines de ces techniques ont été testées par Woolley et al. (2008). Elles ne sont efficaces que dans des configurations particulières. De plus, il n'existe aucun critère statistique permettant de valider les réseaux ainsi obtenus et de choisir le réseau qui représente le mieux l'histoire évolutive des espèces étudiées.

Les réseaux d'hybridation que nous obtenons dans cette thèse ne correspondent pas à une classe de réseaux déjà identifiée. Ils ne sont pas enracinés, mais la présence des hybrides créent des orientations sur certaines branches. De plus, un hybride ne peut pas avoir deux parents qui sont descendants l'un de l'autre.

1.7 Biolinguistique

1.7.1 Phylogénie et linguistique

De nombreuses ressemblances entre le processus de l'évolution des langues et celui de l'évolution des espèces ont été observées depuis la publication des travaux de Darwin (1888). Même avant cette publication, en 1863, August Schleicher discutait, dans une lettre envoyée à Ernst Haeckel, de certaines de ces similarités, comparant, par exemple, les langues mixtes aux plantes hybrides en botanique. Atkinson & Gray (2005) ont présenté un tableau des parallèles conceptuels les plus importants entre l'évolution linguistique et l'évolution biologique. Ils comparent notamment la sélection sociale en linguistique à la sélection naturelle des espèces, les emprunts de mots aux transferts horizontaux de gènes, les langues créoles aux hybrides de plantes, les textes anciens aux fossiles, les cognats aux homologies. Précisons ces deux derniers termes. Deux caractères observés chez deux espèces distinctes sont dits homologues s'ils sont hérités d'un ancêtre commun (Fitch, 2000). Ils sont dits, au contraire, analogues s'ils sont semblables mais se sont développés indépendamment les uns des autres, comme les ailes des oiseaux et celles des chauve-souris. En linguistique, un ensemble de cognats est un groupe de mots apparentés dans différentes langues, ayant une racine commune (Trask, 2000). Il existe une différence principale entre ces deux concepts : le concept d'homologie inclut la possibilité de transferts latéraux, tandis que le concept de cognat exclut tout processus d'emprunt. Les cognats jouent un rôle fondamental dans l'étude phylogénétique des langues (Gray & Atkinson, 2003), comme nous le verrons par

la suite. D'autres différences sont à noter (Geisler & List, 2013). Par exemple, l'alphabet biologique (p. ex., l'ADN) est universel, tandis que l'ensemble des sons utilisés pour former des mots est propre à chaque langue. De plus, les séquences sont généralement beaucoup plus longues en biologie moléculaire qu'en linguistique, et la sélection d'une liste de mots de base pour l'application des méthodes phylogénétiques reste une tâche très délicate.

Néanmoins, de nombreux chercheurs ont utilisé des méthodes phylogénétiques pour étudier des données linguistiques (Gray & Atkinson, 2003; Rexová et al., 2003; Atkinson & Gray, 2005), notamment dans le cadre de l'évolution des langues indo-européennes (IE) (Diamond & Bellwood, 2003). En effet, un arbre phylogénétique représentant l'évolution lexicale d'un ensemble de langues est équivalent à un arbre d'espèces représentant les principaux événements de spéciation parmi ces espèces (Atkinson & Gray, 2005; Gray & Atkinson, 2003; Rexová et al., 2003). La figure 1.13 reprend, par exemple, la topologie et les principaux groupes linguistiques de l'arbre enraciné de 84 langues IE inféré par Gray & Atkinson (2003). Cet arbre est représenté dans son ensemble dans la figure 1.14 (Gray & Atkinson, 2003; Atkinson & Gray, 2006).

L'origine des langues IE est un des problèmes fondamentaux de l'histoire des langues. Deux théories principales s'affrontent. La première (Gimbutas, 1973a,b) a été développée sur des bases archéologiques et linguistiques. Elle fait remonter l'origine des langues IE à 3000 ou 4000 ans avant JC au sein de la culture Kourgane, au sud de la Russie et de l'Ukraine. La deuxième (Renfrew, 1988) est basée essentiellement sur des considérations archéologiques et situe l'origine des langues IE aux environs de 7000 ans avant JC en Anatolie. Gray & Atkinson (2003), ainsi que Bouckaert et al. (2012), ont permis de renforcer cette dernière hypothèse grâce à des méthodes phylogénétiques bayésiennes.

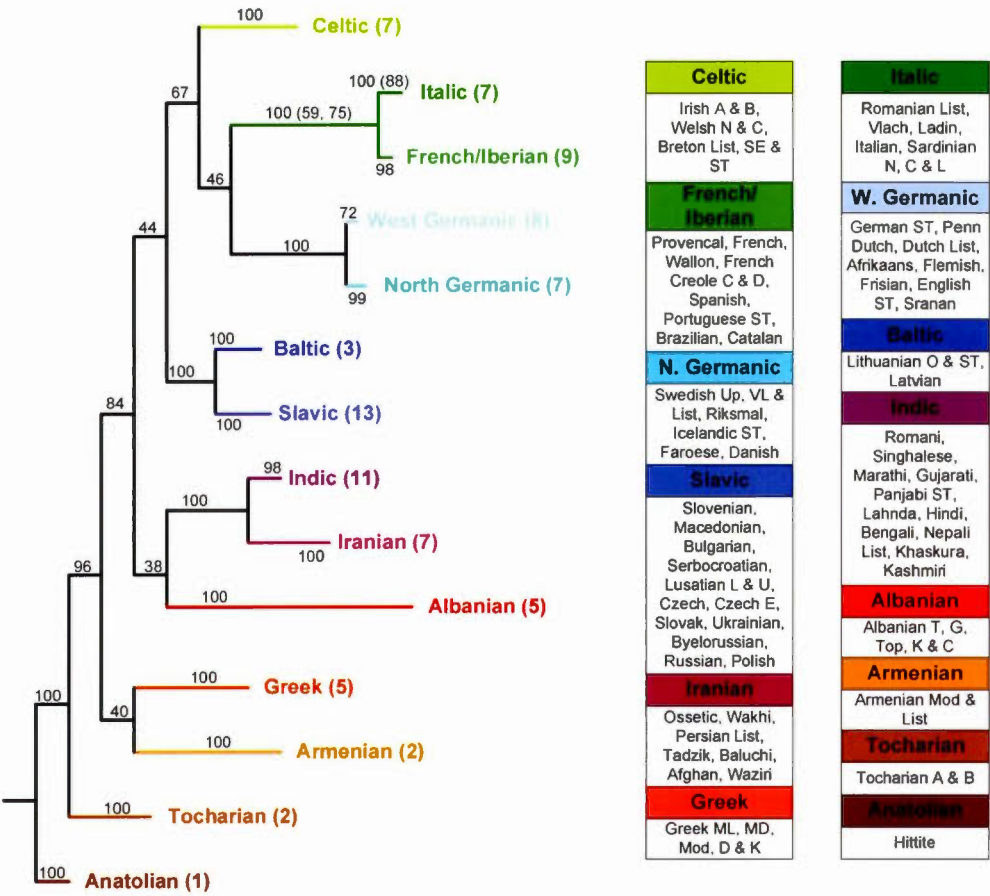


Figure 1.13 L'arbre d'évolution des langues IE pour 14 des principaux groupes linguistiques. Les nombres sur les branches représentent leurs scores de bootstrap. Le nombre de langues appartenant à chaque groupe est indiqué entre parenthèses.

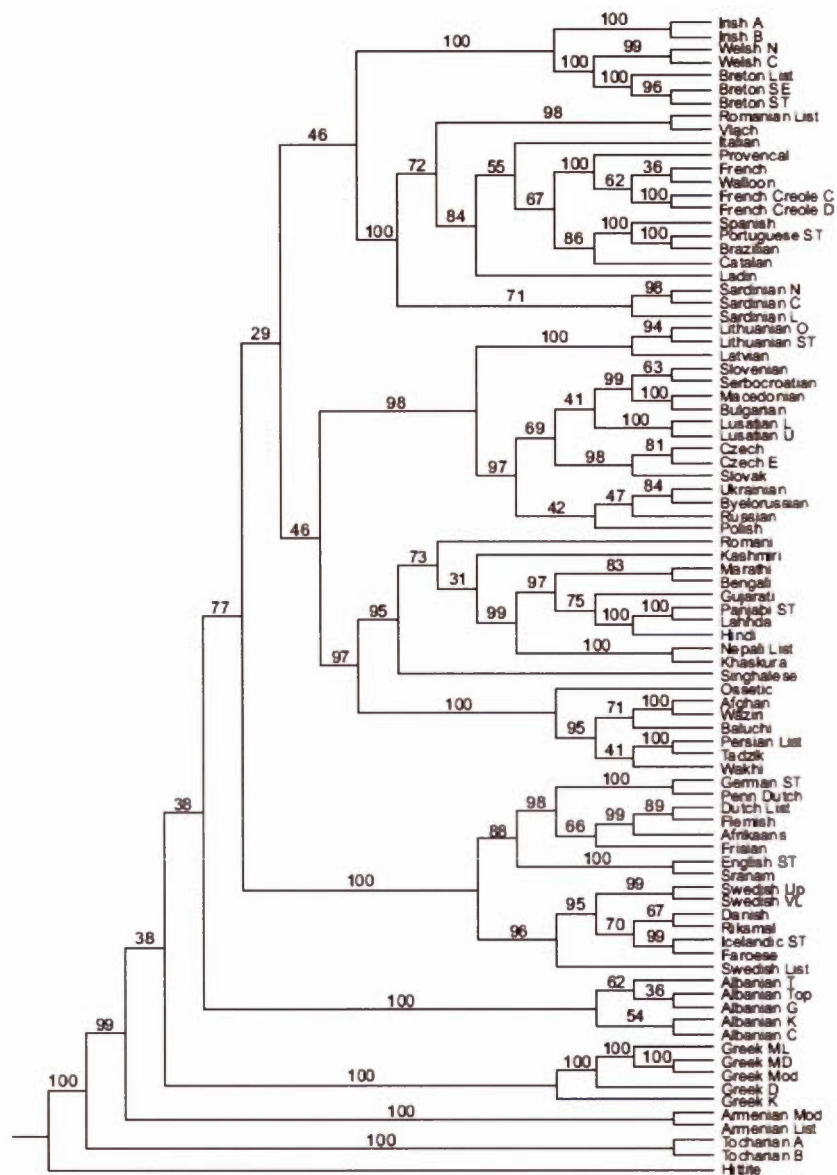


Figure 1.14 L'arbre complet d'évolution de 84 langues IE selon Gray & Atkinson (2003); Atkinson & Gray (2006). Les nombres sur les branches représentent leurs scores de bootstrap.

Notons qu'il n'est pas possible d'appliquer directement les méthodes de biologie computationnelle, sans un souci d'adaptation au contexte de la linguistique. Cet effort d'adaptation est toujours essentiel dans la recherche interdisciplinaire. Les algorithmes phylogénétiques existants doivent donc être modifiés et les flux de travaux adaptés, afin d'obtenir des résultats linguistiques qui ont une réelle signification. Bromham (2017) démontre malgré tout que de nombreux problèmes anciens de linguistique historique pourraient être résolus à l'aide d'outils développés en bioinformatique.

1.7.2 Utilisation des réseaux phylogénétiques pour représenter l'évolution des langues

Le modèle de représentation en arbre suppose que l'évolution des langues a été strictement divergente, que chaque langue a été transmise dans son intégralité et que la fréquence d'emprunts de mots entre les langues a été relativement faible. Cependant, de nombreux phénomènes comme l'emprunt de mots et les langues hybrides ne peuvent pas être représentés par un arbre. Par exemple, une étude de 80000 mots montre que l'anglais, qui est une langue germanique, a emprunté 56,5% de son lexique au vieux français (langue d'oïl) et au latin, 5,3% au grec, 13,2% à d'autres langues, et a hérité seulement 25% de son lexique actuel de son ancêtre direct, le vieux germanique (Finkenstaedt & Wolff, 1973; Pagel, 2000). Les emprunts de mots sont également à l'origine des langues hybrides. Il existe plusieurs types de telles langues : pidgins, créoles et hybrides lexicales (Velupillai, 2015). Dans le cas des pidgins, le lexique vient d'une langue et la syntaxe d'une autre langue. Une langue créole est, quant à elle, une langue stable, qui a dérivé d'un pidgin et qui est parlée comme une langue maternelle.

Les réseaux phylogénétiques apparaissent ainsi comme un outil essentiel pour modéliser l'évolution des langues. Déjà en 1872, certains inconvénients de l'utilisation

du modèle arborescent en linguistique ont été mis en lumière par Schmidt (1872). Nakhleh et al. (2005) sont parmi les premiers à avoir utilisé des réseaux phylogénétiques pour identifier les contacts lexicaux entre 24 langues IE. D'autres travaux ont utilisé des split-graphes (Atkinson & Gray, 2005; Bryant et al., 2005; Heggarty et al., 2010), des réseaux latéraux minimaux (Nelson-Sathi et al., 2011, 2013; List et al., 2014a), et des réseaux de transferts horizontaux de mots (Boc et al., 2010a). La plupart de ces réseaux permettent essentiellement de détecter des emprunts de mots. Seuls les split-graphes peuvent être utilisés pour identifier des hybrides, mais de manière implicite. Par exemple, le split-graphe obtenu pour 9 langues germaniques par Bryant et al. (2005) (figure 1.15) permet d'identifier le sranan parlé au Surinam comme une langue hybride de l'anglais et du néerlandais, ce qui est effectivement le cas. Cependant, ces réseaux ne permettent que de représenter de manière implicite les réticulations les plus évidentes. Par exemple, ils ne peuvent pas être utilisés pour identifier explicitement la langue hybride, les parents et le degré d'hybridation correspondant (c.-à-d., la proportion du lexique de la langue hybride qui provient de chacun des deux parents). Ils ne permettent donc pas de quantifier la fréquence des événements d'emprunts de mots. De plus, seulement certaines méthodes offrent l'avantage de détecter le sens des réticulations en linguistique. Parmi les travaux qui ont utilisé de telles méthodes, mentionnons ceux de Van Der Aek et al. (2007), qui utilisent la distance de Levenshtein (Levenshtein, 1966) pour identifier la source et la destination des emprunts de mots, et ceux de Delz (2013), qui a appliqué l'algorithme de détection des transferts horizontaux de gènes (Boc & Makarenkov, 2003; Boc et al., 2010b) du site web T-REX (Boc et al., 2012) à la détection explicite des emprunts de mots.

1.7.3 Description des données

La plupart des études précédentes concernant la classification des langues IE ont utilisé le jeu de données de 84 langues IE collecté par Dyen et al. (1992). Cette

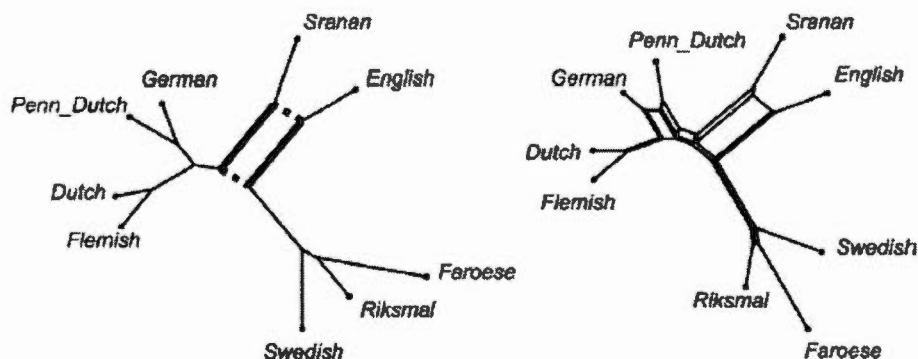


Figure 1.15 Split-graphes pour 9 langues IE.

base de données reprend les 200 mots de la liste Swadesh (1952). Cette liste est une des diverses listes de mots de signification de base, constituées par M. Swadesh dans les années 1940-50. Elle est souvent utilisée en lexicostatistique (évaluation quantitative de la parenté des langues) et en glottochronologie (datation des divergences entre les langues). Pour chacun des 200 mots de la liste Swadesh, la base de données contient les traductions utilisées dans 84 langues IE. Les traductions de chacun de ces mots ont été regroupées par Dyen et al. (1992) en différents ensembles de cognats. Deux traductions, dans deux langues différentes, ont été identifiées comme apparentées (c.-à-d., appartenant au même ensemble de cognats) si, pour ces deux langues, elles avaient la même signification, et si elles avaient une histoire évolutive ininterrompue depuis une forme ancestrale commune. Les traductions pour lesquelles il était connu qu'elles étaient reliées par des emprunts ou par des similarités accidentelles (p. ex., le mot "bad" existe en anglais et en farsi, mais les linguistes considèrent généralement cette similarité comme accidentelle) ont été mises dans une classe séparée. Il a été parfois difficile de différencier les traductions appartenant à un même ensemble de cognats, de celles résultant d'emprunts ou de similarités accidentelles ; de telles traductions (peu nombreuses) ont

été classifiées comme cognats douteux. La base de données de Dyen et al. (1992) a été utilisée par Gray & Atkinson (2003) et Atkinson & Gray (2006) pour inférer l'arbre d'évolution de 84 langues IE (figure 1.14). Pour reconstruire nos réseaux d'hybridation, nous avons consulté d'autres ressources linguistiques (le dictionnaire d'étymologie Douglas Harper's Online (Harper, 2001) et la base de données IE Lexical Cognacy (IELex) (Dunn, 2012) notamment), qui contiennent certaines informations pertinentes sur les emprunts de mots et les similarités accidentelles. Nous avons ainsi modifié certains des ensembles de cognats créés par Dyen et al. (1992). Plus précisément, les mots empruntés, mis de côté par Dyen et al. (1992), ont été ajoutés aux ensembles de cognats correspondants (c.-à-d., les ensembles de cognats contenant la source du mot emprunté). Dans certains cas exceptionnels, certains ensembles contenant des cognats douteux ont été soit supprimés soit réunis. Au total, notre base de données contient 1315 ensembles de cognats et est disponible à l'adresse URL suivante :

<http://www.info2.uqam.ca/~makarenv/BL/index.html>.

Nous avons testé différentes stratégies explicitées dans le chapitre 3 pour calculer une matrice de distances entre les 84 langues IE considérées dans notre étude. Nous avons finalement retenu la distance de Hamming, qui donnait les meilleurs résultats. Nous partons de la matrice de présence/absence de taille (84×1315) . L'élément (i, j) de cette matrice est égal à 1 si un mot de la langue i est présent dans l'ensemble de cognats j , et égal à 0 sinon. La distance entre deux langues est alors calculée comme la distance de Hamming entre les deux colonnes correspondant à ces deux langues (c.-à-d., le nombre d'ensembles de cognats qui contiennent des mots de seulement une des deux langues comparées). Les autres distances testées sont présentées en détails dans le chapitre 2.

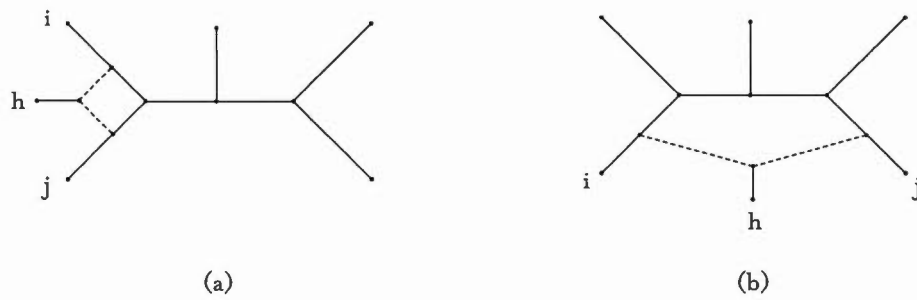


Figure 1.16 Hybrides entre branches terminales.

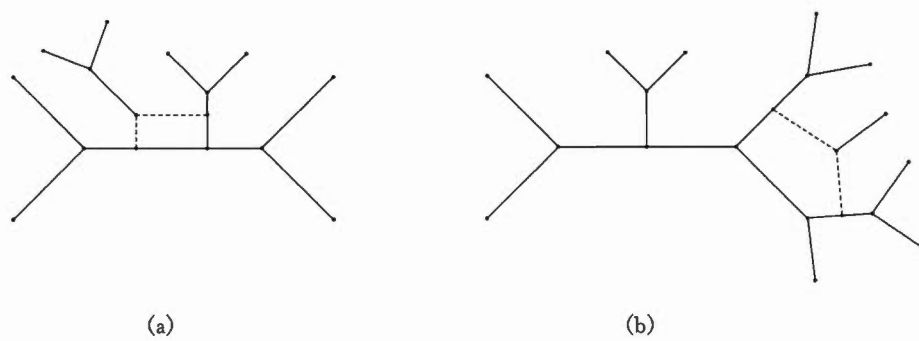


Figure 1.17 Hybrides entre branches ancestrales.

1.8 Présentation des différentes parties de la thèse

Dans le chapitre 2, nous présentons les résultats publiés dans (Willems et al., 2014). Nous avons adapté l'algorithme neighbor-joining (Saitou & Nei, 1987) au cas des réseaux d'hybridation. Notons que dans notre modèle, les événements d'hybridation peuvent être situés entre des branches terminales comme dans la figure 1.16, aussi bien qu'entre des branches ancestrales comme dans la figure 1.17. Dans les deux cas, les deux branches parentes peuvent avoir (ou ne pas avoir) d'ancêtre commun direct (remarquez la différence entre les réseaux (a) et (b) dans les deux figures). Dans toutes les figures, les branches d'hybridation (c.-à-d., de réticulation) sont représentées en pointillés.

Notre algorithme prend en entrée une matrice de distances de taille n correspondant aux n espèces étudiées. La sortie de notre algorithme est soit un arbre phylogénétique classique à n feuilles, soit un réseau d'hybridation avec les mêmes n nœuds terminaux. Dans ce réseau, chaque hybride est clairement identifié avec ses deux parents et le degré d'hybridation correspondant à chacun des parents. La complexité temporelle de notre algorithme est $O(n^3)$ comme pour l'algorithme NJ standard.

Nous avons tout d'abord démontré que si la matrice de distances en entrée est une distance d'arbre, notre algorithme retrouve l'unique arbre phylogénétique correspondant à cette matrice. Puis, nous avons testé notre méthode sur des données réelles et simulées, et nous avons trouvé de très bons taux de détection des hybrides.

Dans le chapitre 3, nous présentons les résultats obtenus (Willems et al., 2016) en appliquant notre nouvel algorithme dans le cadre de la biolinguistique afin d'identifier explicitement des langues hybrides (hybrides lexicales dans notre étude) et

des événements d'emprunt de mots. Nous avons ainsi obtenu un réseau d'hybridation explicite pour 84 langues IE. Nous avons également montré que les split-graphes (Bandelt & Dress, 1992a; Bryant & Moulton, 2004) et les réseaux de niveau 1 (Huson et al., 2009) peuvent être utilisés pour analyser des données linguistiques. Cependant, ces réseaux sont, dans certains cas, très difficiles à interpréter. L'avantage essentiel de notre méthode est d'identifier explicitement la langue hybride (ou réceptrice dans le cas des emprunts de mots) et ses parents, et de quantifier l'influence de chaque parent sur la langue hybride.

Nous présentons également des données linguistiques et historiques qui soutiennent les résultats de notre analyse.

Dans le chapitre 4, nous présentons une méthode d'inférence qui utilise une approche mixte entre l'algorithme NJ et le principe du maximum de vraisemblance. Il s'agit d'un algorithme itératif, qui détermine à chaque pas les meilleurs voisins selon le critère de NJ. On détermine ensuite l'hybride le plus vraisemblable et on compare sa vraisemblance à celle des deux voisins trouvés par NJ. On choisit entre les voisins et l'hybride selon un critère statistique appelé BIC pour Bayesian Information Criterion (Schwarz & Gideon, 1978). Nous nous plaçons ici dans le cadre de données binaires. Ce nouvel algorithme a été comparé à celui du chapitre 2 sur différents types de données. Il s'est avéré plus efficace sur les données simulées, notamment pour les arbres de petite taille (8 et 16 feuilles), malgré des temps de calcul plus importants.

Dans l'annexe A, nous présentons le code en *C++* du programme présenté dans le chapitre 2. Dans l'annexe B se trouvent les arbres phylogénétiques de certains mots utilisés pour les analyses biolinguistiques du chapitre 3.

1.9 Glossaire

ADN (acide désoxyribo nucléique) : Macromolécule constituée de deux chaînes enroulées en double hélice. Ses deux brins sont assemblés à partir de nucléotides. Chaque nucléotide comprend un sucre, le désoxyribose, un phosphate et une des quatre bases azotées (adénine, guanine, cytosine et thymine). L'ADN est le support de l'information génétique des organismes vivants.

Alignement : Opération qui consiste à disposer les unes en dessous des autres des portions de séquences similaires en minimisant leurs différences (on peut aligner entre eux des gènes d'une même famille multigénique ou des gènes d'espèces différentes). Si ces gènes sont homologues, les différences d'acides aminés ou d'acides nucléiques entre les séquences actuelles sont le témoignage de mutations qui ont eu lieu dans le passé.

Aminoacide (acide aminé) : Unité constitutive des protéines. Il existe vingt acides aminés communs : alanine, arginine, asparagine, aspartate, cystéine, glutamine, glycine, histidine, isoleucine, leucine, lysine, méthionine, phénylalanine, proline, glutamate, sérine, thréonine, tryptophane, tyrosine et valine.

Archaea : Les Archées ou Archaea (anciennement appelés archéobactéries, du grec *archaios*, « ancien » et *backterion*, « bâton ») sont un groupe majeur de microorganismes. Elles constituent un taxon du vivant caractérisé par des cellules sans noyau et se distinguant des Eubactéries (vraies bactéries) par certains caractères biochimiques, comme la constitution de la membrane cellulaire ou le mécanisme de réplication de l'ADN.

ARN (acide ribonucléique) : Polymère linéaire dont la sous-unité de base, un ribonucléotide, contient le sucre ribose.

Bacteria : Les bactéries appartiennent au vaste ensemble des microbes qui com-

prennent également les virus, les champignons et les parasites. Microorganismes invisibles à l'œil nu, les bactéries sont constituées d'une seule cellule dépourvue d'un vrai noyau. Elles contiennent un seul chromosome formé d'un long filament d'ADN.

Clade : Vient du grec *clados* qui signifie « arête ». Taxon strictement monophylétique, c'est-à-dire contenant un ancêtre et tous ses descendants.

Cognat : Un ensemble de cognats est un groupe de mots apparentés dans différentes langues et ayant une racine commune.

Configuration : Ensemble de topologies d'arbres ayant des caractéristiques communes.

Créole : On parle de créole lorsqu'un pidgin devient la langue maternelle d'une portion ou de l'ensemble d'une communauté linguistique qui l'a adopté.

Eucarya : Les Eucaryotes (du grec *eu*, vrai et *karuon*, noyau) comprennent quatre grands règnes du monde vivant : les animaux, les champignons, les plantes et les protistes. Ils constituent donc un très large groupe d'organismes, unis et pluricellulaires, définis par leur structure cellulaire (noyau, ADN, cytosquelette, etc).

Extragroupe (outgroup) : On dit aussi groupe extérieur ou encore « outgroup » tiré de l'anglais. Groupe que l'on sait a priori placé en dehors d'un ensemble de taxons dont on cherche les relations de parenté.

Horloge moléculaire (hypothèse) : Hypothèse selon laquelle les molécules d'une même classe fonctionnelle évoluent régulièrement dans le temps à un rythme égal dans différentes lignées. Ainsi la quantité des différences moléculaires constatées de nos jours dans des séquences homologues d'espèces distinctes peut être utilisée pour estimer le temps écoulé depuis le dernier ancêtre commun à ces espèces (ou

le temps de divergence).

Pidgin : Système linguistique utilisant généralement le lexique d'une langue et la syntaxe d'une autre langue, doté de structures rudimentaires (lexique réduit, structures grammaticales élémentaires) et de fonctions sociales limitées. Le pidgin n'est la langue maternelle d'aucun des locuteurs qui l'utilisent.

Racine : Segment de l'arête en amont du nœud du rang le plus important, définissant le groupe extérieur (voir Extragroupe). En d'autres termes, c'est la position dans l'arbre du groupe extérieur. La racine peut être considérée comme un point de référence pour l'interprétation des caractères : les états de caractères de l'extragroupe (outgroup) sont des états plésiomorphes, les états qui en diffèrent sont apomorphes. Remarque : pour pouvoir comparer aisément deux arbres, il faut les enraciner chacun avec la même espèce ou avec le même taxon.

Site de restriction : Séquence de nucléotides qui est reconnue par certaines enzymes spécifiques comme un site de coupure dans la molécule d'ADN.

Taxon : Ensemble des organismes reconnus et définis dans chacune des catégories de la classification biologique hiérarchisée. En d'autres termes : contenu concret d'une catégorie. Exemple : *Canis lupus*, le loup, est un taxon de rang spécifique (catégorie : espèce); les canidés (Chien, Loup, Renard) constituent un taxon de rang familial (catégorie : famille).

CHAPITRE II

UN NOUVEL ALGORITHME EFFICACE POUR L'INFÉRENCE DE RÉSEAUX D'HYBRIDATION EXPLICITES EN SUIVANT LE PRINCIPE DE NEIGHBOR-JOINING

RÉSUMÉ

Ce chapitre est une reproduction de l'article suivant :

Willems, M., Tahiri, N. & Makarenkov, V. (2014). A new efficient algorithm for inferring explicit hybridization networks following the neighbor-joining principle. *Journal of Bioinformatics and Computational Biology*, 12(05), 1450024.

Depuis une cinquantaine d'années, plusieurs méthodes ont été développées pour inférer des arbres phylogénétiques, et de nombreux programmes sont disponibles dans ce contexte. Cependant, il existe des phénomènes biologiques, comme l'hybridation, la recombinaison, le transfert horizontal de gènes, qui ne peuvent pas être représentés sous la forme d'un arbre. Nous avons alors besoin des réseaux phylogénétiques pour représenter adéquatement ces phénomènes évolutifs fondamentaux. Dans ce chapitre, nous présentons un nouvel algorithme efficace pour inférer un réseau d'hybridation à partir d'une matrice qui contient les distances évolutives entre les espèces considérées. Il s'agit d'une heuristique basée sur le concept de neighbor-joining et le critère des moindres carrés. À chaque étape de l'algorithme, avant de joindre deux nœuds, on vérifie si un événement d'hybridation impliquant l'un des ces deux nœuds pourrait être identifié. Si la matrice de départ correspond à une métrique d'arbre, notre algorithme retrouve systématiquement l'arbre phylogénétique associé à cette matrice. Des simulations avec des arbres de grande taille (32 et 64 feuilles) ont permis de trouver de très bons taux d'identification des hybrides, aussi bien avec des données séquentielles que des matrices de distances. Tous les résultats obtenus par notre nouvel algorithme pour des données réelles et simulées sont présentés en détails dans ce chapitre.

Contributions : J'ai rédigé l'article, conçu l'algorithme et mené l'essentiel des simulations. Nadia Tahiri a fait certaines des simulations. Vladimir Makarenkov a participé à la conception de l'algorithme et a relu l'ensemble de l'article.

2.1 Abstract

Several algorithms and software have been developed for inferring phylogenetic trees. However, there exist some biological phenomena such as hybridization, recombination, or horizontal gene transfer which cannot be represented by a tree topology. We need to use phylogenetic networks to adequately represent these important evolutionary mechanisms. In this article, we present a new efficient heuristic algorithm for inferring hybridization networks from evolutionary distance matrices between species. The famous neighbor-joining concept and the least-squares criterion are used for building networks. At each step of the algorithm, before joining two given nodes, we check if a hybridization event could be related to one of them or to both of them. The proposed algorithm finds the exact tree solution when the considered distance matrix is a tree metric (i.e., it is representable by a unique phylogenetic tree). It also provides very good hybrids recovery rates for large trees (with 32 and 64 leaves in our simulations) for both distance and sequence types of data. The results yielded by the new algorithm for real and simulated datasets are illustrated and discussed in detail.

2.2 Introduction

The evolution of species is commonly modeled by means of phylogenetic (i.e., additive) trees, in which leaves represent contemporary species, internal nodes - their ancestors, branches - the ancestor-descendant relationships between species and branch lengths - the evolutionary time. There exist many algorithms for inferring phylogenetic trees from distance or sequence datasets. Distance-based methods take as input a distance matrix between species, whereas sequence-based methods take as input some DNA or protein sequences. Distance-based methods are usually much faster than the sequence-based methods. The most popular of these methods are neighbor-joining (Saitou & Nei, 1987) and UPGMA (Sneath

& Sokal, 1973) (Unweighted Pair Group Method with Arithmetic mean). If we have to analyze datasets with relatively small number of species, we can also use sequence-based approaches such as maximum parsimony (Fitch, 1971) and maximum likelihood (Felsenstein, 1981). More recently, several Bayesian methods (Rannala & Yang, 1996) have been developed to extend maximum-likelihood estimations to a larger number of species.

Some well-known reticulate evolution phenomena, such as horizontal (lateral) gene transfer, hybridization, homoplasy and genetic recombination cannot be modeled by a tree structure (Sneath et al., 1975; Leclerc & Makarenkov, 1998; Makarenkov & Leclerc, 2000). However, reticulate evolution has long been neglected in phylogenetic analyses. The first methods for studying the mechanisms of reticulate evolution started to appear in the mid-1970s (Sneath et al., 1975; Sonea & Panisset, 1983). Several tentative methods have been proposed for the identification of reticulate evolution in nucleotide sequences. They include displays of compatibility (Sneath et al., 1975), tests for clustering (Stephens, 1985), a randomization approach (Sawyer, 1989), and an extension of the parsimony method of phylogenetic reconstruction that allows recombination (Hein, 1993). Rieseberg & Morefield (1995) developed a computer program, RETICLAD, allowing one to identify hybrids based on the expectation that they would combine the characters of their parents. However, the latter program could only find reticulation events between terminal branches of a tree. The popular method of split decomposition enables the representation of data in the form of a split graph revealing conflicting signals contained in the data (Bandelt & Dress, 1992a,b). A split graph is an implicit phylogenetic network, that is used to show conflicting placements of taxa. Bryant & Moulton (2004); Huson & Bryant (2006) introduced a network-inferring method, Neighbor-Net, allowing the reconstruction of planar phylogenetic split networks. The latter method usually provides several decompositions of the set of species

but it is very difficult to deduce explicit reticulation events (e.g., horizontal gene transfers with their directions) from these decompositions. Gambette & Huson (2008) have then improved the visualization of these decompositions. Huson and Bryant have developed the SplitsTree (Huson & Bryant, 2006) software, which has become the most commonly used tool for inferring implicit phylogenetic networks. Legendre & Makarenkov (2002); Makarenkov & Legendre (2004, 2000) proposed to use reticulograms for detecting reticulation events in evolutionary data. They developed a distance-based method to infer reticulate phylogenies. The latter method uses the topology of a phylogenetic tree as backbone structure for building a reticulogram by adding reticulation branches to this tree according to an optimization criterion. Hallett & Lagergren (2001) showed how horizontal gene transfer events can be detected by evaluating topological differences between species and gene trees. Makarenkov et al. (2006a), Boc et al. (2010b), and Boc & Makarenkov (2011) have proposed algorithms for identifying and validating statistically horizontal gene transfer events from species and gene trees defined on the same set of species. These methods have been implemented in the software T-REX (Boc et al., 2012). In the case of a set of individuals of the same population, Bandelt et al. (1999) have used a parsimony criterion to construct a network from several minimal cover trees. Doyon et al. (2010) have also proposed a parsimony method to reconcile a species tree and several gene trees, by taking into account horizontal transfers, gene losses and duplications. Huson & Rupp (2008) and Van Iersel et al. (2010) have used the notion of cluster networks to reconcile several contradictory phylogenetic trees. In the case of two contradictory trees, Albrecht et al. (2012) have proposed a partially parallel algorithm to find a minimum hybridization network from two input trees, but this algorithm remains very slow even when it is executed on a computer with multiple cores (Chen et al., 2012). Wu (2010) and Chen & Wang (2012) have presented algorithms for constructing hybridization networks from more than two input trees. Van Iersel & Kelk (2011)

have developed a polynomial-time algorithm for inferring a phylogenetic network from a dense set of triplets. Some quartet-based methods have been also introduced for tree reconstruction by Schmidt et al. (2002), and then implemented by Strimmer & von Haeseler (1996) in the program TREE-PUZZLE. These methods have been extended to network inference in the software QNet (Grünwald et al., 2007) and Quartet-Net (Yang et al., 2013). Huson & Kloepper (2005); Huson & Klöpper (2007) have designed two methods to detect recombination events from binary sequences by using general reticulate networks and galled trees (i.e., reticulate networks in which all reticulations are independent from each other). Many of the discussed techniques were tested by Woolley et al. (2008). Mention that all of them are plausible only under specific evolutionary constraints and assumptions. It is worth noting that all of the existing methods for building explicit hybridization networks take as input either a set of contradictory trees, or a set of clusters, or a set of triplets (see (Semple, 2007) for a more detailed description of all these structures). The main novelty of our work is that our algorithm takes as input a single distance matrix to infer from it an explicit hybridization network using the well-known principle of minimum evolution (Saitou & Nei, 1987).

The main goal of our article is to present a new efficient algorithm for inferring phylogenetic networks from a distance matrix between species. Our method can be seen as a generalization of the very popular neighbor-joining algorithm (Saitou & Nei, 1987) to hybridization networks. The paper is organized as follows: in Section 2.3, we describe the phenomenon of hybridization, in Section 2.4, we recall the main features of the NJ algorithm, in Section 2.5, we define a hybridization network and prove some of its important general properties, in section 2.6, we present the necessary least-squares formulas and discuss the network building strategy, Section 2.7 is dedicated to our algorithm, in Section 2.8, we provide some experimental results obtained on additive, non-additive and real datasets.

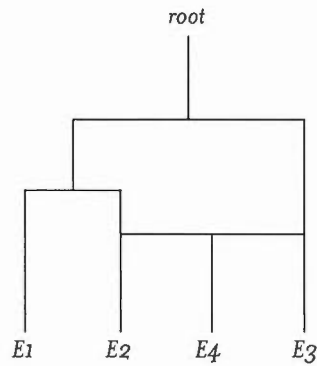


Figure 2.1 An example of hybridization.

2.3 Hybridization

Hybridization is a very common mechanism of reticulate evolution. In Fig. 2.1, two lineages (Root-Species E2 and Root-Species E3) recombine to create a new species (Species E4). If the new species have the same number of chromosomes as the parent species, the process is called diploid hybridization. When it has the sum of the number of its parents' chromosomes, it is called polyploid hybridization. The three main mechanisms of hybridization are the following:

- Autopolyploidization is a speciation event involving the doubling of the chromosomes within a single species (intraspecific hybridization). It produces a bifurcating speciation event in a phylogenetic tree.
- Allopolyploidization is a type of hybridization between two species, when an offspring acquires the complete diploid chromosome complements of the two parents. In this case, the parents do not need to have the same number of chromosomes. Allopolyploidization results in instantaneous speciation because any backcrossing to the diploid parents is likely to produce a sterile triploid offspring.

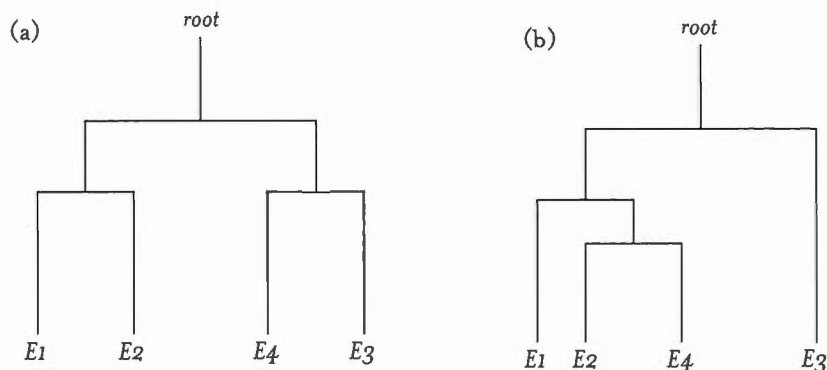


Figure 2.2 Two different trees for representing the same hybridization phenomenon of Fig. 2.1.

- Diploid hybrid speciation is a normal sexual event taking place between parents from different but related species. In nearly all cases, the two parents need to have the same number of chromosomes. In this case, successful backcrossing to the parents is possible, so the hybrids have to be isolated from the parents to become new species.

Consider the problem of modeling reticulate evolution after diploid hybrid speciation. In normal diploid organisms, each chromosome consists of a pair of homologs. In the process of diploid hybridization, the hybrid inherits one of the two homologs for each chromosome from each of its two parents. Since the genes from both parents contributed to the hybrid, the evolution of genes inherited from each parent can be represented on separate trees inside a network model. Classical phylogenetic analysis of the four species involved in a hybrid speciation event (Fig. 2.1) will produce one of the two trees in Fig. 2.2.

Hybridization is very common in plants. There exist more than 70000 natural hybrid plants (Stace, 1991), and some hybrid plants can be created by humans

to introduce some specific characteristics into cultivated species (Judd, 2008). Hybridization is also very common among fish, amphibians and reptiles (Dawley, 1989), and is rare in other groups, particularly in birds, mammals, and most arthropods. The latter groups are only occasionally affected by hybrid speciation. They usually produce triploids which can only reproduce by asexual modes.

The main goal of our article is to model and infer phylogenetic networks taking into account possible hybridization events.

2.4 Neighbor-joining for trees

This section starts with some basic definitions concerning phylogenetic trees (Barthélemy & Guénoche, 1991). The distance $d(x, y)$ between two vertices x and y in a phylogenetic tree T is defined as the sum of all branch lengths of the unique path connecting x and y in T .

Définition 2.1. *Let X be a set of n species. A dissimilarity d on X is a non-negative function on $X \times X$ such that for all x, y in X :*

$$(1) \ d(x, y) = d(y, x), \text{ and}$$

$$(2) \ d(x, y) = d(y, x) \geq d(x, x) = 0.$$

Définition 2.2. *A dissimilarity d on X is said to satisfy the four point condition (Buneman, 1974) if for all x, y, z , and w in X : $d(x, y) + d(z, w) \leq \text{Max}\{d(x, z) + d(y, w); d(x, w) + d(y, z)\}$.*

Définition 2.3. *For any finite set X , an X -tree is an ordered pair (T, ϕ) consisting of a tree T , with a set of vertices V and a relation $\phi : X \rightarrow V$, such that, for all $v \in V$ with a degree at most equal to 2, $v \in \phi(X)$. An X -tree is a phylogenetic tree if ϕ is a bijection from X to the set of all leaves of T . It is said to be binary if each internal vertex has a degree equal to 3.*

The main theorem relating the four point condition and phylogenetic trees is as follows:

Théorème 2.1. (*Zaretskii, Buneman, Patrinos and Hakimi, Dobson*) *Any dissimilarity satisfying the four point condition can be represented as a phylogenetic tree such that for all x, y in X , $d(x, y)$ is equal to the length of the path connecting leaves x and y in T .*

This dissimilarity is called an additive distance, a tree distance or a tree metric. This tree is unique.

The neighbor-joining (Saitou & Nei, 1987) (NJ) algorithm is the most popular distance-based method for inferring phylogenetic trees. Atteson (1999) proved that this algorithm finds the correct phylogeny if the input distances between species are sufficiently close to the real evolutionary distances.

Throughout this article, we take as input a distance matrix $D = \{D[i][j]\}_{\substack{1 \leq i \leq n; \\ 1 \leq j \leq n}}$ on a set of n species, and we obtain as output a network corresponding to the evolutionary history of these species. Obviously, $D[i][i] = 0$ for all $1 \leq i \leq n$, and $D[i][j] = D[j][i]$ for all $1 \leq i \leq n$ and $1 \leq j \leq n$.

NJ is a clustering algorithm which starts with a bush composed of n leaves and n branches, where n is the number of current species. This tree is gradually transformed into an unrooted phylogenetic tree with the same n leaves and with $2n - 3$ branches. The i -th step consists in choosing two neighbors among $n - i + 1$ candidates. We consider all the $\frac{(n-i+1)(n-i)}{2}$ configurations similar to the one represented in Fig. 2.3. For each of these configurations, we calculate the branch lengths which minimize a least-squares criterion, in which we compare the input dissimilarities with the tree metric distances.

Saitou & Nei (1987) showed that the sum of the branch lengths of the tree topology

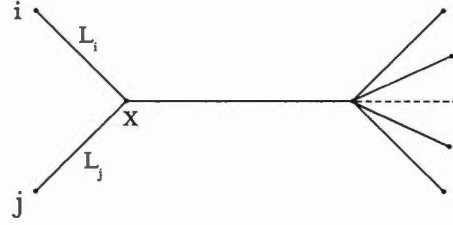


Figure 2.3 Configuration where nodes i and j are chosen as neighbors.

in Fig. 2.3 is equal to:

$$S_{i,j} = \frac{1}{2}D[i][j] + \frac{\sum_{1 \leq k \leq n, k \neq i,j} [D[i][k] + D[j][k]]}{2(n-2)} + \frac{\sum_{1 \leq k < l \leq n, k,l \neq i,j} D[k][l]}{n-2}. \quad (2.1)$$

We connect nodes i and j that minimize the total evolution, i.e., the sum of branch lengths $S_{i,j}$. We replace the selected nodes i and j by node X (their direct common ancestor) and obtain a distance matrix of size $n-1$. We compute the new distances from X to the remaining leaves of the tree by using the following formula:

$$d(X, k) = \frac{1}{2}(D[i][k] + D[j][k]), \quad k \neq i, j. \quad (2.2)$$

After $n-3$ steps, we obtain an unrooted phylogenetic tree whose branch lengths are calculated at each step by using the following equations:

$$L_i = \frac{1}{2}D[i][j] + \frac{1}{2(n-2)}(P - Q), \quad L_j = \frac{1}{2}D[i][j] - \frac{1}{2(n-2)}(P - Q), \quad (2.3)$$

where

$$P = \sum_{1 \leq k \leq n, k \neq i,j} D[i][k], \text{ and } Q = \sum_{1 \leq k \leq n, k \neq i,j} D[j][k].$$

We adapt this algorithm to the case of hybridization networks. Note that in our model hybridization events may occur between terminal branches such as shown in Fig. 2.4 as well as between ancestral branches such as shown in Fig. 2.5. In both

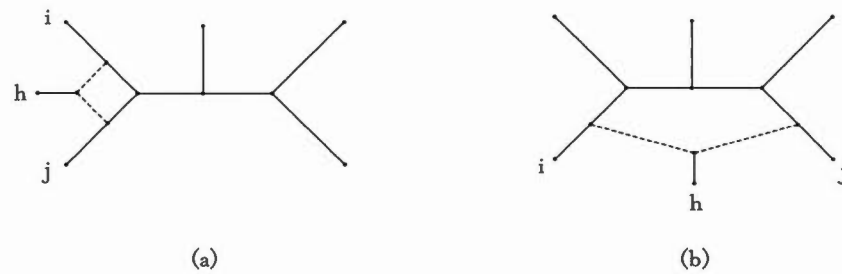


Figure 2.4 Hybrids between terminal branches.

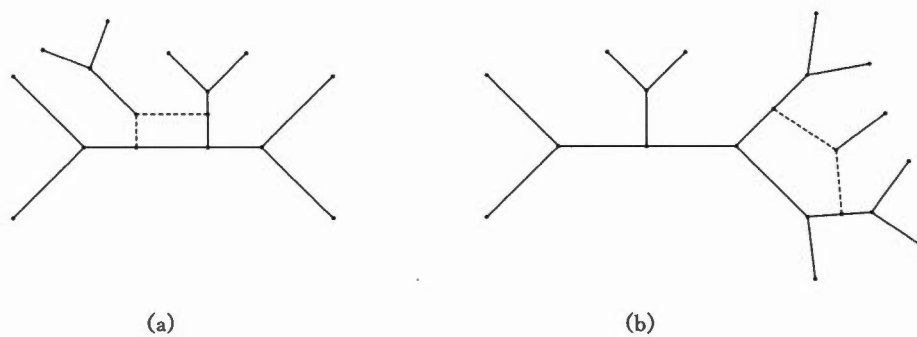


Figure 2.5 Hybrids between ancestral branches.

cases, the two parent branches may (or may not) have a direct common ancestor (see the difference between networks (a) and (b) in both figures). In all figures, hybridization (i.e., reticulation) branches are depicted by dashed lines.

2.5 Some properties of hybridization networks

In this section, we consider hybrids between terminal branches. We describe how distances between species are defined in such a network.

2.5.1 Hybrids between neighbor parents

We take the network (a) in Fig. 2.4 as an example where species h is the hybrid of species i and j . We denote by X the common ancestor of species i and j (see Fig. 2.6). Obviously, if we remove species h , we obtain a traditional additive tree. For hybrid species h , we define a real value α between 0 and 1, which is the proportion of the hybrid's genetic inheritance coming from species i (see Fig. 2.6). We also need to know lengths L_i^0 , L_j^0 and L_h shown in Fig. 2.6. The dashed reticulation branches labeled α and $1 - \alpha$ have branch lengths equal to 0. The distances between hybrid h and other species in the network are defined as follows:

$$D[i][h] = L_h + \alpha(L_i - L_i^0) + (1 - \alpha)(L_j^0 + L_i), \text{ and then:}$$

$$D[i][h] = L_h + L_i - \alpha L_i^0 + (1 - \alpha)L_j^0, \quad (2.4)$$

$$D[j][h] = L_h + (1 - \alpha)(L_j - L_j^0) + \alpha(L_i^0 + L_j), \text{ and then:}$$

$$D[j][h] = L_h + L_j + \alpha L_i^0 - (1 - \alpha)L_j^0, \quad (2.5)$$

$$D[k][h] = L_h + \alpha L_i^0 + (1 - \alpha)L_j^0 + d(X, k), \quad (2.6)$$

for any species k different from i, j and h , where the distance $d(X, k)$ between nodes X and k are computed as in a traditional additive tree.

Since the only terms containing the hybridization parameter α , L_i^0 or L_j^0 are αL_i^0 and $(1 - \alpha)L_j^0$, α is not uniquely defined. We can increase α and L_j^0 and decrease L_i^0 in such a way that we obtain exactly the same distances between species. It is no longer the case for hybrids between non-neighbor parents.

2.5.2 Hybrids between non-neighbor parents

We take the network (b) in Fig. 2.4 as an example where h is the hybrid of species i and j . We denote by X (respectively Y) the closest ancestor of species

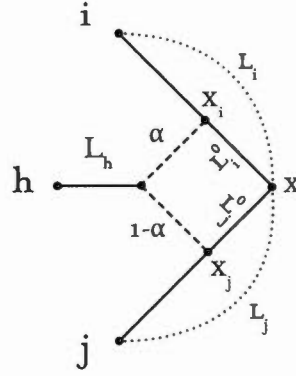


Figure 2.6 Network configuration in which species h is a hybrid of two neighbor species i and j .

i (respectively j). We use the notation indicated in Fig. 2.7. Thus, the distances between hybrid h and other species in the network are defined as follows:

$$D[i][h] = L_h + \alpha(L_i - L_i^0) + (1 - \alpha)(L_j^0 + d(Y, X) + L_i), \text{ and then:}$$

$$D[i][h] = L_h + L_i - \alpha L_i^0 + (1 - \alpha)(L_j^0 + d(Y, X)), \quad (2.7)$$

$$D[j][h] = L_h + (1 - \alpha)(L_j - L_j^0) + \alpha(L_i^0 + d(Y, X) + L_j), \text{ and then:}$$

$$D[j][h] = L_h + L_j + \alpha(L_i^0 + d(Y, X)) - (1 - \alpha)L_j^0, \quad (2.8)$$

$$D[k][h] = L_h + \alpha(L_i^0 + d(X, k)) + (1 - \alpha)(L_j^0 + d(Y, k)), \quad (2.9)$$

for all species k different from i, j and h , where the distances $d(X, k)$, $d(Y, X)$ and $d(Y, k)$ are computed as in a traditional additive tree. If we set $Y = X$ in these equations, we obtain the equations for a hybrid between neighbors. As we will show later, if species i and j are not neighbors, α is uniquely defined and can be calculated directly from the distance matrix.

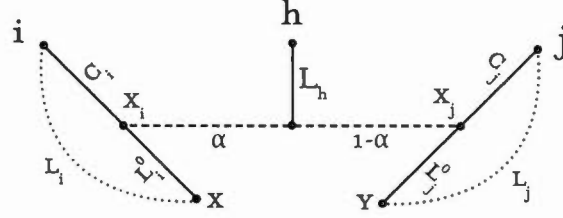


Figure 2.7 Network configuration in which species h is a hybrid of two non-neighbor species i and j .

2.5.3 Two important properties

In this section, we will describe two very important properties of hybridization networks that will be used in our algorithm.

Proposition 2.1. *If species h is the hybrid of species i and j , then for all species k different from species i, j and h :*

$$\begin{aligned} D[i][j] + D[k][h] - D[i][h] - D[k][j] &> 0, \\ D[i][j] + D[k][h] - D[j][h] - D[k][i] &> 0. \end{aligned} \quad (2.10)$$

Moreover, if species i and j are neighbors, then for all species k different from species i, j and h :

$$\begin{aligned} D[i][j] + D[k][h] - D[i][h] - D[k][j] &= 2\alpha L_i^0, \\ D[i][j] + D[k][h] - D[j][h] - D[k][i] &= 2(1 - \alpha) L_j^0. \end{aligned} \quad (2.11)$$

Proof. Let k be a species different from species i, j and h . Using Eqs. 2.7 and 2.9, we obtain:

$$\begin{aligned} D[i][j] + D[k][h] - D[i][h] - D[k][j] &= D[i][j] + L_h + \alpha(L_i^0 + d(X, k)) \\ &+ (1 - \alpha)(L_j^0 + d(Y, k)) - (L_h + L_i - \alpha L_i^0 + (1 - \alpha)(L_j^0 + d(Y, X))) - D[k][j]. \end{aligned}$$

If we replace $D[i][j]$ by $L_i + d(Y, X) + L_j$ and $D[k][j]$ by $d(Y, k) + L_j$, we obtain:

$$\begin{aligned} D[i][j] + D[k][h] - D[i][h] - D[k][j] &= L_i + d(Y, X) + L_j + L_h + \alpha(L_i^0 + d(X, k)) \\ &\quad + (1 - \alpha)(L_j^0 + d(Y, k)) - (L_h + L_i - \alpha L_i^0 + (1 - \alpha)(L_j^0 + d(Y, X))) \\ &\quad - (d(Y, k) + L_j) = \alpha (d(Y, X) + d(X, k) - d(Y, k) + 2L_i^0) > 0. \end{aligned}$$

Indeed, $d(Y, X) + d(X, k) - d(Y, k)$ is non-negative according to the triangle inequality.

In the same way, we have:

$$\begin{aligned} D[i][j] + D[k][h] - D[j][h] - D[k][i] \\ = (1 - \alpha) (d(Y, X) + d(Y, k) - d(X, k) + 2L_j^0) > 0. \end{aligned}$$

Moreover, if species i and j are neighbors, $X = Y$, then we find Eqs. 2.11 by replacing X by Y in the formulas above. $\square$

Définition 2.4. For all triplets of species i , j and h , we define $\text{MIN}_{i,j}^h$ as the minimum of all values $D[i][j] + D[k][h] - D[i][h] - D[k][j]$ and $D[i][j] + D[k][h] - D[j][h] - D[k][i]$, for all species k different from species i , j and h .

It is worth noting that in an additive tree (without a hybrid), for all triplets of species i , j , h , at least one of the values $D[i][j] + D[k][h] - D[i][h] - D[k][j]$ or $D[i][j] + D[k][h] - D[j][h] - D[k][i]$ is non-positive according to the four point condition. Then all the values $\text{MIN}_{i,j}^h$ are non-positive.

In the same way, we can prove the following proposition:

Proposition 2.2. If species h is the hybrid of species i_1 and j , and if $i_2 \neq j$ is the direct neighbor of i_1 (see Fig. 2.8), then for all species k different from i_1, i_2 ,

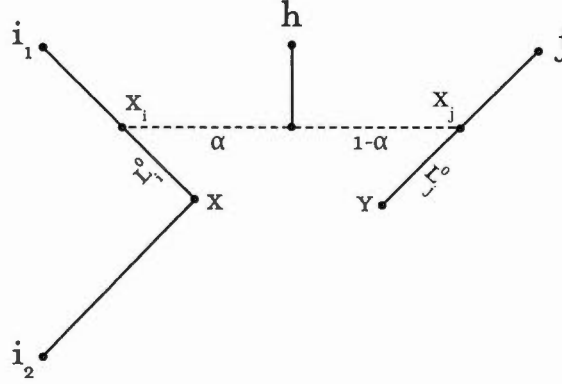


Figure 2.8 Hybrid h whose parent i_1 has a direct neighbor i_2 .

h and j :

$$\begin{aligned} D[i_2][h] + D[i_1][k] - D[k][h] - D[i_1][i_2] &> 0, \\ D[i_2][h] + D[i_1][k] - D[i_1][h] - D[k][i_2] &> 0, \end{aligned} \quad (2.12)$$

and also $\text{MIN}_{i_2, h}^{i_1} > 0$.

2.6 Identification of hybrids

If species h is the hybrid of species i and j in an additive network, we have the following system of equations, where we use the length C_i (respectively C_j) between species i and X_i (respectively j and X_j), as shown in Fig. 2.7:

$$D[i][h] = L_h + \alpha C_i - (1 - \alpha) C_j + (1 - \alpha) D[i][j], \quad (2.13)$$

$$D[j][h] = L_h - \alpha C_i + (1 - \alpha) C_j + \alpha D[i][j], \quad (2.14)$$

$$D[k][h] = L_h - \alpha C_i - (1 - \alpha) C_j + \alpha D[i][k] + (1 - \alpha) D[j][k], \quad (2.15)$$

for all species k different from i, j and h .

In a general network, we computed the least square solutions to this system of

equations, and we found the following formulas:

$$L_h = \frac{1}{2}(D[i][h] + D[j][h] - D[i][j]), \quad (2.16)$$

$$\begin{aligned} \alpha C_i = -\frac{1}{2(n-3)} \left(\sum_{k' \neq i, j, h} (D[k'][h] - \alpha D[k'][i] - (1-\alpha)D[k'][j]) \right) \\ + \frac{1}{2} (D[i][h] - (1-\alpha)D[i][j]), \end{aligned} \quad (2.17)$$

$$\begin{aligned} (1-\alpha)C_j = -\frac{1}{2(n-3)} \left(\sum_{k' \neq i, j, h} (D[k'][h] - \alpha D[k'][i] - (1-\alpha)D[k'][j]) \right) \\ + \frac{1}{2} (D[j][h] - \alpha D[i][j]). \end{aligned} \quad (2.18)$$

If we set $A = -\frac{1}{2(n-3)} \left(\sum_{k' \neq i, j, h} (D[k'][h] - \alpha D[k'][i] - (1-\alpha)D[k'][j]) \right)$, we obtain:

$$\begin{aligned} \alpha C_i &= A + \frac{1}{2} (D[i][h] - (1-\alpha)D[i][j]), \\ (1-\alpha)C_j &= A + \frac{1}{2} (D[j][h] - \alpha D[i][j]). \end{aligned}$$

If we replace L_h , αC_i and $(1-\alpha)C_j$ by these formulas in Eqs. 2.13 to 2.15, we obtain:

$$\begin{aligned} D[i][h] &= \frac{1}{2}(D[i][h] + D[j][h] - D[i][j]) \\ &+ \frac{1}{2} (D[i][h] - (1-\alpha)D[i][j] - D[j][h] + \alpha D[i][j]) + (1-\alpha)D[i][j], \\ D[j][h] &= \frac{1}{2}(D[i][h] + D[j][h] - D[i][j]) \\ &- \frac{1}{2} (D[i][h] - (1-\alpha)D[i][j] - D[j][h] + \alpha D[i][j]) + \alpha D[i][j], \\ D[k][h] &= \frac{1}{2}(D[i][h] + D[j][h] - D[i][j]) - 2A \\ &- \frac{1}{2} (D[i][h] - (1-\alpha)D[i][j] + D[j][h] - \alpha D[i][j]) + \alpha D[i][k] + (1-\alpha)D[j][k]. \end{aligned}$$

After simplification, we obtain twice $0 = 0$, and

$$\begin{aligned}
D[k][h] &= -2A + \alpha D[i][k] + (1 - \alpha) D[j][k]. \\
D[k][h] &= \frac{1}{n-3} \left(\sum_{k' \neq i, j, h} (D[k'][h] - \alpha D[k'][i] - (1 - \alpha) D[k'][j]) \right) \\
&\quad + \alpha D[i][k] + (1 - \alpha) D[j][k]. \\
D[k][h] &= \frac{1}{n-3} \left(\sum_{k' \neq i, j, h} (D[k'][h] - D[k'][j]) \right) \\
&\quad + \frac{\alpha}{n-3} \left(\sum_{k' \neq i, j, h} (D[k'][j] - D[k'][i]) \right) + \alpha D[i][k] + (1 - \alpha) D[j][k].
\end{aligned}$$

If we set

$$S_h = \frac{\sum_{k' \neq i, j, h} D[k'][h]}{n-3}, S_i = \frac{\sum_{k' \neq i, j, h} D[k'][i]}{n-3}, S_j = \frac{\sum_{k' \neq i, j, h} D[k'][j]}{n-3}, \quad (2.19)$$

we obtain:

$$D[k][h] = S_h - S_j + \alpha (S_j - S_i) + \alpha D[i][k] + (1 - \alpha) D[j][k], \text{ and then:}$$

$$D[k][h] - D[j][k] = S_h - S_j + \alpha (S_j - S_i + D[i][k] - D[j][k]).$$

Now, we have to find an optimal value of α allowing us to minimize the following function:

$$\sum_{k \neq i, j, h} (Y_k - S_h + S_j - \alpha X_k)^2,$$

where

$$Y_k = D[k][h] - D[j][k], \quad X_k = S_j - S_i + D[i][k] - D[j][k]. \quad (2.20)$$

If we differentiate according to α , we obtain:

$$\sum_{k \neq i, j, h} X_k (Y_k - S_h + S_j - \alpha X_k) = 0,$$

and then:

$$\alpha^* = \frac{\sum_{k \neq i, j, h} X_k (Y_k - S_h + S_j)}{\sum_{k \neq i, j, h} X_k X_k}.$$

Définition 2.5. For all triplets of species i, j, h , we define the degree of hybridization, $\alpha_{i,j}^h$, of h as a hybrid of i and j , by the following formula:

$$\alpha_{i,j}^h = \frac{\sum_{k \neq i,j,h} X_k (Y_k - S_h + S_j)}{\sum_{k \neq i,j,h} X_k X_k},$$

where S_h, S_j, X_k and Y_k are defined by Eqs. 2.19 and 2.20.

If $\sum_{k \neq i,j,h} X_k X_k = 0$ (which is the case when h is the hybrid of neighbor species i and j in an additive network), the optimal value of α cannot be determined and we set $\alpha_{i,j}^h = 0.5$.

We also define $L_{i,j}^h$ by the following formula:

$$L_{i,j}^h = \frac{\sum_{k \neq i,j,h} (Y_k - S_h + S_j - \alpha_{i,j}^h X_k)^2}{(n-3)}.$$

Remarque 2.1. The closer $L_{i,j}^h$ is to 0, the more likely species h is the hybrid of species i and j .

Our strategy to identify hybrids is the following. First, we determine the couple (i_1, i_2) , which minimizes $S_{i,j}$ according to the NJ classical criterion.

Then, we identify the species h that is the most likely to be a hybrid between i_1 or i_2 and any other species j . We notice that in an additive tree, if i_1 and i_2 are true neighbors, we have the following equations:

$$D[i_1][i_2] = L_{i_1} + L_{i_2},$$

$$D[i_1][k] - D[i_2][k] = L_{i_1} - L_{i_2},$$

for all species k different from i_1 and i_2 . Then, we obtain the following equations:

$$D[i_2][k] + D[i_1][k'] - D[i_2][k'] - D[i_1][k] = 0,$$

for all species k and k' different from i_1 and i_2 . However, if species h is the hybrid of species i_1 (respectively i_2) and any species j , according to Prop. 2.2, we have:

$$D[i_2][h] + D[i_1][k'] - D[i_2][k'] - D[i_1][h] > 0,$$

$$(D[i_2][h] + D[i_1][k'] - D[i_2][k'] - D[i_1][h] < 0, \text{ respectively}),$$

for all species k' different from i_1 , i_2 and h . Then, we define Σ_{i_1, i_2}^h as follows:

$$\Sigma_{i_1, i_2}^h = \frac{\sum_{k \neq i_1, i_2} (D[i_2][h] + D[i_1][k] - D[i_2][k] - D[i_1][h])}{n - 2},$$

and we choose the species h_H that maximizes the absolute value of Σ_{i_1, i_2}^h .

If $\Sigma_{i_1, i_2}^{h_H} > 0$ ($\Sigma_{i_1, i_2}^{h_H} < 0$, respectively), we consider i_1 (respectively i_2) as one of the two parents of h_H . We set $i_H = i_1$ (respectively $i_H = i_2$).

Then, we need to determine the second parent of h_H . We compute the smallest value $L_{i_H, j_H}^{h_H}$, of all possible values $L_{i_H, j}^{h_H}$ for all species j such that: $\text{MIN}_{i_H, j}^{h_H} > 0$ and $\alpha_{min} < \alpha_{i_H, j}^{h_H} < \alpha_{max}$, where the thresholds α_{min} and α_{max} ($0 < \alpha_{min} < \alpha_{max} < 1$) are some parameters which can be suggested by the user.

If $L_{i_H, j_H}^{h_H} < (\Sigma_{i_1, i_2}^{h_H})^2$, then h_H is identified as the hybrid of species i_H and j_H . The detailed scheme of our algorithm is given below.

2.7 Algorithm for inferring hybridization networks

In this section, we introduce a new algorithm for inferring hybridization networks based on the neighbor-joining principle. This algorithm takes as input a distance matrix $D = \{D[i][j]\}_{1 \leq i \leq n; 1 \leq j \leq n}$ on a set of n species and two real values α_{min} and α_{max} such that $0 < \alpha_{min} < \alpha_{max} < 1$.

ALGORITHM

- $n_A = n$
- $D_A = D$
- **While** ($n_A > 4$)
 1. We determine the couple (i_1, i_2) , that minimizes $S_{i;j}$.
 2. We choose the species h_H that maximizes the absolute value of Σ_{i_1, i_2}^h .
 3. **If** $\Sigma_{i_1, i_2}^{h_H} > 0$, **then** $i_H = i_1$. **Else if** $\Sigma_{i_1, i_2}^{h_H} < 0$, **then** $i_H = i_2$.
 4. We compute the smallest value $L_{i_H, j_H}^{h_H}$ of all values $L_{i_H, j}^{h_H}$ for all species j such that: $\text{MIN}_{i_H, j}^{h_H} > 0$ and $\alpha_{\min} < \alpha_{i_H, j}^{h_H} < \alpha_{\max}$.
 5. **If** ($L_{i_H, j_H}^{h_H} \leq (\Sigma_{i_1, i_2}^{h_H})^2$), **then** h_H is identified as the hybrid of i_H and j_H .
We remove from D_A the row and the column corresponding to h_H . We keep in memory the length

$$L_{h_H}^H = \frac{1}{2} (D[i_H][h_H] + D[j_H][h_H] - D[i_H][j_H]), \quad (2.21)$$
 6. **Else**, species i_1 and i_2 are considered as neighbors. We replace rows and columns corresponding to i_1 and i_2 in D_A by a row and a column corresponding to their ancestor X . Distances from X to remaining leaves are calculated using Eq. 2.2. We keep in memory the lengths L_{i_1} and L_{i_2} given by Eq. 2.3.
 7. $n_A \leftarrow n_A - 1$
- **End(While)**
- At the end of this loop, we have one quadruplet of nodes remaining. We use the standard NJ algorithm to determine the tree structure involving these 4 species.

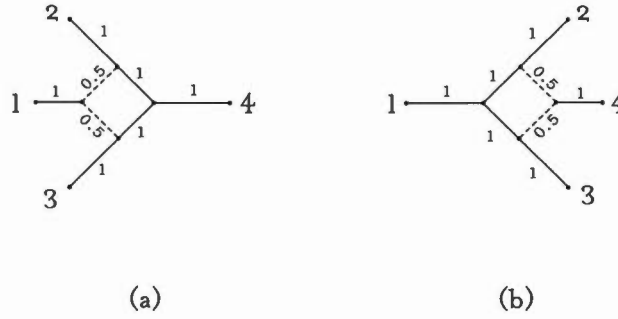


Figure 2.9 Two networks corresponding to the same distance matrix of size 4.

The output of our algorithm is either a classical phylogenetic tree with n leaves, or a hybridization network with the same n terminal nodes. Its time complexity is $O(n^3)$ as in the standard NJ algorithm.

Remarque 2.2. *We don't include the iteration $n_A = 4$ in the loop since in this case the hybridization network corresponding to a distance matrix is not unique, as it is illustrated by the example of Fig. 2.9, where networks (a) and (b) correspond*

to the same distance matrix $D = \begin{pmatrix} 0 & 3 & 3 & 3 \\ 3 & 0 & 4 & 3 \\ 3 & 4 & 0 & 3 \\ 3 & 3 & 3 & 0 \end{pmatrix}$. In these networks, $\alpha = 0.5$

and all branch lengths are equal to 1. The length of the dashed branches is equal to 0.

This algorithm has been implemented in the C++ programming language and a Monte Carlo simulation study was carried out to assess its performances.

2.8 Results of simulations

In our simulations, we consider as a true positive any hybrid that is identified as a hybrid, even if its parents are not correctly identified. We give more precise results about the identification of hybrids' parents in some particular cases. We consider as a false positive any identified hybrid that is not a true hybrid, even if one or two of the identified parents are true hybrids. The true positive rate is computed as the number of true positives divided by the number of true hybrids. The false positive rate is computed as the number of false positives divided by the number of non-hybrid species in the tree.

In our simulations, we set $\alpha_{min} = 1 - \alpha_{max}$. Then, we defined $DIFFMAX = \alpha_{max} - 0.5$.

2.8.1 Simulations with additive networks

A theoretical result for trees

Proposition 2.3. *If the input distance matrix is a tree metric, our algorithm finds the exact phylogenetic tree corresponding to this matrix.*

Proof. If the input distance matrix is a tree metric, then for all triplets of species we have $\text{MIN}_{i,j}^h \leq 0$. Consequently, our algorithm does not find any hybrid, and the identification of neighbors is exactly the same as in the standard NJ algorithm. Thus, the correct phylogenetic tree is recovered (Studier & Keppler, 1988).

□

Simulation with hybrids between terminal branches

In our simulations, we used an algorithm generating random phylogenetic trees available on the T-REX website (Boc et al., 2012). This algorithm takes as input the size of the tree n and the average branch length $\bar{l}$, and gives as output a random binary phylogenetic tree with n leaves constructed according to the procedure described by Kuhner & Felsenstein (1994). In this way, we generated 1000 unrooted trees for each considered tree size $n = 8$, $n = 16$, $n = 32$ and $n = 64$, with $\bar{l} = 0.1$. Then, the following procedure was used to add a hybrid: we randomly selected two integers $1 \leq i < j \leq n$, two real values β and γ between 0 and 1, and a real value x from an exponential distribution with mean 0.1. We added a species $h = n + 1$ by using Eqs. 2.13 to 2.15 with $C_i = \beta L_i$, $C_j = \gamma L_j$ and $L_h = x$. We added from 1 to 5 hybrids to each of the trees. We carried out three series of simulations for $\alpha = 0.3$, $\alpha = 0.4$ and $\alpha = 0.5$, with $DIFFMAX = 0.25$. We computed the average true positive and false positive rates for all sizes of trees and all numbers of hybrids. These results are shown in Fig. 2.10.

The best results were obtained for greater values of n , for smaller numbers of hybrids, and for the values of α close to 0.5. We can also observe that the false positive rate is very close to 0, and that the true positive rates are close to 100% for $n = 64$. Most false negatives are hybrids between neighbors or between very close parents. That is the reason why the identification of hybrids is more complicated for smaller values of n . In the case of neighbor parents, the following issue can appear. If species h is the hybrid of neighbor species i and j , the NJ algorithm can identify i and h (or j and h) as neighbors. Thus, species h is identified as a parent of a hybrid and not as a hybrid.

It is worth mentioning that all true positives in this simulation were identified with both correct parents and with the correct value of α .

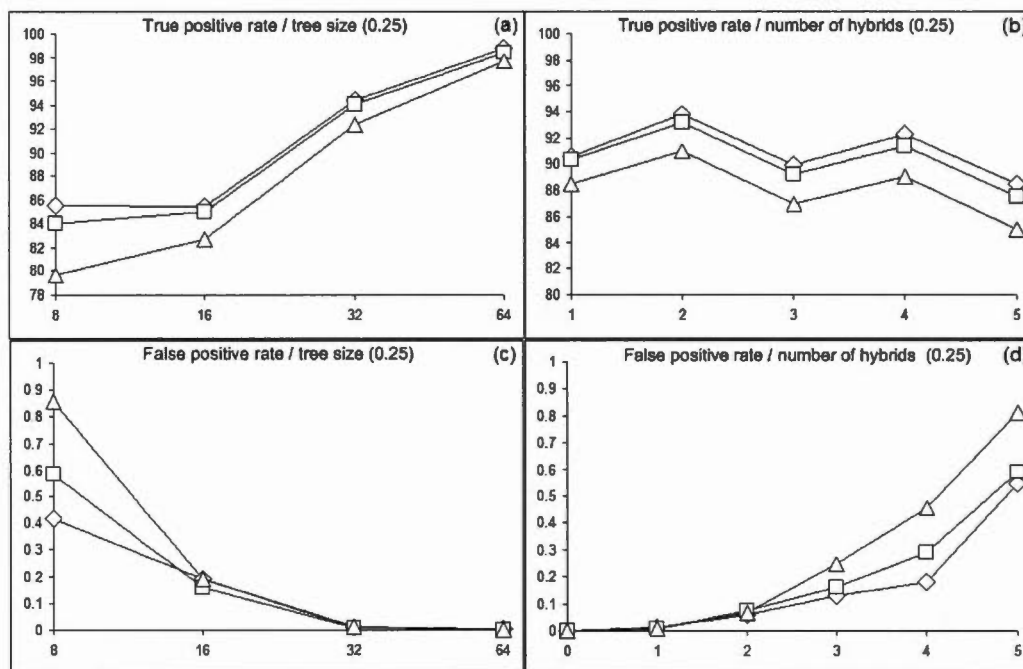


Figure 2.10 Average simulation results for additive networks with hybridization level $\alpha = 0.3$ (Δ), $\alpha = 0.4$ ($\square$) and $\alpha = 0.5$ ($\diamond$), and with $DIFFMAX = 0.25$. Figure (a) (respectively (c)) shows the true (respectively false) positive rate as a function of the tree size. Figure (b) (respectively (d)) shows the true (respectively false) positive rate as a function of the number of hybrids.

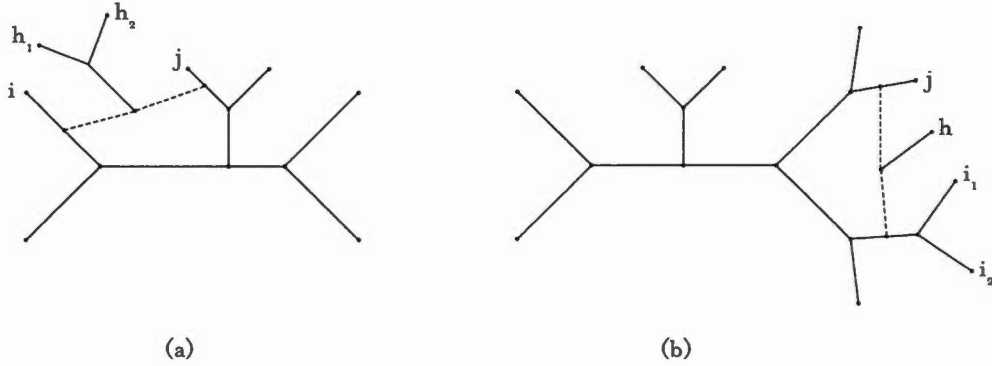


Figure 2.11 Two network topologies used in our simulations with additive networks.

Simulation with hybrids having two descendants

We also carried out a simulation with a hybrid having two descendants as in the configuration (a) in Fig. 2.11. The obtained true positive and false positive rates were very similar to those shown in Fig. 2.10. However, the detection of correct parents was not systematic. Table 2.1 reports the identification of hybrids' parents for $\alpha = 0.5$. If the algorithm identifies a set of at least two species as a hybrid, we can solve this problem by applying the algorithm once more, replacing this set of species by their common ancestor. In this case, we will find both correct parents like in the previous simulation.

Simulation with hybrids between non-terminal branches

We also run a simulation involving a hybrid with one parent having two descendants as shown in Fig. 2.11 (b). The true positive rates in this simulation were lower than in the two previous simulations, as it is shown in Tab. 2.2 for $n = 32$ and $\alpha = 0.5$. These results are due to the fact that the condition $\text{MIN}_{i_H, j}^{h_H} > 0$ at

Tree size	$n = 8$	$n = 16$	$n = 32$	$n = 64$
True positives with both correct parents	75%	75%	69%	49%
True positives with only one correct parent	3%	11%	25%	49%
True positives with no correct parents	0%	0%	2%	0%

Table 2.1 Identification of hybrids' parents for additive networks with one hybrid having two descendants.

Tree size	$n = 8$	$n = 16$	$n = 32$	$n = 64$
Hybrids between terminal branches	81%	85%	98%	99%
Hybrids with one parent having two descendants	52%	64%	66%	79%

Table 2.2 True positive rates for $n = 32$ and one hybrid with hybridization level $\alpha = 0.5$ in additive networks.

step 4 of our algorithm does not hold for hybrids between non-terminal branches. Then, hybrid h is detected only if species i_1 and i_2 are joined before possible detection of h . If we replace the condition $\text{MIN}_{i_H,j}^{h_H} > 0$ by $\text{MIN}_{i_H,j}^{h_H} > -0.01$, for example, we will obtain almost the same true positive rates as in the two previous simulations. However, the false positive rate will be higher.

2.8.2 Simulations with non-additive networks (i.e., with sequence-based networks)

As previously, we generated 1000 unrooted trees for each size $n = 8$, $n = 16$, $n = 32$ and $n = 64$, with $\bar{l} = 0.1$. Then, using Seq-Gen (Rambaut & Grass, 1997) we simulated the evolution of nucleotide sequences of length $N = 1000$ along these trees, by using the Kimura-2-parameter substitution model (Kimura, 1980). Thus, for each generated tree we obtained n sequences (one sequence per species) of size N . Then, the hybrids were added to the data as follows. We randomly chose two

integers $1 \leq i < j \leq n$. Let α be the selected degree of hybridation. We created a new hybrid sequence with the first $\alpha \times N$ nucleotides of sequence i to which we added the last $(1 - \alpha) \times N$ nucleotides of sequence j . This sequence was added to the n original sequences. In our simulations, we considered $\alpha = 0.3$, $\alpha = 0.4$ and $\alpha = 0.5$, and we added to trees 0 to 5 hybrid species. Then, we used the Phylip package (Felsenstein, 1989) to obtain a distance matrix from each set of sequences, by using the Kimura-2-parameter substitution model. Thus, for each size n , we obtained 1000 matrices corresponding to the original trees and 15000 matrices corresponding to networks having 1 to 5 hybrids with 3 different values of α . The obtained results are shown in Fig. 2.12 for $DIFFMAX = 0.25$ and in Fig. 2.13 for $DIFFMAX = 0.35$. We used two different values of $DIFFMAX$ because there was a significant difference in the true positive rates in this simulation (this difference was much smaller in the simulations with additive data).

The greatest true positive rates and the lowest false positive rates were provided by the new algorithm for $n = 32$ and $n = 64$. The identification of hybrids was much more difficult for $\alpha = 0.3$, even though the results are much better for $DIFFMAX = 0.35$.

Table 2.3 reports the results concerning the identification of hybrids' parents for $\alpha = 0.5$. We can observe that the identification of both parents is more difficult for greater values of n .

Table 2.4 reports the number of iterations after which each hybrid was found in trees and in networks with one hybrid and with hybridization level $\alpha = 0.5$. Mention that true positives are generally detected after a much smaller number of iterations than false positives. Thus, the number of iterations could be an interesting criterion to consider to distinguish between true positives and false positives.

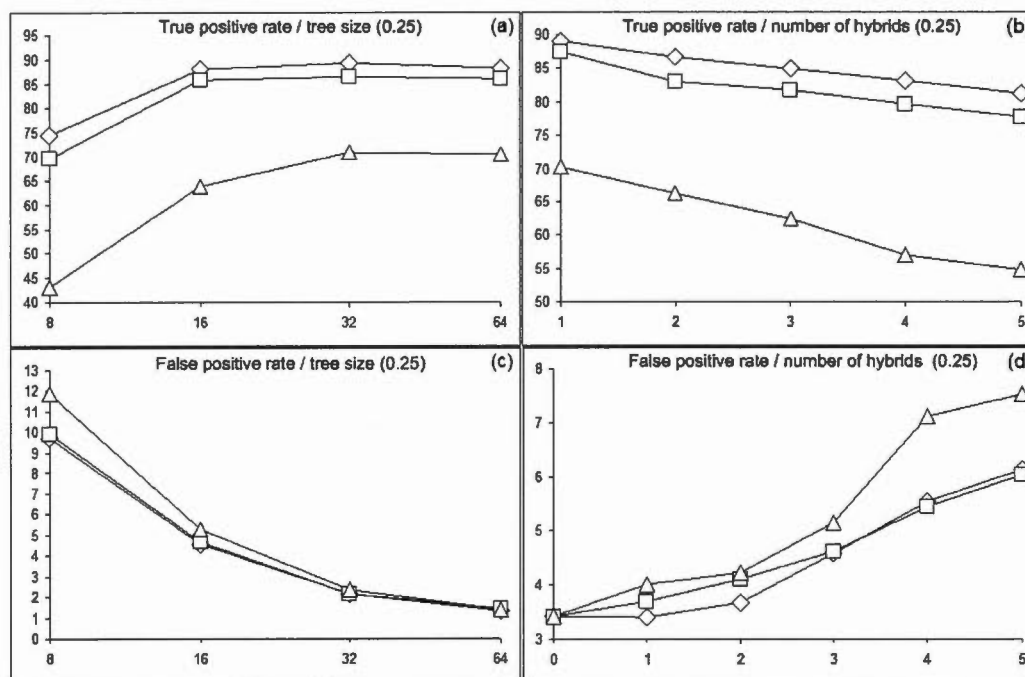


Figure 2.12 Average simulation results for non-additive networks with hybridization level $\alpha = 0.3$ (Δ), $\alpha = 0.4$ ($\square$) and $\alpha = 0.5$ ($\diamond$), and with *DIFFMAX* = 0.25. Figure (a) (respectively (c)) shows the true (respectively false) positive rate as a function of the tree size. Figure (b) (respectively (d)) shows the true (respectively false) positive rate as a function of the number of hybrids.

Tree size	$n = 8$	$n = 16$	$n = 32$	$n = 64$
True positives with both correct parents	86%	88%	77%	53%
True positives with only one correct parent	1%	3%	11%	30%
True positives with no correct parents	0%	0%	2%	6%

Table 2.3 Identification of hybrids' parents in non-additive networks with one hybrid and hybridization level $\alpha = 0.5$.

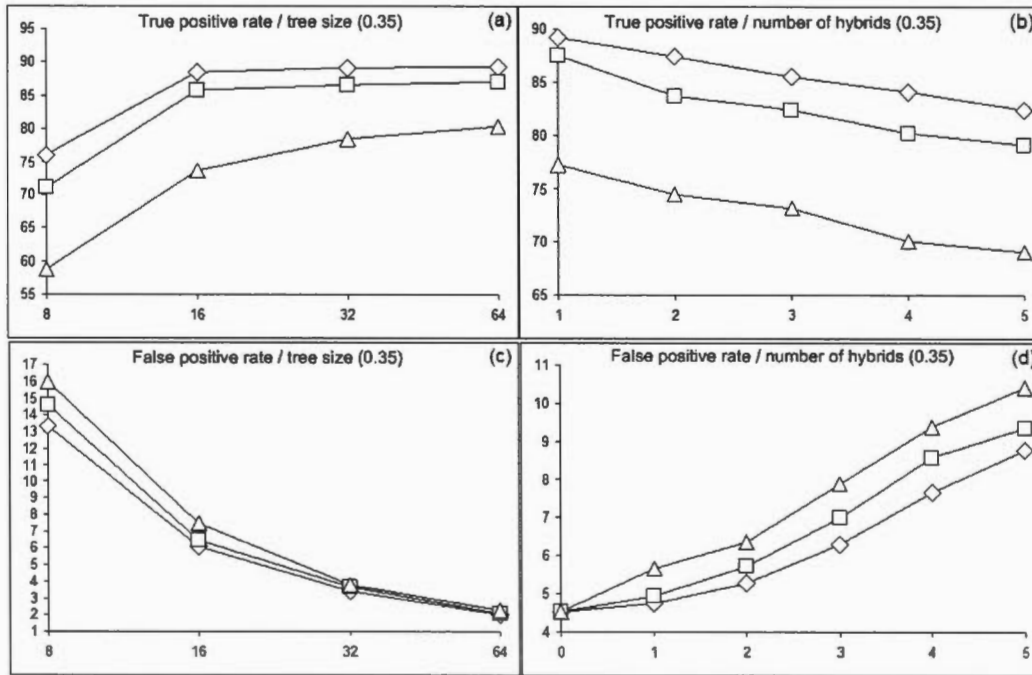


Figure 2.13 Average simulation results for non-additive networks with hybridization level $\alpha = 0.3$ (Δ), $\alpha = 0.4$ ($\square$) and $\alpha = 0.5$ ($\diamond$), and with $DIFFMAX = 0.35$. Figure (a) (respectively (c)) shows the true (respectively false) positive rate as a function of the tree size. Figure (b) (respectively (d)) shows the true (respectively false) positive rate as a function of the number of hybrids.

Tree size	True positives with both parents	False positives
$n = 8$	$\mu = 1.5, \sigma = 1.4$	$\mu = 2.3, \sigma = 0.9$
$n = 16$	$\mu = 4.1, \sigma = 3.0$	$\mu = 8.9, \sigma = 2.6$
$n = 32$	$\mu = 10.0, \sigma = 7.6$	$\mu = 22.1, \sigma = 6.7$
$n = 64$	$\mu = 22.7, \sigma = 13.7$	$\mu = 50.7, \sigma = 13.7$

Table 2.4 Average number of iterations μ (and the corresponding standard deviation σ) after which hybrids were detected in networks with one hybrid and hybridization level $\alpha = 0.5$.

Notice that the values of α determined by our algorithm were very close to the simulated values. For example, for the case of $n = 32$ and one hybrid, we found $\mu_\alpha = 0.493$ and $\sigma_\alpha = 0.034$ for $\alpha = 0.5$, $\mu_\alpha = 0.398$ and $\sigma_\alpha = 0.034$ for $\alpha = 0.4$, $\mu_\alpha = 0.302$ and $\sigma_\alpha = 0.033$ for $\alpha = 0.3$, where μ_α (respectively, σ_α) is the mean (respectively, the standard deviation) of α found for true positive hybrids with both correct parents.

2.8.3 Experiments with real data

We tested our new algorithm on real data. We considered the dataset of restriction maps of the rDNA cistron of 16 species of mosquitoes constructed using 8 recognition restriction enzymes (Kumar et al., 1998). A total of 26 sites were scored. This dataset is presented in Tab. 2.5.

Huson & Klöpper (2007) have constructed the split graph and the galled network associated to this data set (see Fig. 2.14)

We computed the Hamming distances from the sequences of Tab. 2.5 to obtain a distance matrix of size 16 between these species. Then, we applied our algorithm (with $DIFFMAX = 0.1$) to this matrix and obtained the hybridization network

<i>Aedes albopictus</i>	1 1 1 1 0 1 0 1 0 1 0 1 0 0 0 1 0 1 0 1 0 1 0 0 1 0
<i>Aedes aegypti</i>	1 1 1 1 0 1 0 1 0 0 0 1 0 0 0 1 0 1 0 1 0 0 0 0 1 0
<i>Aedes seatoi</i>	1 1 1 1 0 1 0 1 0 1 0 1 0 0 0 1 0 1 0 1 0 1 0 0 0 0
<i>Aedes flavopictus</i>	1 1 1 1 0 1 0 1 0 1 0 1 0 0 0 1 0 1 0 1 0 1 0 0 1 0
<i>Aedes alcasidi</i>	1 1 1 1 0 1 0 1 0 1 0 1 0 0 0 1 0 1 0 1 0 1 0 0 0 0
<i>Aedes katherinensis</i>	1 1 1 1 0 1 0 1 0 1 0 1 0 0 0 1 0 1 0 1 0 1 0 0 0 0
<i>Aedes polynesiensis</i>	1 1 1 1 0 1 0 1 0 0 0 1 0 0 0 1 0 1 0 1 0 1 0 0 1 0
<i>Aedes triseriatus</i>	1 0 1 1 0 1 0 1 0 0 0 1 1 0 0 1 0 1 0 1 0 0 0 0 0 0
<i>Aedes atropalpus</i>	1 0 1 1 0 1 0 1 0 0 0 1 0 0 0 1 0 1 1 1 0 0 0 0 1 0
<i>Aedes epactius</i>	1 0 1 1 0 1 0 1 0 0 0 1 0 0 0 1 0 1 1 1 0 0 0 0 1 0
<i>Haemagogus equinus</i>	1 0 1 1 0 1 0 1 0 0 0 1 1 0 0 1 0 1 0 1 0 1 0 0 0 0
<i>Armigeres subalbatus</i>	1 0 1 1 0 1 0 1 0 0 0 1 0 0 0 1 0 1 0 1 0 0 0 0 0 0
<i>Culex pipiens</i>	1 1 1 1 0 1 1 1 0 0 0 1 0 0 0 1 1 1 0 1 0 0 1 0 1 1
<i>Tripteroides bambusa</i>	1 1 1 1 0 1 1 1 0 0 0 1 0 0 0 1 0 1 0 1 0 0 0 0 1 0
<i>Sabethes cyaneus</i>	1 1 1 1 0 1 0 1 0 0 1 1 0 0 0 1 0 1 0 1 0 1 0 0 0 0
<i>Anopheles albimanus</i>	1 1 0 1 1 1 0 1 1 0 0 1 0 1 1 1 0 1 0 1 1 1 0 1 0 0

Table 2.5 Dataset of the restriction maps of the rDNA cistron of 16 species of mosquitoes constructed using 8 recognition restriction enzymes.

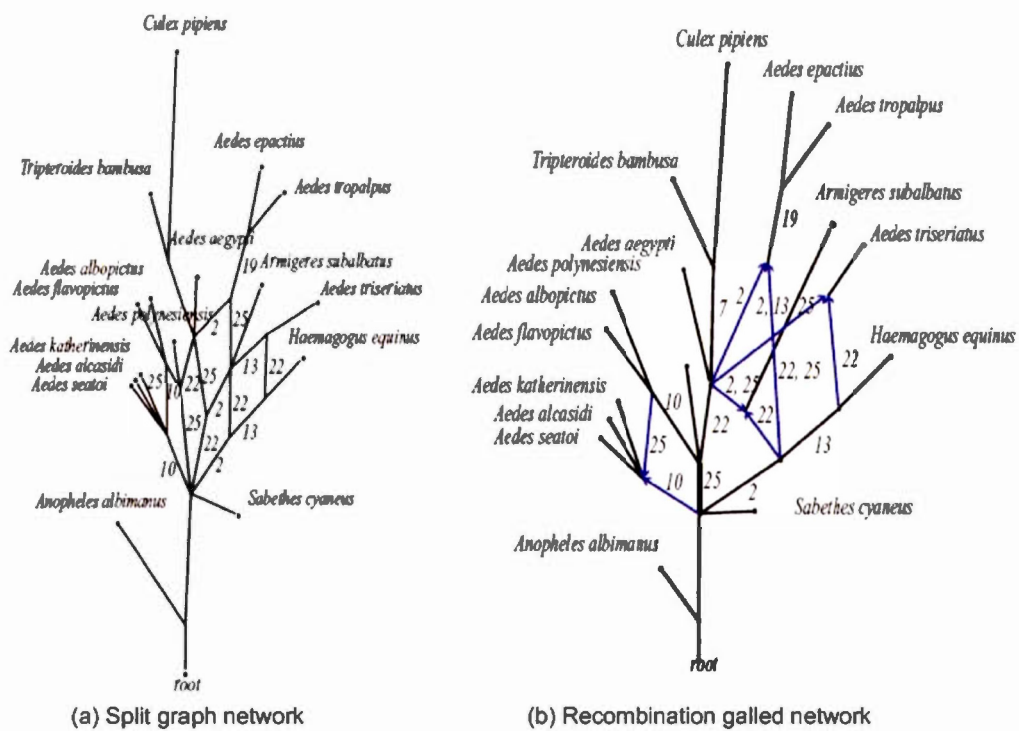


Figure 2.14 Split graph and galled network obtained for the rDNA cistron dataset in Tab. 2.5.

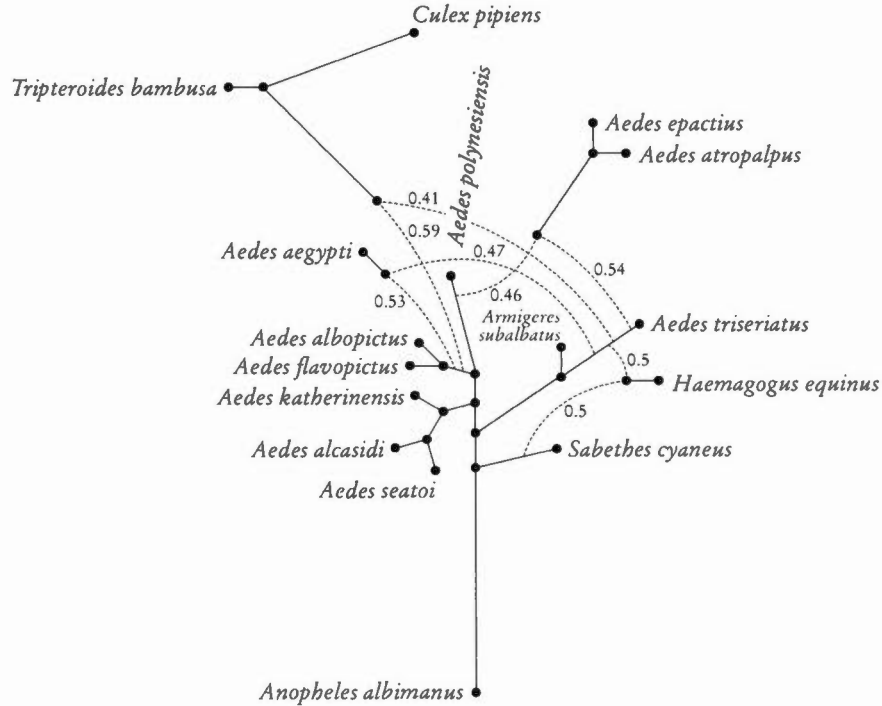


Figure 2.15 Hybridation network obtained with our new algorithm. The values of α are indicated on the reticulation branches (depicted by dashed lines).

shown in Fig. 2.15.

We obtained the same number of reticulations (4) as Huson and Klöpper, and the general structure of our network is quite similar to the split graph and galled tree topologies presented in Fig. 2.14. For example, species *Aedes epactius* and *Aedes atropalpus* are located at the extremity of a reticulation in both networks. However, some significant differences can also be observed. For example, species *Aedes triseriatus* could be considered as a hybrid in the split graph and galled tree topologies, whereas it is identified as a potential parent of 4 hybrids in our network.

In the same way, species *Sabethes cyaneus* is not involved in any reticulation in the split graph and galled tree topologies, whereas it is a parent of a hybrid in our network. The main advantage of our network representation over split graphs and galled trees is that it identifies hybrids and their parents explicitly. Moreover, our algorithm provides the exact hybridization levels α , and these levels are compatible with the sequences of hybrids and their parents (see Tab. 2.5). For example, the sequence of species *Haemagogus equinus* can be obtained by concatenation of the first half of the sequence of species *Aedes triseriatus* and the second half of the sequence of species *Sabethes cyaneus*.

It is worth noting that in this example we changed the condition $\text{MIN}_{i_H,j}^{h_H} > 0$ to $\text{MIN}_{i_H,j}^{h_H} > -0.01$ at step 4 of our algorithm. Indeed, we had $\text{MIN}_{i_H,j}^{h_H} = 0$ for all potential hybrids. This kind of adaptation could be used when the number of detected hybrids is too small (or too large, in the latter case, we should replace 0 by a small positive threshold). This threshold is one of the parameters of our program.

2.9 Conclusion

We have described a novel fast algorithm for inferring hybridization networks from distance matrices based on the neighbor-joining principle. These distance matrices, assumed to encompass contradictory evolutionary signals, could be obtained from the concatenation of genetic sequences or directly from the comparison of genomes of the observed species. The new algorithm provides a good practical solution to the complex problem of the identification of hybridization events. The algorithm's time complexity of $O(n^3)$ makes it applicable for the analysis of large genomic datasets. Moreover, the quality of the obtained results improves as the numbers of considered species grows. The new algorithm finds the exact tree solution when the input distance matrix is a tree metric (or a distance close to a

tree metric). The true positive detection rate is very high and the correct hybrids parents are always recovered for additive networks when the hybrids are located between terminal branches. We also provide a way of recovering the correct additive networks when the hybrids are located at any place in the network. The simulation study carried out with sequence data provided very good detection rates as well, even though the false positive rates were a little bit higher in this case.

The execution of our algorithm on the rDNA cistron data (Kumar et al., 1998) allowed us to infer an explicit hybridization network which was compared to the split graph (Huson & Klöpper, 2007) and galled tree (Huson & Klöpper, 2007) topologies. Mention that both split graph and galled tree algorithms infer only implicit phylogenetic networks and are not capable of determining the precise levels of hybridization.

In the future, it would be also important to investigate in more detail how the new technique copes with the tree reconstruction artifacts which generally affect phylogenetic analysis; the main of them are long-branch attraction and unequal evolutionary rates.

The software implementing the discussed algorithm was implemented in the *C++* language. It is freely available to the research community at the following URL address:

http://www.info2.uqam.ca/~makarenkov_v/makarenv/hybrids_detection.zip.

CHAPITRE III

UTILISATION DES RÉSEAUX D'HYBRIDATION POUR RETRACER L'ÉVOLUTION DES LANGUES INDO-EUROPÉENNES

RÉSUMÉ

Ce chapitre est une reproduction de l'article suivant :

Willems, M., Lord, E., Laforest, L., Labelle, G., Lapointe, F.-J., Di Sciullo, A. M. & Makarencov, V. (2016). Using hybridization networks to retrace the evolution of indo-european languages. *BMC Evolutionary Biology*, 16(1), 180.

Plusieurs parallèles entre le processus de l'évolution des espèces et l'histoire des langues ont été observés par de nombreux chercheurs. Retracer l'évolution des langues indo-européennes (IE) demeure un des problèmes fondamentaux de la linguistique historique. Dans ce contexte, la plupart des études utilisent le modèle traditionnel de l'arbre phylogénétique pour représenter l'évolution des langues naturelles, et ne prennent donc pas en compte les événements d'évolution réticulée, comme les langues hybrides et les emprunts de mots, qui peuvent être associés, respectivement, à l'hybridation d'espèces et au transfert horizontal de gènes, en biologie. Plus récemment, des réseaux implicites comme les split-graphes, ont été utilisés pour tenir compte de l'évolution réticulée en linguistique.

Le but principal de ce chapitre est d'appliquer une version adaptée de l'algorithme présenté dans le chapitre précédent pour obtenir un réseau explicite représentant l'évolution des langues IE. On utilise pour cela la base de données lexicale de Dyen, qui contient certaines traductions dans 84 langues IE de 200 mots d'une des listes constituées par Swadesh dans les années 1940-50. De plus, nous comparons nos résultats avec ceux obtenus à l'aide de split-graphes et de réseaux de niveau 1. Contrairement aux autres méthodes, notre nouvel algorithme permet d'identifier explicitement la (les) langue(s) d'origine et la langue réceptrice des emprunts de mots (ou des phénomènes d'hybridation linguistique) et de quantifier l'influence de chacune des langue d'origine sur la langue réceptrice. Nous présentons également des justifications historiques et linguistiques pour chacune des réticulations identifiées. Nos résultats pourraient être raffinés si nous pouvions analyser des données syntaxiques, phonologiques et morphologiques en plus des données lexicales.

Contributions : J'ai rédigé l'article et mené les expérimentations avec l'aide de Vladimir Makarencov. Étienne Lord s'est occupé des comparaisons entre la distance de Levenshtein et la distance SCA. Les autres auteurs ont participé à la

conception de l'étude et à la recherche de justifications linguistiques et historiques des résultats obtenus.

3.1 Abstract

3.1.1 Background

Curious parallels between the processes of species and language evolution have been observed by many researchers. Retracing the evolution of Indo-European (IE) languages remains one of the most intriguing intellectual challenges in historical linguistics. Most of the IE language studies use the traditional phylogenetic tree model to represent the evolution of natural languages, thus not taking into account reticulate evolutionary events, such as language hybridization and word borrowing which can be associated with species hybridization and horizontal gene transfer, respectively. More recently, implicit evolutionary networks, such as split graphs and minimal lateral networks, have been used to account for reticulate evolution in linguistics.

3.1.2 Results

Striking parallels existing between the evolution of species and natural languages allowed us to apply three computational biology methods for reconstruction of phylogenetic networks to model the evolution of IE languages. We show how the transfer of methods between the two disciplines can be achieved, making necessary methodological adaptations. Considering basic vocabulary data from the well-known Dyen's lexical database, which contains word forms in 84 IE languages for the meanings of a 200-meaning Swadesh list, we adapt a recently developed computational biology algorithm for building explicit hybridization networks to study the evolution of IE languages and compare our findings to the results provided by the split graph and galled network methods.

3.1.3 Conclusion

We conclude that explicit phylogenetic networks can be successfully used to identify donors and recipients of lexical material as well as the degree of influence of each donor language on the corresponding recipient languages. We show that our algorithm is well suited to detect reticulate relationships among languages, and present some historical and linguistic justification for the results obtained. Our findings could be further refined if relevant syntactic, phonological and morphological data could be analyzed along with the available lexical data.

3.2 Background

Many curious similarities between the processes of species and language evolution have been observed since Darwin's *The Descent of Man* (Darwin, 1888). But even earlier, in 1863, August Schleicher (Schleicher, 1873) sent a letter to Ernst Haeckel in which he discussed some of these similarities, comparing, for example, mixed languages to hybridized plants in botany. Atkinson & Gray (2005) presented a table that highlights the most important conceptual parallels which can be drawn between these evolutionary phenomena. In particular, the latter study compares the process of social selection in linguistics to natural selection of species, borrowing of words across languages to horizontal transfer of genes, creole languages to plant hybrids, ancient texts to fossils, and cognates to homologies. There are also a few differences between these processes (Geisler & List, 2013). For instance, the biological alphabet (e.g., DNA) is universal, whereas the set of sounds used to form words is specific to each language. Moreover, the sequence data are usually much longer in molecular biology than in linguistics, and the selection of a perfect list of basic meanings suitable for the application of phylogenetic methods in the context of language evolution remains a challenging task. Nevertheless, the similarities and parallels between the two disciplines make it possible

for researchers to use several well-developed computational biology methods for studying the evolution of species, and in particular reticulate evolution, in the field of linguistics. Obviously, it's not possible to apply these computational biology methods directly, without an appropriate adaptation, which is critical in interdisciplinary research. Thus, the existing phylogenetic algorithms should be modified and workflows adapted in order to obtain meaningful linguistic results and interpretations.

Two nucleotide sequences observed in two distinct species are said to be homologous if they have evolved from a common ancestral sequence (Fitch, 2000). Similarly, in linguistics, a group of cognates is a group of word forms in different languages that have been inherited from a common ancestral word form (Trask, 2000). The main difference between these concepts is that the concept of homology includes the possibility of lateral transfers, whereas the concept of cognacy excludes all potential processes of borrowing. Cognates and phylogenetic trees play a fundamental role when studying the evolution of natural languages using phylogenetic methods (Gray & Atkinson, 2003). For instance, a phylogenetic tree representing the main traits of lexical evolution is equivalent to a species phylogeny depicting the key speciation events (Atkinson & Gray, 2005; Gray & Atkinson, 2003; Rexová et al., 2003). Several linguistic studies used phylogenetic methods to better understand the evolution of Indo-European (IE) languages (Gray & Atkinson, 2003; Rexová et al., 2003; Bouckaert et al., 2012; Nelson-Sathi et al., 2011, 2013). Discovering the origin and main evolutionary trends characterizing the IE language family is one of the most recalcitrant intellectual challenges in historical linguistics (Gray & Atkinson, 2003; Diamond & Bellwood, 2003). Two opposing theories, Kurgan and Anatolian, concerning early Indo-European origins are generally considered (Gray & Atkinson, 2003). The Kurgan theory (Gimbutas, 1973a,b) postulates that IE languages originate from the Kurgan culture dated

around 3000 to 4000 BC, whereas the Anatolian theory (Renfrew, 1988) dates the origin of IE languages around 7000 BC. For example, the works of Gray & Atkinson (2003) and Bouckaert et al. (2012), which focus on inferring and dating the divergence times of the contemporary and extinct IE languages using Bayesian phylogenetic methods, support the Anatolian theory of IE origin.

Phylogenetic tree model widely considered in linguistics assumes that the frequency of lateral word exchanges across languages has been relatively low. For example, Gray & Atkinson (2003) and Bouckaert et al. (2012) removed known loanwords from the basic vocabulary data before inferring their IE language trees. Obviously, linguistic phenomena such as word borrowing (Nelson-Sathi et al., 2011) and birth and evolution of hybrid languages (Thomason & Kaufman, 1988), resulting from languages in contact, cannot be adequately represented by a tree model. For instance, a study of 80000 words of the old Shorter Oxford Dictionary points out that English, which is a Germanic language, has borrowed 56.5% of its total lexicon from Old French (*Langue d'oïl*) and Latin, 5.3% from Greek, 13.2% from other languages, and has inherited only 25% of its current lexicon from its direct ancestor, Old Germanic (Finkenstaedt & Wolff, 1973; Pagel, 2000). In this work, we analyzed basic vocabulary data from a 200-meaning Swadesh list (Swadesh, 1952). While the use of this list may lead to a certain decrease in the number of loanwords (Bower et al., 2011), it remains helpful for detecting the most important word borrowing trends (List et al., 2014a). For example, the traditional 200-meaning English Swadesh list includes 33 confirmed loanwords (16.5%) (Donohue et al., 2012) and 10 additional “irregular phylogenetic patterns” which might be suggestive of unrecognized borrowings (List et al., 2014a). Moreover, in a recent revision of the Albanian Swadesh list 31.8% of its entries were identified as probable borrowings (Holm, 2011).

Word borrowing can be viewed as one of the main development mechanisms lead-

ing to the emergence of hybrid (i.e., mixed or contact) languages. There exists a variety of hybrid languages, including pidgins, creoles, and lexical hybrids (Velupillai, 2015). In a pidgin, the lexicon and minimum hybridization score usually comes from one parent language and the syntax comes from another one. A creole language, which arises from a pidgin, is a stable natural language spoken as a mother tongue. There are however many other types of lexical and grammatical transmission that produce a variety of linguistic outcomes. For example, Michif, which is the language of the Métis people of Canada and the United States, combines noun phrase phonology, lexicon, syntax and morphology from Métis French and verb phrase phonology, lexicon, syntax and morphology from Cree. As our analysis is based on lexical data only, here we address the problem of detection of lexical hybrids and word borrowing events.

Clearly, phylogenetic networks, and not phylogenetic trees, should be used to represent hybrid languages and word borrowing events. In fact, some drawbacks of the tree model in historical linguistics were already pointed out by Schmidt (1872). Nakhleh et al. (2005) were among the first to use directed phylogenetic networks to identify lexical contacts among 24 IE languages. These contacts have been represented by bidirectional reticulations, but the donor languages were not clearly distinguished from the recipient languages in the presented “perfect linguistic networks”. The study of Nakhleh and colleagues was restricted to the earliest attested languages of 12 subgroups of the IE family. Some other works that address the topic of modeling reticulate evolution in linguistics rely on the use of split graphs (Atkinson & Gray, 2005; Bryant et al., 2005; Heggarty et al., 2010), minimal lateral networks (MLN) (Nelson-Sathi et al., 2011, 2013; List et al., 2014a), and horizontal word transfer networks (HWTN) (Boc et al., 2010a). While the MLN and HWTN methods can be applied to detect word borrowing events, split graphs can be used to identify hybrid-like features of certain natural languages. For example, the split

graph topology obtained for nine Germanic languages (Bryant et al., 2005) allows one to identify Sranan, a language spoken in Suriname, as a hybrid of English and Dutch. However, split graphs were not specifically designed to detect and explicitly represent network relationships among languages. For instance, they cannot be used to identify explicitly the hybrid language, its parent languages and the corresponding hybridization/reticulation degree (i.e., percentage of lexical material transferred from each of the parent languages). Split graphs cannot be used to quantify the frequency of word borrowing events either. Furthermore, Wichmann et al. (2011) proposed to infer reticulations based on distances retrieved from the Levenshtein metric (Levenshtein, 1966) scores. Wang & Minett (2005) used maximum parsimony to detect language contacts. The test, they designed, is based on the distribution of lexical similarities between languages. Köllner (2016) proposed an ancestral state reconstruction method, which is specific to linguistics. The latter authors used the dissimilarities between a given node and its immediate ancestor in the tree in order to identify potential word borrowing events. In all these methods, the exact source and destination of the detected word borrowings cannot be identified explicitly. Only a few methods offer the advantage of finding the direction of reticulation events in linguistics. Mention here the work of Van Der Aek et al. (2007), who used the Levenshtein distance (Levenshtein, 1966) to identify the source and the destination of word borrowing events, and that of Delz (2013), who applied the horizontal gene transfer algorithm (Boc & Makarenkov, 2003) from the T-Rex web server (Boc et al., 2012) to detect loanwords and the corresponding word borrowings.

In this study, we adapt a recently developed computational biology method (Saitou & Nei, 1987), which was originally designed to detect hybrid species, their parents and the corresponding hybridization degrees, to identify explicitly hybrid languages (i.e., lexical hybrids in this study) and word borrowing events. One of

the main advantages of our method over the MLN (Nelson-Sathi et al., 2011, 2013; List et al., 2014a) and perfect networks (Nakhleh et al., 2005) approaches is that it allows for determining the direction of reticulation events (e.g., word borrowing events) in addition to the quantification of influence of each of the donor languages on the corresponding recipient languages. For a more complete description of the benefits and shortcomings of the MLN approach, the reader is referred to (List et al., 2014b; List, 2014, 2015). We compare our explicit hybridization networks to the corresponding split graphs (Bandelt & Dress, 1992a; Bryant & Moulton, 2004) and galled networks (Huson et al., 2009). Finally, we present some historical evidence that supports the results of our analysis.

3.3 Methods

3.3.1 Data description

Several important studies dedicated to the classification of IE languages (Gray & Atkinson, 2003; Rexová et al., 2003; Nelson-Sathi et al., 2011; Boc et al., 2010a) have examined the data from the 84 IE language database organized by Dyen et al. (1992). The Dyen database contains word forms for the meanings of the 200-meaning Swadesh list (Swadesh, 1952). This list is one of a few lists of fundamental meanings collected by M. Swadesh in the 1940s and 50s. It is often used in lexicostatistics, which focuses on quantitative evaluation of lexical cognates, and in glottochronology, which focuses on dating divergence times of natural languages. Swadesh lists have been used by linguists to test the level of chronological separation of languages by comparing words, as they contain universal stable items with low levels of borrowing (Gray & Atkinson, 2003; Rexová et al., 2003; Nelson-Sathi et al., 2011; Boc et al., 2010a). However, it has been noticed that even though the use of Swadesh lists may decrease the level of borrowings to a certain degree, it cannot exclude all of them (List et al., 2014a). For each of the 200 basic mean-

ings of the Swadesh list, the Dyen database contains their word forms in 84 IE languages. These word forms have been regrouped in cognate sets (Dyen et al., 1992). Two word forms were identified as cognate if they share an uninterrupted evolutionary history characterized by the presence of a common ancestral form. The word forms resulting from word borrowing (e.g., English word *fruit* which was borrowed from Old French) and those related by accidental similarity (e.g., the word form *bad* exists in both English and Farsi, but this is rather considered as an accidental similarity by linguists) were placed in a separate class. When it was difficult to differentiate between cognates and word forms resulting from borrowing or accidental similarities, the corresponding word forms, albeit not numerous, were categorized as doubtful cognates. For instance, this database was used by Gray & Atkinson (2003) and Atkinson & Gray (2006) to infer evolutionary trees of IE languages. In order to reconstruct our hybridization networks, we also considered some additional linguistic resources (Douglas Harper’s Online Etymology Dictionary (Harper, 2001), the IE Lexical Cognacy Database (IELex) (Dunn, 2012) and the IE etymological dictionaries collection (Lubotsky, 2005)), which include relevant etymological information regarding loanwords and accidental similarities. Using these resources, we modified some of the original cognate sets created by Dyen et al. (1992). Precisely, the loanwords, put aside by Dyen and colleagues, were added to the corresponding cognate sets (i.e., cognate sets containing the donor forms for these loanwords). In some rare cases, the original cognate sets including doubtful cognates were either merged or eliminated. In total, our modified database included 1315 cognate sets. It is available at: <http://www.trex.uqam.ca/biolinguistics>.

3.3.2 Reconstruction of explicit linguistic hybridization networks

In (Willems et al., 2014), we presented a new algorithm for inferring explicit hybridization networks from distance data. This algorithm takes as input a matrix

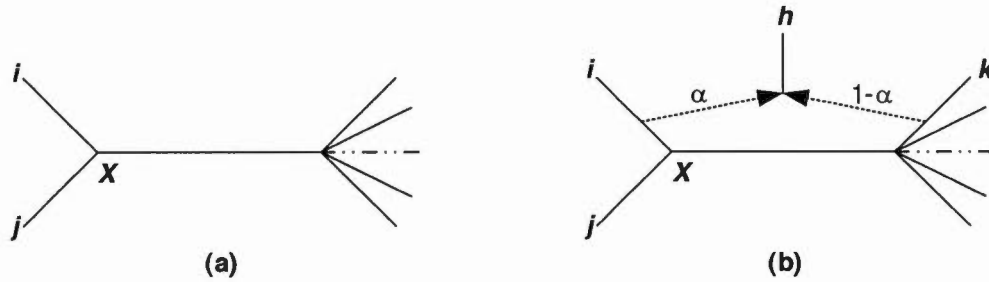


Figure 3.1 (a) Configuration in which languages i and j are selected as neighbors by the NJ algorithm, and (b) configuration in which language h is identified as a recipient of lexical material from languages i and k by our algorithm for inferring explicit hybridization networks (here, the parameters α and $1 - \alpha$ represent the hybridization (i.e., reticulation) degree of donor languages i and k , respectively).

of evolutionary distances between species of size $(n \times n)$ and the three following user-defined parameters: minimum and maximum levels of hybridization (the value of these parameters varies between 0 and 1), and minimum hybridization score. The output of this algorithm, based on a famous neighbor-joining (NJ) principle (Saitou & Nei, 1987), is either a traditional phylogenetic tree with n leaves or a hybridization network with n terminal nodes. It is worth noting that NJ remains by far the most popular distance-based method in phylogenetics, even though in linguistics Bayesian framework is also frequently used (Gray & Atkinson, 2003). NJ is specifically well suited for the inference of large phylogenies. It takes as input a distance matrix $D = \{d(i, j)\}_{1 \leq i, j \leq n}$ defined on a set of n species (i.e., taxa or languages) and gives as output a phylogenetic tree representing their evolutionary history. NJ starts with a star tree including n leaves, one internal node and n branches. This tree is progressively transformed into an unrooted binary phylogeny with n leaves and $2n - 3$ branches. The p -th step of NJ consists of selecting and connecting the two most appropriate neighbors

among $(n - p + 1)$ candidates. For all of the $\frac{(n-p+1)(n-p)}{2}$ tree configurations equivalent to that shown in Fig. 3.1(a), the branch lengths are calculated according to the least-squares criterion. The configuration that minimizes the sum of all branch lengths of the tree is then selected and the two nodes i and j , which are neighbors in this configuration, are connected as shown in Fig. 3.1(a). The nodes i and j are then replaced by the node X (their direct common ancestor; Fig. 3.1(a)) and the distance matrix D is updated by computing the new distances $d(X, k)$, from X to each remaining leaf k of the tree, by means of the following formula $d(X, k) = \frac{1}{2} (d(i, k) + d(j, k))$. We used the NJ criterion (Saitou & Nei, 1987) to infer explicit hybridization networks between species (Willems et al., 2014) and adapted it here to the identification of hybrids and word borrowings among natural languages. Note that in our networks both terminal and ancestral branches can be involved in hybridization. Obviously, the two parent branches (i.e., languages or groups of languages) are not necessarily neighbors. Each hybrid language (or recipient of lexical material) is explicitly identified along with its parent languages (or donors) and the degree of hybridization (or reticulation) corresponding to each of them. In the case of word borrowing, this degree of hybridization represents the proportion of the relative influence of each of the two donors on the recipient (Fig. 3.2(c)). As we will see later, it can also take into account the direct inheritance part of the recipient's lexicon (Fig. 3.2(b, d)).

Here we present some important computational details of our algorithm. We use the following formula to determine the level of hybridization, $\alpha_{i,j}^h$ for each possible triplet of languages, (h, i, j) , assuming that h is a hybrid of i and j :

$$\alpha_{i,j}^h = \frac{\sum_{k \neq i,j,h} X_k (Y_k - S_h + S_j)}{\sum_{k \neq i,j,h} X_k X_k}, \quad (3.1)$$

where $S_l = \frac{\sum_{k \neq i,j,h} d(k,l)}{n-3}$ (for $l = h, l = i$ or $l = j$), $Y_k = d(k, h) - d(k, j)$ and $X_k = S_j - S_i + d(k, i) - d(k, j)$. Formula 3.1 was obtained by minimizing the

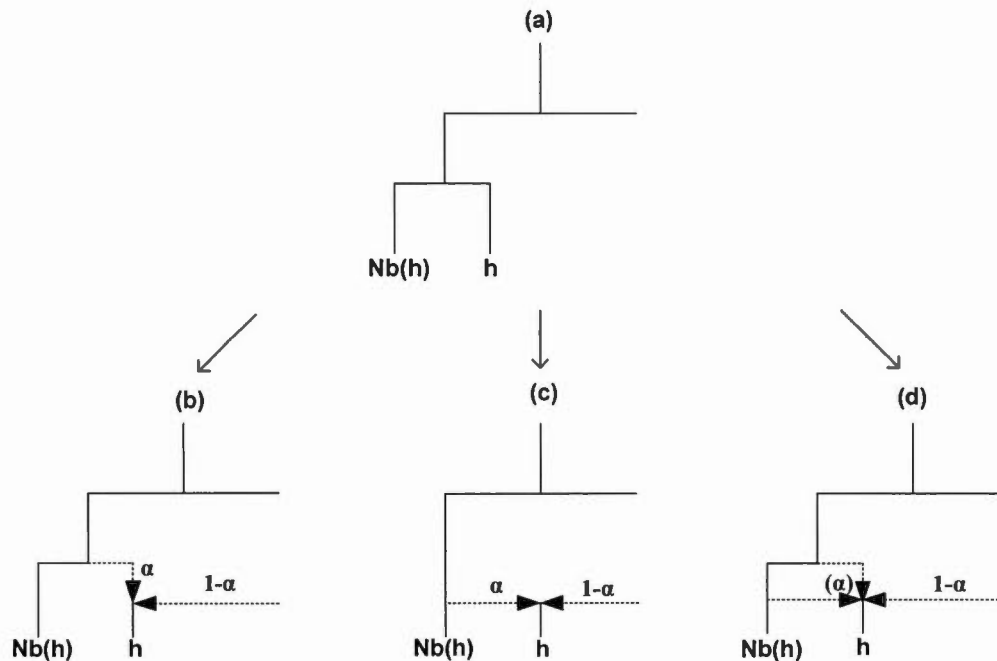


Figure 3.2 This figure illustrates three possible network configurations (b-d), when our algorithm detects a hybrid, h , which is neighbor of one of its parents, $Nb(h)$, in the phylogenetic tree (a), e.g., in the IE language phylogeny inferred by Gray and Atkinson (see Fig. 3.1 in (Gray & Atkinson, 2003)). In configuration (b), language h receives the proportion, α , of its lexicon from its closest ancestor in the tree via direct inheritance and the remaining part of its lexicon, $(1 - \alpha)$, from a distant parent via word borrowing (e.g., see the case of Penn Dutch in Figs. 3.4 and 3.5(b)). In configuration (c), language h is a lexical hybrid of $Nb(h)$ and a distant parent (e.g., see the case of Sranan in Figs. 3.4 and 3.5(b)). In configuration (d), language h receives the proportion α (indicated, in this case, in parentheses) of its lexicon from both its closest ancestor via direct inheritance and from its neighbor $Nb(h)$ via word borrowing, and the remaining part, $(1 - \alpha)$, of its lexicon from a distant parent via word borrowing (e.g., see the case of Old Armenian in Fig. 3.4).

following least-squares function of α (its minimum is attained with $\alpha = \alpha_{i,j}^h$):

$$LS_{i,j}^h = \sum_{k \neq i,j,h} (Y_k - S_h + S_j - \alpha X_k)^2. \quad (3.2)$$

The hybridization (reticulation) score, $Sc_{i,j}^h$, is defined as follows for all triplets of languages (h, i, j) :

$$Sc_{i,j}^h = \text{Min}_{k \neq i,j,h} \{d(i, j) + d(k, h) - d(i, h) - d(k, j); \\ d(i, j) + d(k, h) - d(j, h) - d(k, i)\}. \quad (3.3)$$

Formula 3.3 is related to the four point condition, which is satisfied in an additive tree (i.e., phylogenetic tree), but not in a phylogenetic network. We restrict the search of hybrids to the triplets of languages satisfying the following constraints: $Sc_{i,j}^h \geq MIN_{Sc}$ and $\alpha_{MIN} \leq \alpha_{i,j}^h \leq \alpha_{MAX}$, where the parameters $0 < \alpha_{MIN} < \alpha_{MAX}$ and MIN_{Sc} are selected by the program's user depending on the desired number of hybridization events (see (Willems et al., 2014) for more details about parameter selection). Our network reconstruction algorithm can be defined as follows. First, we determine the languages i and j that should be connected at the current step by the traditional NJ algorithm. Prior to connecting i and j , we identify the language h that is the best candidate for being a hybrid of either i or j (Parent 1 of h) and any other remaining language k (Parent 2 of h ; see Fig. 3.1(b)). We search for the language h_0 that maximizes the absolute value of the following function:

$$\Delta_{i,j}^h = \sum_{k \neq i,j} (d(j, h) + d(i, k) - d(j, k) - d(i, h)). \quad (3.4)$$

Note that $\Delta_{i,j}^h$ equals 0 if i and j are true neighbors in an additive tree.

Then, we select the triplet (h_0, i_0, k) , here $i_0 = i$ or $i_0 = j$, that provides the minimum of the least-squares function $LS_{i,j}^h$ and satisfies the above-mentioned constraints. If $LS_{i,j}^h < (\Delta_{i,j}^h)^2$, we consider that h_0 is a hybrid of i_0 and k , and

remove from the distance matrix the row and the column corresponding to h_0 . Otherwise, we connect the languages i and j as in the conventional NJ algorithm (Saitou & Nei, 1987). The time complexity of our network building algorithm is $O(n^3)$, which is equivalent to the time complexity of NJ.

It's important to mention that hybrid languages identified by our algorithm should not be always interpreted as real lexical hybrids or real mixed languages. In some cases, the detected parent-hybrid relationship may also represent the processes of word borrowing or even inheritance from the closest ancestor in the tree (see Fig. 3.2). This figure illustrates three possible network configurations which reflect the case where our algorithm detects a hybrid, h , which is a direct neighbor, or a very close neighbor, of one of its parents, $Nb(h)$, in the phylogenetic tree (Fig. 3.2(a)). This tree is assumed to be inferred by a traditional tree reconstruction algorithm (e.g., NJ). For instance, language h may receive the proportion, α , of its lexicon either from its closest ancestor in the tree via direct inheritance (Fig. 3.2(b)), or from its neighbor $Nb(h)$ in the tree as its lexical hybrid (Fig. 3.2(c)), or from both its closest ancestor via inheritance and from its neighbor $Nb(h)$ via word borrowing (Fig. 3.2(d); α is indicated in parentheses in this case).

We tested several strategies of computing the distance matrix D between the 84 IE languages considered in our study. As Dyen's database (Dyen et al., 1992) does not contain any word form from the Hittite and Tocharian languages, these ancient languages were discarded from our analysis. The first strategy, which provided the most plausible experimental results, used a binary presence-absence matrix of languages over the established cognate sets (1315 cognate sets in total). It is worth noting that our binary encoding concerned language presence-absence data only (e.g., as in (Gray & Atkinson, 2003)). The presence-absence matrix D had 84 rows and 1315 columns. The element (i, j) of this matrix was equal to 1 if

a word form of language i was present in cognate set j , otherwise it was equal to 0. In total, 19.69% of the data were missing in our database. Missing data were mostly due to the presence of the corresponding word forms in the special “non-cognate” class of the Dyen’s database; such word forms that were neither cognate with any other word form of the given meaning nor related to any word form by the way of borrowing were excluded from our database. The distance between any pair of languages was then calculated as the Hamming distance between the rows corresponding to these languages in the presence-absence matrix (i.e., it was equal to the number of cognate sets that contained word forms of only one of these languages). Two data encoding strategies were tested. The first, when the missing data were encoded by 0’s, and the second, when the Hamming distance between two languages was normalized by the number of meanings for which the corresponding word forms existed in both languages. As these two strategies provided very similar hybridization networks, only the results of the first strategy will be presented. The workflow chart of our method is presented in Fig. 3.3(a), and a simple example of its application is shown in Fig. 3.3(b). Here we consider a dataset with 8 languages, $L1, L2, \dots, L8$, 4 meanings and 16 cognate sets (i.e., 4 cognate sets for each meaning). According to the language content in these 16 cognate sets, language $L4$ can be seen as a hybrid of languages $L3$ and $L5$. Language $L8$ is used as an outgroup. In the first (respectively, second) step of our algorithm, languages $L1$ and $L2$ (respectively, $L6$ and $L7$) are joined, following the NJ principle. Then, before joining $(L1, L2)$ and $L3$, language $L4$ is identified as a hybrid of $L3$ and $L5$ with the degree of hybridization, α , equal to 0.5 for both of its parents. Language $L4$ is then removed from the dataset, and the remaining steps of the algorithm correspond to the steps of traditional NJ. The obtained explicit hybridization network is presented in Fig. 3.3(b).

We also conducted the analysis using the Levenshtein distance (Levenshtein, 1966)

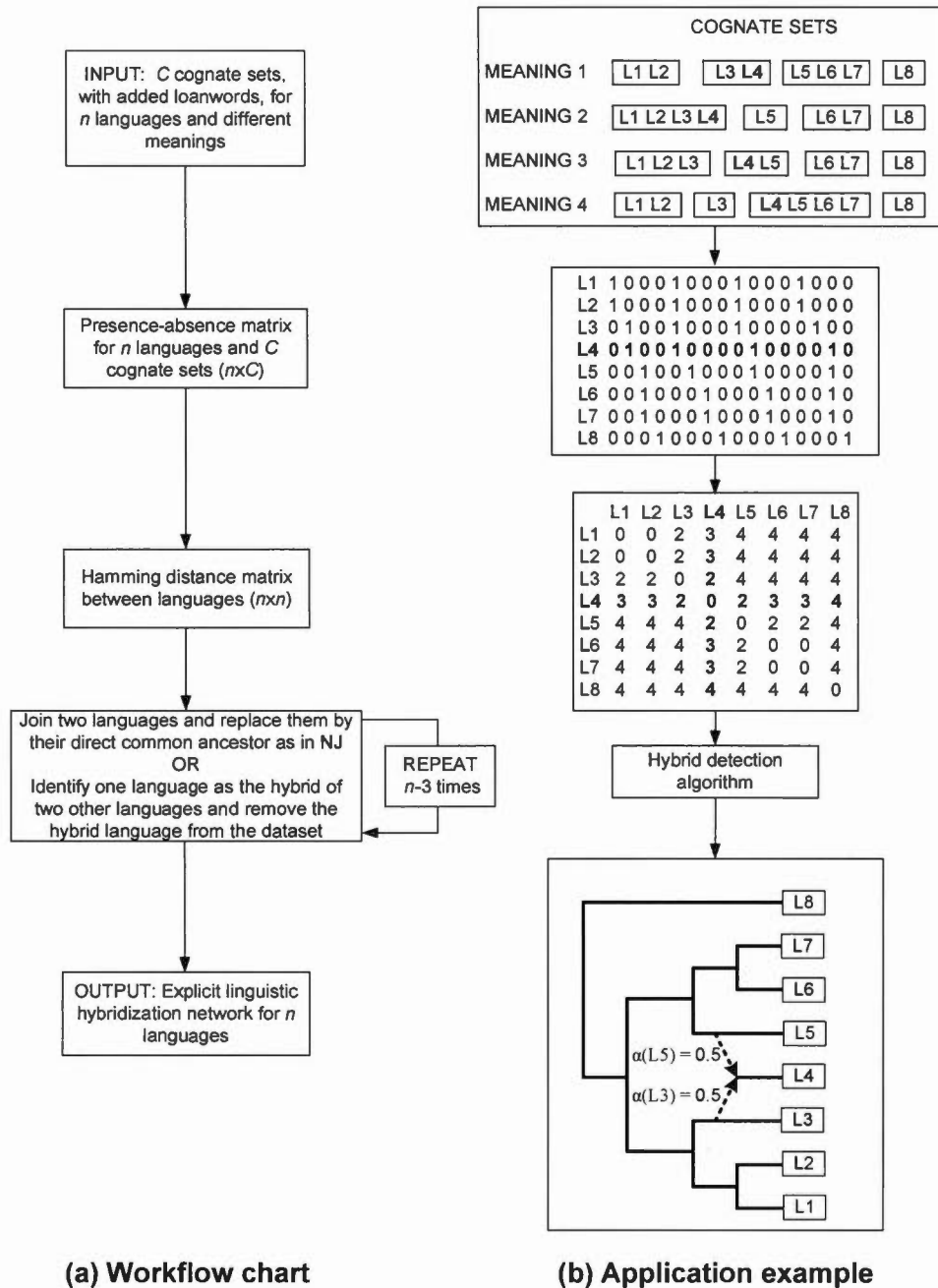


Figure 3.3 (a) Workflow chart of the new method for inferring explicit hybridization networks, and (b) an example of its application to a dataset consisting of 8 languages (including the hybrid language L4), 4 meanings and 16 cognate sets.

between words of the same meaning but did not obtain convincing results using such an approach. This should be due to the fact that this distance tends to reflect chance similarity when the compared word forms are not cognate (Greenhill, 2011). The Levenshtein distance will be further used for inferring galled networks from word trees, but its application will be restricted to word forms belonging to the same cognate set.

We applied our hybridization network inferring algorithm to the entire Hamming distance matrix of 84 IE languages, denoted here by D_{84} , as well as to its submatrices corresponding to each of the 11 considered IE language groups. In particular, some plausible lexical hybrids and word borrowing donors and recipients were found when the submatrices of the 5 following language groups were analyzed: Germanic, Latin (including the Italic and French/Iberian groups), Slavic, Sanskrit and Persian (see Figs. 3.4, 3.5(b), 3.6(b) and 3.7(b) for the detailed results). The tree topology in these networks correspond to the IE language phylogeny inferred by Gray and Atkinson (see Fig. 3.1 in (Gray & Atkinson, 2003)). Furthermore, the analysis of two submatrices corresponding to the union of the West Germanic and French/Iberian groups and the union of the Celtic and French/Iberian groups also provided very relevant results. We did not find additional reticulations within the other IE groups. We needed a distance matrix of size greater than 4 to be able to apply our algorithm. It is worth noting that the recovery of hybrid languages and word borrowing events seemed to be more complicated within smaller linguistic groups (i.e., groups with 5 or 6 taxa here).

The input parameters of our algorithm, MIN_{Sc} , α_{MIN} and α_{MAX} , were chosen according to the size of the considered distance matrix (see (Willems et al., 2014) for a detailed discussion on the parameter selection). For smaller distance matrices corresponding to particular language groups, the following set of input parameters: ($MIN_{Sc} = 0$, $\alpha_{MIN} = 0.1$ and $\alpha_{MAX} = 0.9$) was used. To avoid an excessive

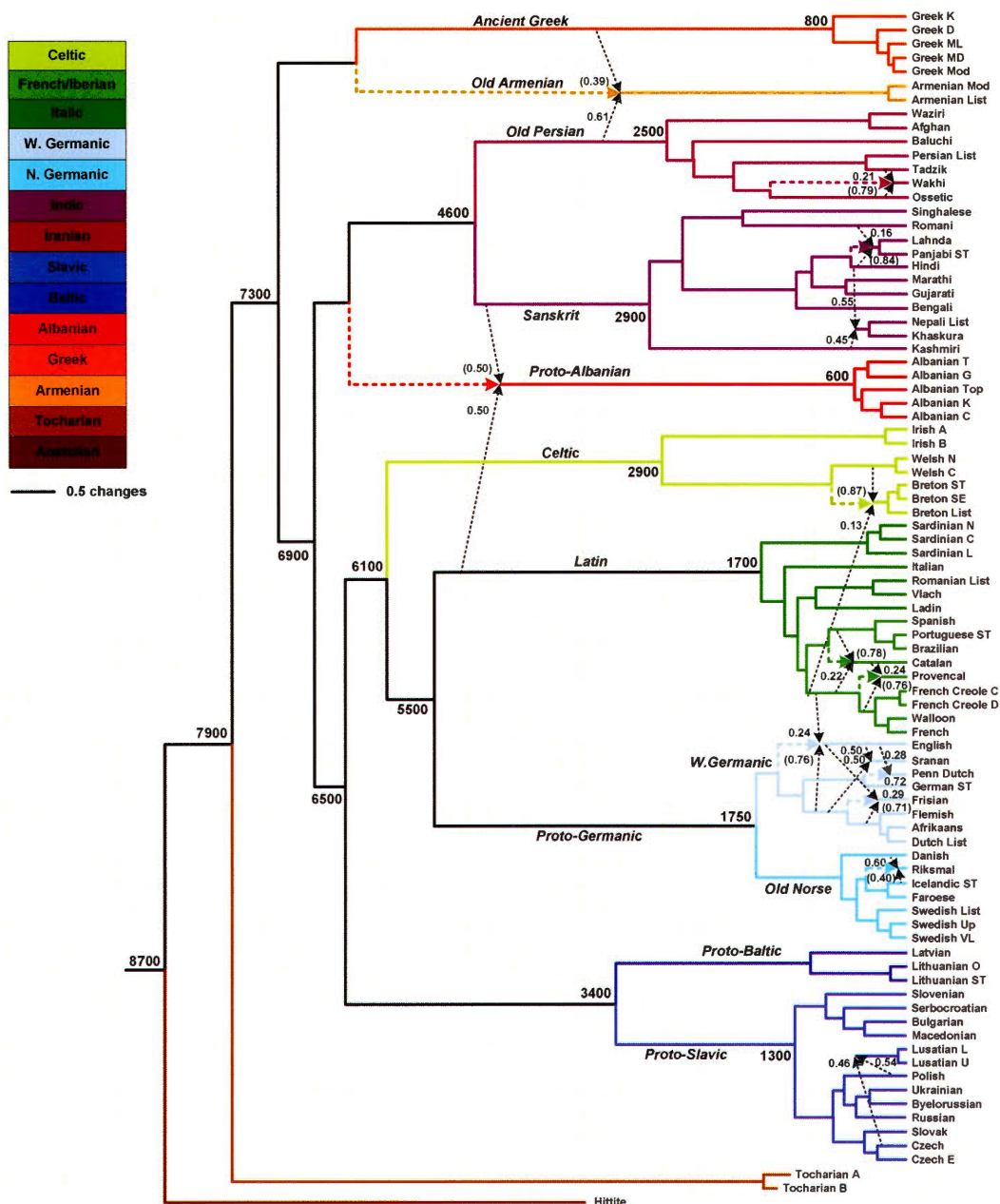


Figure 3.4 Explicit hybridization network given by our algorithm for the group of 84 IE languages originally considered by Dyen et al. (1992). Language groups are indicated on the left. The numbers at the arrows are the reticulation degrees corresponding to each of the donor languages and the numbers at the internal tree nodes are their age estimates.

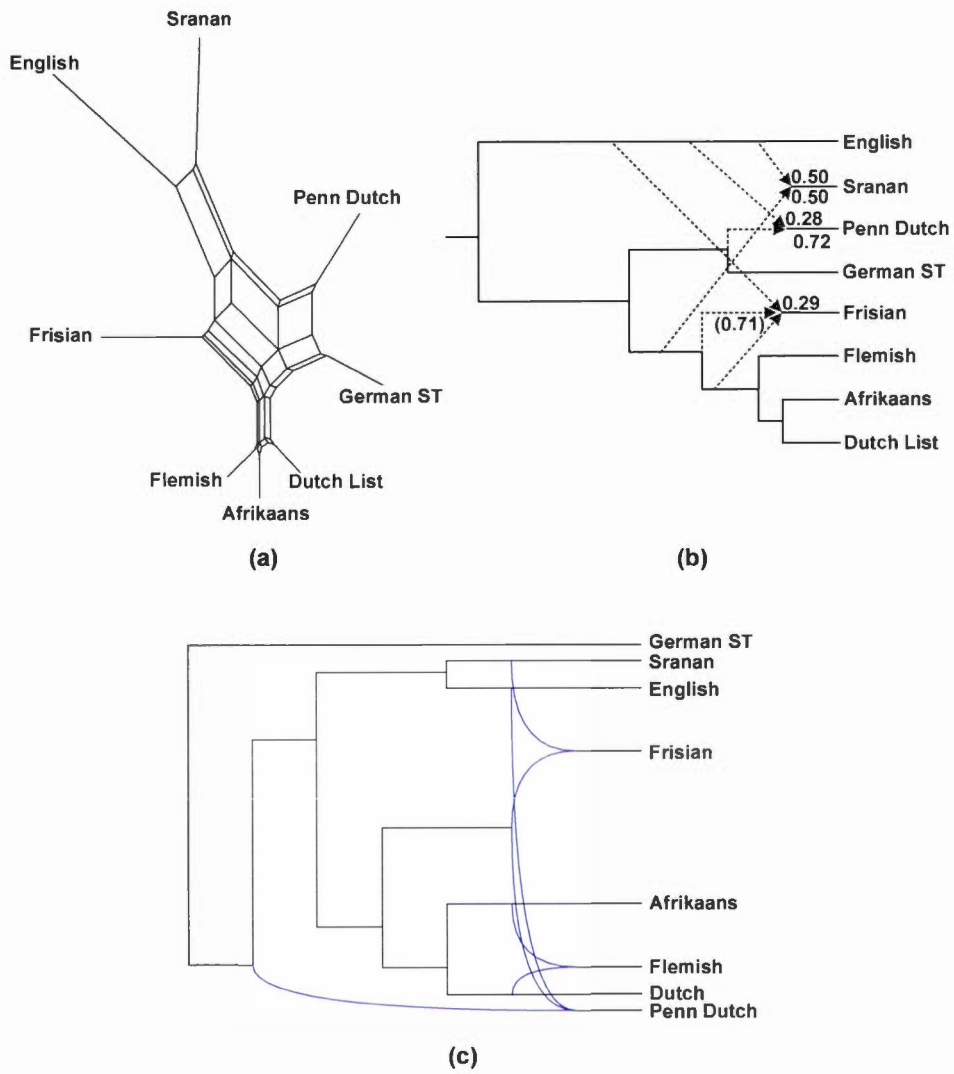


Figure 3.5 (a) Split graph, (b) explicit hybridization network and (c) galled network, obtained for 8 languages of the West-Germanic group.

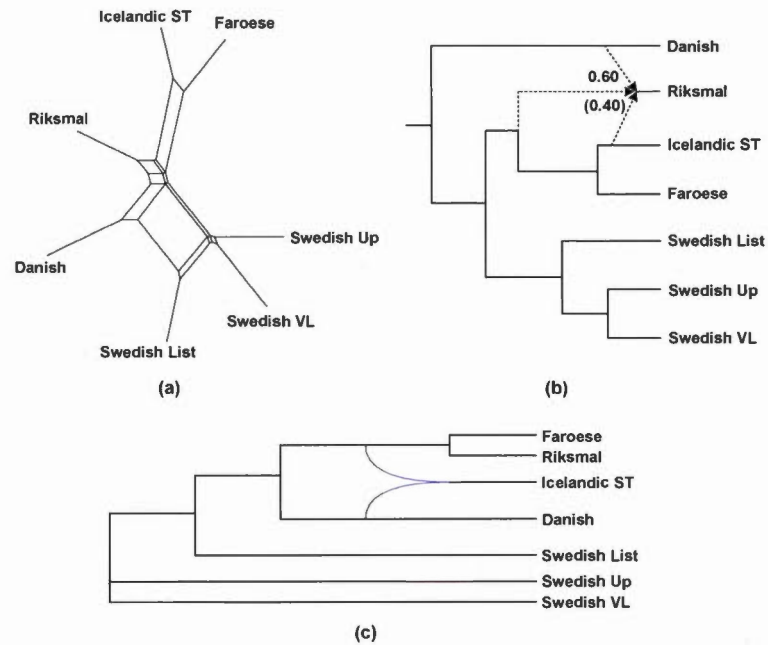


Figure 3.6 (a) Split graph, (b) explicit hybridization network and (c) galled network, obtained for 7 languages of the North-Germanic group.

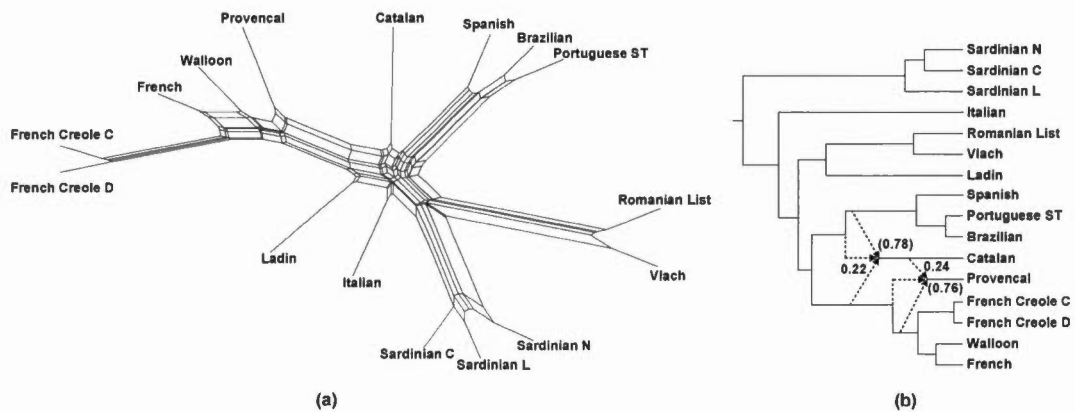


Figure 3.7 (a) Split graph and (b) explicit hybridization network, obtained for 16 languages of the Latin group.

number of false positives, more restrictive parameters: ($MIN_{Sc} = 0.1$, $\alpha_{MIN} = 0.25$ and $\alpha_{MIN} = 0.75$) were used for the entire distance matrix D_{84} . For the representation of our hybridization networks (Figs. 3.4, 3.5(b), 3.6(b) and 3.7(b)), we used the backbone IE phylogenetic tree inferred by Gray and Atkinson (Fig. 1 in (Gray & Atkinson, 2003)), mapping into it the detected lexical hybrids and word borrowing events with their respective reticulation degrees. Our program for inferring explicit hybridization networks is available at:

www.info2.uqam.ca/~makarenkov_v/makarenv/hybrids_detection.zip.

The data used in our study can be found at:

www.trex.uqam.ca/biolinguistics/Biolinguistic_networks_data.zip.

3.3.3 Reconstruction of split graph-based linguistic networks

The split decomposition method introduced by Bandelt & Dress (1992a) decomposes the given distance matrix into simple components based on weighted splits (i.e., bipartitions of taxa, species or languages). These splits can then be represented using a split graph, a particular type of phylogenetic network that simultaneously represents both clusters in the data and evolutionary distances between taxa. The Neighbor-Net method introduced by Bryant & Moulton (2004) and implemented in the SplitsTree program (Huson & Bryant, 2006) works in a similar way, but constructs phylogenetic networks that are much more resolved than those given by split decomposition. Split graphs have been widely used in phylogenetic studies to depict phylogenetic relationships between species, but several works have also considered their applications in historical linguistics (Atkinson & Gray, 2005; Bryant et al., 2005). We used SplitsTree (Huson & Bryant, 2006) to infer the split graphs corresponding to the West Germanic, North Germanic and Latin groups of IE languages, with the same submatrices of D_{84} as mentioned

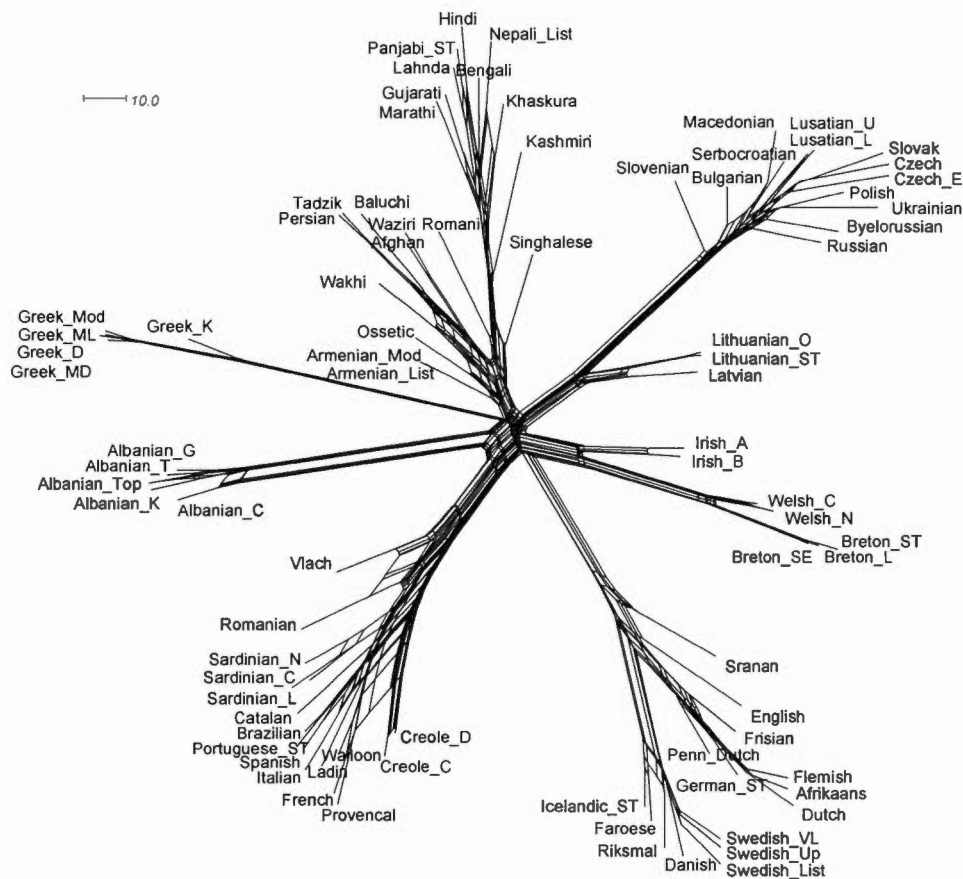


Figure 3.8 Split graph obtained for the entire set of 84 IE languages.

above. A total of 22 (respectively, 16 and 51) splits were identified for the West Germanic (respectively, North Germanic and Latin) language groups. These split graphs will be compared to our hybridization networks and galled phylogenetic networks inferred for the same groups of languages (Figs. 3.5, 3.6 and 3.7). Figure 3.8 shows the split graph, with 371 splits, obtained for the entire set of 84 IE languages examined in our study.

3.3.4 Reconstruction of galled linguistic networks from word trees

Several methods have been developed for inferring consensus phylogenetic networks from contradictory sets of two or more phylogenetic trees. They include, among others, cluster networks (Huson & Bryant, 2006), galled networks (Huson et al., 2009) and level- k networks (Van Iersel et al., 2010). A cluster network is a rooted phylogenetic network obtained from a given set of clusters (i.e., set of bipartitions). In such a network, every branch represents exactly one input cluster. A galled network is a rooted phylogenetic network in which each reticulation has a tree cycle. A tree cycle is an undirected cycle consisting of two disjoint tree paths between a tree node and a reticulation node. A level- k network is a rooted phylogenetic network, such that the maximum number of reticulations contained in a biconnected component equals k . A given set of clusters can always be represented by a galled network, but not necessarily by a level- k network (Van Iersel et al., 2010). These three methods have been implemented in the Dendroscope software (Huson & Scornavacca, 2012). We conducted our analyses with all of them but present here only the results of the galled network method which provided the “most interpretable” linguistic networks (i.e., networks in which the obtained reticulations correlate the best with known contacts between natural languages). Since for running this method we needed a set of phylogenetic trees, we reconstructed word phylogenies for each of the 200 meanings of the Swadesh list. We used the normalised Levenshtein metric (Levenshtein, 1966), denoted here by d_L , to calculate the distances between the cognate word forms of the same meaning; the distance between any two non cognate word forms was set to 1 (see below for more details). The Levenshtein distance between two words is defined as the minimum number of editing operations, consisting of insertions, deletions and substitution of a single letter, necessary to transform one word into the other. This distance was normalized by the maximum length of two words. The Leven-

shtein distance has been criticized as a poor distance for building language trees because of its reflection of chance similarity when the compared words are not cognate (Greenhill, 2011). Our comparative study presented below suggests that this distance can be used for building word trees from cognate data.

Several recent linguistic studies argued that accurate comparisons between words should also incorporate likely changes to pronunciation and phonological system (Greenhill, 2011; Atkinson, 2013). Thus, we decided to compare, in terms of reconstruction word trees and word borrowing events, the normalized Levenshtein distance with the SCA (Sound-Class-based phonetic Alignment) distance recently introduced by List (2012a). While the Levenshtein distance applies to orthographic data, the SCA distance is based on the comparison of phonological forms. Note that phonological forms are still not available for many word forms of Dyen's database. Thus, among 42 cognate sets that were found to be suggestive of borrowing into English according to the modified MLN approach (List et al., 2014a), we selected the 28 cognate sets (Tab. 3.1) for which at least 4 cognates with available phonological forms were present in the IELex database (Dunn, 2012). Trees with less than 4 leaves have identical topologies and thus cannot be used to recognize word borrowings (Boc et al., 2010a). It is important to note that the MLN approach is an automatic approach based on tree topology and the 42 suggestive cases of borrowing recovered by MLN, which include 33 English loanwords identified by Donohue et al. (2012), cannot be considered as crystal-clear borrowings. They may comprise some false positives, which can be due to parallel semantic development (List et al., 2014a), for example.

In Tab. 3.1, cases of confirmed borrowing, according to Donohue et al. (2012) are not colored; cases of possible borrowing, which could have been unrecognized according to List et al. (2014a), are shaded in gray (note that List and colleagues point out that the grayed words can also represent cases of parallel semantic de-

Word	Status	Borrowing inference based on the normalized Levenshtein distance	Borrowing inference based on the SCA distance
<i>animal</i>	loan F	donor group identified but exact WBE source not identified	donor group identified but exact WBE source not identified
<i>bark</i>	loan ON	donor group identified	donor group identified
<i>belly</i>	loan ON	donor group identified	donor group identified
<i>count</i>	loan F	donor group and exact WBE source identified	donor group and exact WBE source identified
<i>die</i>	loan ON	donor group identified	donor group identified
<i>dust</i>	loan ON	donor group identified	donor group identified
<i>egg</i>	loan ON	donor group identified	donor group identified
<i>few</i>	loan ON	donor group identified	donor group identified
<i>flower</i>	loan F	donor group and exact WBE source identified	donor group identified but exact WBE source not identified
<i>fruit</i>	loan F	donor group and exact WBE source identified	neither donor group nor exact WBE source identified
<i>give</i>	loan ON	donor group identified	donor group identified
<i>husband</i>	loan ON	donor group identified	donor group identified
<i>lake</i>	loan F	donor group and exact WBE source identified	donor group and exact WBE source identified
<i>leg</i>	loan ON	donor group identified	donor group identified
<i>many</i>	loan ON	donor group identified	donor group identified
<i>mountain</i>	loan F	donor group identified but exact WBE source not identified	donor group identified but exact WBE source not identified
<i>person</i>	loan F	neither donor group nor exact WBE source identified	neither donor group nor exact WBE source identified
<i>river</i>	loan F	donor group and exact WBE source identified	donor group and exact WBE source identified
<i>root</i>	loan ON	donor group identified	donor group identified
<i>skin</i>	loan ON	donor group identified	donor group identified
<i>smoke</i>	loan CELT	donor group identified	donor group identified
<i>snake</i>	loan ON	donor group identified	donor group identified
<i>split</i>	loan LG	donor group and exact WBE source identified	neither donor group nor exact WBE source identified
<i>stick</i>	loan ON	donor group identified	donor group identified
<i>they</i>	loan ON	donor group not identified	donor group identified
<i>tree</i>	loan ON	donor group not identified	donor group identified
<i>vomit</i>	loan F	donor group and exact WBE source identified	donor group and exact WBE source identified
<i>wing</i>	loan ON	donor group identified	donor group identified

Abbreviations: F - French, ON - Old North, CELT -Celtic group, WBE - word borrowing event

Table 3.1 This table reports the results provided by the word borrowing event detection algorithm (Boc et al., 2010a) applied to the normalized Levenshtein (Levenshtein, 1966) and SCA (List, 2012a) distance matrices.

velopment (List et al., 2014a)); an additional suggestive case of borrowing (i.e., dust) is in yellow. When the Old North (ON) group was assumed to be the word donor, the exact source of borrowing in the ON group was not considered since it remains unknown in most of these cases. The neighbor-joining algorithm (Saitou & Nei, 1987) was carried out to reconstruct phylogenies from distance matrices. Here, we considered all of the confirmed or suggested loanwords of the basic English vocabulary associated with the 200-word Swadesh list. The loanword information was taken from the studies of Donohue et al. (2012) (Supplementary Material) and List et al. (2014a) (modified MLN approach). To the 33 English loanwords identified by Donohue et al. (2012) and 8 possible additional English borrowings suggested by List et al. (2014a) (Supplementary Material II; Tab. C), we added the word dust, which might have been borrowed by English from Old North according to the results given by our algorithm for detecting word borrowing events (Boc et al., 2010a). This English word was also identified as belonging to an irregular pattern, which may be suggestive of borrowing, by the modified MLN approach (List et al., 2014a). To compare the results yielded by the normalized Levenshtein distance (Levenshtein, 1966) and the SCA phonetic distance (List, 2012a), we examined 28 of the 42 above-mentioned cases of borrowing into English, i.e., all the cases for which at least 4 cognate phonological forms were available in the IELex database (see <http://ielex.mpi.nl> (Dunn, 2012) for the cognate sets that included the considered English word forms). Thus, 28 of 42 suggested English borrowings (see the table) were examined along with their donor word form and all of the donor's cognates that had phonological entries in IELex (the data were taken in September 2015). The detailed results have been included in the archive file *Biolinguistic_networks_data.zip* available at: <http://www.trex.uqam.ca/biolinguistics>. This data archive comprises word tree representations, lexeme and phonological distance matrices, and the Perl script for computing the normalised Levenshtein distance

We applied our algorithm for inferring word borrowing events (Boc et al., 2010a) to the word trees obtained with the normalized Levenshtein and the SCA distances (the inferred word tree topologies are available in Appendix B). The results provided by using these distances can be considered as equivalent. The normalized Levenshtein distance allowed us to identify 23 of 28 suggested borrowings, while the SCA approach was able to detect 22 of them. For instance, the SCA-based algorithm was unable to recover the correct borrowings into English for the words *flower*, *fruit* and *split* (Tab. 3.1). The results of this analysis as well as the fact that orthographic cognate data are much more complete than phonological ones are the main reasons that justify the use of the normalized Levenshtein distance for inferring word trees. It is worth noting that one of the most significant differences between language history and biological evolution is that in the case of natural languages our alphabet systems change, while biological sequences change via mutation. Thus, methods using the Levenshtein distance as well as the more historically-oriented SCA distance may have shortcomings, since both of these distances are based on the idea that similarities and differences are due to mutations. For example, the distance between French *tête* and Latin *testa* should be 0 in linguistic terms, since the sound change was completely regular. Moreover, some words may contain cognate material, but only in parts. For example, the French word *soleil* is different from Italian *sole*, since it stems from a suffixed form of Latin *sol*, namely Latin *soliculus*. This case cannot be handled successfully by the Levenshtein and SCA distances, and the use of any of them will lead to the addition of noise to the distance matrix.

Borrowings can be seen as mutations in some parts, since they are not produced by regular sound change. Thus, methods based on sequence similarity, like those using the Levenshtein distance, may have advantages in identifying borrowings over methods that seek to ignore regular dissimilarities between words, like those

using the SCA distance. Furthermore, the presented method could be modified to account for language-specific distances, which could be measured by other algorithms, as for example, the LexStat algorithm by List (2014) or the algorithm proposed by Steiner et al. (2011).

For each considered meaning m of the 200-meaning Swadesh list, we denoted by L_m the set of languages for which we had at least one word form of m in our database, and by C_m the collection of cognate sets available for the meaning m . Let n_m be the cardinality of L_m . Note that for most of the meanings, the value of n_m was lower than 84 since our database, as well as its original version created by Dyen, had some missing word forms for almost all the meanings. Mention that in some, rather rare, cases multiple word forms of the same language existed for a given meaning m . For each meaning m , a distance matrix D_m of size $(n_m \times n_m)$ was computed by applying the following formula to each pair of languages, l_1 and l_2 , in L_m :

$$d_m(l_1, l_2) = \frac{\sum_{c \in C_m} d_c(l_1, l_2)}{n_{l_1, l_2}}, \quad (3.5)$$

where $d_c(l_1, l_2)$ was equal to 0 if neither word forms of l_1 nor those of l_2 were present in c ; $d_c(l_1, l_2)$ was equal to 1 if word forms of either only l_1 or only l_2 were present in c ; and, it was equal to the minimum value of $d_L(i, j)$, over all cognates i representing l_1 and all cognates j representing l_2 in c , if word forms of both l_1 and l_2 were present in c . The integer n_{l_1, l_2} was the number of cognate sets of the meaning m that included at least one word form of either l_1 or l_2 . Thus, we obtained 200 distance matrices D_m of different sizes. For each such a matrix, we then inferred the corresponding unrooted word phylogeny T_m using the NJ algorithm (Saitou & Nei, 1987). The obtained word phylogenies were given as input to the galled network algorithm (Huson et al., 2009). Since these word trees did not contain the same sets of languages (i.e., tree leaves), we used the Z-closure method, available in Dendroscope 3 (Huson & Scornavacca, 2012), to

merge partial data (Huson et al., 2004). Figures 3.5(c), 3.6(c) and 3.9 present the most plausible networks provided by the galled network algorithm (Huson et al., 2009). First, we inferred networks from the trees restricted to the languages of the West Germanic (Fig. 3.5(c)) and North Germanic (Fig. 3.6(c)) groups. The trees including at least 4 Germanic languages (West or North) were analyzed. Here we considered splits that were present in at least 30% of the input trees. In the case of the West Germanic group, we examined 190 input trees with 207 input splits, 237 splits after Z-closure, and 76 remaining splits after the removal of partial splits. The consensus galled network obtained for the West Germanic group (Fig. 3.5(c)) contains 20 splits and 3 putative lexical recipients (i.e., *Frisian*, *Flemish* and *Pennsylvania Dutch*). In the case of the North Germanic group, we examined 188 input trees with 109 input splits, 112 splits after Z-closure, and 49 remaining splits after the removal of partial splits. The consensus galled network for the North Germanic group (Fig. 3.6(c)) contains 12 splits and 1 putative lexical recipient (i.e., *Icelandic ST*). We also inferred a galled network for a total of 84 IE languages. In this case, we considered only the splits that were present in at least 75% of the input trees to avoid false positive reticulations. Figure 3.9 illustrates a sub-network of 12 IE languages that contains all the reticulations identified in the complete galled network of 84 languages. Here we considered 200 input trees with 6176 input splits, 11299 splits after Z-closure and 5124 remaining splits after the removal of partial splits to obtain a consensus galled network with 101 splits and 3 putative recipient languages (i.e., *Armenian List*, *Armenian Mod* and *Ossetic*). The presented network correctly identifies the influence of the languages of the Iranian group and that of Ancient Greek on Armenian, but also includes false positive reticulations reflecting, for example, the influence of Frisian on Armenian.

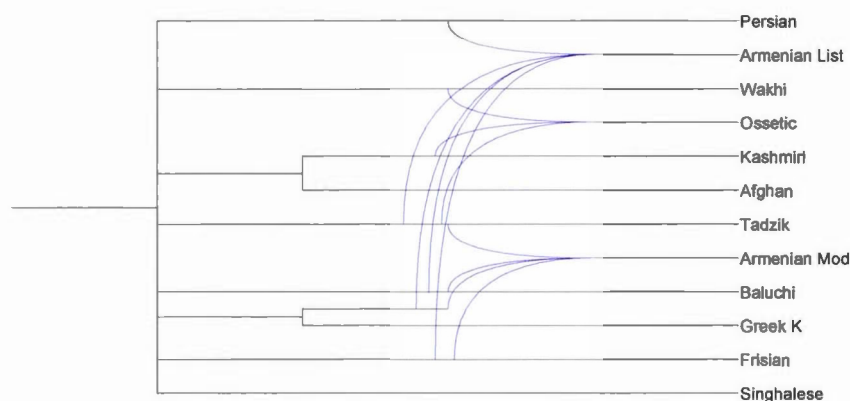


Figure 3.9 Partial galled network obtained for 12 IE languages. This is a maximum sub-network that includes reticulations of the complete galled network built for the entire set of 84 IE languages.

3.4 Results and Discussion

Hybrid languages emerge in a few generations as a new means of communication between two (or more) populations not sharing a common language. In many cases, e.g., when we found that Old Armenian is a lexical hybrid of Old Persian and Old Greek, we should interpret the results of our algorithm as the identification of the influence, e.g., cultural, political or military, which the two parent languages (i.e., the donors) had on their lexical hybrid (i.e., the recipient), at possibly different periods of time, and which could last over several centuries.

As known from evolutionary biology, the position of hybrid species in a phylogenetic tree or network is often uncertain (Legendre & Makarenkov, 2002). Furthermore, some of the hybrids added to the data can influence the position of their parents when a phylogenetic tree or network is inferred. Often a hybrid is placed as a direct neighbor of one of its parents in a phylogenetic tree or network, and the parents' location may change when this hybrid is removed from the data set.

Thus, some of the results presented in this section were obtained after rerunning our algorithm on the distance matrices from which the detected lexical hybrids, identified at the first run of this algorithm, were removed.

Here we present the most important reticulation events characterizing the evolution of IE languages which were identified by the three competing algorithms for inferring split graphs, galled networks and our explicit hybridization networks, respectively. The related historical facts and justifications are also discussed. Since only lexical data were considered in our study, the presented phylogenetic networks represent interactions between languages which are mainly based on lexical borrowings. They do not account for other language interactions, such as contact-induced syntactic restructuring, for example.

3.4.1 Network relationships within the Germanic group

We carried out our algorithm independently for the languages of the West Germanic group, the North Germanic group, and finally, the entire Germanic group. Four putative lexical hybrids were discovered in this analysis (Figs. 3.5(b) and 3.6(b)):

- *Pennsylvania Dutch* as a recipient of lexical material from *English* (by word borrowing) and *German* (by inheritance): Pennsylvania German or Pennsylvania Dutch (Penn Dutch) is a variant of German developed by the descendants of German, French (from Alsace and Lorraine) and Swiss emigrants to the East Coast of the United-States (Buffington & Preston, 1965). These migrants settled in the United-States in the 17th and 18th centuries. Pennsylvania Dutch borrowed many words from English, particularly in the 19th century.
- *Frisian* as a recipient of lexical material from *Old English* (by word borrowing) and the ancestor of *Flemish*, *Afrikaans* and *Dutch* (by word borrowing,

but the inheritance from a close common ancestor is also possible here): The Frisian dialects are spoken in the northern parts of the Netherlands and Germany (Rolf Jr, 2009). They are the closest living languages to English, after Scots. Due to the long lasting influence of Old Dutch (since the Middle Ages), Frisian is now more similar to Dutch than to English (see a greater reticulation degree obtained for Old Dutch than for Old English in Fig. 3.5(b), i.e., 0.71 vs. 0.29).

- *Sranan* as a recipient of lexical material from *English* (by word borrowing) and *Old Dutch* (by word borrowing): *Sranan* is an English-based creole language spoken in Suriname (Carlin & Arends, 2002). After the invasion of Suriname by the Dutch in 1667, *Sranan*'s vocabulary was greatly influenced by Dutch. *Sranan* also borrowed some Portuguese and African words.
- *Riksmål* as a recipient of lexical material from *Danish* (by word borrowing) and *Icelandic* (by word borrowing, but the inheritance from a close common ancestor is also possible here): Historically, the North Germanic languages were divided into three main branches: East Scandinavian (Danish and Swedish), West Scandinavian (Icelandic, Faroese and Norwegian) and Old Gutnish (Bandle, 2005). *Riksmål* (or *Bokmål*) is now the most widely-used written standard of contemporary Norwegian. It was strongly influenced by Danish, because of the political domination of Denmark over Norway during several centuries. Nowadays, *Riksmål* is closer to Danish than to Icelandic and Faroese (see the corresponding reticulation degrees in Fig. 3.6(b)).

The following common features can be observed when comparing the explicit networks provided by our algorithm to those given by the split graphs (Figs. 3.5(a) and 3.6(a)) and galled networks (Figs. 3.5(c) and 3.6(c)) methods. In the case of the North Germanic group, the split graph (Fig. 3.6(a)) allows us to identify *Riksmål* as a potential lexical hybrid of Danish and the ancestor of Icelandic and

Faroese. Very similar reticulations were found by our method (Fig. 3.6(b)). However, the split graph does not yield any quantitative estimation of the influence of donor languages on recipient languages. In the case of the West Germanic group, the identification of network relationships in the split graph is more sophisticated (Fig. 3.5(a)). For example, we could implicitly identify in this graph the same lexical recipients as in our explicit network, but we could also see German as a recipient of lexical material from (Flemish, Afrikaans and Dutch) and Pennsylvania Dutch, or the ancestor of (Flemish, Afrikaans and Dutch) as a lexical recipient of German and Frisian. The galled network method yielded more explicit linguistic networks than split graphs. However, the galled network obtained for the North Germanic languages (Fig. 3.6(c)) incorrectly identifies Icelandic as a recipient of lexical material from Danish and the ancestor of (Faroese and Ríkismál). For the West Germanic group (Fig. 3.5(c)), the reconstructed galled network was able to depict two correct recipients languages: Frisian and Pennsylvania Dutch. Nevertheless, Flemish was wrongly identified as a recipient of lexical material from Dutch and Afrikaans, and Sranan was not detected as a lexical hybrid but rather as one of the donors of Frisian.

3.4.2 Network relationships within the Latin group

Only two possible lexical hybrids were identified by our algorithm in the Latin group (including Italian and French/Iberian subgroups; Fig. 3.7(b)):

- *Catalan* as a recipient of lexical material from the ancestor of *Spanish*, *Portuguese* and *Brazilian* (by word borrowing, but the inheritance from a close common ancestor is also possible here) and from Old French (by word borrowing).
- *Provençal* as a recipient of lexical material from *Catalan* (by word borrowing) and *Old French* (by word borrowing, but the inheritance from a close

common ancestor is also possible here).

The detected reticulation events reflect the history of the Occitan language, which is a Romance language spoken in Southern France, Northern Italy and Eastern Spain (Pierre, 1973). There have been many interactions between Occitan and French since the Middle Ages. For instance, “Langue d’Oïl”, from which evolved the modern French, was spoken in the North, and “Langue d’Oc”, the ancestor of Occitan, was spoken in the South. Catalan, which is the closest relative of Occitan, is sometimes considered as one of its dialects (Pierre, 1973; Smith & Bergin, 1984). After the union of Aragon and Castile in 1479, the influence of the Iberian languages, in particular that of Spanish, on Catalan became more noticeable. Provençal is a dialect of Occitan spoken in Southern France (Pierre, 1973).

The split graph obtained for the entire Latin group (Fig. 3.7(a)) represents a highly implicit linguistic network, which is not easy to interpret. For example, we could identify here Provençal as a lexical recipient with donors Catalan and the ancestor of French, Walloon and French Creole, as well as Italian as a lexical recipient with donors Ladin and Sardinian. No interpretable galled network has been obtained for the Latin language group.

3.4.3 Network relationships within the Slavic group

Here we identified *Lusatian* as a lexical hybrid of *Polish* and *Czech* (both by word borrowing). The Sorbian (or Lusatian) languages are Slavic languages spoken in North East Germany (Vogt & Geis, 2007). These languages have been strongly influenced by Czech and Polish, since Lusatia is located at the border between Germany, the Czech Republic and Poland.

3.4.4 Network relationships within the Persian and Sanskrit groups

Here we identified three possible lexical hybrids in two different program runs, i.e., one run for each of these groups:

- *Wakhi* as a recipient of lexical material from *Tadzik* (by word borrowing) and *Ossetic* (by word borrowing, but the inheritance from a close common ancestor is also possible here). *Wakhi* is an Iranian language spoken in Pamir, a mountain region between Pakistan, Afghanistan, China and Tajikistan. For the small nations of Pamir the language of oral and written communication is *Tadzik*. Moreover, the *Wakhi* oral tradition is bilingual (*Wakhi* and *Tadzik*), and most *Wakhs* speak *Tadzik* quite fluently (Kolga, 2001).
- Ancestor of *Nepali* and *Khaskura* as a recipient of lexical material from *Hindi* (by word borrowing) and *Kashmiri* (by word borrowing). *Nepali* and *Khaskura* are spoken mainly in Nepal, India and Bhutan. They share about 80% of their lexicon with *Hindi* (Hodgson, 1874).
- Ancestor of *Lahnda* and *Panjabi* as a recipient of lexical material from *Hindi* (by word borrowing, but the inheritance from a close common ancestor is also possible here) and *Romani* (by word borrowing). *Lahnda* and *Panjabi* are the languages spoken in Pakistan and India (Kachru et al., 2008). The *Romani* migrated from Northern India to Europe between the 6th and 11th centuries (Kenrick, 2007). They had numerous interactions with Northern Indian, Iranian and European languages during their migrations.

3.4.5 Network relationships within the Celtic and French/Iberian groups

We applied our algorithm to the union of the Celtic and French/Iberian groups excluding from our analysis the lexical hybrids that we had already identified when examining the Latin group alone, i.e., Catalan and Provençal. This way, we found that the *Breton* subgroup was a recipient of lexical material from *Old Welsh* (by word borrowing, but the inheritance from a close common ancestor is also possible here) and *Old French* (by word borrowing). The former reticulation shows a close etymological relationship between Welsh and Breton, whereas the latter accounts for the important number of words that Breton borrowed from Old French, namely in the 15th and 16th centuries (Piette, 1973).

3.4.6 Network relationships within the West Germanic and French/Iberian groups

We also applied our algorithm to the union of the West Germanic and French/Iberian groups ruling out the lexical hybrids we had already detected in these groups, i.e., Catalan, Provençal, Sranan, Pennsylvania Dutch and Frisian. This allowed us to identify *English* as a recipient of lexical material from the *Old French* (by word borrowing) and *Old Dutch* (by word borrowing, but the inheritance from a close common ancestor is also possible here) subgroups. Mention that these two reticulations do not exclude the direct inheritance of Old English from the Anglo-Frisian and North Germanic dialects originally spoken by Germanic tribes, traditionally known as the Angles, Saxons and Jutes (Baugh & Cable, 2002). Moreover, the relationship between Dutch and English originates in Old Saxon, which was spoken in North West Germany and in the Netherlands by Saxon peoples. Old Saxon was closely related to both Old English and Old Dutch (Robinson, 1947). After the Norman conquest of England in the 11th century, many French words were borrowed by Middle English. Furthermore, English was replaced as the language

of the upper classes by Anglo-Norman, a relative of Old French, and Old English developed into the next historical form of English, known as the Middle English language (Baugh & Cable, 2002).

3.4.7 Network relationships between IE language groups

In our final analysis, we removed from our data set the 12 lexical hybrids already identified in the original set of 84 IE languages, thus obtaining a reduced distance matrix D_{72} of size (72×72) . We applied our algorithm to this reduced matrix and limited the search of recipient and donor languages to the ancestor branches of the 11 main IE language groups (Armenian, Albanian, Baltic, Celtic, Greek, Latin, North Germanic, Persian, Sanskrit, Slavic and West Germanic).

First, we identified the Armenian group as a recipient of lexical material from the Albanian and Persian groups, and, second, the Albanian group as a recipient of lexical material from the Sanskrit and Latin groups. Since the reticulation (hybridization) score, which reflect the likelihood of a reticulation event (see Formula 3.3), of Albanian was much higher than that of Armenian, we applied our algorithm once again after removing from the distance matrix the data corresponding to the 5 languages of the Albanian group. It is worth noting that the position of the Albanian group in the IE language tree has been found to be unstable by many authors (Gray & Atkinson, 2003; Rexová et al., 2003; Bouckaert et al., 2012; Nakhleh et al., 2005). The following application of our method to the reduced distance matrix D_{67} of size (67×67) allowed us to identify Old Armenian as a recipient of lexical material from Old Persian and Old Greek (Fig. 3.4). A similar network pattern was found by the galled network method (Fig. 3.9). Thus, we could identify here:

- *Old Albanian* as a recipient of lexical material from *Sanskrit* (by word borrowing, but the inheritance from a close common ancestor is also possible

here) and *Latin* (by word borrowing). Albanian borrowed many words from Latin, in particular between the 2nd century B.C. and the 5th century A.D. (Bonnet, 1998). The Albanian group is also a close relative of the union of the Sanskrit and Persian in the IE language tree (see for example Fig. 1 in (Gray & Atkinson, 2003)).

- *Old Armenian* as a recipient of lexical material from *Old Greek* (by word borrowing, but the inheritance from a close common ancestor is also possible here) and *Old Persian* (by word borrowing). The Armenians stayed under Persian rule for long periods of time from the 5th century BC to the 19th century AC and the Armenian language includes a large number of Iranian loanwords in its vocabulary (Bournoutian, 2012). Moreover, the well-known “Graeco-Armenian” hypothesis postulates that Armenian is the closest relative of Greek (Clackson, 1994).

3.5 Conclusion

The application of computational biology methods presented here in the context of historical linguistic can be viewed as a step towards a better understanding of the evolution of natural languages (Lightfoot, 1979, 2006; Roberts, 2007; Di Sciullo, 2011). In this paper, we adapted a recently developed bioinformatics method for inferring explicit hybridization networks (Willems et al., 2014) to identify reticulate relationships between languages. We also showed how the well-known split graph (Bandelt & Dress, 1992a; Bryant & Moulton, 2004) and galled network (Huson et al., 2009) algorithms can be applied to analyze linguistic data. While all the three competing methods can be used to reconstruct evolutionary relationships between natural languages, our method has the important advantage of identifying these relationships explicitly. It also allows one to establish the extent of influence of each of the donor languages on the corresponding recipient languages through the computation of the reticulation degree parameter. Some recent stud-

ies have used syntactic distances to infer phylogenies of IE languages (Colonna et al., 2010; Longobardi et al., 2013). Syntactic parameters reveal complementary relationships between languages which are often not reflected by lexicon (Colonna et al., 2010). This type of syntactic distances could be further used to refine the inference of linguistic networks along with plausible phonological and morphological data. It would be also interesting to extend our method to infer the exact timing of the obtained reticulation events. This will allow us to discover new historical events that have shaped the evolution of natural languages.

3.6 Declarations

3.6.1 Acknowledgements

We thank Dr. QD. Atkinson and two anonymous reviewers for their helpful comments and suggestions.

3.6.2 Funding

This work was supported by Natural Sciences and Engineering Research Council of Canada, Fonds de Recherche sur la Nature et Technologies of Québec, and Fonds de Recherche sur la Société et la Culture of Québec.

3.6.3 Availability of data and materials

All the data presented in this article, including linguistic and phonetic data, distance matrices, methods' parameters, reconstructed trees and networks are available at: www.trex.uqam.ca/biolinguistics.

3.6.4 Authors' contributions

MW and VM wrote the article and carried out the experimental study. EL carried out the experimental study. VM, LL, GL, FJL and AMD participated in the

design of the study and in the search of linguistic and historical justifications for the results obtained. All authors gave final approval for publication.

CHAPITRE IV

CONSTRUCTION DE RÉSEAUX D'HYBRIDATION EXPLICITES EN UTILISANT NEIGHBOR-JOINING ET LE PRINCIPE DU MAXIMUM DE VRAISEMBLANCE

RÉSUMÉ

Ce chapitre est une reproduction de l'article suivant :

Willems, M., Tahiri, N. & Makarenkov, V. (2017). Building explicit hybridization networks using the maximum likelihood and neighbor-joining approaches. *European conference on data analyses 2017 (numéro spécial qui sera publié dans Archives of Data Science Series A)*, soumis pour publication.

Les topologies arborescentes sont les structures les plus simples qui peuvent être utilisées pour représenter l'évolution des espèces. Depuis une vingtaine d'années, des structures plus complexes, appelées réseaux phylogénétiques, ont été introduites pour prendre en compte les mécanismes d'évolution réticulée, comme l'hybridation d'espèces et les transferts horizontaux de gènes chez les bactéries et les virus. De nombreux algorithmes et plusieurs programmes ont été développés dans ce contexte, mais la plupart d'entre eux infèrent des réseaux implicites, qui peuvent s'avérer difficiles à interpréter. Dans ce chapitre, nous introduisons une nouvelle méthode pour inférer un réseau d'hybridation explicite à partir de données binaires. Notre algorithme est basé sur une approche de maximum de vraisemblance appliquée aux configurations d'arbres de neighbor-joining. Il prend en entrée un ensemble de n séquences binaires, p. ex., présence/absence de sites de restriction, ou présence/absence de certain gènes dans les génomes des n espèces considérées. On obtient en sortie un réseau d'hybridation, dans lequel les nœuds terminaux correspondent aux n espèces considérées, et où les espèces hybrides sont explicitement identifiées. Nous avons testé notre nouvel algorithme sur différentes données réelles et simulées, et nous avons comparé son efficacité avec la méthode de distances présentée dans le chapitre 2. Sur l'ensemble de nos simulations, l'algorithme présenté dans ce chapitre a fourni de meilleurs résultats (en termes de taux de faux positifs et de vrais positifs pour la détection des hybrides) que l'algorithme décrit dans le chapitre 2.

Contributions : J'ai rédigé l'article et mené les expérimentations. Vladimir Makarenkov a participé à la conception de l'algorithme et a relu l'ensemble de l'article. Nadia Tahiri a présenté ces travaux à la conférence ECDA 2017.

4.1 Abstract

Tree topologies are the simplest structures, which can be used to represent the evolution of species. Over the two last decades, more complex structures, called phylogenetic networks, have been introduced to take into account the mechanisms of reticulate evolution, such as species hybridization and horizontal gene transfer among bacteria and viruses. Several algorithms and software have been developed in this context, but most of them yield as output only an implicit network, which can be difficult to interpret. In this paper, we introduce a new algorithm for inferring explicit hybridization networks from binary data. In order to build our explicit hybridization networks, we use a maximum likelihood approach applied to neighbor-joining tree configurations. Our algorithm takes as input a set of n binary sequences, e. g., presence/absence of restriction sites or presence/absence of certain genes in genomes, corresponding to a set of n species. We obtain as output a hybridization network in which terminal nodes correspond to the n input species and the hybrids are explicitly identified among them. The new algorithm was tested on various simulated and real datasets, and its efficiency was compared to a distance-based method, developed in our previous study. Overall, the new algorithm provided better hybrid recovery results in terms of false positive and true positive rates than the distance-based method.

4.2 Introduction

Topological discordance among gene trees representing the evolution of a given set of species is commonly attributed to different reticulate evolutionary processes, including species hybridization, horizontal gene transfer, ancient gene duplication, gene loss and incomplete lineage sorting (Huson & Bryant, 2006). Interspecific gene exchange has been well documented and is very frequent across many groups of animals, plants and bacteria. Reticulate evolutionary processes, not following

vertical inheritance of genetic material, cannot be adequately represented by traditional phylogenetic trees such as a species tree or the tree of life. Phylogenetic networks should be used instead to represent these evolutionary events.

Many efforts in the field of phylogenetics have been dedicated to the inference and statistical validation of phylogenetic trees, while effective methods and user-friendly software for reconstructing phylogenetic networks still remain limited or under development (Solís-Lemus & Ané, 2016). The main disadvantage of existing hybridization networks building methods is that the networks they provide are rather implicit than explicit, thus rarely allowing for an accurate identification of hybrid species and their parents (Willems et al., 2014). Several attempts to model reticulate evolutionary relationships using phylogenetic networks have been made by a number of research groups around the world. Bandelt & Dress (1992b) were among the first authors to do it, by describing the split decomposition method that allows for data representation under the form of a split graph revealing conflicting signals hidden in the data. Bryant & Moulton (2004) continued the work of Bandelt and Dress by proposing the Neighbor-Net network-building algorithm that reconstructs planar phylogenetic split networks. Despite the popularity of the split decomposition and Neighbor-Net methods, the networks they generate usually contain a very important number of splits from which explicit reticulation events cannot be deduced easily. Huson & Bryant (2006) reviewed the terminology and interpretations used when defining different types of phylogenetic networks. The authors showed how a split network can depict confidence sets of trees and introduced a statistical test for determining whether the conflicting signal in a network is tree-like. Huson and Bryant have also developed the popular SplitTree program allowing for inferring different types of phylogenetic networks from sequences, distances and trees. Huson & Klöpper (2007) devised an algorithm to infer recombination events from binary sequences by using general reticulation

networks and galled trees. Huson & Scornavacca (2012) developed the Dendroscope 3 program to study rooted phylogenetic trees and networks. This program includes a number of algorithms for drawing and comparing rooted phylogenetic networks as well as for inferring them from a set of rooted trees. Chen & Wang (2012) described an algorithm for constructing hybridization networks from multiple conflicting gene trees. Albrecht et al. (2012) developed a fast parallel algorithm to infer a minimum hybridization network from two input trees.

Most of the existing network-building algorithms rely on the distance-based approach, but recently some new methods based on the maximum likelihood approach, which usually provides more accurate results, have started to appear. For instance, Solís-Lemus & Ané (2016) presented a new method for inferring pseudolikelihood phylogenetic networks from multi-locus genetic data. The main advantage of their method is that it accounts for incomplete lineage sorting in the framework of a coalescent model as well as for horizontal inheritance of genes through reticulation nodes in the network. Olave et al. (2017) have recently proposed a new fast method to detect hybridization in the presence of incomplete lineage sorting. The authors presented a new way to detect gene flow by evaluating the likelihood of various models with different levels of gene flow and assessing the expected gene tree discrepancy. In this article, we extend our previous distance-based algorithm for inferring explicit hybridization networks (Willems et al., 2014) by considering a maximum likelihood approach which will be applied to neighbor-joining (Saitou & Nei, 1987) tree configurations.

4.3 Methods

In this section, we recall the terminology which is necessary to define our new algorithm.

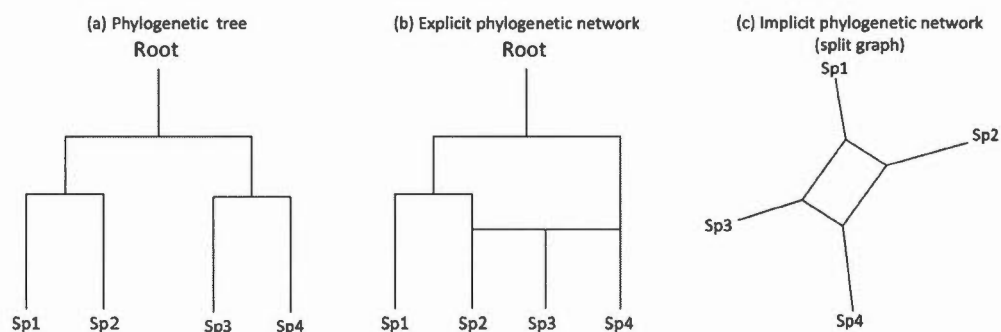


Figure 4.1 (a) A rooted phylogenetic tree; (b) A rooted phylogenetic (hybridization) network - here, Species 3 is a hybrid of Species 2 and 4; (c) An implicit phylogenetic network (split graph).

Trees and networks:

Figure 4.1 presents examples of a traditional phylogenetic tree defined on a set of 4 species (case a), of an explicit phylogenetic (hybridization) network (here Species 3 is a hybrid of Species 2 and 4)(case b), and of a split graph, which is an implicit phylogenetic network (case c).

Neighbor-joining:

Neighbor-joining (Saitou & Nei, 1987) is the most popular distance-based algorithm for inferring phylogenetic trees. Starting from a star tree, this clustering algorithm selects at each step the best neighbors i and j (according to the minimum evolution criterion) and replaces them by their common ancestor X . The minimum evolutionary criterion states that the optimal tree is the tree with the shortest total sum of edge lengths. The NJ algorithm requires as input a matrix of evolutionary distances between species at hand and returns as output a tree metric matrix (i.e., a metric that can be perfectly represented by a tree). NJ infers the correct phylogenetic tree if the input distances between species are sufficiently

close to the true evolutionary distances.

Maximum likelihood:

The maximum likelihood in a phylogenetic tree context can be defined as follows.

Input: n binary sequences of size L .

Let $\Pr(t, N_1, N_2)$ be the probability that character N_1 (DNA, amino acid or binary) changes into character N_2 during evolutionary time t . There exist several evolutionary models based on Markov processes. The likelihood of a tree can be given by the following formula:

$$\mathcal{L}(T) = \prod_{l=1}^L \mathcal{L}_l(T), \quad (4.1)$$

where $\mathcal{L}_l(T)$ is the sum of the probabilities of all possible evolutionary scenarios at position l .

F81 Model (Felsenstein (1981)): We will use the F81 evolutionary model originally defined by Felsenstein because it perfectly suits for describing the evolution of binary sequences (e. g., presence/absence of restriction sites or presence/absence of certain genes in genomes). Let π_0 (respectively π_1) be the proportion of 0's (respectively, 1's) in the input data. If $\beta = \frac{1}{1-\pi_0^2-\pi_1^2}$, then the transition probabilities are given by the following formulas:

$$\Pr(t, i, j) = \begin{cases} e^{-\beta t} + \pi_j (1 - e^{-\beta t}) & , \text{ if } i = j, \\ \pi_j (1 - e^{-\beta t}) & , \text{ if } i \neq j. \end{cases} \quad (4.2)$$

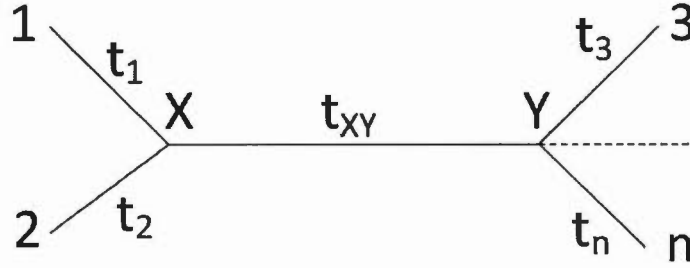


Figure 4.2 An intermediate NJ tree configuration used to compute the likelihood that Species 1 and 2 are neighbors. X and Y are the internal nodes of the presented intermediate NJ tree with n leaves and $n + 1$ edges.

Probability vectors:

For each species i , we consider its binary sequence as a probability vector of dimension L (probability of having 1 at position i): $P(i, l) = 0$ (respectively, 1), if the l th character of sequence i is equal to 0 (respectively, 1).

Likelihood of an NJ tree:

The maximum likelihood function to be used to estimate the likelihood of a given intermediate NJ tree configuration (i.e., configuration considered at a certain step of the NJ algorithm) can be defined as follows:

$$\mathcal{L}_{1,2}^T = \prod_{l=1}^L \left(\sum_{(\epsilon_X, \epsilon_Y) \in \{0,1\}^2} \left(P_{1X} P_{2X} P_{XY} \prod_{k=3}^n P_{Yk} \right) \right), \quad (4.3)$$

where:

- $P_{1X} = (1 - P(1, l))\Pr(t_1, 0, \epsilon_X) + P(1, l)\Pr(t_1, 1, \epsilon_X)$,
- $P_{2X} = (1 - P(2, l))\Pr(t_2, 0, \epsilon_X) + P(2, l)\Pr(t_2, 1, \epsilon_X)$,
- $P_{XY} = \Pr(t_{XY}, \epsilon_X, \epsilon_Y)$,
- $P_{Yk} = (1 - P(k, l))\Pr(t_k, 0, \epsilon_Y) + P(k, l)\Pr(t_k, 1, \epsilon_Y)$,

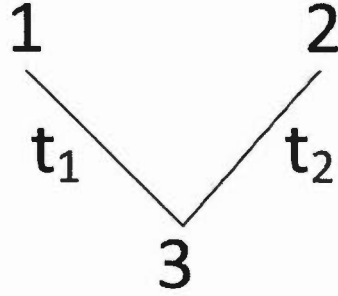


Figure 4.3 Configuration used to compute the likelihood that Species 3 is a hybrid of Species 1 and 2.

where all the notations correspond to Fig. 4.2.

Hybrid Likelihood:

Now, we can define the likelihood function to estimate the likelihood that Species 1, 2 and 3 are linked by a parents-hybrid relationship. In this case we define the probability that Species 3 is a hybrid of Species 1 and 2.

$$\mathcal{L}_{3,1,2}^H = \prod_{l=1}^L \left(\sum_{i \in \{1,2\}} \sum_{(\epsilon_i, \epsilon_3) \in \{0,1\}^2} P_i P_3 \Pr(t_i, \epsilon_i, \epsilon_3) \right), \quad (4.4)$$

where $P_i = (1 - \epsilon_i)(1 - P(i, l)) + \epsilon_i P(i, l)$, for $1 \leq i \leq 3$, and all the notations correspond to Fig. 4.3.

4.4 Main algorithm

In this section we describe the main algorithm allowing us to infer an implicit hybridization network, or a phylogenetic tree if no hybrids are present in the

data, from a set of n binary sequences (e. g., encoding presence-absence of certain genes in species genomes).

Algorithm:

Input: n binary sequences of size L , corresponding to n species.

1. $n_A = n$.
2. While ($n_A > 3$)
 - Determine the best pair of neighbors (i^*, j^*) according to NJ, over all possible pairs of current species i and j .
 - Compute the likelihood $\mathcal{L}_{i^*, j^*}^T$ of the NJ tree topology in which i^* and j^* are neighbors (see Fig. 4.2 and Formula 4.3).
 - Determine the best triplets h', i', j' and compute its maximum likelihood $\mathcal{L}_{h', i', j'}^H$ (where h' is a hybrid of i' and j' ; see Fig. 4.3 and Formula 4.4), over all possible triplets of current species h, i and j .
 - If $(\mathcal{L}_{h', i', j'}^H > \mathcal{L}_{i^*, j^*}^T)$, then species h' is considered as a hybrid of i' and j' , and we remove it from the data.
 - Else, species i^* and j^* are considered as neighbors, and we replace them by their direct common ancestor X determined by NJ.
 - $n_A = n_A - 1$.
3. We join the three last species.

Output: An explicit hybridization network with terminal nodes corresponding to the input species. Some of these terminal nodes will be identified as hybrids.

Remarks:

When we replace i and j by their common ancestor X , the l th component of the probability vector of X is computed by dividing $\mathcal{L}(X, l, 1)$ (the likelihood of having 1 at position l) by $\mathcal{L}(X, l, 1) + \mathcal{L}(X, l, 0)$.

Degree of hybridization: for each position, we compute the probability that the hybrid character comes from each of the parents.

Edge lengths are optimized using the Newton-Raphson method in the maximum likelihood computations.

We also corrected the likelihood values using a Bayesian Information Criterion (Schwarz & Gideon, 1978) in the following way:

$$- 2\mathcal{L}_{i,j}^T + (n+1)\ln(2L),$$

$$- 2\mathcal{L}_{h,i,j}^H + 2\ln(3L),$$

since there are $n+1$ degrees of freedom and $2L$ data in an NJ tree with n leaves (see Fig. 4.2), and 2 degrees of freedom and $3L$ data in the hybrid configuration (see Fig. 4.3). This refined version of our algorithm was also tested in our simulations (see the next section).

4.5 Results of simulations

In our simulations, we first generated random phylogenetic trees using the tree generation algorithm available on the T-REX website (Boc et al., 2012). This algorithm requires as input the size of the tree, n , as well as the average edge length, and returns as output a random binary phylogenetic tree with n leaves that is built according to the method of Kuhner & Felsenstein (1994). In total, we generated 1000 unrooted phylogenetic trees for each of the following tree sizes: $n = 8$, $n = 16$ and $n = 32$. Then, we simulated the evolution of binary sequences of length $L = 1000$ along these trees using the F81 evolutionary model (Felsenstein, 1981). For each phylogenetic tree generated this way, we obtained n binary sequences (corresponding to the tree leaves) of size L . The hybrids were added to the data afterwards. To add the hybrids, we first randomly selected two integers $1 \leq i < j \leq n$. Let α be the selected degree of hybridization (see Willems et al.

(2014) for more details on this hybridization parameter). We generated a new hybrid sequence with the first $\alpha \times L$ binary characters of sequence i to which we added the last $(1 - \alpha) \times L$ binary characters of sequence j . This new hybrid sequence was added to the n original sequences. In our simulations, we considered the following values of the hybridization parameter: $\alpha = 0.3$, $\alpha = 0.4$ and $\alpha = 0.5$ (for each tree considered in our simulations, the value of α was selected randomly, following the uniform distribution). The number of hybrids added to trees ranged from 0 to 5. Thus, for each considered tree size, n , we obtained 1000 binary sequence alignments corresponding to the original trees and 6000 matrices corresponding to phylogenetic networks having 0 to 5 hybrids. Our first simulation was carried out using: (1) our previous distance-based method (Willems et al., 2014), (2) our new algorithm described in this paper, and (3) its refined version in which we corrected the likelihood function by means of a Bayesian Information Criterion (see the previous section). The results of this first simulation are shown in Figs. 4.4.

Moreover, for the trees with 8 leaves (i.e., species), we also simulated the data with different lengths of the binary sequences. Specifically, the sequences of the following sizes: 20, 50, 100, 200, 500 and 1000, were considered. This second simulation was also conducted with phylogenetic networks including 0 to 5 hybrids. The results of this additional simulation are presented in Figs. 4.5.

The results of the first simulation demonstrate that the new hybrid detection method presented here clearly outperformed our previous distance-based method (Willems et al. (2014)) in terms of the false positive rate. However, the two methods showed very close results in terms of the true positive rate. Furthermore, the refined version of our new method, in which we corrected the likelihood with a Bayesian Information Criterion, provided the best overall results in terms of the true positive rate and the second best overall results, after our new method,

in terms of the false positive rate. Our second simulation confirmed the trends observed in the first one: The refined version of our new method was much better than the new method itself in terms of the true positive rate, but was not good enough in terms of the false positive rate. These trends are especially noticeable when short binary sequences (i.e., when L ranged between 20 and 200) are considered.

4.6 Analysis of the mosquitoes dataset

To test our new method on real data, we considered the dataset of restriction maps of the rDNA cistron of 16 species of mosquitoes constructed using eight recognition restriction enzymes (Kumar et al., 1998). A total of 26 sites were scored. The original binary sequence data are reported in Tab. 2.5.

Huson & Klöpper (2007) have constructed a galled network (i.e., a specific type of recombination phylogenetic network) for this dataset and found that these data encompass 4 hybrids (see Fig. 2.14). In Chapter 2, we used the Hamming distances to transform the original mosquitoes data into a distance matrix before applying our distance-based network reconstruction method (see Fig. 2.15). We found 4 hybrids (i.e., 4 reticulations in the network) as that by Huson and Klöpper.

After the application of the new method described in this chapter, we obtained an explicit phylogenetic network with 5 hybrids (see Fig. 4.6). It is worth noting that the general structure of our network is very similar to that of the galled network by Huson and Klöpper as well as to that of the reticulation network found with our distance-based algorithm. However, we can also observe some permutations between hybrids and parents in these networks. For example, species *Aedes polynesiensis* and *Aedes triseriatus* have been identified as parent species in the network found in Chapter 2, but have been identified as hybrids in the network provided by our new method. Also, species *Tripteroides bambusa* has

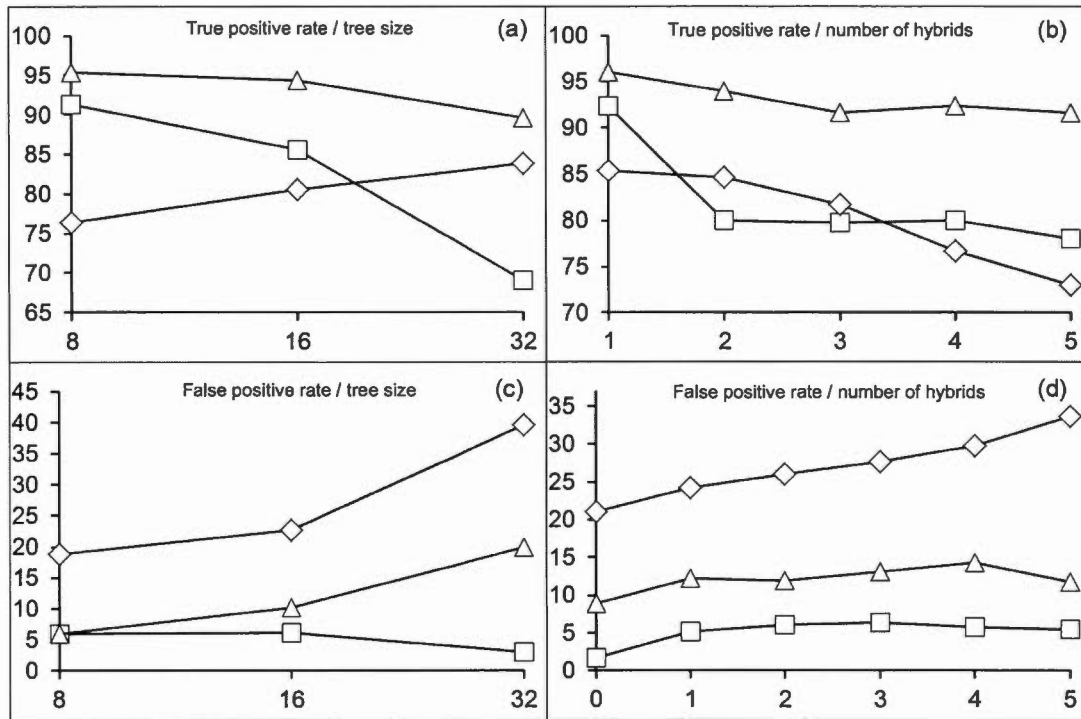


Figure 4.4 True positive and false positive rates with respect to the tree size (cases a and c) and the number of hybrids (cases b and d) obtained in simulations with 0 to 5 hybrids for trees with 8, 16 and 32 leaves and binary sequences of size 1000 using: (1) our previous distance-based method (Willems et al. (2014); $\diamond$), (2) our new method ($\square$), and (3) its refined version in which we corrected the likelihood with a Bayesian Information Criterion ($\triangle$). The averages over all parameter combinations except the fixed one (tree size or number of hybrids) are shown.

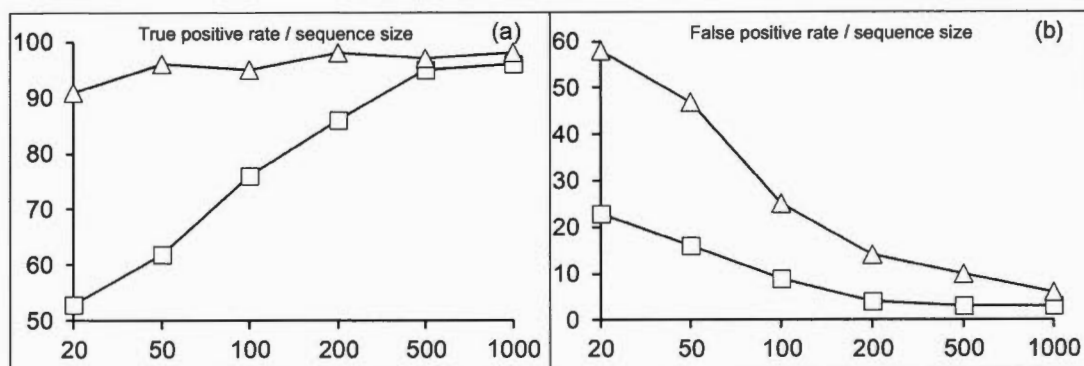


Figure 4.5 True positive (a) and false positive (b) rates shown with respect to the sequence size. These results were obtained from simulations with 0 to 5 hybrids and trees with 8 leaves. The sequences of sizes: 20, 50, 100, 200, 500 and 1000 were analyzed. The simulations were conducted using: (1) our new method (□) and (2) its refined version in which we corrected the likelihood with a Bayesian Information Criterion (△). The averages over all parameter combinations except the fixed one (sequence size) are shown.

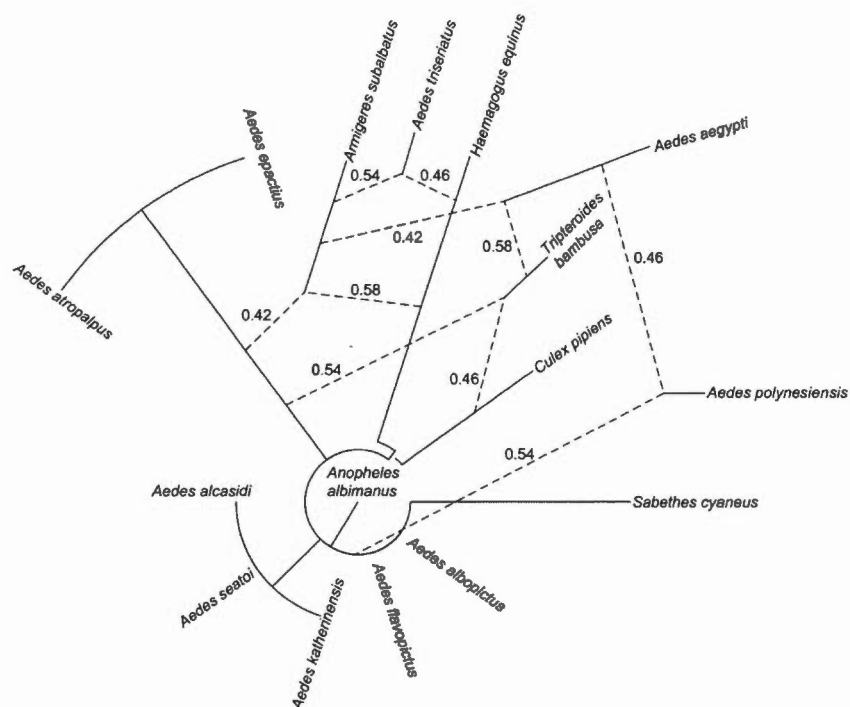


Figure 4.6 Network obtained for the restriction map of the 16-species mosquitoes dataset (Kumar et al., 1998) using our new method. Five hybrids, linked to the rest of the network by dashed lines, were identified. The numbers on dashed network edges represent the respective contributions of parents to hybrids.

been identified as a hybrid in the phylogenetic network of Fig. 4.6, but not in that of Fig. 2.15.

The key advantage of our network representation, as well as of that described in Chapter 2, over galled networks is that our methods identify hybrids and their parents explicitly.

4.7 Conclusion

In this paper, we have introduced a new accurate method for inferring explicit hybridization networks from presence-absence data (i.e., binary sequences). These networks can be used to represent adequately reticulate phylogenetic relationships between species, including for example, hybridization events and horizontal gene transfers between species. Most of the existing algorithms developed in this context return as output only an implicit phylogenetic network, which is often very difficult to interpret. Our new method infers explicit hybridization networks using both the maximum likelihood and neighbor-joining approaches. This method can be helpful for building and interpreting phylogenetic networks for different types of binary sequences associated with species at hand, e. g., presence/absence of certain genes in genomes or presence/absence of restriction sites. Our simulations showed that the new method, and especially its modified version that uses a maximum likelihood correction by a Bayesian Information Criterion, outperforms the distance-based procedure by Willems et al. (2014) in terms of both the true positive and false positive rates, regardless the number of species and hybrids in the dataset. Another advantage of the method presented here, compared to the distance-based method by Willems et al., is that our new method does not use the additional hybrid selection threshold parameter which should be specified by the user in the distance-based method. The main drawback of the new method is that it is slower than the distance-based method by Willems et al. (2014). This is a general disadvantage of maximum-likelihood-based methods which are usually more accurate, but also slower, than their distance-based counterparts.

CHAPITRE V

CONCLUSION

Le modèle classique de l'arbre phylogénétique ne permet pas de représenter adéquatement certains mécanismes réticulés de l'évolution des espèces (Doolittle, 1999; Legendre, 2000), comme l'hybridation ou le transfert horizontal de gènes, par exemple. Depuis une vingtaine d'années, de nombreuses méthodes ont ainsi été développées pour inférer des réseaux phylogénétiques censés représenter l'évolution réticulée des espèces (Huson et al., 2010). Cependant, la plupart de ces méthodes produisent des réseaux implicites, qui sont souvent très difficiles à interpréter. Par ailleurs, de nombreuses analogies entre l'évolution des espèces et l'histoire des langues permettent d'utiliser des approches phylogénétiques en linguistique historique (Atkinson & Gray, 2005). Il est donc pertinent d'utiliser des algorithmes d'inférence de réseaux phylogénétiques pour représenter l'histoire de certaines langues, tout en prenant certaines précautions méthodologiques.

Dans cette thèse, nous avons développé deux nouveaux algorithmes pour inférer des réseaux d'hybridation explicites. Le premier prend en entrée une matrice de distances entre n espèces données, tandis que le deuxième prend en entrée n séquences binaires associées aux n espèces étudiées. Dans les deux cas, on obtient en sortie un réseau phylogénétique explicite dont les feuilles sont en correspondance avec les n espèces de départ. Dans les réseaux que nous construisons, chaque

hybride est clairement identifié avec ses deux parents et le degré d'hybridation correspondant à chacun des deux parents. Les nombreuses simulations que nous avons effectuées démontrent l'efficacité de nos nouveaux algorithmes pour la détection de phénomènes d'hybridation.

Notre méthode de distances est beaucoup plus rapide que notre méthode probabiliste, et peut être utilisée avec n'importe quel type de séquences, à condition qu'on puisse les comparer à l'aide d'une distance. Notre deuxième méthode ne peut, quant à elle, être utilisée que dans un cadre beaucoup plus restreint, mais donne des résultats sensiblement meilleurs, malgré des temps de calcul plus importants.

Nous avons appliqué notre méthode de distances dans le contexte de la biolinguistique, en l'adaptant préalablement. Nous avons ainsi obtenu un réseau d'hybridation représentant l'évolution de 84 langues indo-européennes. Un des intérêts principaux de notre méthode est de quantifier l'apport lexical entre les langues hybrides et chacun de leurs deux parents. De plus, les réseaux obtenus avec d'autres méthodes se sont avérés beaucoup plus difficiles à interpréter que notre réseau d'hybridation explicite. Nous avons également présenté des justifications historiques et linguistiques pour la plupart des réticulations identifiées. Nos résultats pourraient être raffinés si nous pouvions analyser, en plus, certaines données syntaxiques et morphologiques disponibles. Quelques études récentes ont en effet démontré que de telles données pouvaient révéler des relations entre certaines langues, qui ne sont pas détectables à partir du lexique seul (Longobardi et al., 2013). Il serait par ailleurs pertinent d'utiliser notre méthode pour dater certains des événements identifiés dans notre réseau.

De nombreux prolongements de ces travaux peuvent être envisagés, notamment dans le cadre de notre méthode basée sur les caractères, que nous avons présentée

dans le chapitre 4.

Tout d'abord, nous pourrions utiliser cette méthode dans le cadre de la biolinguistique, puisque les données initiales sont des séquences binaires (c.-à-d., présence/absence des langues dans les ensembles de cognats). Il serait ainsi intéressant de construire un réseau d'hybridation pour les 84 langues indo-européennes disponibles et de le comparer avec celui obtenu avec notre méthode de distances, qui est présenté dans le chapitre 3. Les séquences étant cependant très divergentes d'un groupe linguistique à un autre, il faudra probablement travailler sur des sous-ensembles de langues, puis recoller les différents réseaux obtenus.

Par ailleurs, nous pourrions adapter notre algorithme probabiliste pour prendre en entrée des séquences d'ADN ou d'acides aminés. Il semble cependant difficile de calculer la vraisemblance qu'une séquence d'ADN (par exemple) soit l'hybride de deux autres séquences. En effet, les caractères ne peuvent pas être considérés comme indépendants dans ce contexte, puisqu'ils sont transmis par blocs. Il faudrait donc introduire un score d'hybridation pour chaque triplet de séquences. Le calcul du score d'un triplet pourrait, par exemple, être basé sur la recherche de blocs ayant un degré de similarité différent, selon que l'on compare la troisième séquence avec la première ou avec la deuxième. Une fois le meilleur hybride ainsi identifié, il faudrait également définir un critère statistique pour choisir, à chaque étape de l'algorithme, entre un événement de spéciation et un événement d'hybridation.

Il serait également intéressant d'étudier la qualité des arbres phylogénétiques construits par notre programme. On pourrait par exemple effectuer une simulation dans laquelle on inférerait un réseau d'hybridation, puis on retirerait les hybrides identifiés, et on générerait un arbre phylogénétique à l'aide de RAxML (Stamatakis et al., 2012). On pourrait alors comparer la vraisemblance des arbres

ainsi obtenus.

Par ailleurs, notre algorithme probabiliste détermine à chaque pas où on joint deux voisins, un vecteur de probabilités correspondant à la séquence de l'ancêtre de ces deux voisins. On obtient ainsi une reconstruction des séquences ancestrales sous la forme d'une distribution de probabilités. Ces résultats pourraient être comparés, par exemple, avec ceux du programme Ancestor (Diallo et al., 2010).

Notons que pour comparer la vraisemblance d'un couple de voisins avec celle d'un triplet d'hybridation, puisque la taille des données ainsi que le nombre de degrés de liberté ne sont pas identiques, nous avons essayé plusieurs critères statistiques, notamment les critères BIC (Schwarz & Gideon, 1978) et AIC (Akaike, 1974). Nous avons finalement retenu le premier, qui donnait de meilleurs résultats. Nous comptons en essayer d'autres (voir par exemple (Burnham & Anderson, 2003) pour une revue exhaustive des critères existants) afin d'optimiser notre algorithme.

De plus, dans le cadre des deux algorithmes que nous avons développés, nous pourrions rajouter comme paramètre d'entrée le nombre maximal d'hybrides souhaité dans le réseau d'hybridation inféré. Il existe en effet plusieurs algorithmes pour déterminer un nombre d'hybrides à partir de différents types de données (Piovesan & Kelk, 2013; Iersel et al., 2014).

ANNEXE A

CODE EN C++ DU PROGRAMME QUI IMPLÉMENTE L'ALGORITHME PRÉSENTÉ DANS LE CHAPITRE 2

Nous présentons dans cette annexe le code du programme qui implémente l'algorithme présenté dans le chapitre 2 (Willems et al., 2014), disponible gratuitement à l'adresse URL suivante :

http://www.info2.uqam.ca/~makarenikov_v/makarenv/hybrids_detection.zip

Ce programme permet de calculer un réseau d'hybridation explicite (un arbre d'espèces auquel on rajoute des hybrides ayant chacun deux parents, ainsi qu'un degré d'hybridation α associé à un de ces deux parents) à partir d'une matrice de distances entre espèces. Notons qu'un hybride peut consister en un groupe d'espèces.

Notre algorithme requiert $O(n^3)$ opérations pour déterminer un réseau d'hybridation à partir d'une matrice de distances de taille $(n \times n)$.

Le programme est distribué sous la forme d'un code source en C++. Il prend en entrée deux fichiers qui doivent être placés dans le même dossier que le fichier exécutable. Le premier fichier "Matrix.txt" doit contenir la matrice de distances au format Phylip (sans limites pour la longueur des noms d'espèces, chaque nom devant finir par un espace), et le fichier "Parameters.txt" contient les trois para-

mètres suivants : MIN - niveau d'acceptation des hybrides défini dans la proposition 2.2, ALPHAMIN - valeur minimale du degré d'hybridation α et ALPHAMAX - valeur maximale de α).

Le fichier de sortie contient la liste des hybrides (avec leurs parents et le degré d'hybridation correspondant au premier parent) et une matrice de distances correspondant à l'arbre phylogénétique des espèces non hybrides.

```

/*
 * File:    main.cpp
 * Author:  wilm29107705
 *
 * Created on 28 janvier 2014, 13:01
 */
#include <stdlib.h>
#include <stdio.h>
#include <math.h>
#include <time.h>
#include <string.h>
#include <vector>
#include<iostream>
#include<fstream>
#include <cstdlib>
using namespace std;
/*
 *
 */
int main()
{
FILE * input1;
FILE * input2;
FILE * output;
input1 = fopen ("Parameters.txt","r");
input2 = fopen ("Matrix.txt","r");
output = fopen ("Network.txt","w");
double MINMIN,alphaMIN,alphaMAX;
int n;

```

```

char s[10];

fscanf(input2, "%d", &n);
fscanf(input1, "%s", &s);
fscanf(input1, "%lf", &MINMIN);
fscanf(input1, "%s", &s);
fscanf(input1, "%lf", &alphaMIN);
fscanf(input1, "%s", &s);
fscanf(input1, "%lf", &alphaMAX);

double **D,**DA,*T1,*S,*LP,Som,Smin,Sij,L;
double Lii,Ljj,L1,D1,D2,EX,EY,EXY,EX2;
double LSMAX2,LS2,LSN,LS2HV,LSMIN,alphaH,b,a;
double SigmaI,SigmaJ,SigmaH,Lhh,LH,l1,l2,l3;
double C1,C2,AA,MINL,MINH,MAX1;
int *T,*H,i,j,h,k,l,ii,jj,n1,nA,iH,jH,hH,nHyb,nbSP;
bool Hybrid,Tree;

nA=n;
nHyb=0;

D=(double **) malloc((n+1)*sizeof(double*));
DA=(double **) malloc((n+1)*sizeof(double*));
T1=(double *) malloc((n+1)*sizeof(double));
S=(double *) malloc((n+1)*sizeof(double));
LP=(double *) malloc((n+1)*sizeof(double));
T=(int *) malloc((n+1)*sizeof(int));
H=(int *) malloc((n+1)*sizeof(int));

char **SP = NULL;

```

```

SP = (char**) malloc ((n+1)* sizeof(char*));
for(i=0 ; i<=n; i++)
    SP[i] = (char*) malloc (30 * sizeof(char));

/* Input Matrix */
for (i=0;i<=n;i++)
{
    D[i]=(double*) malloc ((n+4)*sizeof(double));
    if (D[i]==NULL)
    {
        printf("probleme_de_memoire");
        exit(1);
    }
}

/* Output Matrix */
for (i=0;i<=n;i++)
{
    DA[i]=(double*) malloc ((n+1)*sizeof(double));
    if (DA[i]==NULL)
    {
        printf("probleme_de_memoire");
        exit(1);
    }
}

for (i=1;i<=n;i++)
{
    for (j=1;j<=n;j++)

```

```

        {
            D[i][j]=0;
        }
    }

    for (i=1;i<=n;i++)
    {
        for (j=1;j<=n;j++)
        {
            DA[i][j]=0;
        }
    }

    for (i=1;i<=n;i++)
    {
        fscanf(input2, "%s", &*SP[i]);
        for (j=1;j<=n;j++)
        {
            fscanf(input2, "%lf", &D[i][j]);
        }
    }

    /*      Initialisation      */

    L=0; Som=0;

    for (i=1;i<=n;i++)
    {
        S[i]=0; LP[i]=0;
    }

```

```

    for (j=1;j<=n;j++)
    {
        S[i]=S[i]+D[i][j];
    }
    Som=Som+S[i]/2.0;
    T[i]=i;
    H[i]=0;
    T1[i]=0;
}

/* Main procedure */
n1=n;
while (n1>3)
{
    /* Research of neighbors */

    Smin=2*Som;
    for (i=1;i<=n1-1;i++)
    {
        for (j=i+1;j<=n1;j++)
        {
            Sij=2*Som-S[i]-S[j]+D[i][j]*(n1-2);
            if (Sij<Smin)
            {
                Smin=Sij;
                ii=i;
                jj=j;
            }
        }
    }
}

```

```
}
```

```
/* Research of a potential hybrid */
```

```
MAX1=-1000000;
```

```
Tree=1;
```

```
Hybrid=0;
```

```
for (h=1;h<=n1;h++)
```

```
{
```

```
    if ((h!=ii)&&(h!=jj))
```

```
    {
```

```
        LS2=0;
```

```
        k=1;
```

```
        while (k<=n1)
```

```
        {
```

```
            if ((k!=ii)&&(k!=jj)&&(k!=h))
```

```
            {
```

```
                D1=D[h][jj]+D[k][ii];
```

```
                D1=D1-D[ii][h]-D[k][jj];
```

```
                LS2=LS2+D1;
```

```
            }
```

```
            k++;
```

```
        }
```

```
    if ((LS2<-0.0000001)&&(LS2<-MAX1))
```

```
    {
```

```
        MAX1=-LS2;
```

```
        Hybrid=1;
```

```

        hH=h;
        iH=jj;
    }

    else if((LS2>0.0000001)&&(LS2>MAX1))
    {
        MAX1=LS2;
        hH=h;
        Hybrid=1;
        iH=ii;
    }
}

LSMAX2=MAX1*MAX1/(n1-2.0)/(n1-2.0);

/* Research of a the second parent of the hybrid */

if(Hybrid)
{
    h=hH;
    i=iH;
    Hybrid=0;
    LSMIN=10000000.0;
    for (j=1;j<=n1;j++)
    {
        if((j!=ii)&&(j!=jj)&&(j!=h))
        {
            k=1;
            MINL=10000000;

```

```

while(k<=n1)
{
    if ((k!=i)&&(k!=j)&&(k!=h))
    {
        D1=D[i][j]+D[k][h]-D[i][h]-D[k][j];
        D2=D[i][j]+D[k][h]-D[j][h]-D[k][i];
        if (D1<MINL)
            MINL=D1;
        if (D2<MINL)
            MINL=D2;
    }
    k++;
}

if ((MINL>MINMIN))
{
    SigmaI=0;SigmaJ=0;SigmaH=0;
    for (k=1;k<=n1;k++)
    {
        if ((k!=i)&&(k!=j)&&(k!=h))
        {
            SigmaI=SigmaI + D[k][i];
            SigmaJ=SigmaJ + D[k][j];
            SigmaH=SigmaH + D[k][h];
        }
    }
    SigmaI=SigmaI/(n1-3.0);
    SigmaJ=SigmaJ/(n1-3.0);
    SigmaH=SigmaH/(n1-3.0);
}

```

```

EX = 0;EXY = 0;EX2 = 0;
for (k=1;k<=n1;k++)
{
if ((k!=i)&&(k!=j)&&(k!=h))
{
D1=D[h][k]-D[j][k];
D2=D[i][k]-D[j][k];
D2=D2-SigmaI+SigmaJ;
EX = EX+D2;
EXY = EXY+D1*D2;
EX2 = EX2+D2*D2;
}
}
a=(EXY+EX*(SigmaJ-SigmaH))/EX2;
b=SigmaH-SigmaJ;
AA=a*SigmaI + (1.0-a)*SigmaJ;
AA=(AA - SigmaH)/2.0;
LH=D[i][h]+D[j][h]-D[i][j];
LH=LH/2.0;
C1=(D[i][h]-(1.0-a)*D[i][j])/2.0;
C1=AA+C1;
C2=(D[j][h]-a*D[i][j])/2.0;
C2=AA+C2;
LS2=0;LSN=0;
for (k=1;k<=n1;k++)
{
if ((k!=i)&&(k!=j)&&(k!=h))
{
D1=D[h][k]-D[j][k];

```

```

D2=D[i][k]-D[j][k];
D2=D2-SigmaI+SigmaJ;
LS2=LS2+(D1-b-a*D2)*(D1-b-a*D2);
}
}
LSN=LS2/(n1-3.0);
if (LSN<LSMIN))
{
if (LSN<(LSMAX2+0.00001)
{
if (a<alphaMAX)
{
if (a>alphaMIN)
{
Hybrid=1;
LSMIN=LSN;
jH=j;
Lhh=LH;
alphaH=a;
MINH=MINL;
}}}}
}
}

/* If we choose a hybrid */

if ((Hybrid)&&(n1>4))
{

```

```

Tree=0;
nHyb++;
nbSP=0;
if (Lhh<0.00001) Lhh=0.00005;

/* Updating of H */

fprintf(output, "%s", "(");
for (i=1; i<=n; i++)
{
    if (T[i]==hH)
    {
        H[i]=nHyb;
        fprintf(output, "%s", SP[i]);
        fprintf(output, "%s", ",_");
        nbSP++;
    }
}

fseek(output, -2, SEEK_CUR);
fprintf(output, "%s", ")");
fprintf(output, "%s", "_");
fprintf(output, "%s", "is_Hybrid_of_");
fprintf(output, "%s", "(");
for (i=1; i<=n; i++)
{
    if (T[i]==iH)
    {
        fprintf(output, "%s", SP[i]);
        fprintf(output, "%s", ",_");
    }
}

```

```

}
fseek(output, -2, SEEK_CUR);
fprintf(output, "%s", " ");
fprintf(output, "%s", "_and_");
fprintf(output, "%s", "(");
for (i=1; i<=n; i++)
{
    if (T[i]==jH)
    {
        fprintf(output, "%s", SP[i]);
        fprintf(output, "%s", ",_");
    }
}
fseek(output, -2, SEEK_CUR);
fprintf(output, "%s", " ");
fprintf(output, "%s", "_");
fprintf(output, "%s", "\r");
fprintf(output, "%s", "with_degree_of_hybridization_");
fprintf(output, "%lf", alphaH);
fprintf(output, "%s", "_and_LH=_");
fprintf(output, "%lf", Lhh);
fprintf(output, "%s", "\r");
fprintf(output, "%s", "\r");

if(nbSP==2)
{
    fprintf(output, "%s", "Branch_lengths_");
    fprintf(output, "%s", "\r");
}

```

```

for (j=1;j<=n;j++)
{
    if (H[j]==nHyb)
    {
        fprintf(output,"%s", SP[j]);
        fprintf(output,"%s", "_");
        fprintf(output,"%lf", T1[j]);
        fprintf(output,"%s", "\r");
    }

}

fprintf(output,"%s", "\r");
fprintf(output,"%s", "\r");
}

if (nbSP>2)
{
    fprintf(output,"%s", "Hybrids_additive_distance_matrix_");
    fprintf(output,"%s", "\r");

    for (j=1;j<=n;j++)
    {
        if (H[j]==nHyb)
        {
            fprintf(output,"%s", SP[j]);
            fprintf(output,"%s", "_");
            for (i=1;i<=n;i++)
            {
                if ((H[i]==nHyb))

```

```

        {
            fprintf(output, "%lf", DA[i][j]);
            fprintf(output, "%s", " ");
        }
    }
    fprintf(output, "%s", "\r");
}
}

fprintf(output, "%s", "\r");
fprintf(output, "%s", "\r");
}

```

/ Updating of D */*

```

Som=Som-S[hH];
for (i=1;i<=n1;i++)
{
    if ((i!=hH))
    {
        S[i]=S[i]-D[i][hH];
    }
}

if (hH!=n1)
{
    for (i=1;i<=(n1-1);i++)
    {
        D[i][hH]=D[i][n1]; D[hH][i]=D[n1][i];
    }
}

```

```

    }
    D[hH][hH]=0;S[hH]=S[n1];LP[hH]=LP[n1];
    }

    for (i=1;i<=n;i++)
    {
        if (T[i]==hH) T1[i]=T1[i]+Lhh;
    }

    /* Updating of DA */

    if (hH!=n1)
    {
        for (j=1;j<=n;j++)
            if (T[j]==hH) T[j]=n+1;

        for (j=1;j<=n;j++)
            if (T[j]==n1) T[j]=hH;
    }

    nA--;
    n1--;
}

}

/* If we choose a tree-joining */
if(Tree)
{
    /* New clustering */

```

```

Lii=(D[ ii ][ jj ]+(S[ ii ]-S[ jj ])/(n1-2))/2-LP[ ii ];
Ljj=(D[ ii ][ jj ]+(S[ jj ]-S[ ii ])/(n1-2))/2-LP[ jj ];

```

```

/* Updating of D */

```

```

    if (Lii < 0.00001) Lii = 0.00005;
    if (Ljj < 0.00001) Ljj = 0.00005;
    L=L+Lii+Ljj;
    LP[ ii ]=0.5*D[ ii ][ jj ];
    Som=Som-(S[ ii ]+S[ jj ])/2;

    for ( i=1;i<=n1;i++)
    {
        if ((i!=ii)&&(i!=jj))
        {
            S[ i ]=S[ i ]-0.5*(D[ i ][ ii ]+D[ i ][ jj ]);
            D[ i ][ ii ]=(D[ i ][ ii ]+D[ i ][ jj ])/2;
            D[ ii ][ i ]=D[ i ][ ii ];
        }
    }
    D[ ii ][ ii ]=0;
    S[ ii ]=0.5*(S[ ii ]+S[ jj ])-D[ ii ][ jj ];

    if (jj!=n1)
    {
        for ( i=1;i<=(n1-1);i++)
        {
            D[ i ][ jj ]=D[ i ][ n1 ];
            D[ jj ][ i ]=D[ n1 ][ i ];
        }
    }

```

```

        D[ jj ][ jj ]=0;
        S[ jj ]=S[ n1 ];
        LP[ jj ]=LP[ n1 ];
    }

```

/ Updating of DA */*

```

    for ( i=1;i<=n;i++)
    {
        if (T[ i]==ii ) T1[ i]=T1[ i]+ Lii ;
        if (T[ i]==jj ) T1[ i]=T1[ i]+ Ljj ;
    }
    for ( j=1;j<=n;j++)
    {
        if (T[ j]==jj )
        {
            for ( i=1;i<=n;i++)
            {
                if (T[ i]==ii )
                {
                    DA[ i ][ j]=T1[ i]+T1[ j ]; DA[ j ][ i]=DA[ i ][ j ];
                }
            }
        }
    }

    for ( j=1;j<=n;j++)
        if (T[ j]==jj ) T[ j]=ii ;

    if ( jj!=n1 )

```

```

        {
            for (j=1;j<=n;j++)
                if (T[j]==n1) T[j]=jj;
        }
        n1--;
    }
}

/*Joining the 3 objects remaining */

l1=(D[1][2]+D[1][3]-D[2][3])/2-LP[1];
l2=(D[1][2]+D[2][3]-D[1][3])/2-LP[2];
l3=(D[1][3]+D[2][3]-D[1][2])/2-LP[3];
if (l1<0.00001) l1=0.00005;
if (l2<0.00001) l2=0.00005;
if (l3<0.00001) l3=0.00005;
L=L+l1+l2+l3;

for (j=1;j<=n;j++)
{
    for (i=1;i<=n;i++)
    {
        if ((T[j]==1)&&(T[i]==2))
        {
            DA[i][j]=T1[i]+T1[j]+l1+l2;
            DA[j][i]=DA[i][j];
        }
        if ((T[j]==1)&&(T[i]==3))
        {
            DA[i][j]=T1[i]+T1[j]+l1+l3;

```

```

        DA[j][i]=DA[i][j];
    }
    if ((T[j]==2)&&(T[i]==3))
    {
        DA[i][j]=T1[i]+T1[j]+l2+l3;
        DA[j][i]=DA[i][j];
    }
}
DA[j][j]=0;
}
fprintf(output, "%s", "\r");
fprintf(output, "%s", "Main_tree_additive_distance_matrix");
fprintf(output, "%s", "\r");

for (j=1;j<=n;j++)
{
    if (H[j]==0)
    {
        fprintf(output, "%s", SP[j]);
        fprintf(output, "%s", "_");
        for (i=1;i<=n;i++)
        {
            if ((H[i]==0))
            {
                fprintf(output, "%lf", DA[i][j]);
                fprintf(output, "%s", "_");
            }
        }

        fprintf(output, "%s", "\r");
    }
}

```

```

        }

    }

    free(T);
    free(H);
    free(T1);
    free(S);
    free(LP);
    for (i=0;i<=n;i++)
    {
        free(D[i]);
    }
    free(D);

    for (i=0;i<=n;i++)
    {
        free(DA[i]);
    }
    free(DA);

    for (i=0;i<=n;i++)
    {
        free(SP[i]);
    }
    free(SP);
    return 0;
}

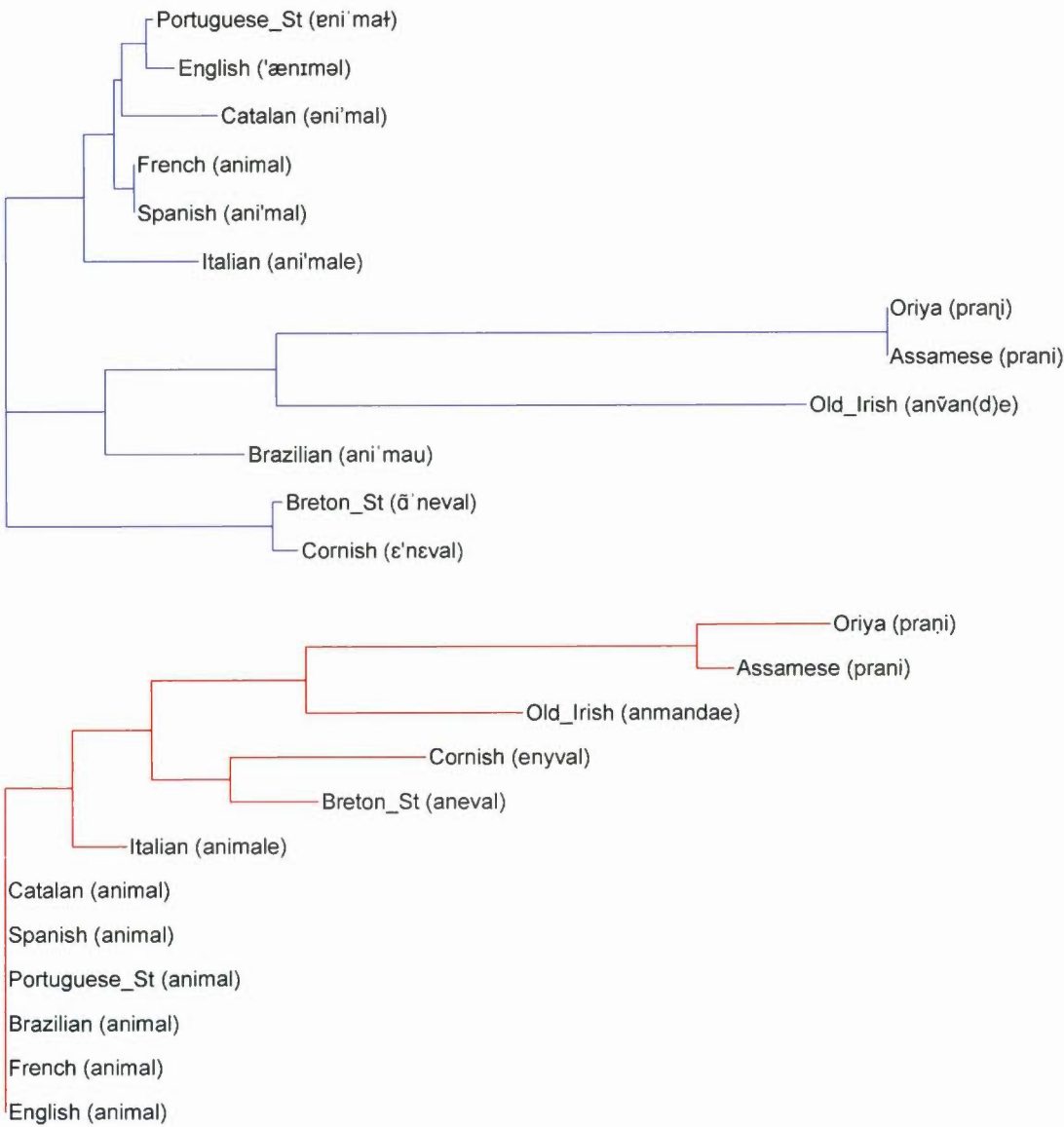
```

ANNEXE B

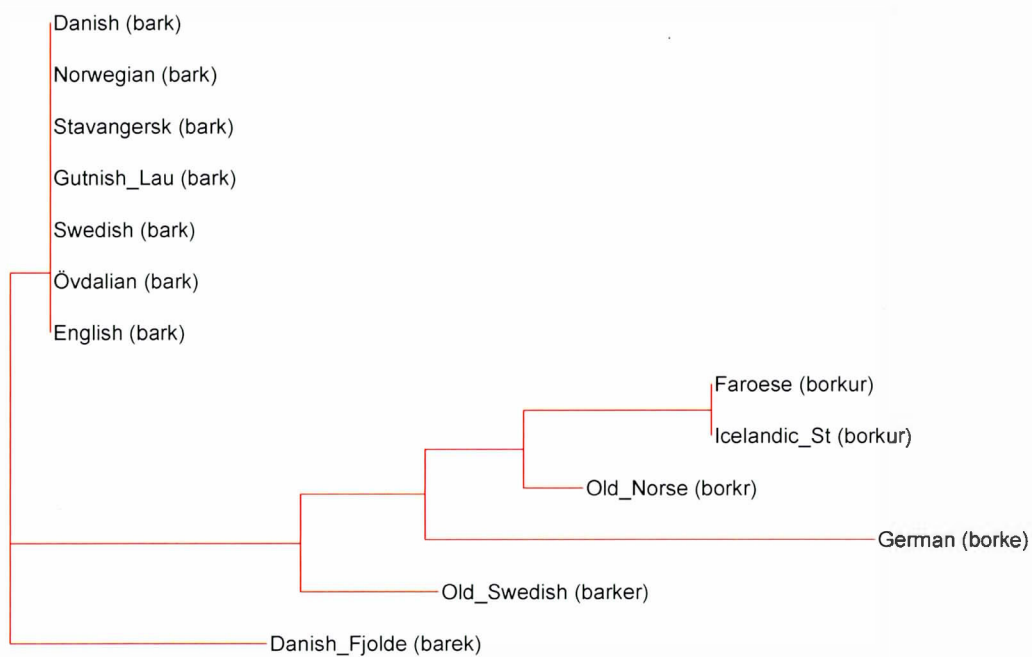
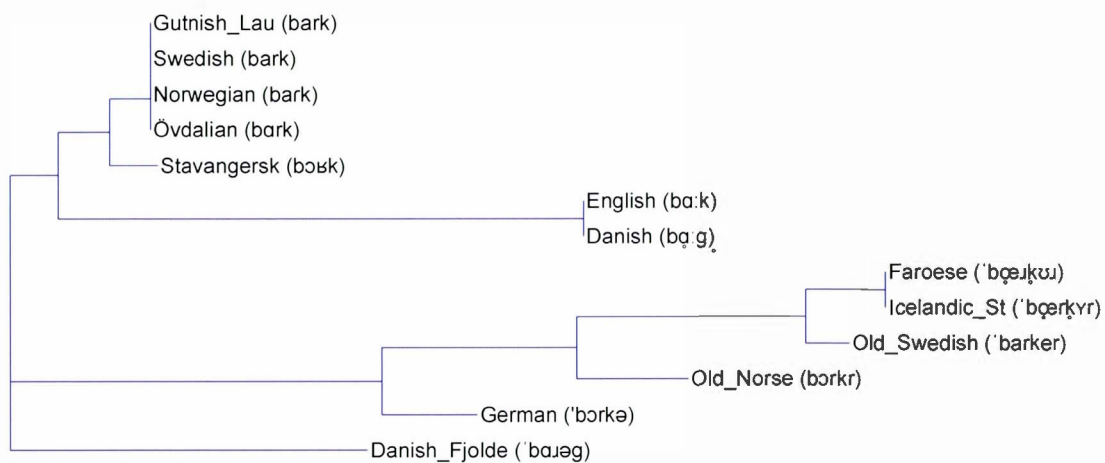
COMPARAISON ENTRE LA DISTANCE SCA ET LA DISTANCE DE LEVENSHTTEIN

Nous présentons dans cette annexe les arbres phylogénétiques de 28 mots parmi ceux utilisés pour les analyses biolinguistiques du chapitre 3. Les arbres en bleu ont été obtenus en utilisant la distance phylogénétique SCA (List, 2012b) et ceux en rouge, en utilisant la distance de Levenshtein normalisée. Ces arbres ont été inférés à partir des matrices de distances avec l'algorithme neighbor-joining (Saitou & Nei, 1987). Toutes les données utilisées proviennent de la base de données IELex (<http://ielex.mpi.nl/>) (Dunn, 2012).

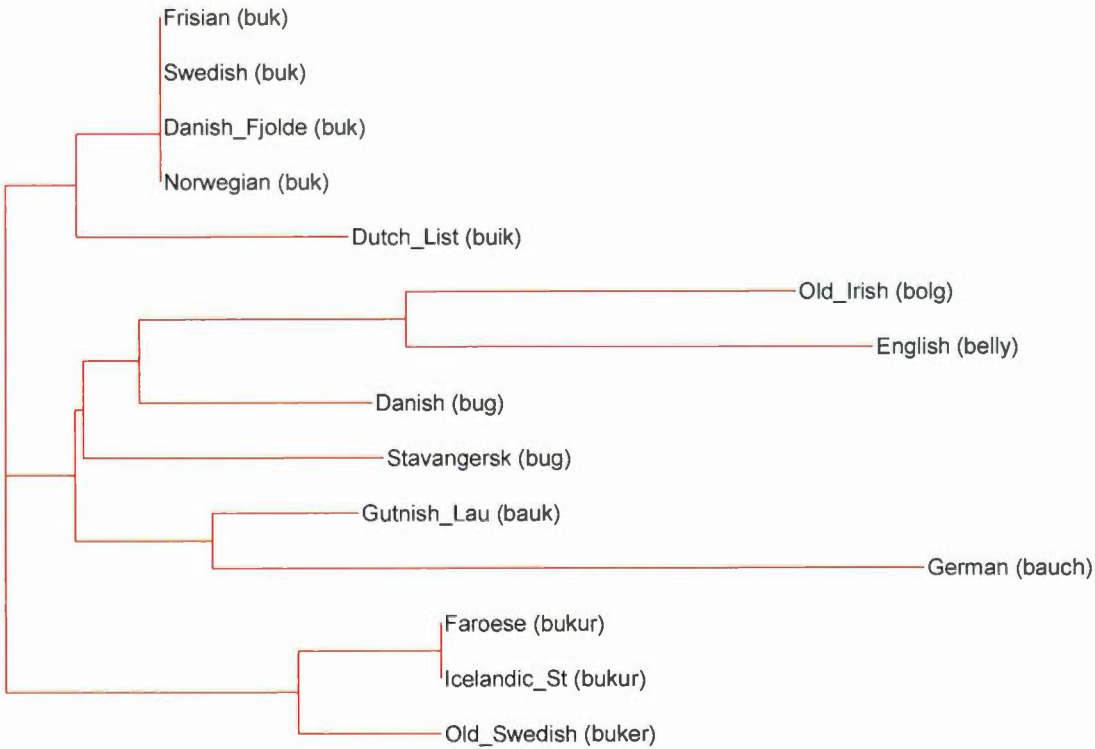
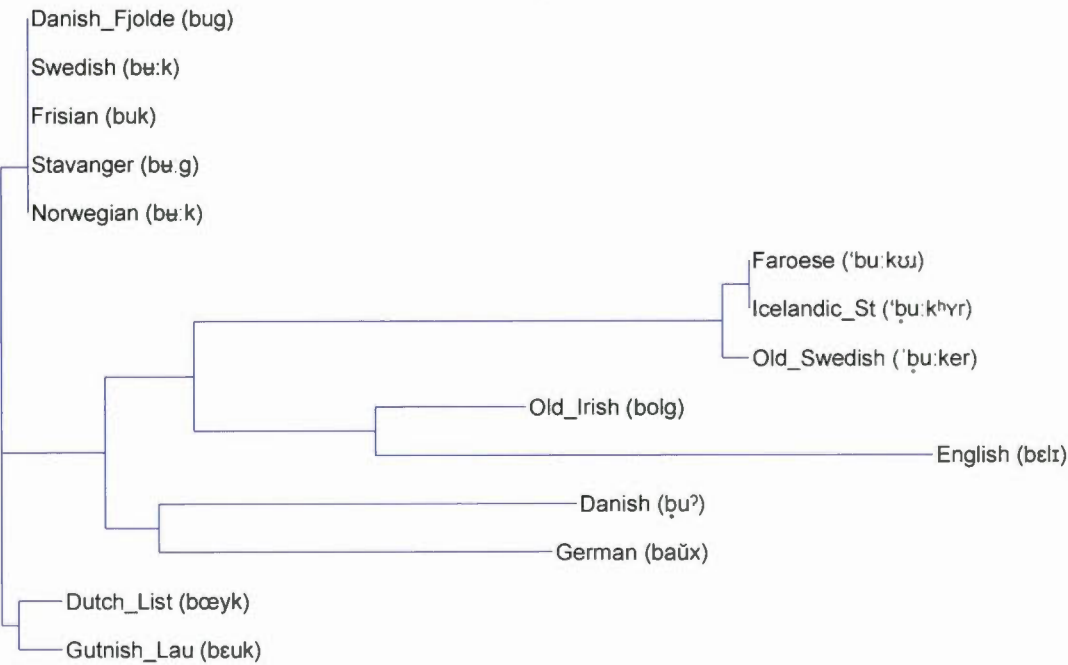
Animal



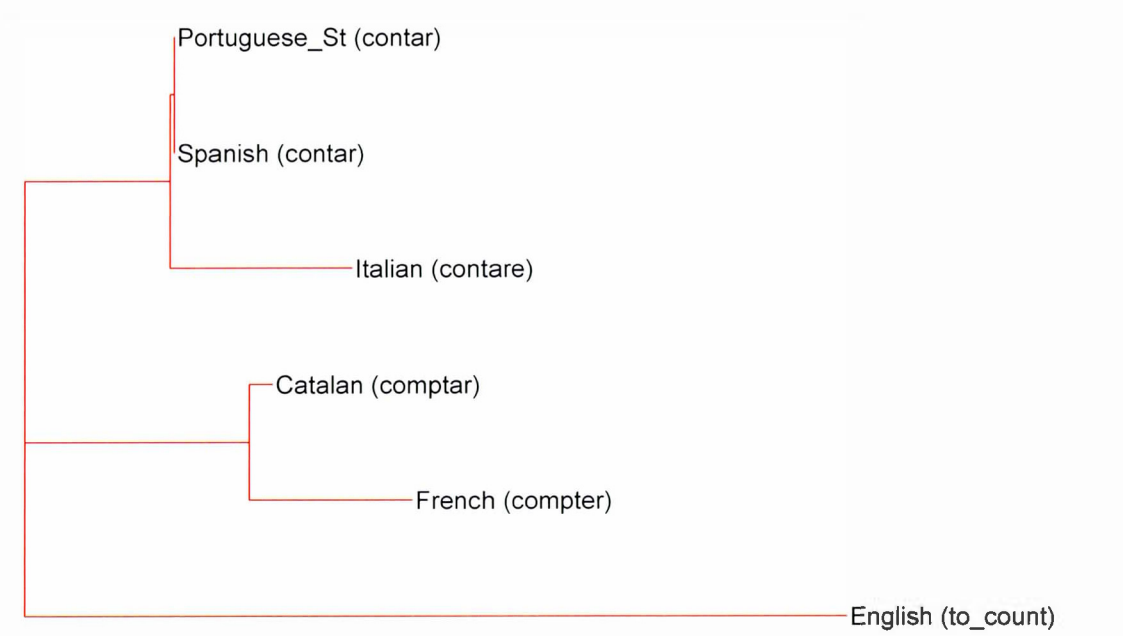
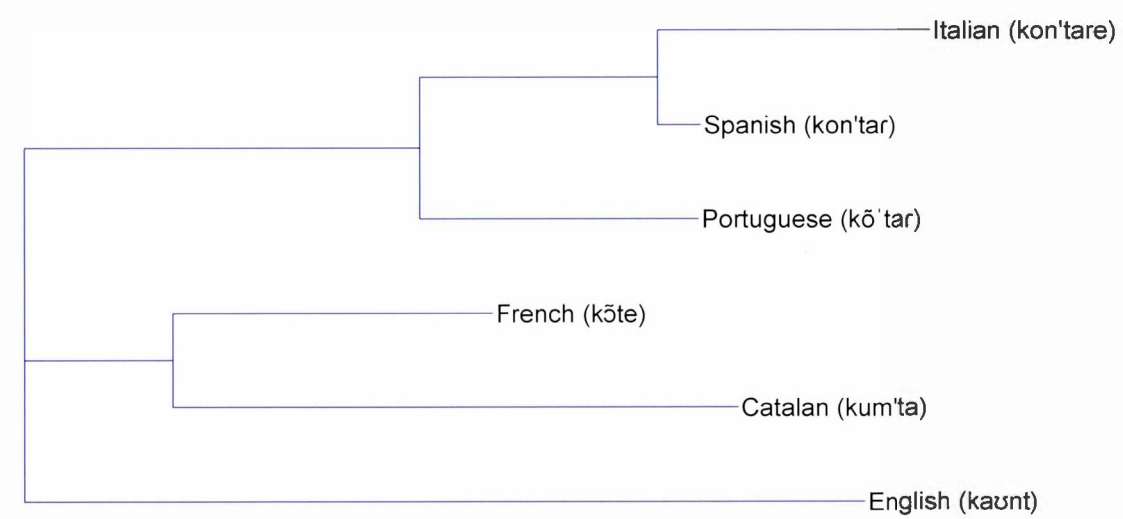
Bark



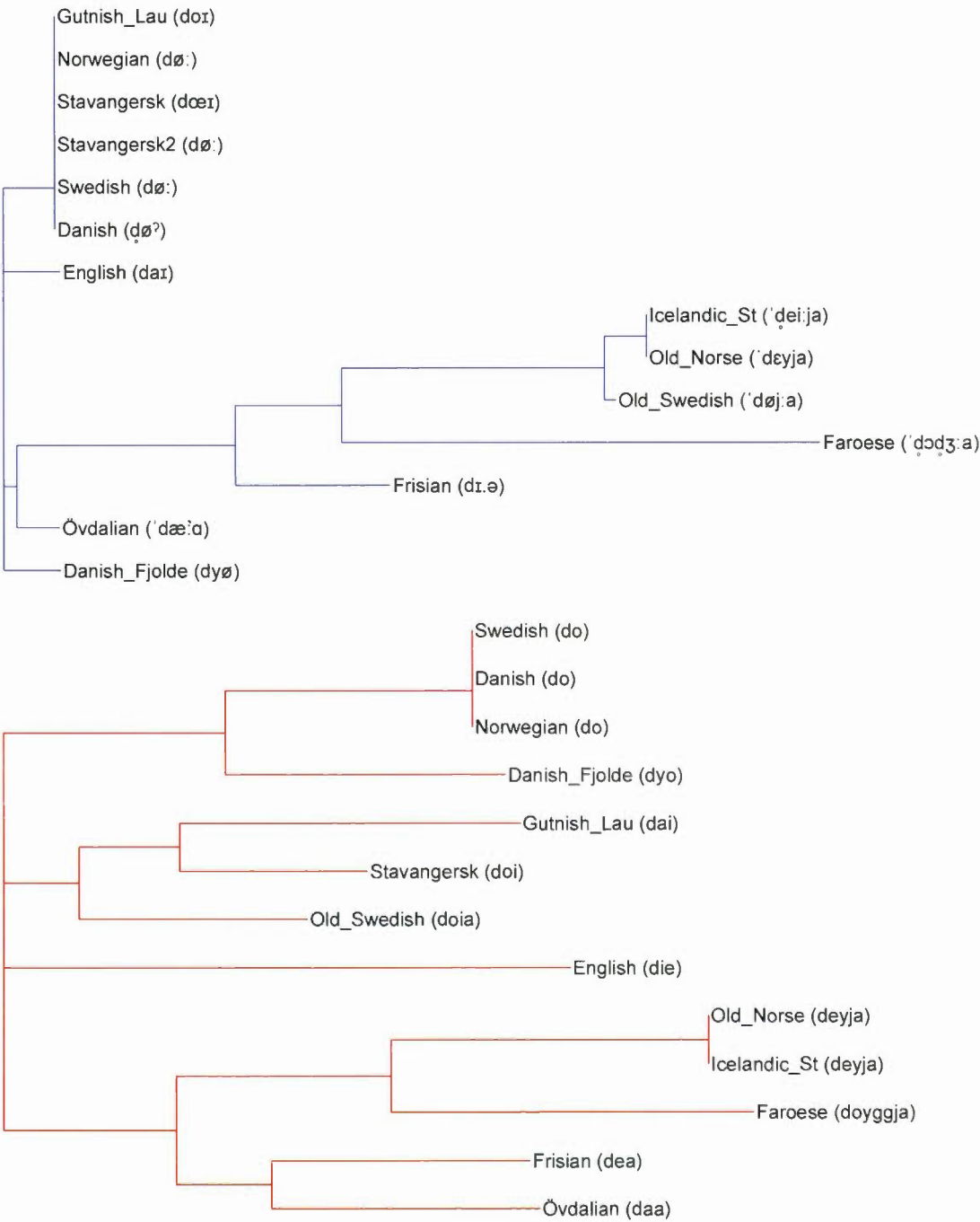
Belly



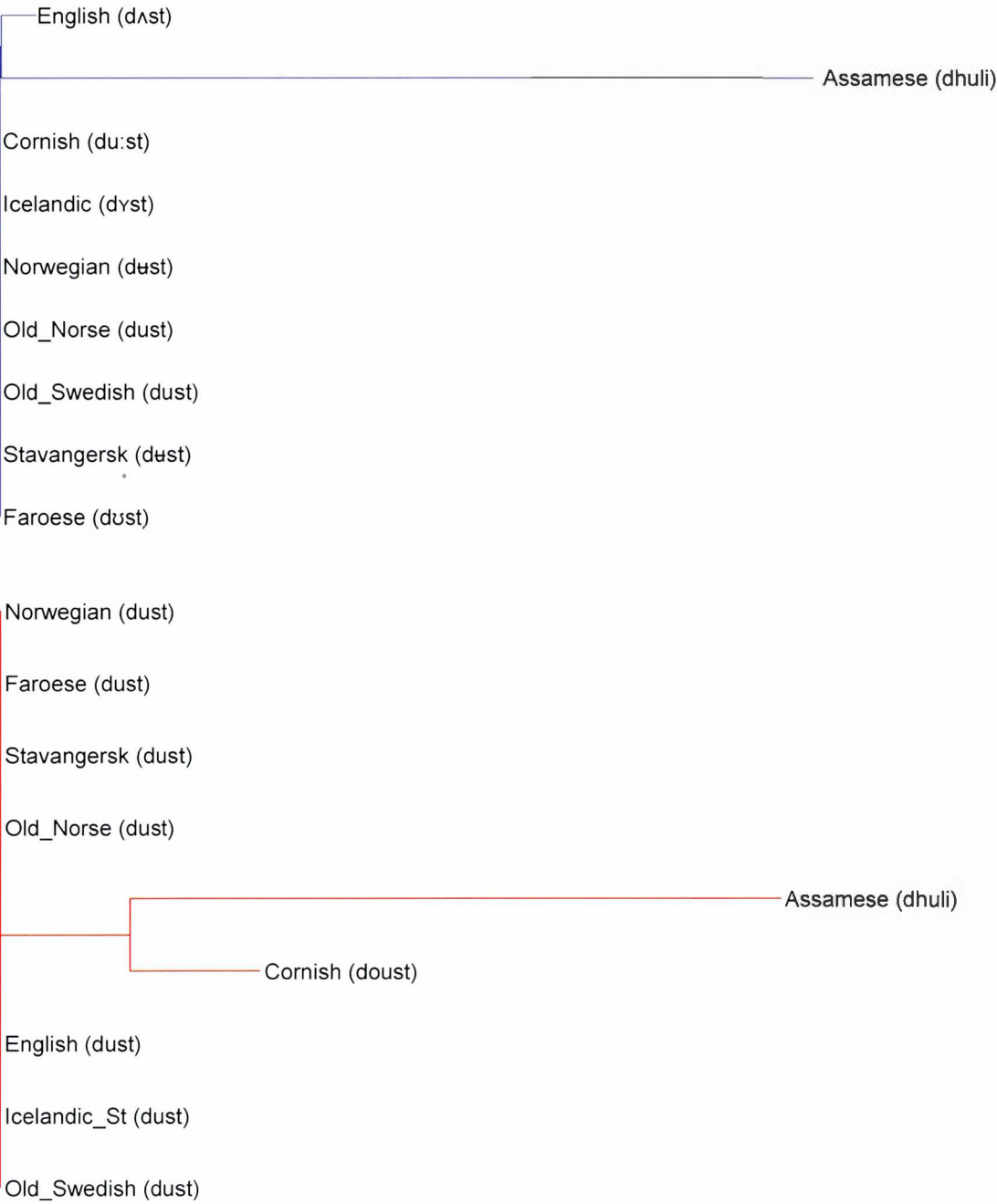
Count



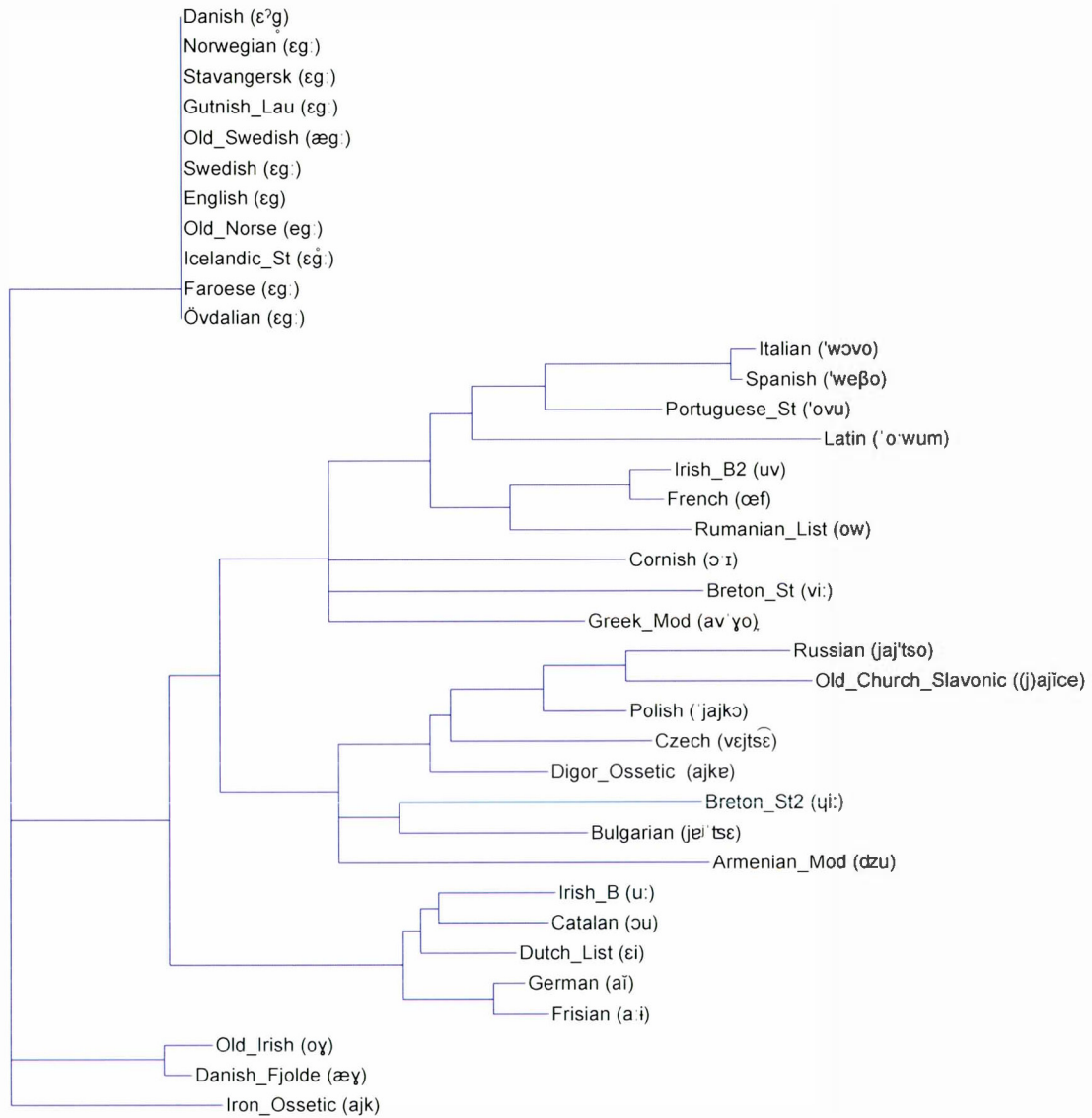
Die



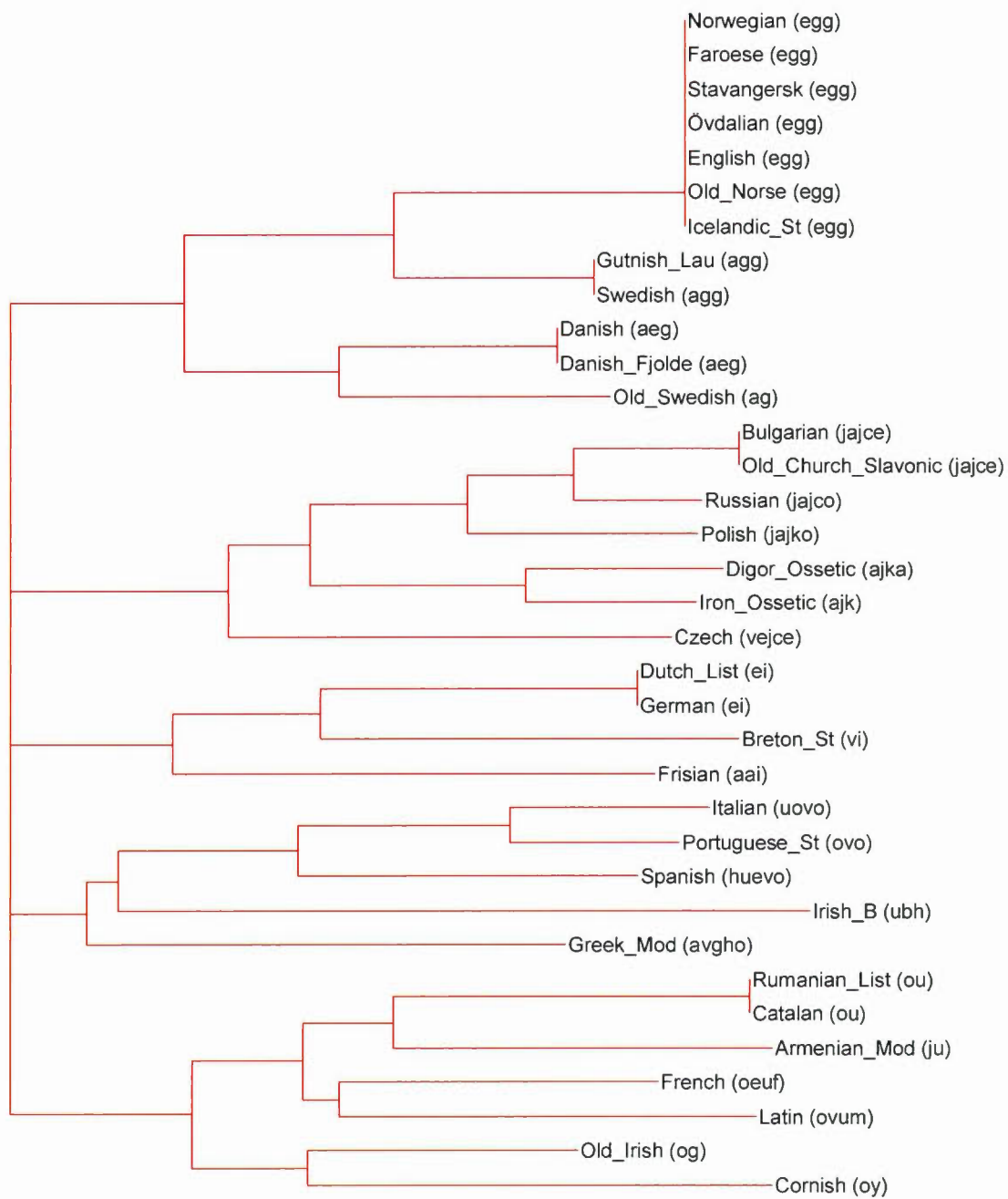
Dust



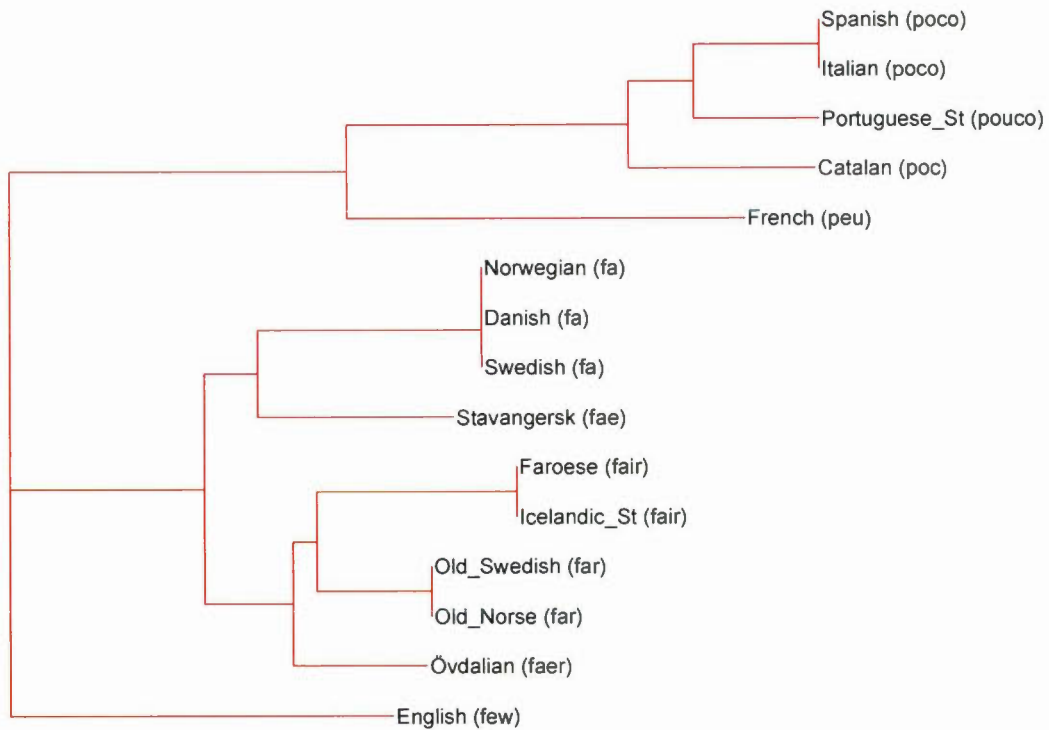
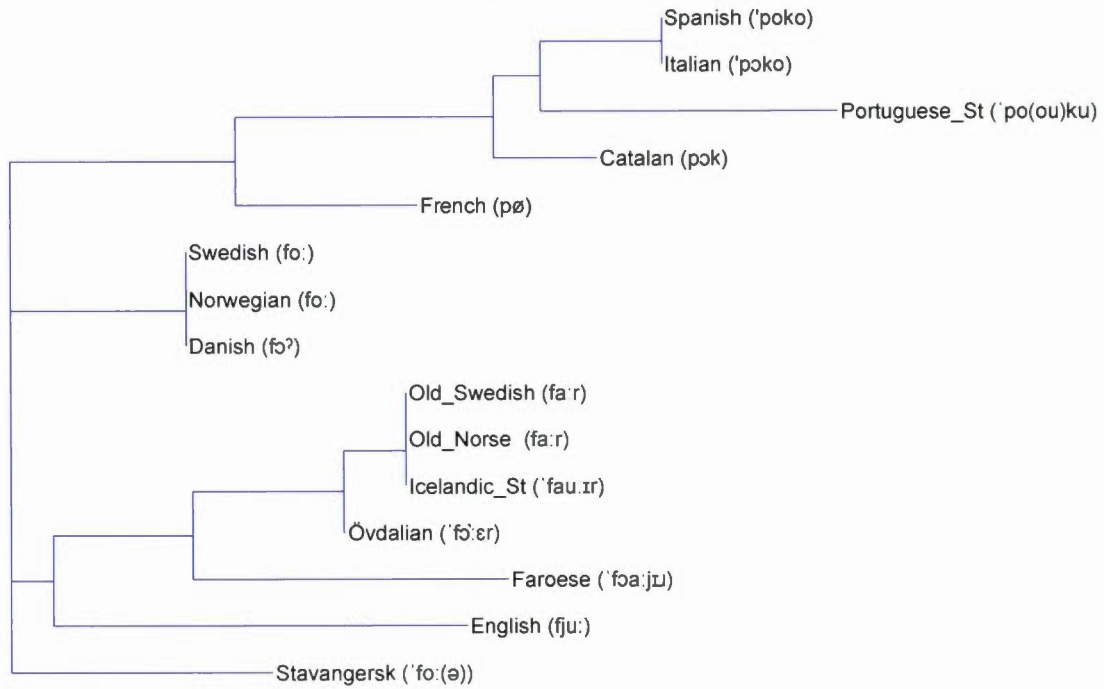
Egg



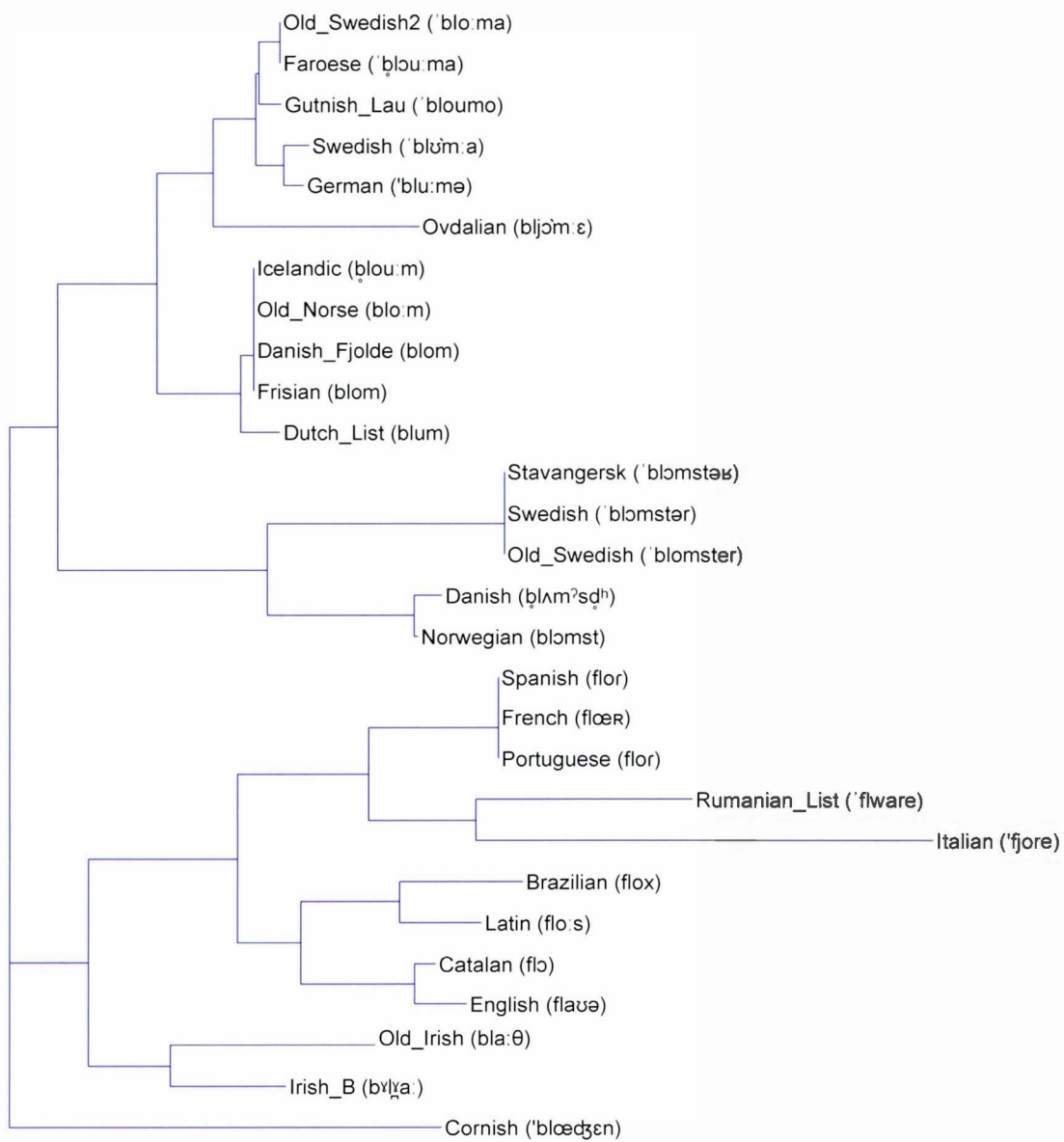
Egg



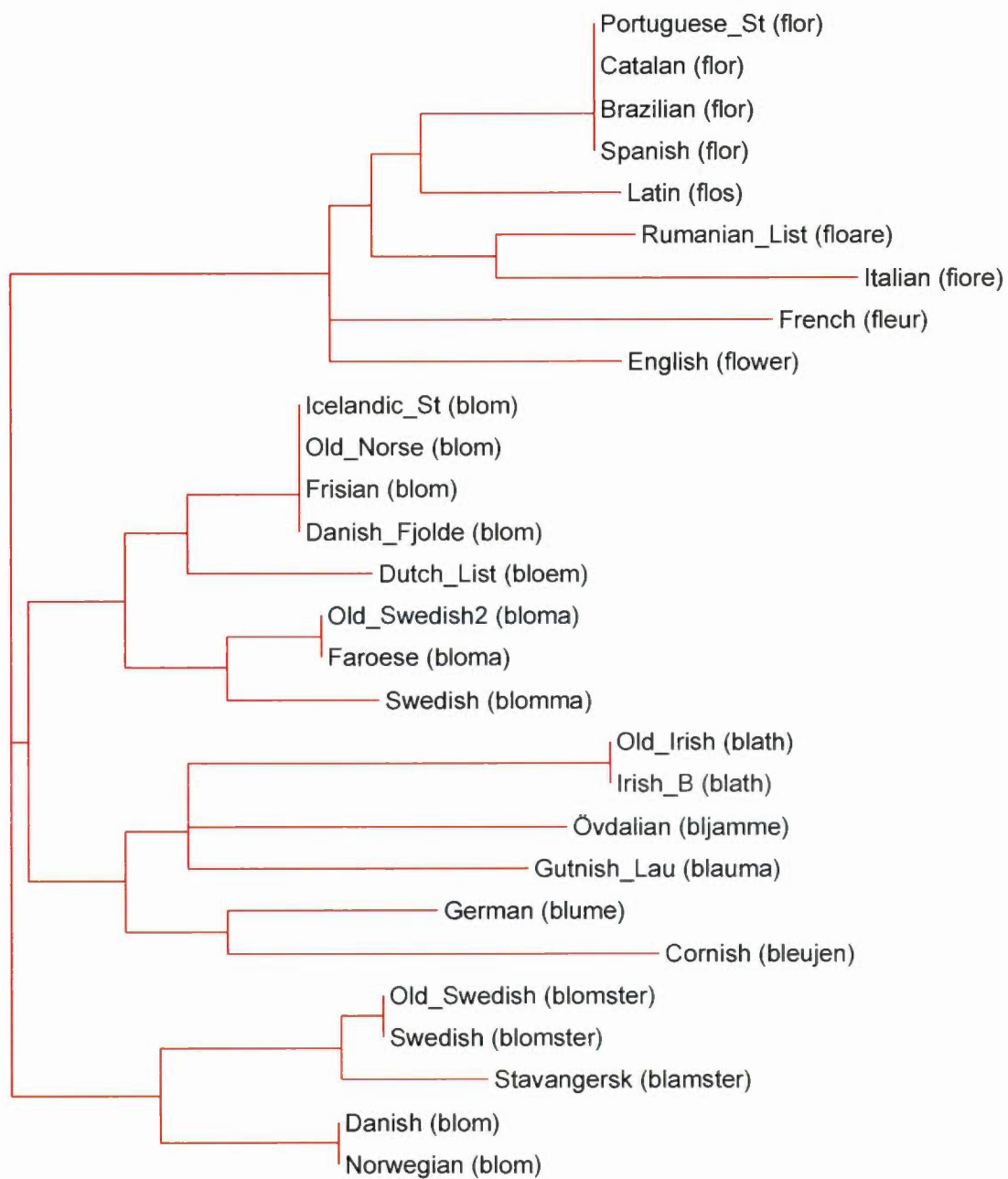
Few



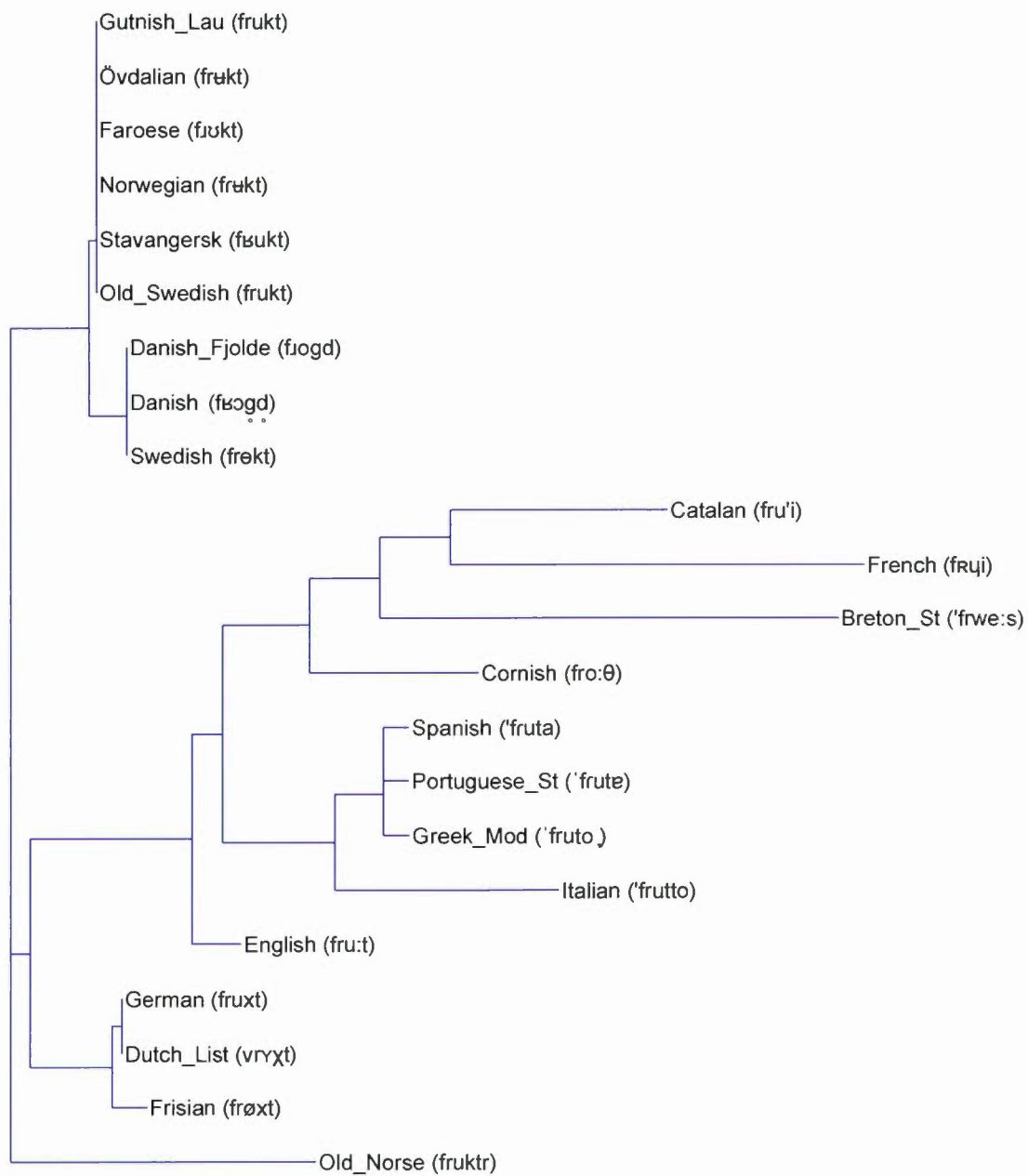
Flower



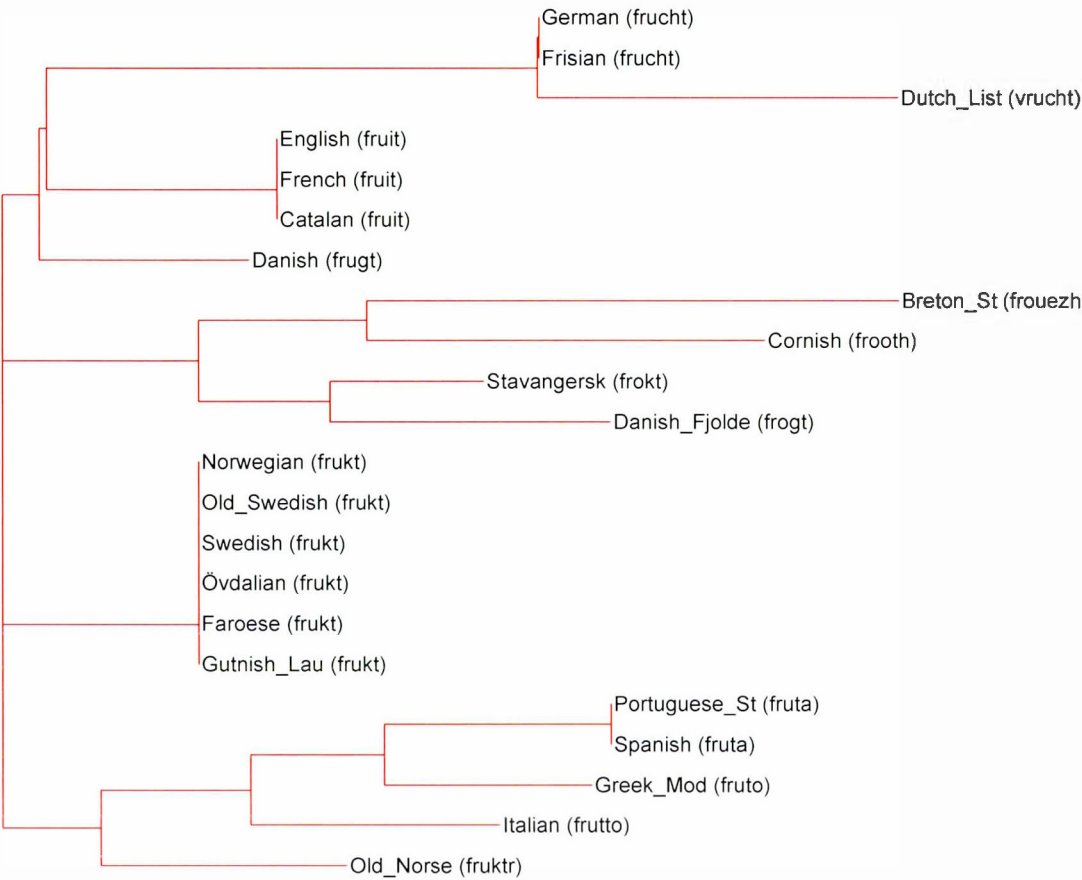
Flower



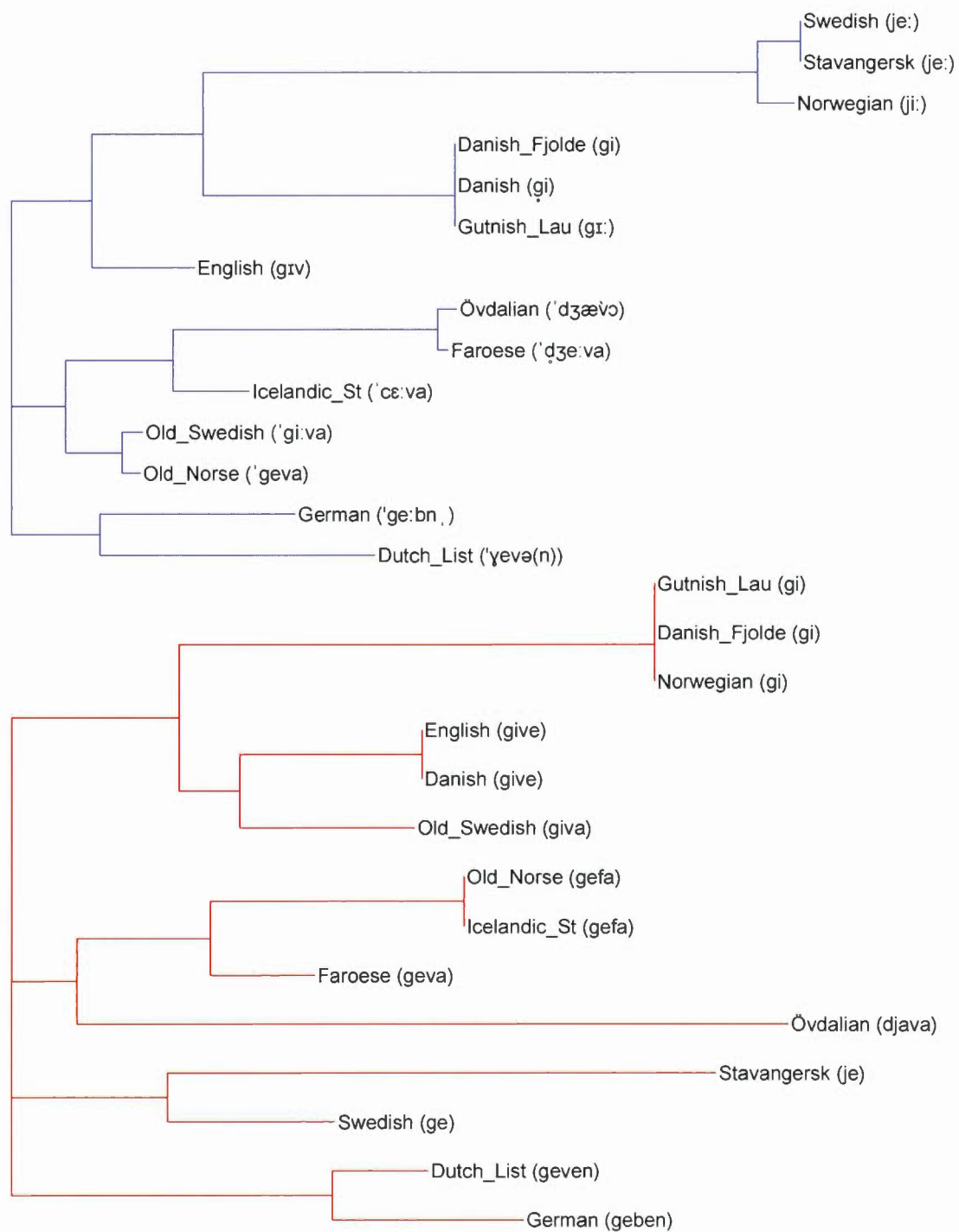
Fruit



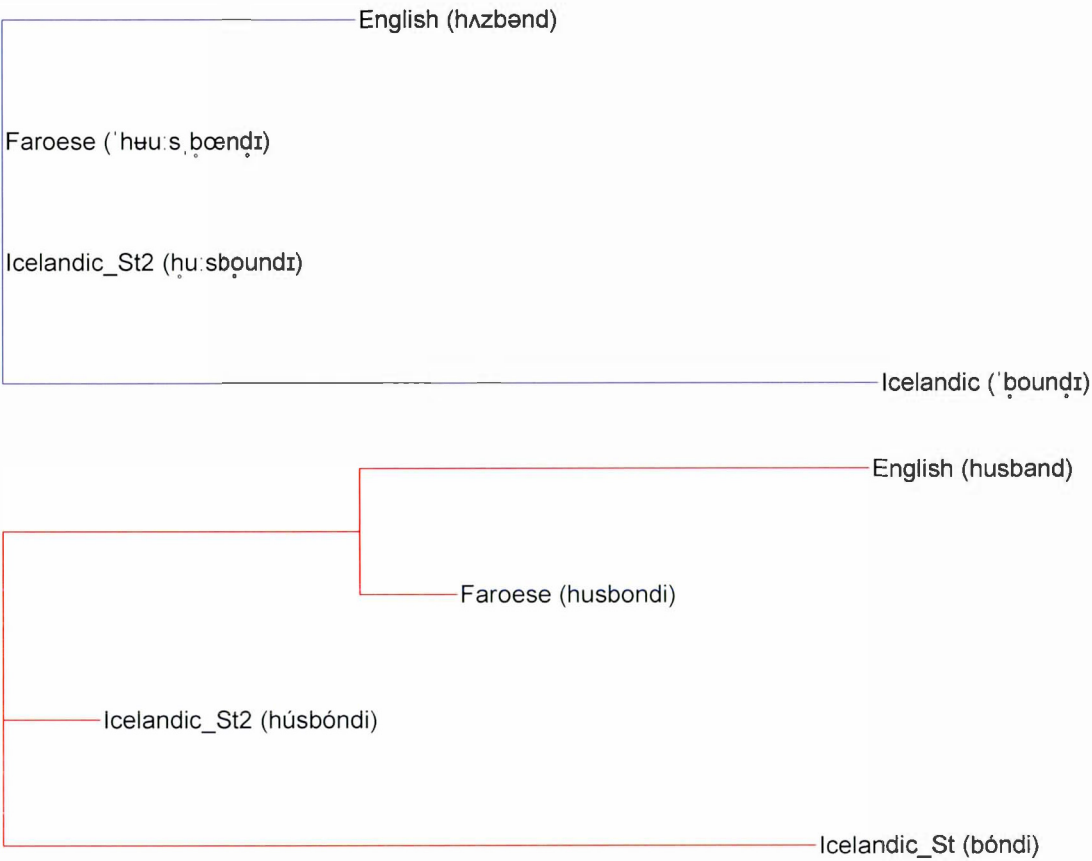
Fruit



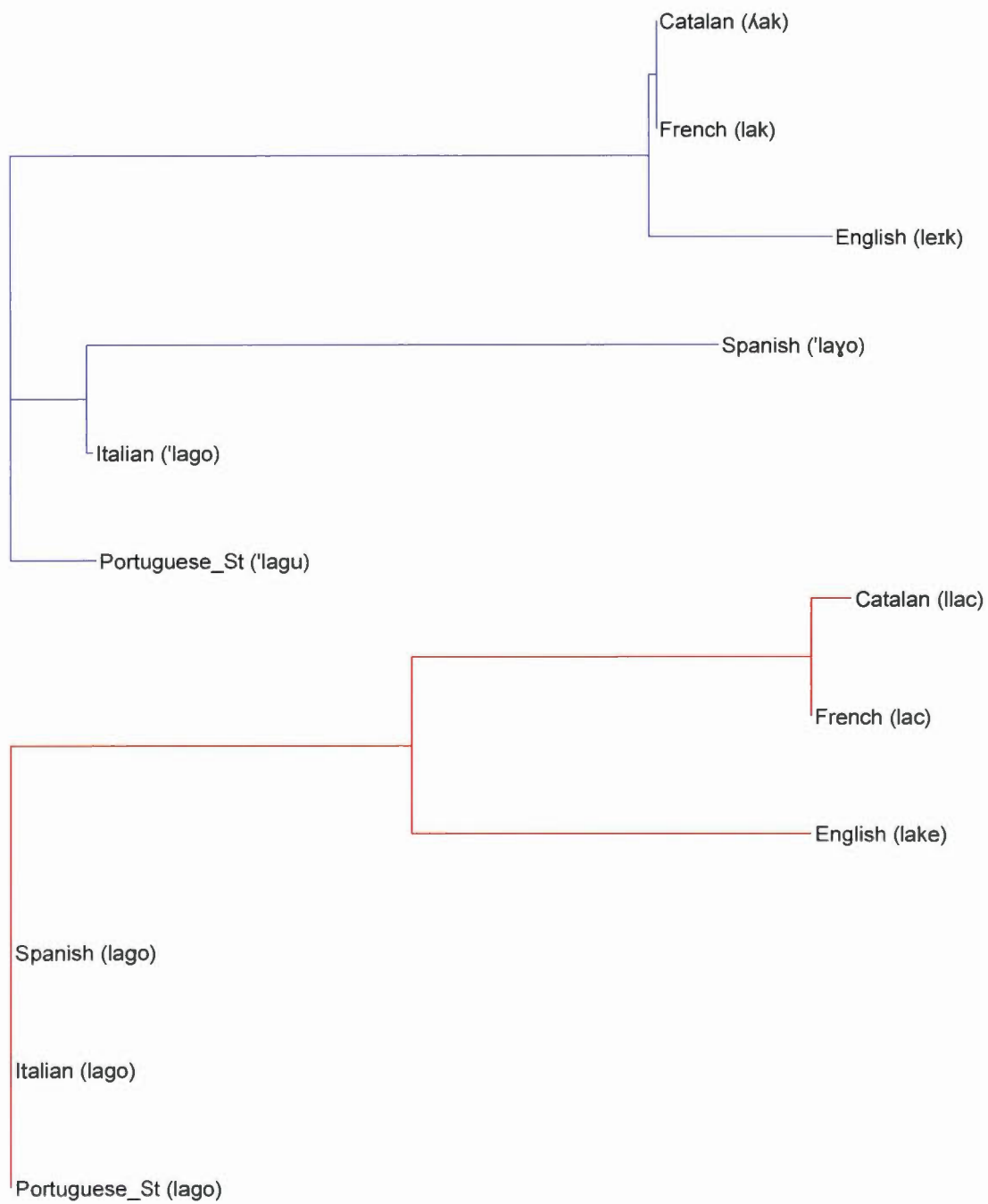
Give



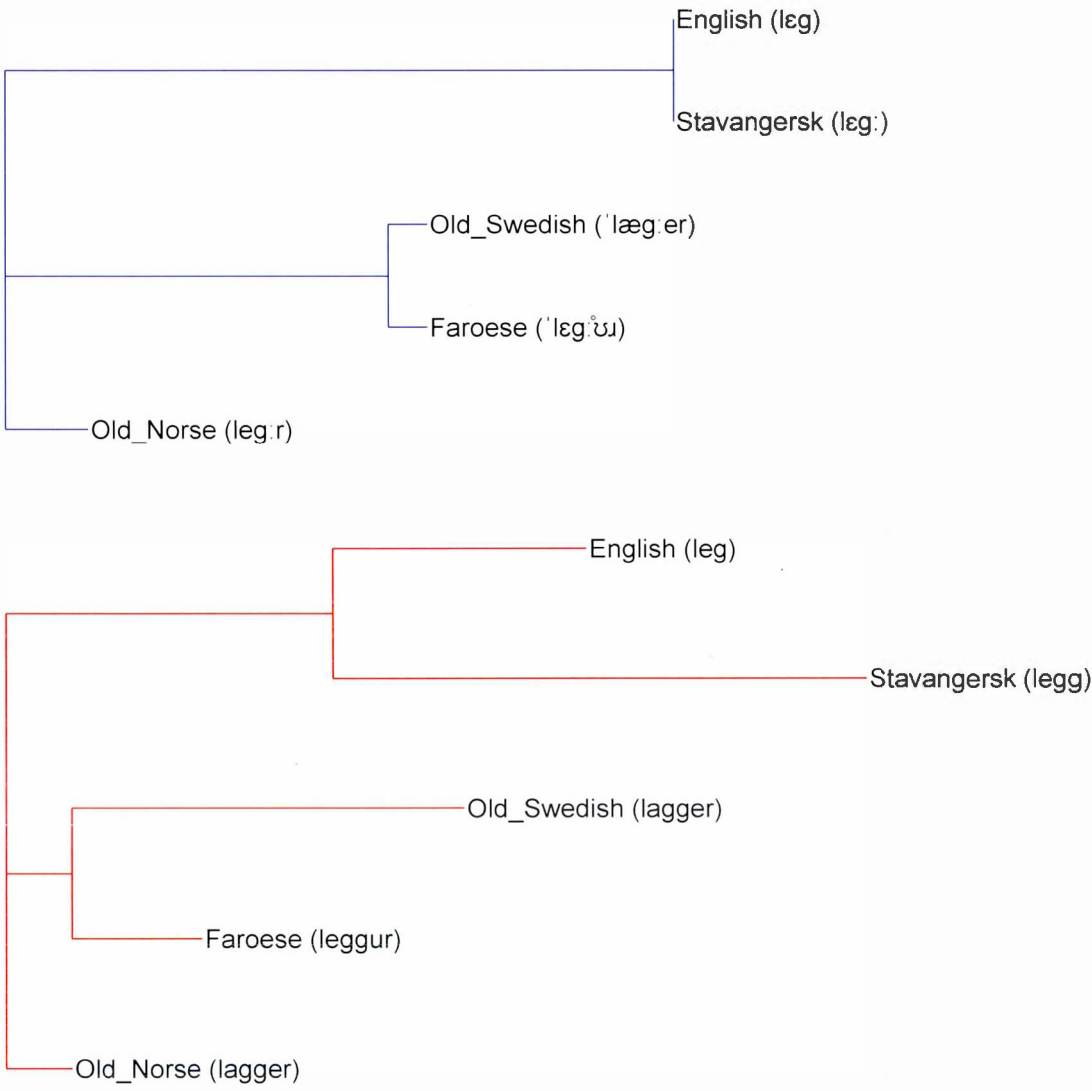
Husband



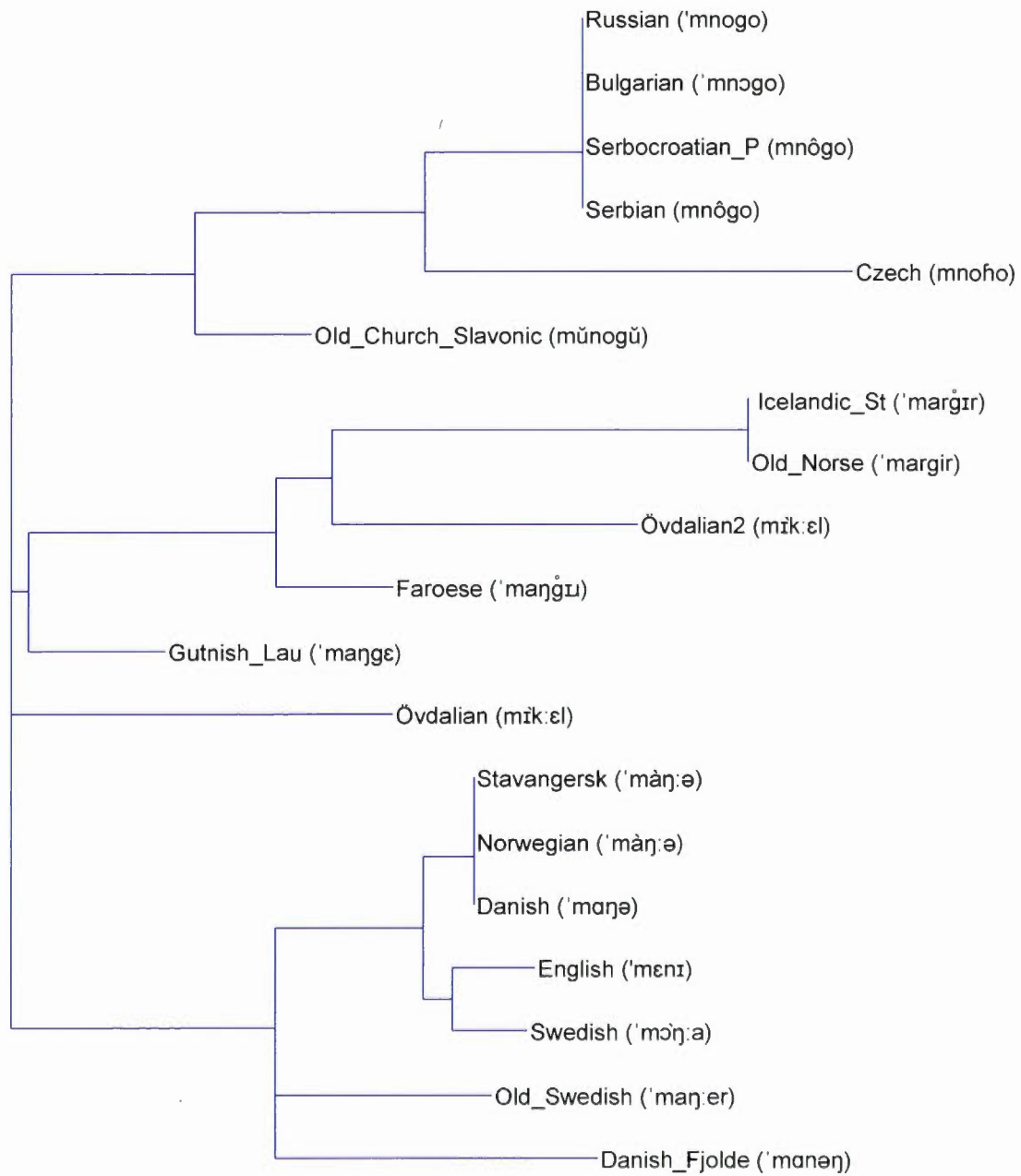
Lake



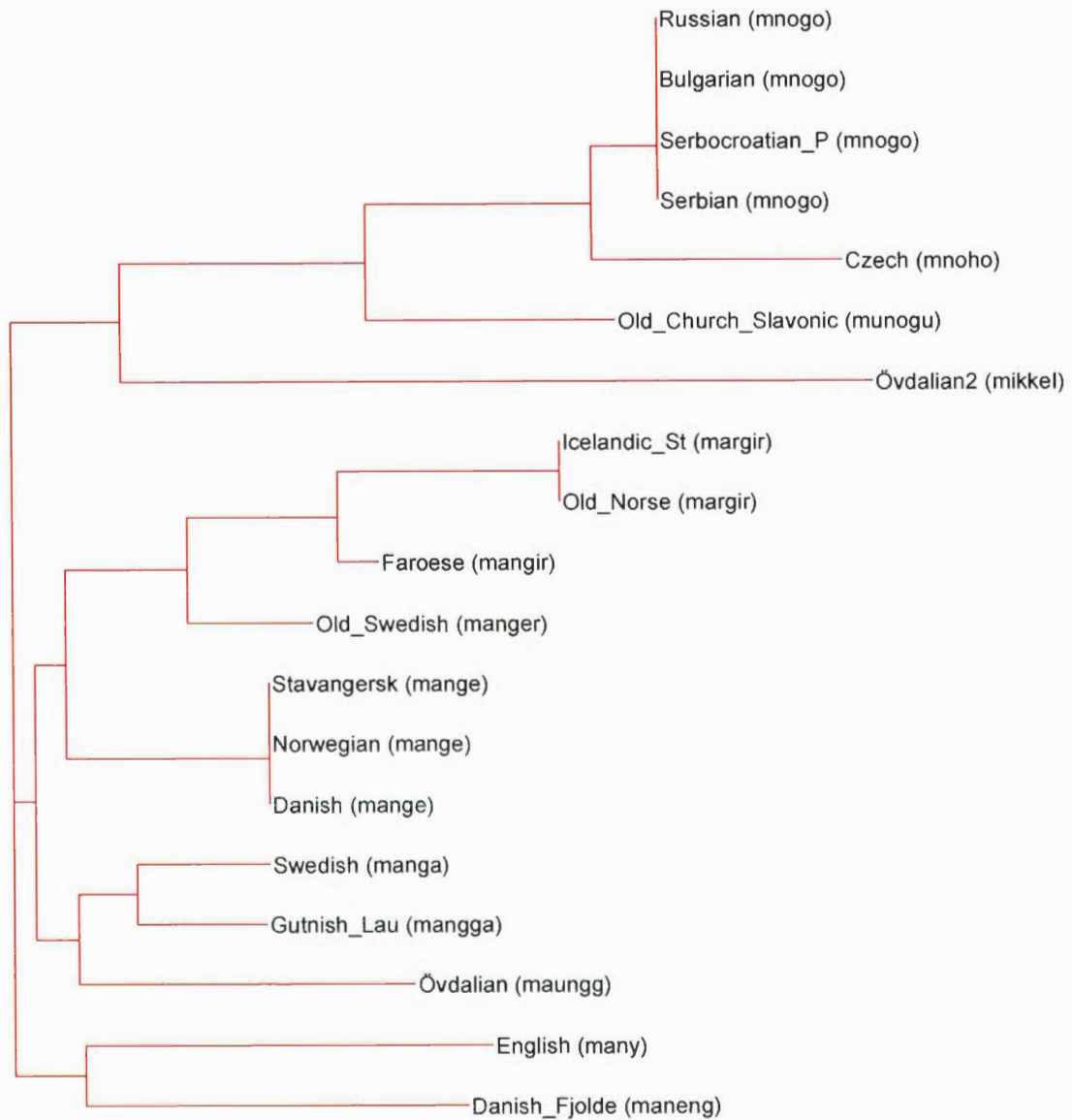
Leg



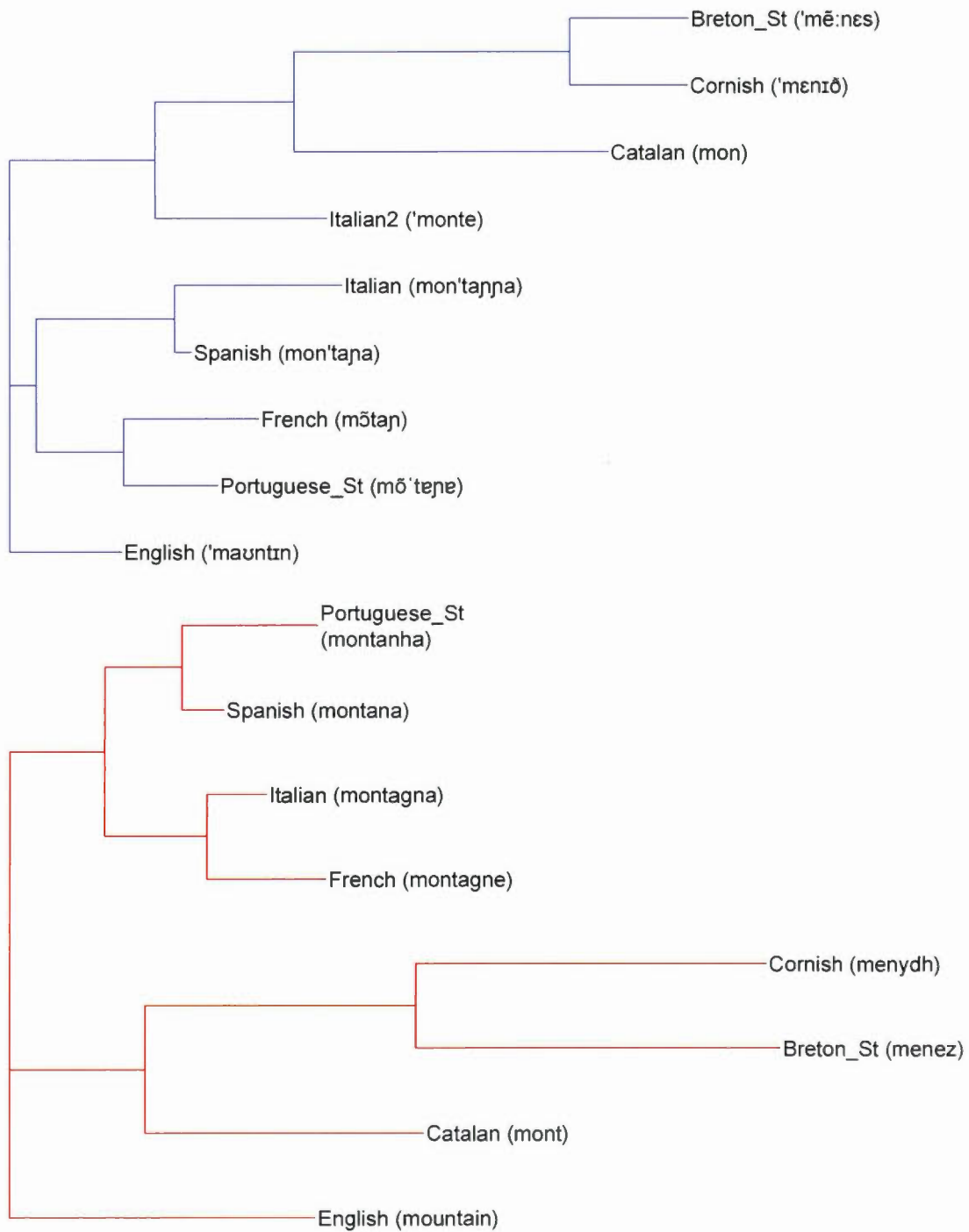
Many



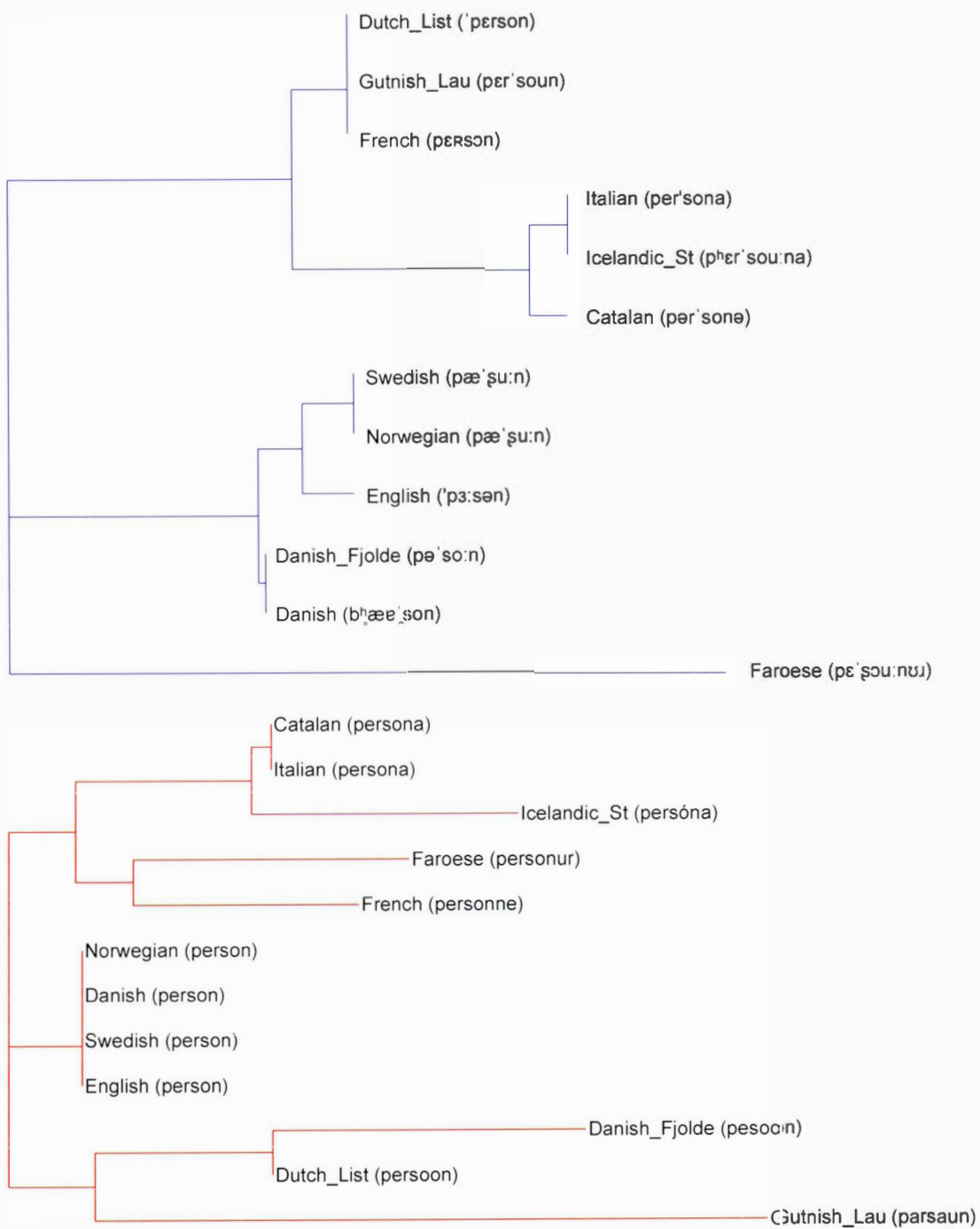
Many



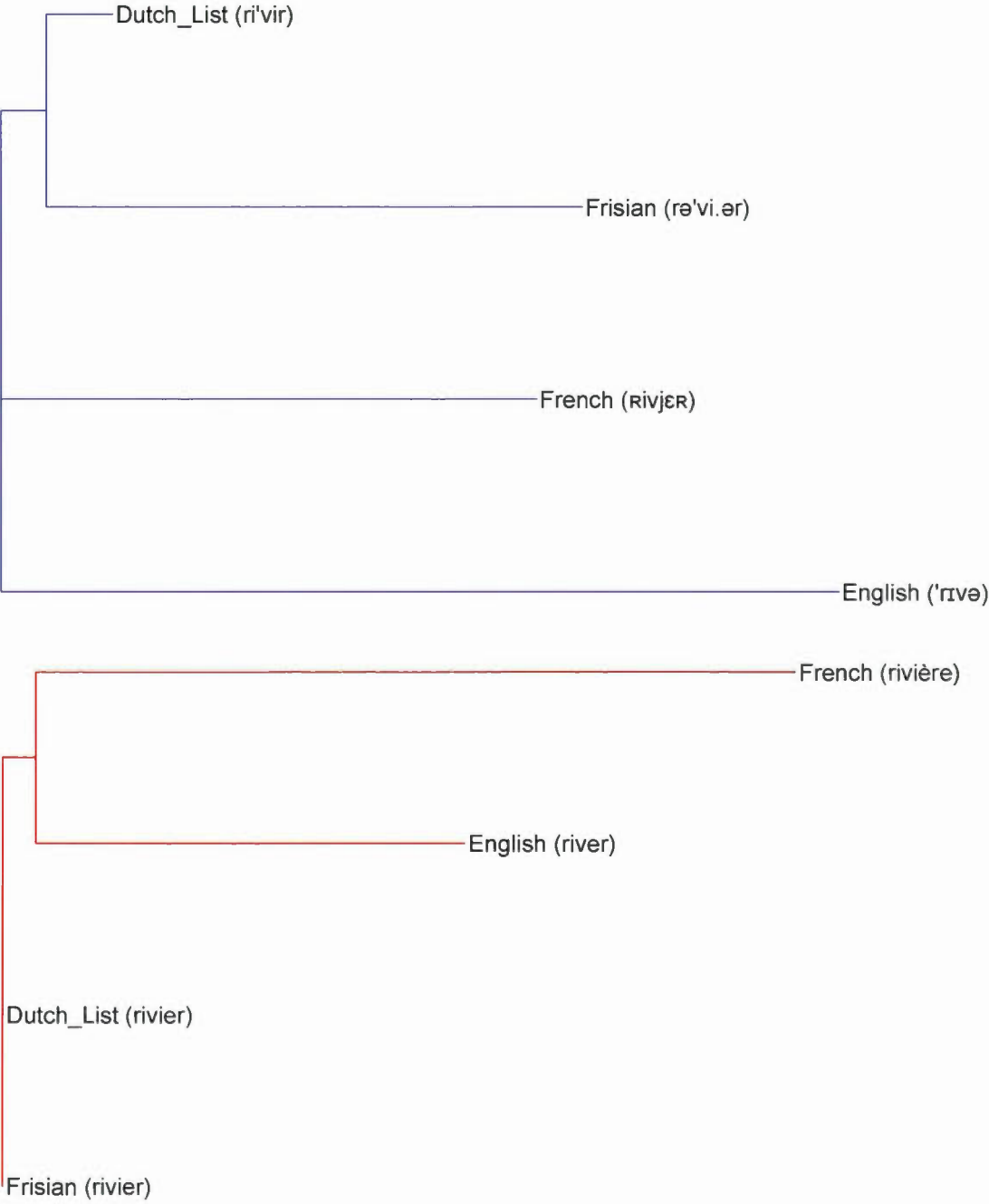
Mountain



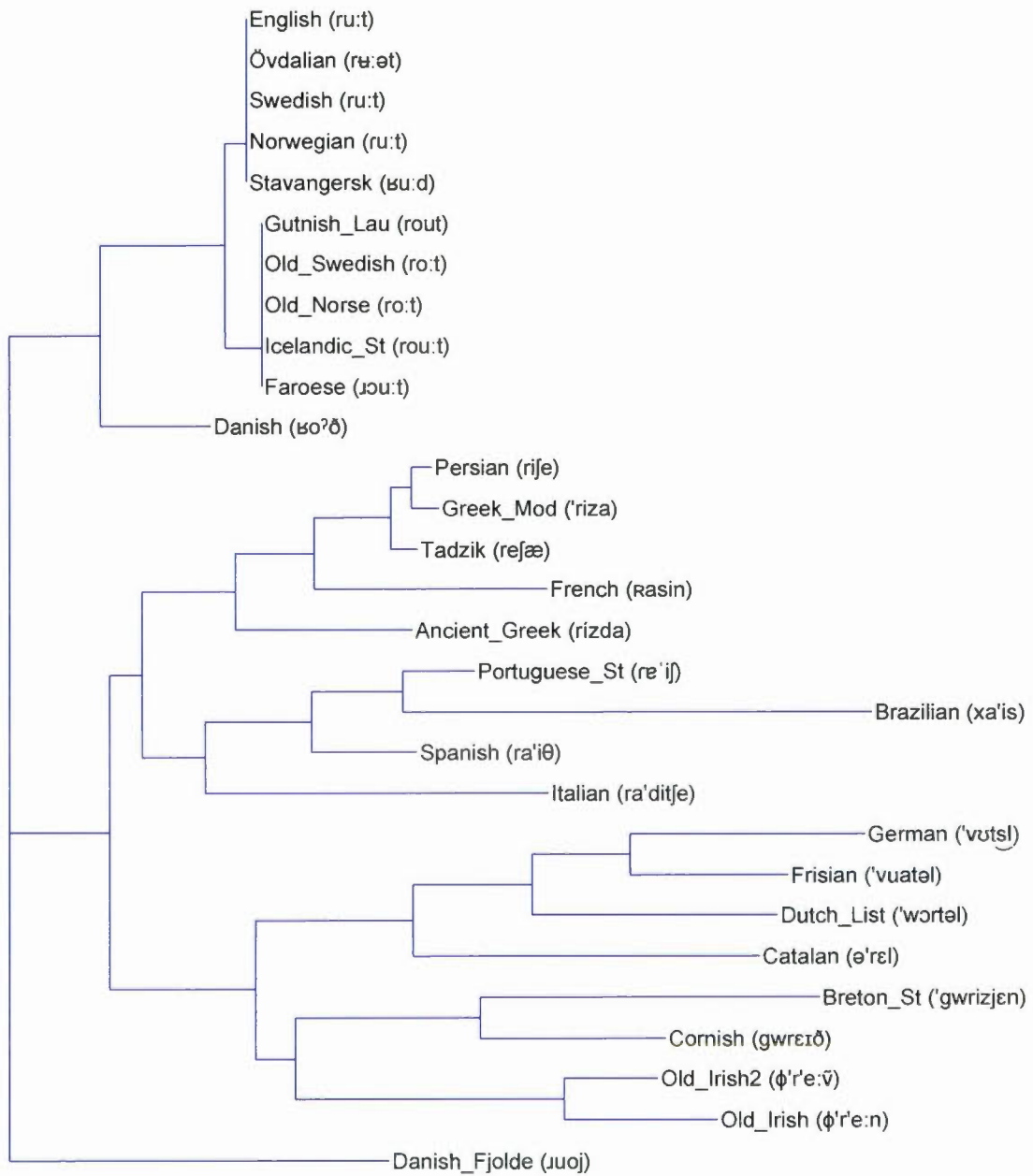
Person



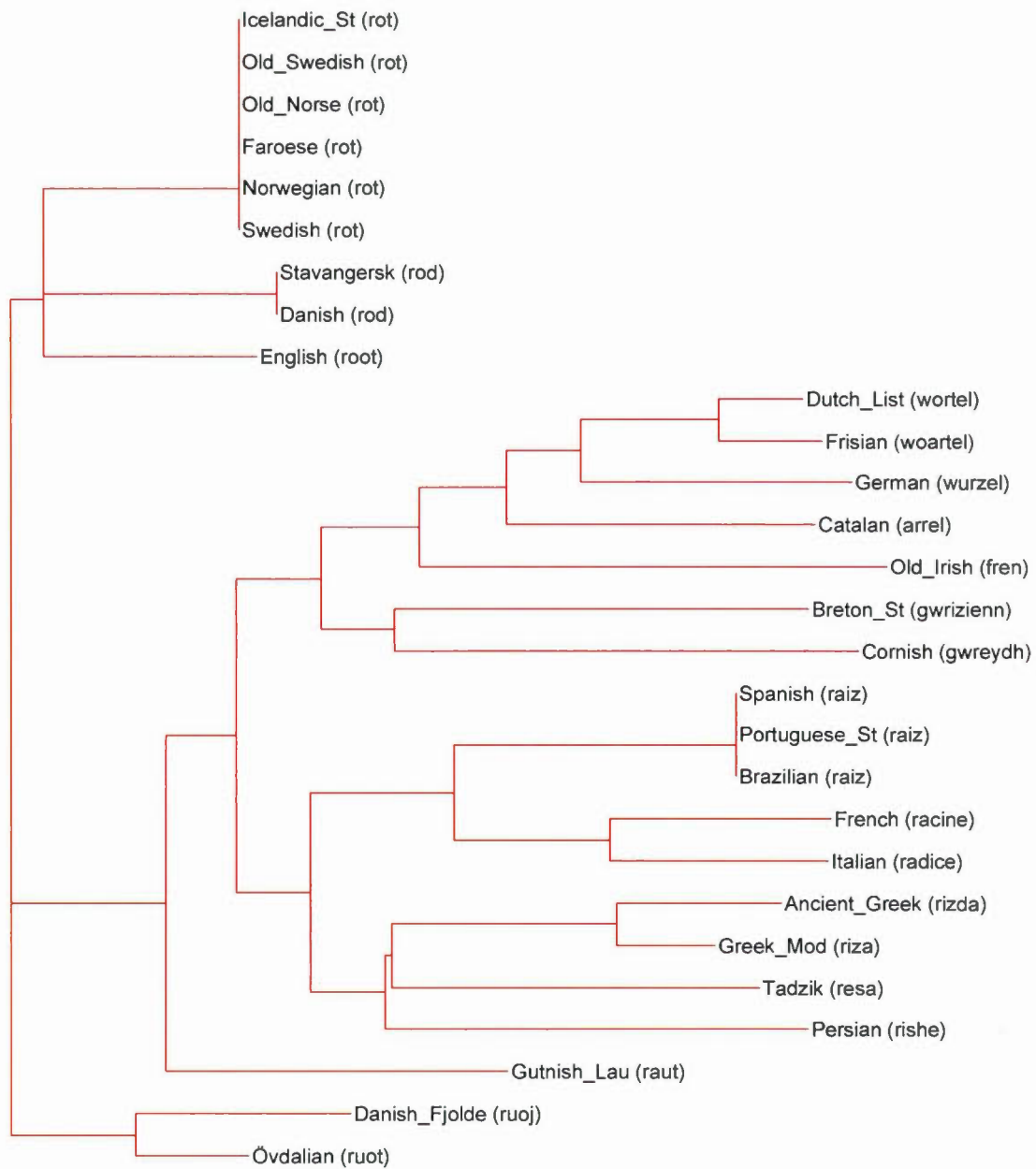
River



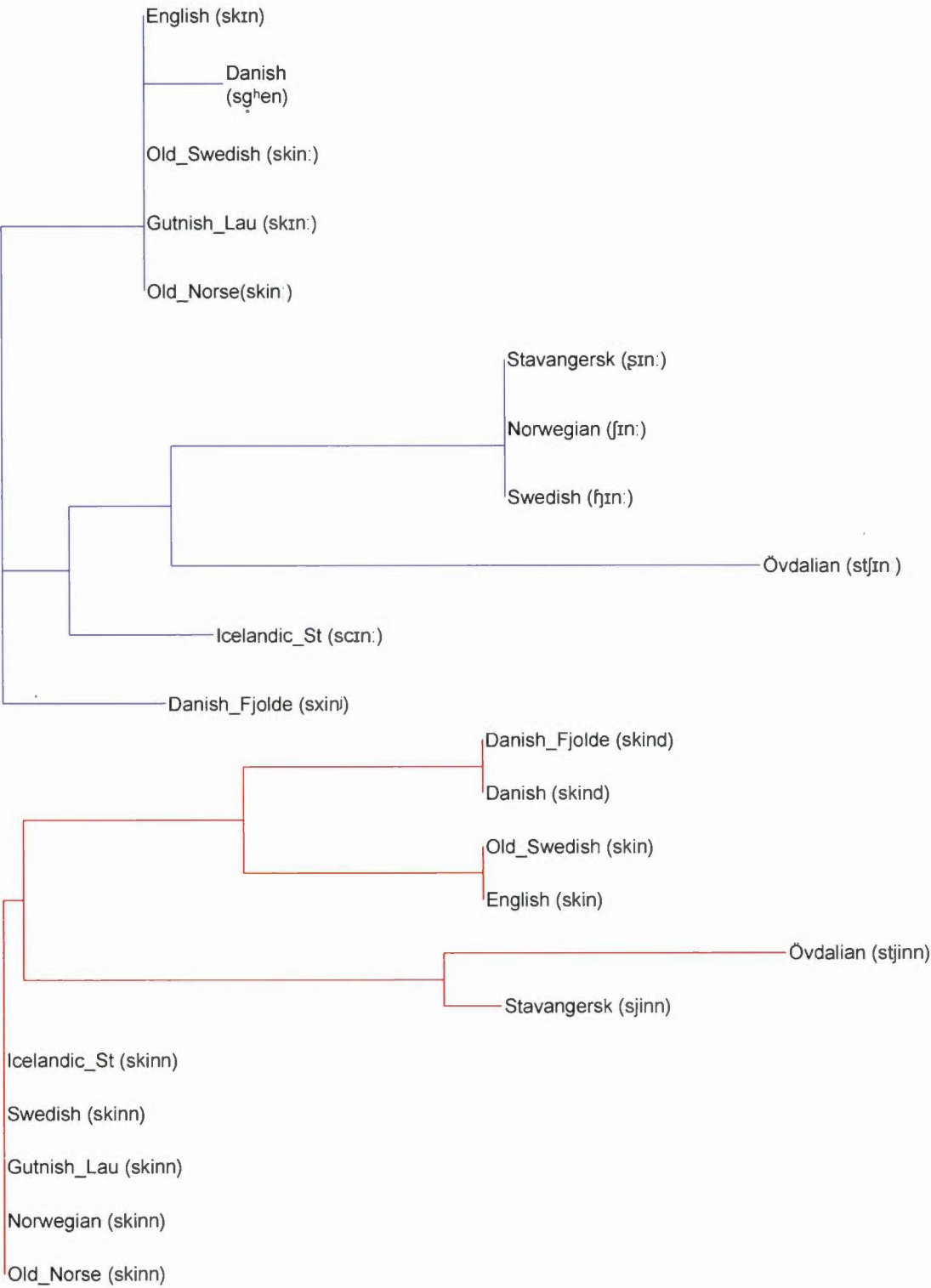
Root



Root

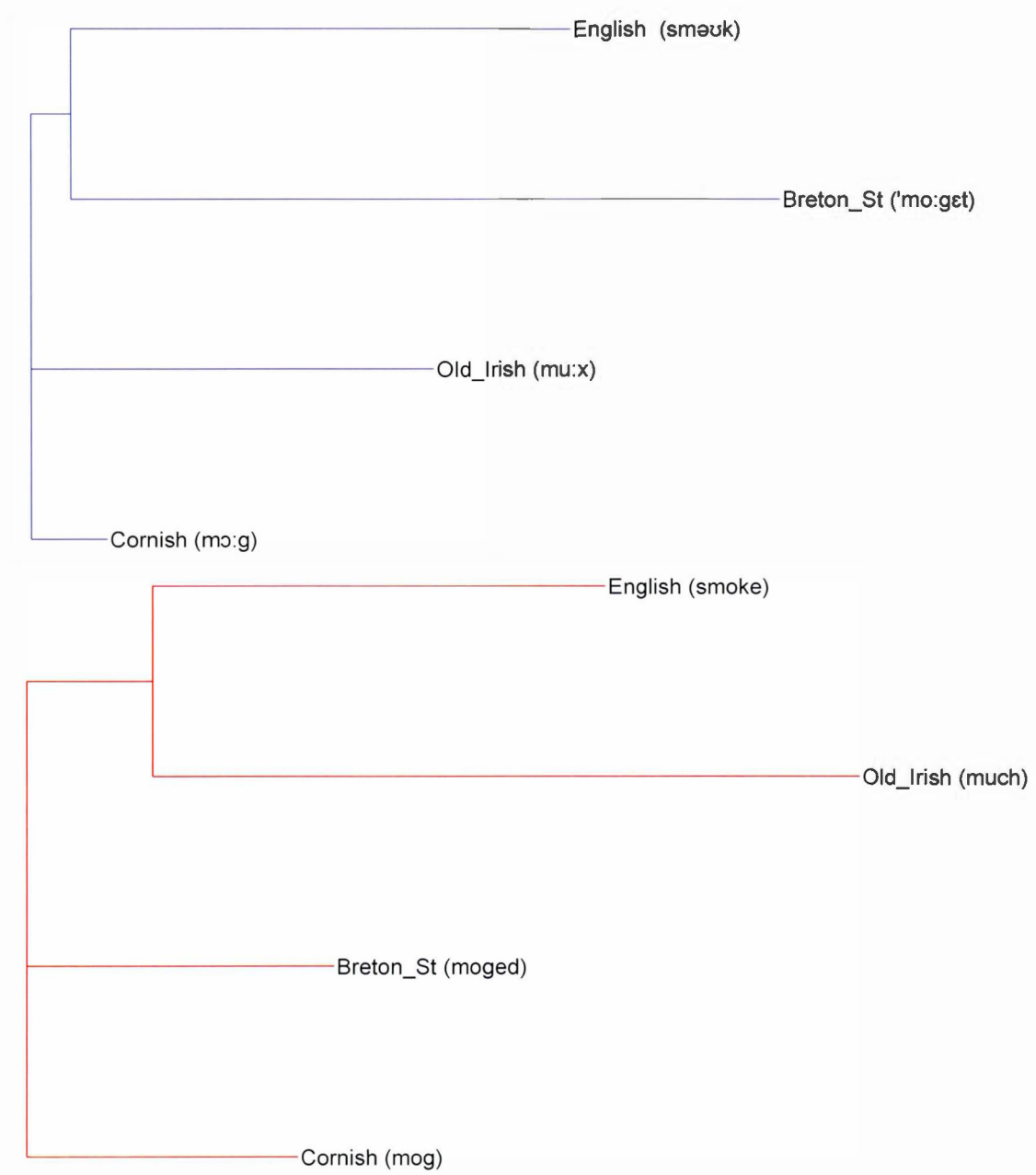


Skin

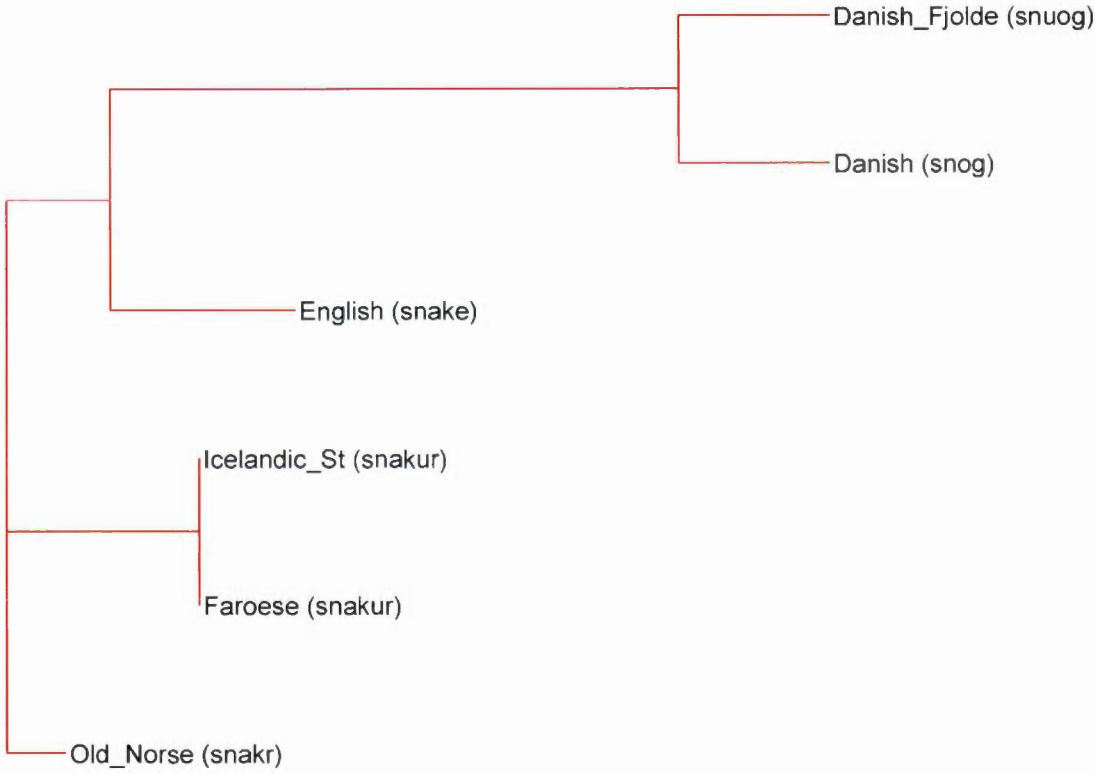
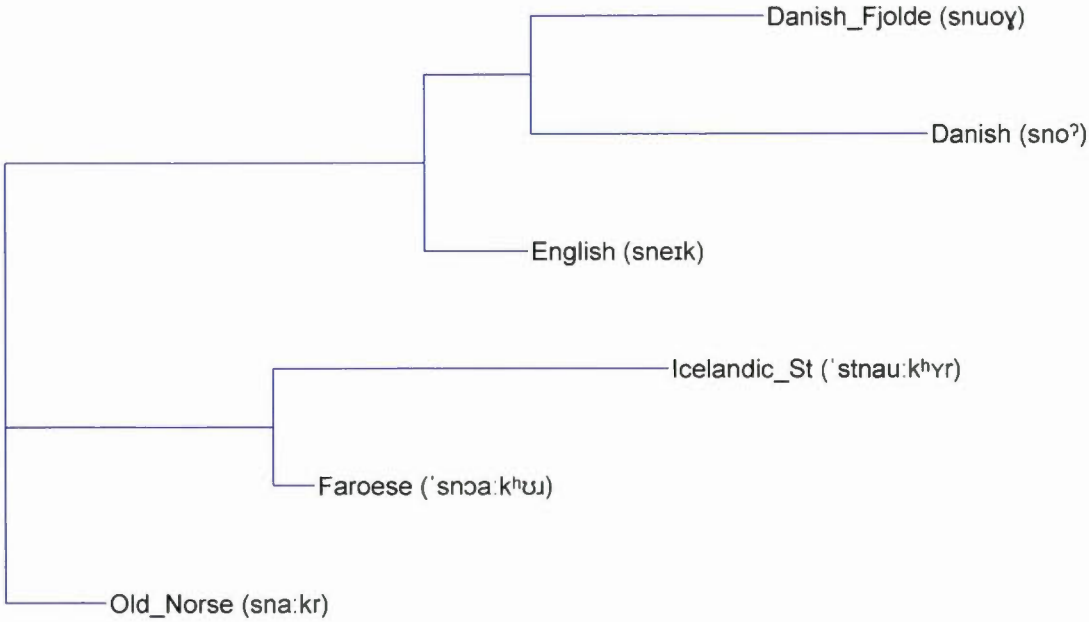


Smoke

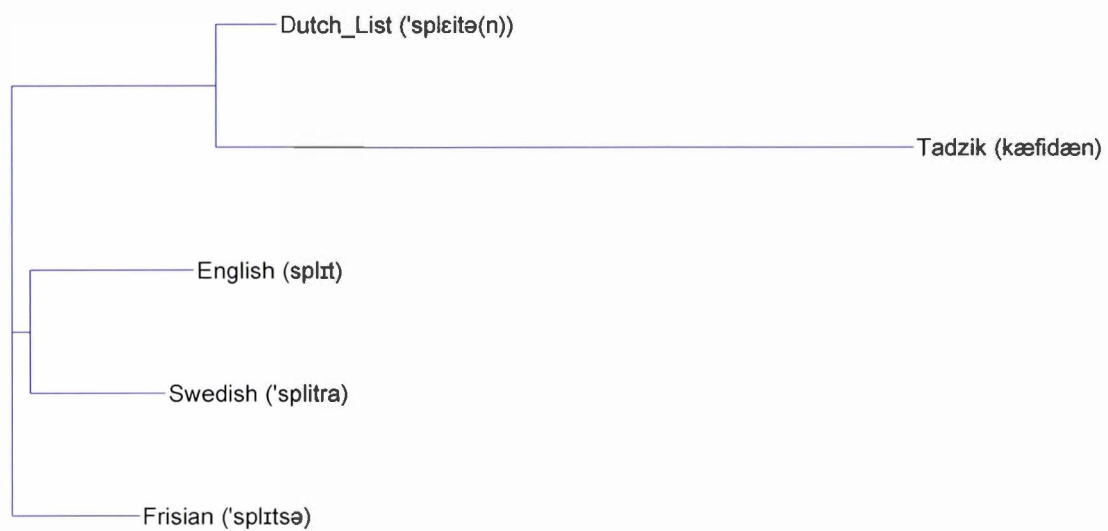
193



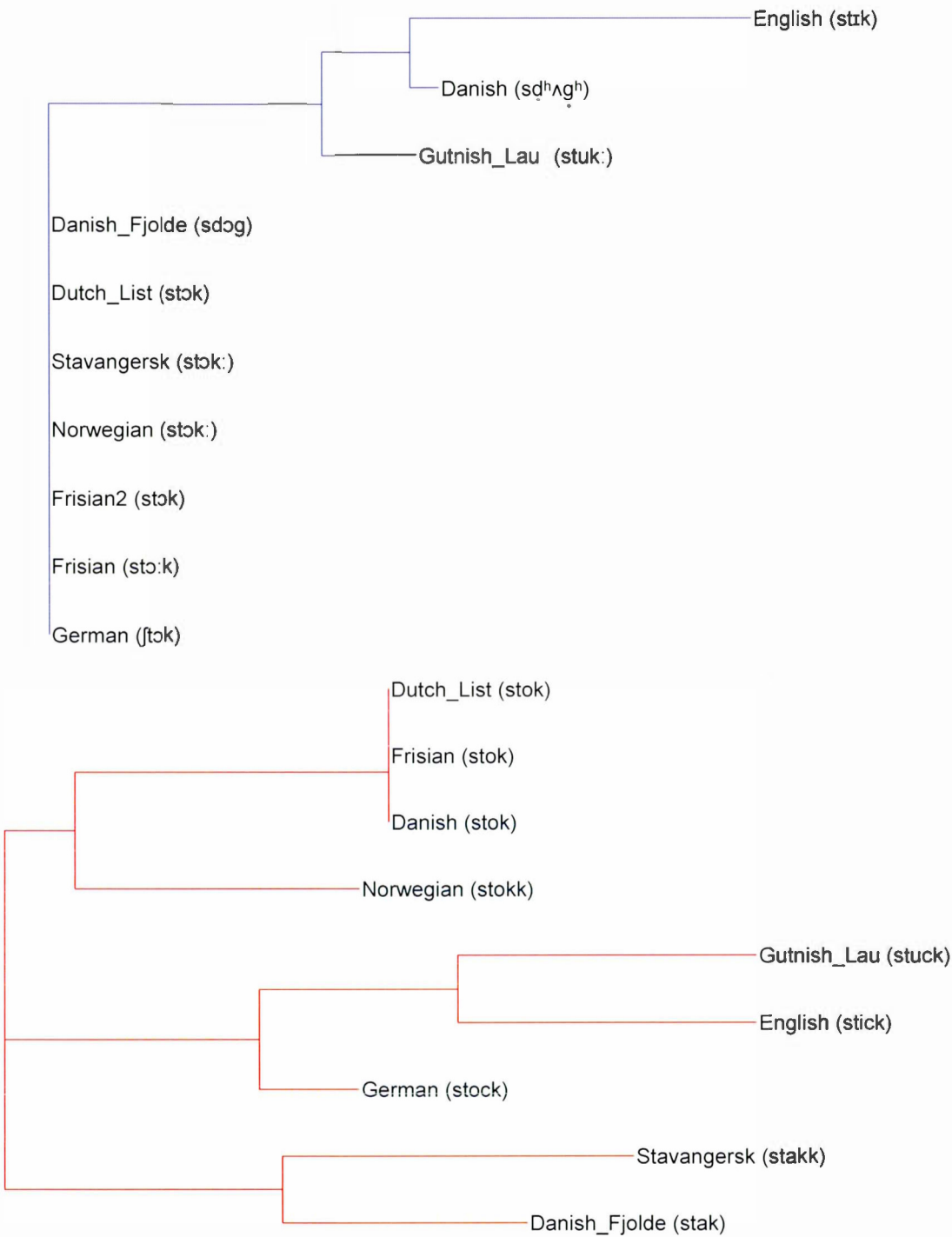
Snake

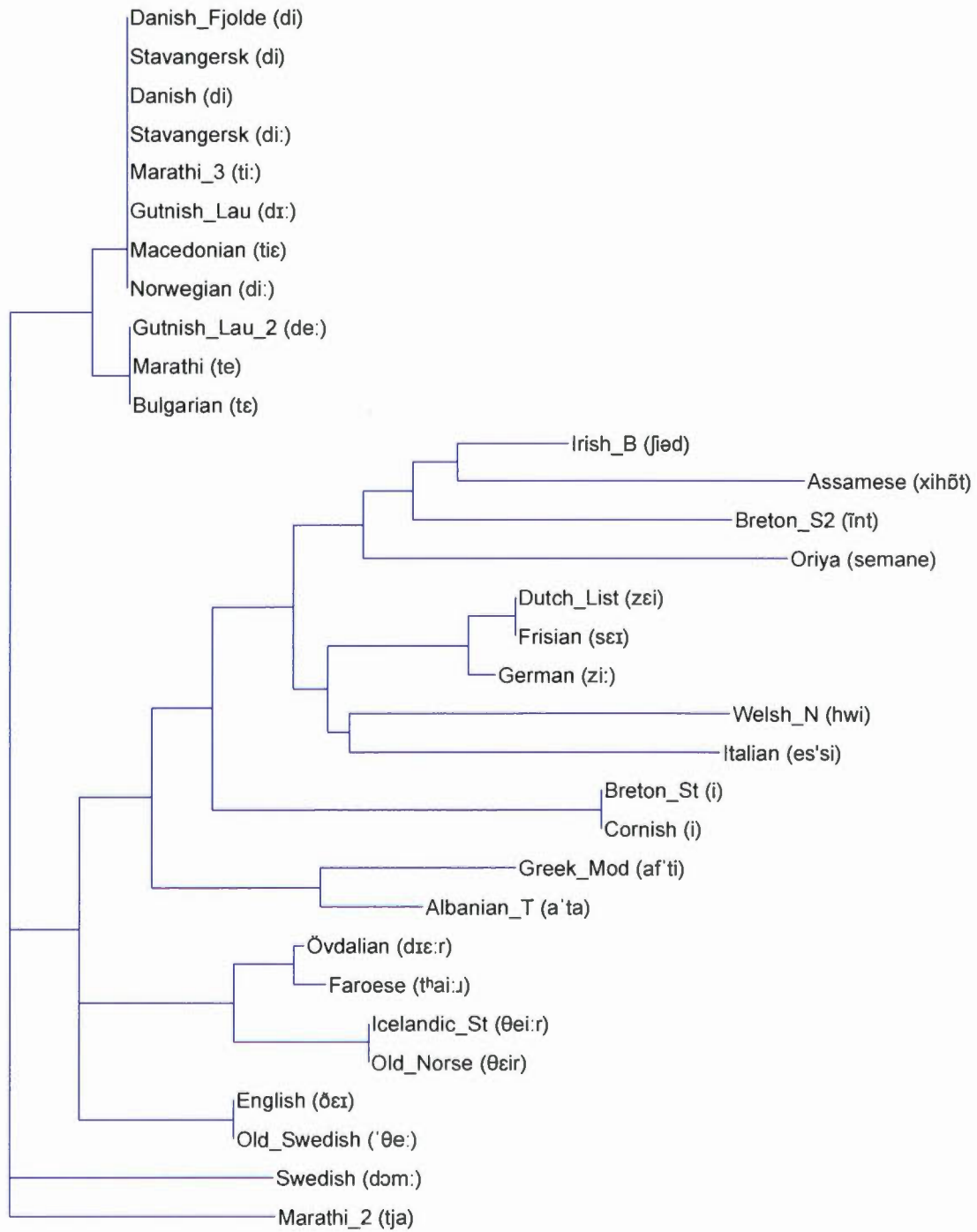


Split

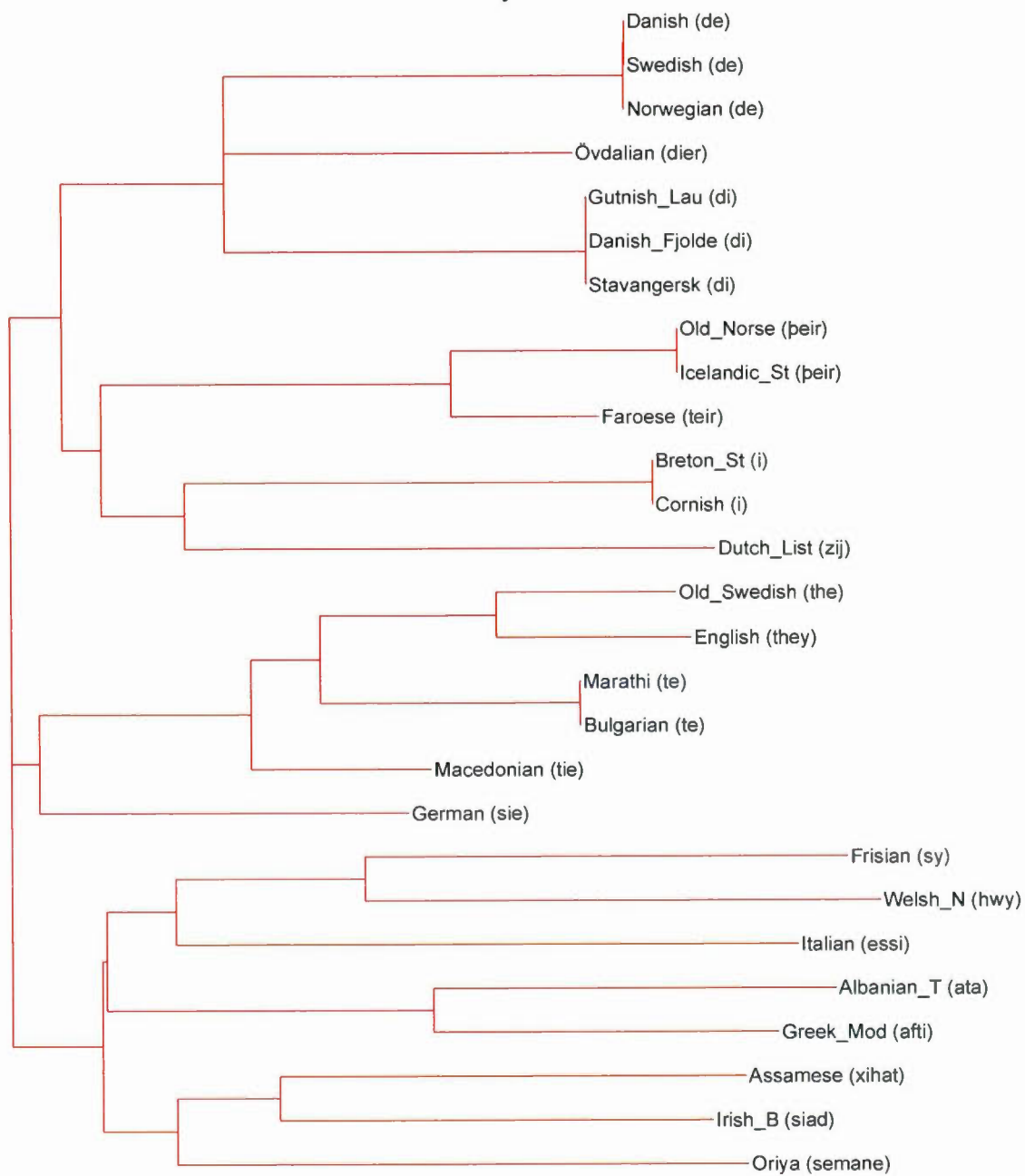


Stick

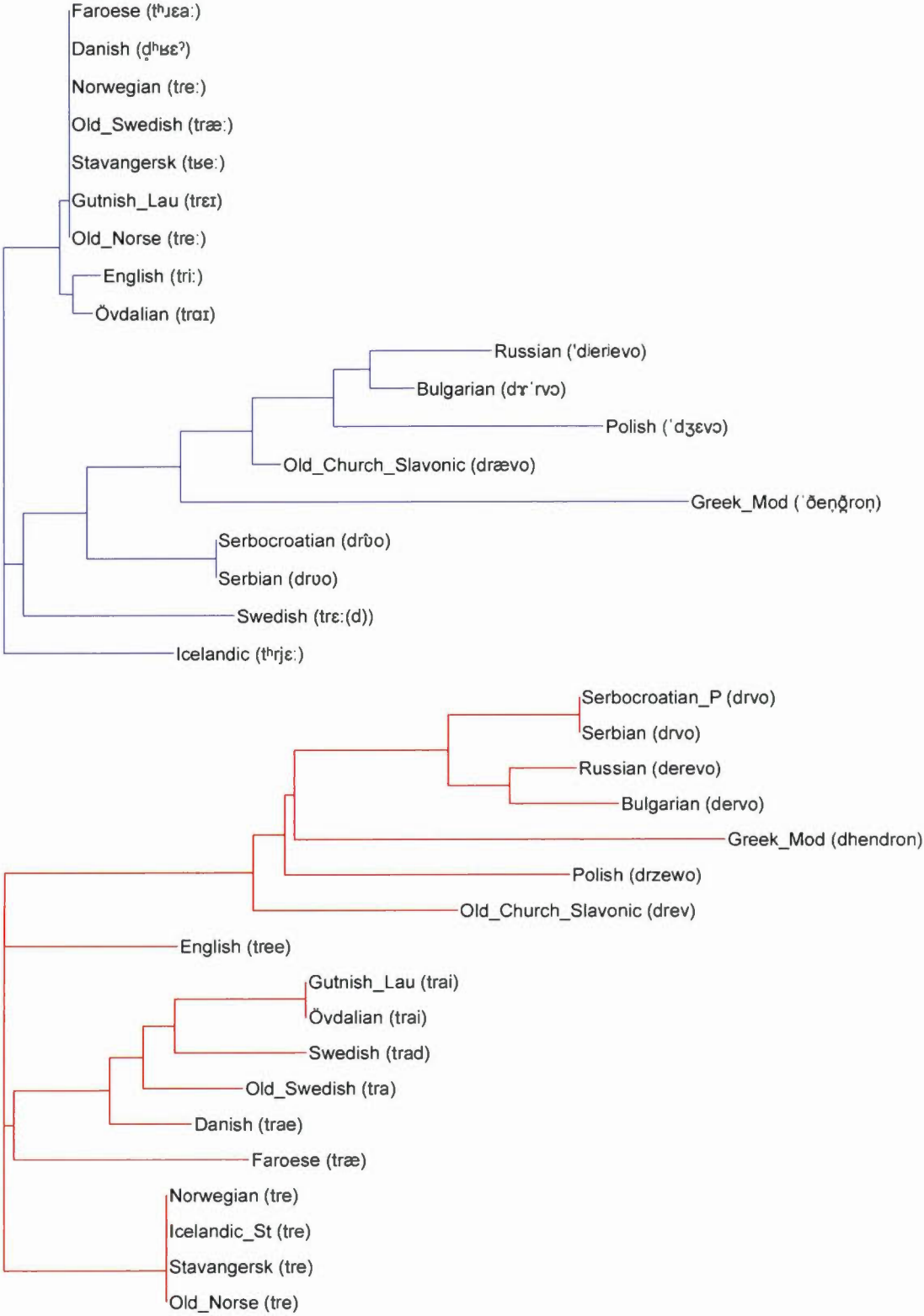




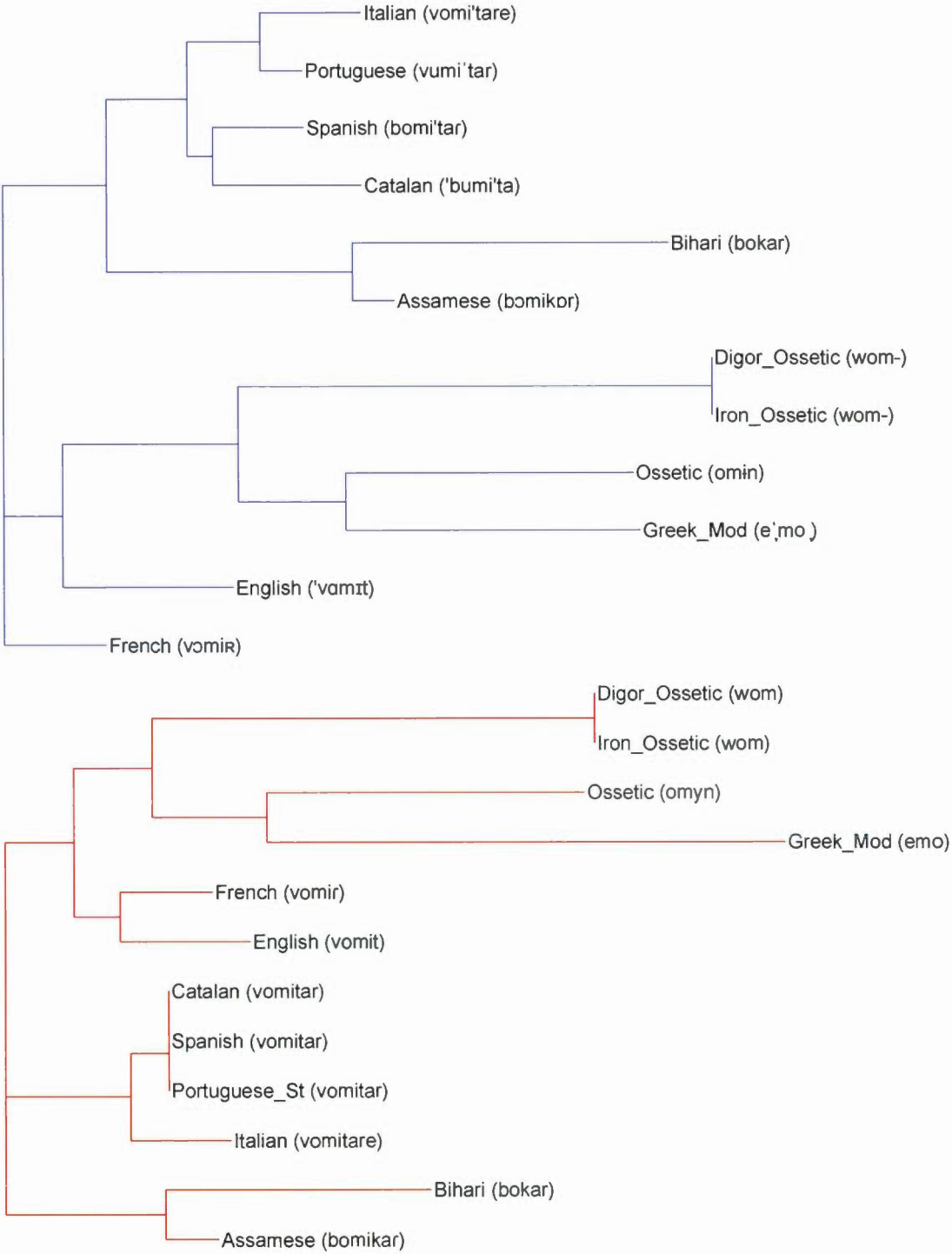
They



Tree

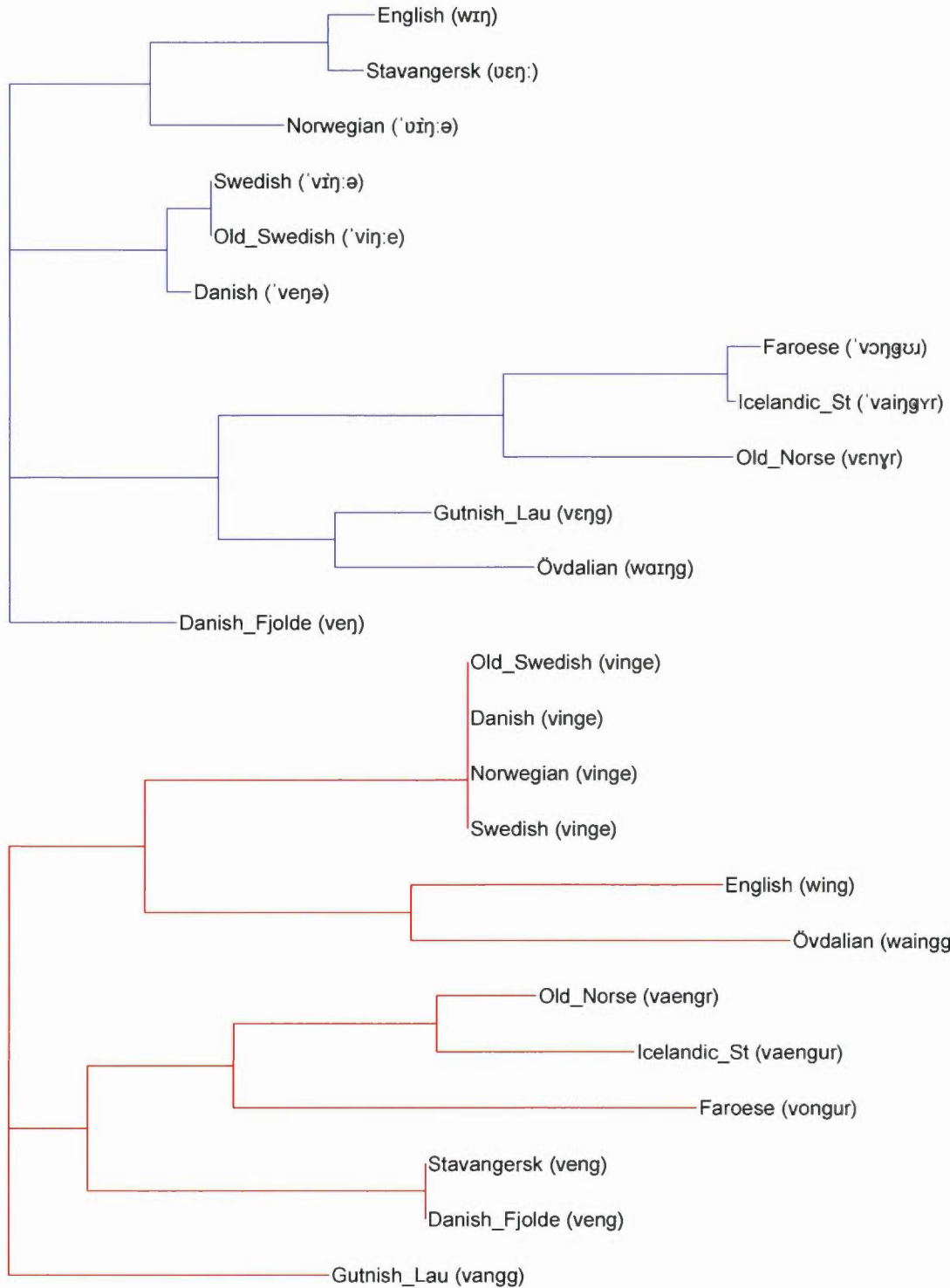


Vomit



Wing

201



BIBLIOGRAPHY

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723.
- Albrecht, B., Scornavacca, C., Cenci, A. & Huson, D. H. (2012). Fast computation of minimum hybridization networks. *Bioinformatics*, 28(2), 191–197.
- Arnold, M. L. (1997). *Natural Hybridization and Evolution*. New York, NY, USA: Oxford University Press.
- Atkinson, Q. D. (2013). The descent of words. *Proceedings of the National Academy of Sciences*, 110(11), 4159–4160.
- Atkinson, Q. D. & Gray, R. D. (2005). Curious Parallels and Curious Connections—Phylogenetic Thinking in Biology and Historical Linguistics. *Systematic Biology*, 54(4), 513–526.
- Atkinson, Q. D. & Gray, R. D. (2006). How Old is the Indo-European Language Family? Illumination or More Moths to the Flame? In P. Forster & C. Renfrew (eds.), *Phylogenetic Methods and the Prehistory of Languages* chapter 8, pp. 91–109. Cambridge, UK: The McDonald Institute for Archaeological Research.
- Atteson, K. (1999). The Performance of Neighbor-Joining Methods of Phylogenetic Reconstruction. *Algorithmica*, 25(2-3), 251–278.
- Ayres, D. L., Darling, A., Zwickl, D. J., Beerli, P., Holder, M. T., Lewis, P. O., Huelsenbeck, J. P., Ronquist, F., Swofford, D. L., Cummings, M. P., Rambaut,

- A. & Suchard, M. A. (2012). BEAGLE: An Application Programming Interface and High-Performance Computing Library for Statistical Phylogenetics. *Systematic Biology*, 61(1), 170–173.
- Bandelt, H.-J. & Dress, A. W. M. (1992a). A canonical decomposition theory for metrics on a finite set. *Advances Math*, 92(1), 47–105.
- Bandelt, H.-J. & Dress, A. W. M. (1992b). Split decomposition: A new and useful approach to phylogenetic analysis of distance data. *Molecular Phylogenetics and Evolution*, 1(3), 242–252.
- Bandelt, H.-J., Forster, P. & Röhl, A. (1999). Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution*, 16(1), 37–48.
- Bandelt, H. J., Forster, P., Sykes, B. C. & Richards, M. B. (1995). Mitochondrial Portraits of Human Populations Using Median Networks. *Genetics*, 141(2), 743–753.
- Bandle, O. (2005). *The Nordic Languages: an international handbook of the history of the North Germanic languages*. Berlin, Germany: Walter de Gruyter.
- Barthélemy, J.-P. & Guénoche, A. (1991). *Trees and proximity representations*. Wiley-Interscience series in discrete mathematics and optimization. Hoboken, NJ, USA: John Wiley & Sons.
- Baugh, A. C. & Cable, T. (2002). *A history of the English language*. 5th ed. London, UK: Routledge.
- Boc, A., Di Sciullo, A. M. & Makarenkov, V. (2010a). Classification of the Indo-European languages using a phylogenetic network approach. In H. Locarek-Junge & C. Weihs (eds.), *Classification as a Tool for Research* pp. 647–655. Berlin Heidelberg, Germany: Springer.

- Boc, A., Diallo, A. & Makarenkov, V. (2012). T-REX: a web server for inferring, validating and visualizing phylogenetic trees and networks. *Nucleic Acids Research*, 40(W1), W573–W579.
- Boc, A. & Makarenkov, V. (2003). *New Efficient Algorithm for Detection of Horizontal Gene Transfer Events*, In *Algorithms in Bioinformatics: Third International Workshop, WABI 2003, Budapest, Hungary, September 15-20, 2003. Proceedings*, (pp. 190–201). Springer: Berlin Heidelberg, Germany.
- Boc, A. & Makarenkov, V. (2011). Towards an accurate identification of mosaic genes and partial horizontal gene transfers. *Nucleic Acids Research*, 39, e144.
- Boc, A., Philippe, H. & Makarenkov, V. (2010b). Inferring and Validating Horizontal Gene Transfer Events Using Bipartition Dissimilarity. *Systematic Biology*, 59(2), 195–211.
- Bonnet, G. (1998). *Les mots latins de l'albanais*. Paris, France: L'Harmattan.
- Bordewich, M., Huber, K. T., Moulton, V. & Semple, C. (2018). Recovering normal networks from shortest inter-taxa distance information. *Journal of Mathematical Biology*.
- Bordewich, M. & Tokac, N. (2016). An algorithm for reconstructing ultrametric tree-child networks from inter-taxa distances. *Discrete Applied Mathematics*, 213, 47–59.
- Bouckaert, R., Lemey, P., Dunn, M., Greenhill, S. J., Alekseyenko, A. V., Drummond, A. J., Gray, R. D., Suchard, M. A. & Atkinson, Q. D. (2012). Mapping the origins and expansion of the indo-european language family. *Science*, 337(6097), 957–960.
- Bournoutian, G. A. (2012). *A concise history of the Armenian people: (From ancient times to the present)*. 6th ed. Costa Mesa, CA, USA: Mazda Publishers.

- Bowern, C., Epps, P., Gray, R., Hill, J., Hunley, K., McConvell, P. & Zentz, J. (2011). Does lateral transmission obscure inheritance in hunter-gatherer languages? *PloS one*, 6(9), e25195.
- Bromham, L. (2017). Curiously the same: swapping tools between linguistics and evolutionary biology. *Biology & Philosophy*, pp. 1–32.
- Bryant, D., Filimon, F. & Gray, R. (2005). Untangling our past: Languages, trees, splits and networks. In R. Mace, S. Holden, & S. Shennan (eds.), *The Evolution of Cultural Diversity: A Phylogenetic Approach*. Walnut Creek, CA, USA: Left Coast Press.
- Bryant, D. & Moulton, V. (2004). Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution*, 21(2), 255–265.
- Buffington, A. F. & Preston, A. B. (1965). *A Pennsylvania German grammar. Revth ed.* Allentown, PA, USA: Schlechter's.
- Buneman, P. (1971). *The Recovery of Trees from Measures of Dissimilarity*, In D. Kendall & P. Tautu (eds.). *Mathematics in the Archeological and Historical Sciences*, (pp. 387–395). Edinburgh University Press: Edinburgh, United Kingdom.
- Buneman, P. (1974). A note on metric properties of trees. *Journal of Combinatorial Theory*, 17(1), 48–50.
- Burnham, K. & Anderson, D. (2003). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. New York, NY, USA: Springer.
- Carlin, E. & Arends, J. (2002). *Atlas of the languages of Suriname*. Leiden, Netherlands: KITLV Press.

- Chan, H.-L., Jansson, J., Lam, T.-W. & Yiu, S.-M. (2006). Reconstructing an ultrametric galled phylogenetic network from a distance matrix. *Journal of Bioinformatics and Computational Biology*, 4(4), 807–832.
- Chen, Z.-Z. & Wang, L. (2012). Algorithms for Reticulate Networks of Multiple Phylogenetic Trees. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(2), 372–384.
- Chen, Z.-Z., Wang, L. & Yamanaka, S. (2012). A fast tool for minimum hybridization networks. *BMC Bioinformatics*, 13, 155.
- Choy, C., Jansson, J., Sadakane, K. & Sung, W.-K. (2005). Computing the maximum agreement of phylogenetic networks. *Theoretical Computer Science*, 335(1), 93–107.
- Clackson, J. (1994). *The linguistic relationship between Armenian and Greek*. Oxford, UK: Philological Society.
- Colonna, V., Boattini, A., Guardiano, C., Dall’Ara, I., Pettener, D., Longobardi, G. & Barbujani, G. (2010). Long-range comparison between genes and languages based on syntactic distances. *Human Heredity*, 70(4), 245–254.
- Darwin, C. (1888). *The descent of man and selection in relation to sex*, volume 1. London, UK: Murray.
- Dawley, R. M. (1989). An introduction to unisexual vertebrates. In R. M. Dawley & J. P. Bogart (eds.), *Evolution and Ecology of unisexual vertebrates*, volume Bulletin 4 pp. 1–18. Albany, NY, USA: New York State Museum.
- Delz, M. (2013). *A theoretical approach to automatic loanword detection*. (PhD thesis). Eberhard-Karls-Universität Tübingen, Germany.

- Di Sciullo, A.-M. (2011). *A biolinguistic approach to variation*, In *The biolinguistic enterprise: new perspectives on the evolution and nature of the human language faculty*, (pp. 305–328). Oxford University Press: Oxford, UK.
- Diallo, A. B., Makarenkov, V. & Blanchette, M. (2010). Ancestors 1.0: a web server for ancestral sequence reconstruction. *Bioinformatics*, 26(1), 130–131.
- Diamond, J. & Bellwood, P. (2003). Farmers and Their Languages: The First Expansions. *Science*, 300(5619), 597–603.
- Dobson, A. J. (1974). Unrooted trees for numerical taxonomy. *Journal of Applied Probability*, 11(1), 32–42.
- Donohue, M., Denham, T. & Oppenheimer, S. (2012). New methodologies for historical linguistics?: Calibrating a lexicon-based methodology for diffusion vs. subgrouping. *Diachronica*, 29(4), 505–522.
- Doolittle, W. F. (1999). Phylogenetic Classification and the Universal Tree. *Science*, 284(5423), 2124–2128.
- Doolittle, W. F., Boucher, Y., Nesbø, C. L., Douady, C. J., Andersson, J. O. & Roger, A. J. (2003). How big is the iceberg of which organellar genes in nuclear genomes are but the tip? *Philosophical Transactions of the Royal Society of London - Series B: Biological Sciences*, 358(1429), 39–58.
- Doyon, J.-P., Scornavacca, C., Gorbunov, K. Y., Szöllosi, G. J., Ranwez, V. & Berry, V. (2010). An Efficient Algorithm for Gene/Species Trees Parsimonious Reconciliation with Losses, Duplications and Transfers. In E. Tannier (ed.). *Comparative Genomic*, volume 6398 of *Lecture Notes in Computer Science*, pp. 93–108., Berlin Heidelberg, Germany. Springer.

- Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. (2012). Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, 29(8), 1969–1973.
- Dunn, M. (2012). The Indo-European Lexical Cognacy Database (IELex). <http://ielex.mpi.nl>. Accessed: 2016-03-14.
- Dyen, I., Kruskal, J. & Black, P. (1992). *An Indoeuropean Classification: A Lexicostatistical Experiment*. American Philosophical Society: Transactions of the American Philosophical Society. Philadelphia, PA, USA: American Philosophical Society.
- Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17(6), 368–376.
- Felsenstein, J. (1989). PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics*, 5, 164–166.
- Felsenstein, J. (2005). PHYLIP (Phylogeny Inference Package). Version 3.6. Distributed by the author.
- Finkenstaedt, T. & Wolff, D. (1973). *Ordered profusion; studies in dictionaries and the English lexicon*. Heidelberg, Germany: Carl Winter.
- Fitch, W. M. (1971). Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Zoology*, 20(4), 406–416.
- Fitch, W. M. (2000). Homology: a personal view on some of the problems. *Trends in genetics*, 16(5), 227–231.
- Flouri, T., Izquierdo-Carrasco, F., Darriba, D., Aberer, A., Nguyen, L.-T., Minh, B., Von Haeseler, A. & Stamatakis, A. (2015). The Phylogenetic Likelihood Library. *Systematic Biology*, 64(2), 356–362.

- Francis, A. R. & Steel, M. (2015). Tree-like reticulation networks - when do tree-like distances also support reticulate evolution? *MBIO*, 259, 12–19.
- Gambette, P. & Huson, D. H. (2008). Improved Layout of Phylogenetic Networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 5(3), 472–479.
- Geisler, H. & List, J.-M. (2013). Do languages grow on trees? the tree metaphor in the history of linguistics. In *Classification and evolution in biology, linguistics and the history of science. concepts–methods–visualization* pp. 111–24. Stuttgart, Germany: Franz Steiner Verlag.
- Gimbutas, M. (1973a). Old Europe c. 7000-3500 B.C.: The Earliest European Civilization before the Infiltration of the Indo-European Peoples. *Journal of Indo-European Studies*, 1(1), 1–20.
- Gimbutas, M. (1973b). The Beginning of the Bronze Age in Europe and the Indo-Europeans: 3500-2500 B. C. *Journal of Indo-European Studies*, 1(2), 163–214.
- Gogarten, J. P., Doolittle, W. F. & Lawrence, J. G. (2002). Prokaryotic evolution in light of gene transfer. *Molecular Biology and Evolution*, 19(12), 2226–2238.
- Gray, R. D. & Atkinson, Q. D. (2003). Language-tree divergence times support the anatolian theory of indo-european origin. *Nature*, 426(6965), 435–439.
- Greenhill, S. J. (2011). Levenshtein distances fail to identify language relationships accurately. *Computational Linguistics*, 37(4), 689–698.
- Grünewald, S., Forslund, K., Dress, A. & Moulton, V. (2007). QNet : an agglomerative method for the construction of phylogenetic networks from weighted quartets. *Molecular Biology and Evolution*, 24(2), 532–8.

- Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W. & Gascuel, O. (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology*, 59(3), 307–321.
- Guindon, S. & Gascuel, O. (2003). A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Systematic Biology*, 52(5), 696–704.
- Gusfield, D., Eddhu, S. & Langley, C. H. (2003). Efficient reconstruction of phylogenetic networks with constrained recombination. In *Proceedings of the IEEE Computer Society conference on Bioinformatics (CSB)*, pp. 363–374., Washington, DC, USA. IEEE Computer Society.
- Gyles, C. & Boerlin, P. (2014). Horizontally transferred genetic elements and their role in pathogenesis of bacterial disease. *Veterinary Pathology*, 51(2), 328–340.
- Hallett, M. T. & Lagergren, J. (2001). Efficient algorithms for lateral gene transfer problems. In *RECOMB '01: Proceedings of the fifth annual international conference on Computational biology*, pp. 149–156., New York, NY, USA. ACM.
- Harper, D. (2001). Douglas Harper's Online Etymology Dictionary. <http://www.etymonline.com>. Accessed: 2016-03-14.
- Heggarty, P., Maguire, W. & McMahon, A. (2010). Splits or waves? trees or webs? how divergence measures and network analysis can unravel language histories. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 365(1559), 3829–3843.
- Hein, J. (1993). A heuristic method to reconstruct the history of sequences subject to recombination. *Journal of Molecular Evolution*, 36(4), 396–406.

- Hodgson, B. H. (1874). *Essays on the languages, literature, and religion of Nepal and Tibet: together with further papers on the geography, ethnology, and commerce of those countries*. London, UK: Trübner & Company.
- Höhna, S., Landis, M. J., Heath, T. A., Boussau, B., Lartillot, N., Moore, B. R., Huelsenbeck, J. P. & Ronquist, F. (2016). RevBayes: Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification Language. *Systematic Biology*, 65(4), 726–736.
- Holland, B. & Moulton, V. (2003). Consensus Networks: A Method for Visualising Incompatibilities in Collections of Trees. In G. Benson & R. D. M. Page (eds.). *WABI*, volume 2812 of *Lecture Notes in Computer Science*, pp. 165–176., Berlin Heidelberg, Germany. Springer.
- Holm, H. J. (2011). "swadesh lists" of albanian revisited and consequences for its position in the indo-european languages. *Journal of Indo-European Studies*, 39(1), 43–99.
- Hotopp, J. C. D. (2011). Horizontal gene transfer between bacteria and animals. *Trends in Genetics*, 27(4), 157 – 163.
- Huber, K. T., van Iersel, L., Moulton, V., Scornavacca, C. & Wu, T. (2017). Reconstructing phylogenetic level-1 networks from nondense binet and trinet sets. *Algorithmica*, 77(1), 173–200.
- Huson, D. H. & Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, 23(2), 254–267.
- Huson, D. H., DeZulian, T., Klopper, T. & Steel, M. A. (2004). Phylogenetic super-networks from partial trees. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(4), 151–158.

- Huson, D. H. & Kloepper, T. H. (2005). Computing recombination networks from binary sequences. *Bioinformatics*, 21(suppl 2), ii159–ii165.
- Huson, D. H. & Klöpper, T. H. (2007). Beyond Galled Trees - Decomposition and Computation of Galled Networks. In T. P. Speed & H. Huang (eds.). *RECOMB*, volume 4453 of *Lecture Notes in Computer Science*, pp. 211–225., Berlin Heidelberg, Germany. Springer.
- Huson, D. H. & Rupp, R. (2008). Summarizing Multiple Gene Trees Using Cluster Networks. In K. A. Crandall & J. Lagergren (eds.). *Algorithms in Bioinformatics*, volume 5251 of *Lecture Notes in Computer Science*, pp. 296–305., Berlin Heidelberg, Germany. Springer.
- Huson, D. H., Rupp, R., Berry, V., Gambette, P. & Paul, C. (2009). Computing galled networks from real data. *Bioinformatics*, 25(12), i85–i93.
- Huson, D. H., Rupp, R. & Scornavacca, C. (2010). *Phylogenetic Networks: Concepts, Algorithms and Applications*. Phylogenetic Networks: Concepts, Algorithms and Applications. Cambridge, UK: Cambridge University Press.
- Huson, D. H. & Scornavacca, C. (2012). Dendroscope 3: An Interactive Tool for Rooted Phylogenetic Trees and Networks. *Systematic Biology*, 61(6), 1061–1067.
- Iersel, L. v., Kelk, S., Lekić, N. & Scornavacca, C. (2014). A practical approximation algorithm for solving massive instances of hybridization number for binary and nonbinary trees. *BMC Bioinformatics*, 15(1), 127.
- Judd, W. S. (2008). *Plant systematics: a phylogenetic approach*. Sunderland, MA, USA: Sinauer Associates.
- Jukes, T. H. & Cantor, C. R. (1969). *Evolution of Protein Molecules*. New York, NY, USA: Academy Press.

- Kachru, B. B., Kachru, Y. & Sridhar, S. N. (2008). *Language in South Asia*. Cambridge, UK: Cambridge University Press.
- Kenrick, D. (2007). *Historical dictionary of the Gypsies (Romanies)*. 2nd ed. Lanham, MD, USA: Scarecrow Press.
- Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16(2), 111–120.
- Kishino, H., Miyata, T. & Hasegawa, M. (1990). Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *Journal of Molecular Evolution*, 31(2), 151–160.
- Kolga, M. (2001). *The red book of the peoples of the Russian Empire*. Tallinn, Estonia: NGO Red Book.
- Köllner, M. (2016). Ancestral state reconstruction and loanword detection. Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics, Tübingen, Germany. Universitätsbibliothek Tübingen.
- Koonin, E. V. (2003). Horizontal gene transfer: the path to maturity. *Molecular Microbiology*, 50(3), 725–727.
- Kuhner, M. & Felsenstein, J. (1994). A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution*, 11(3), 469–468.
- Kumar, A., Black, W. C. & Rai, K. S. (1998). An estimate of phylogenetic relationships among culicine mosquitoes using a restriction map of the rDNA cistron. *Insect Molecular Biology*, 7(4), 367–73.

- Leclerc, B. & Makarenkov, V. (1998). On some relations between 2-trees and tree metrics. *Discrete Mathematics*, 192(1-3), 323–349.
- Legendre, P. (2000). Special section on reticulate evolution. *Journal of Classification*, 17, 153–195.
- Legendre, P. & Makarenkov, V. (2002). Reconstruction of Biogeographic and Evolutionary Networks Using Reticulograms. *Systematic Biology*, 51(2), 199–216.
- Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10(8), 707–710.
- Lightfoot, D. (1979). *Principles of diachronic syntax*. Cambridge, UK: Cambridge University Press.
- Lightfoot, D. (2006). *How new languages emerge*. Cambridge, UK: Cambridge University Press.
- List, J.-M. (2012a). SCA: Phonetic alignment based on sound classes. In M. Slavkovik & D. Lassiter (eds.), *New directions in logic, language, and computation* pp. 32–51. Berlin Heidelberg, Germany: Springer.
- List, J.-M. (2012b). *SCA: Phonetic Alignment Based on Sound Classes*, In *New Directions in Logic, Language and Computation: ESSLLI 2010 and ESSLLI 2011 Student Sessions. Selected Papers*, (pp. 32–51). Springer: Berlin Heidelberg, Germany.
- List, J.-M. (2014). *Sequence comparison in historical linguistics*. Düsseldorf, Germany: Düsseldorf University Press.
- List, J.-M. (2015). Network perspectives on chinese dialect history. *Bulletin of Chinese Linguistics*, 8(1), 27–67.

- List, J.-M., Nelson-Sathi, S., Geisler, H. & Martin, W. (2014a). Networks of lexical borrowing and lateral gene transfer in language and genome evolution. *Bioessays*, 36(2), 141–150.
- List, J.-M., Shijulal, N.-S., Martin, W. & Geisler, H. (2014b). Using phylogenetic networks to model chinese dialect history. *Language Dynamics and Change*, 4(2), 222–252.
- Longobardi, G., Guardiano, C., Silvestri, G., Boattini, A. & Ceolin, A. (2013). Toward a syntactic phylogeny of modern indo-european languages. *Journal of Historical Linguistics*, 3(1), 122–152.
- Lubotsky, A. (2005). IE Etymological Dictionaries Project (Leiden Indo-European Etymological Dictionary Series). <http://dictionaries.brillonline.com/iedo>. Accessed: 2016-03-14.
- Makarenkov, V., Boc, A., Delwiche, C., Diallo, A. B. & Philippe, H. (2006a). New Efficient Algorithm for Modeling Partial and Complete Gene Transfer Scenarios. In *Data Science and Classification*, Studies in Classification, Data Analysis, and Knowledge Organization, pp. 341–349., Berlin Heidelberg, Germany. Springer.
- Makarenkov, V., Kevorkov, D. & Legendre, P. (2006b). Phylogenetic network construction approaches. In D. K. Arora, R. M. Berka, & G. B. Singh (eds.), *Bioinformatics*, volume 6 of *Applied Mycology and Biotechnology* pp. 61 – 97. Amsterdam, Netherlands: Elsevier.
- Makarenkov, V. & Leclerc, B. (2000). Comparison of additive trees using circular orders. *Journal of Computational Biology*, 7(5), 731–744.
- Makarenkov, V. & Legendre, P. (2000). Improving the additive tree representation of a given dissimilarity matrix using reticulations. In H. Kiers, J. Rasson, P. Groenen, & S. M. (eds.). *Data Analysis, Classification, and Related Methods*.

- Studies in Classification, Data Analysis, and Knowledge Organization*, pp. 35–40., Berlin Heidelberg, Germany. Springer.
- Makarenkov, V. & Legendre, P. (2004). From a Phylogenetic Tree to a Reticulated Network. *Journal of Computational Biology*, 11(1), 195–212.
- McGhee, G. (2011). *Convergent Evolution : Limited Forms Most Beautiful*. Cambridge, MA, USA: MIT Press.
- Mirzaei, S. & Wu, Y. (2016). Fast construction of near parsimonious hybridization networks for multiple phylogenetic trees. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 13(3), 565–70.
- Nakhleh, L., Ringe, D. A. & Warnow, T. (2005). Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages. *Language*, 81(2), 382–420.
- Nelson-Sathi, S., List, J.-M., Geisler, H., Fangerau, H., Gray, R. D., Martin, W. & Dagan, T. (2011). Networks uncover hidden lexical borrowing in indo-european language evolution. *Proceedings of the Royal Society of London B: Biological Sciences*, 278(1713), 1794–1803.
- Nelson-Sathi, S., Popa, O., List, J.-M., Geisler, H., Martin, W. F. & Dagan, T. (2013). Reconstructing the lateral component of language history and genome evolution using network approaches. In *Classification and evolution in biology, linguistics and the history of science. Concepts – methods – visualization* pp. 163–80. Stuttgart, Germany: Franz Steiner.
- Olave, M., Avila, L. J., Sites, J. W. & Morando, M. (2017). Detecting hybridization by likelihood calculation of gene tree extra lineages given explicit models. *Methods in Ecology and Evolution*, in press.

- Pagel, M. (2000). Maximum likelihood models for glottochronology and for reconstructing linguistic phylogenies. In C. Renfrew, A. MacMahon, & L. Trask (eds.). *Time depth in historical linguistics*, Papers in the prehistory of languages, pp. 189–207., Cambridge, UK: The McDonald Institute for Archaeological Research.
- Patrinos, A. N. & Hakimi, S. L. (1972). The distance matrix of a graph and its tree realisation. *Quarterly of Applied Mathematics*, 30, 255–269.
- Pierre, B. (1973). *La langue occitane*. 3rd ed. Que sais-je ? Paris, France: PUF.
- Piette, J. R. F. (1973). *French loanwords in Middle Breton*. Cardiff, UK: University of Wales Press.
- Piovesan, T. & Kelk, S. (2013). A simple fixed parameter tractable algorithm for computing the hybridization number of two (not necessarily binary) trees. *Transactions on Computational Biology and Bioinformatics*, 10(1), 18–25.
- Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (2007). *Numerical Recipes : The Art of Scientific Computing* (3 éd.). New York, NY, USA: Cambridge University Press.
- Rambaut, A. & Grass, N. C. (1997). Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer applications in the biosciences : CABIOS*, 13(3), 235–238.
- Rannala, B. & Yang, Z. (1996). Probability Distribution of Molecular Evolutionary Trees: A New Method of Phylogenetic Inference. *Journal of Molecular Evolution*, 43(3), 304–311.
- Renfrew, C. (1988). *Archaeology and language: the puzzle of Indo-European origins*. London, UK: J. Cape.

- Rexová, K., Frynta, D. & Zrzavý, J. (2003). Cladistic analysis of languages: Indo-european classification based on lexicostatistical data. *Cladistics*, 19(2), 120–127.
- Rieseberg, L. H. & Ellstrand, N. C. (1993). What can molecular and morphological markers tell us about plant hybridization? *Critical Reviews in Plant Sciences*, 12(3), 213–241.
- Rieseberg, L. H. & Morefield, J. D. (1995). Character expression, phylogenetic reconstruction, and the detection of reticulate evolution. In P. C. Hoch & A. G. Stephenson (eds.), *Experimental and molecular approaches to plant biosystematics* pp. 333–353. Saint-Louis, MO, USA: Monographs in Systematic Botany at the Missouri Botanical Garden.
- Roberts, I. (2007). *Diachronic syntax*. Oxford, UK: Oxford University Press.
- Robinson, K. M., Sieber, K. B. & Dunning Hotopp, J. C. (2013). A Review of Bacteria-Animal Lateral Gene Transfer May Inform Our Understanding of Diseases like Cancer. *PLOS Genetics*, 9(10), 1–6.
- Robinson, O. W. (1947). *Old English and its closest relatives*. Stanford, CA, USA: Stanford University Press.
- Rolf Jr, H. B. (2009). *An introduction to Old Frisian. History, grammar, reader, glossary*. Amsterdam, Netherlands: John Benjamins.
- Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A. & Huelsenbeck, J. P. (2012). MrBayes 3.2 : Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. *Systematic Biology*, 61(3), 539–542.
- Ruan, Y., House, G. L., Ekanayake, S., Schütte, U., Bever, J. D., Tang, H. & Fox, G. (2014). Integration of Clustering and Multidimensional Scaling to Determine

- Phylogenetic Trees as Spherical Phylograms Visualized in 3 Dimensions. In *2014 14th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*, pp. 720–729., Red Hook, NY, USA. Curran Associates.
- Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4), 406–425.
- Sawyer, S. (1989). Statistical tests for detecting gene conversion. *Molecular Biology and Evolution*, 6(5), 526–538.
- Schleicher, A. (1873). *Die Darwinsche theorie und die sprachwissenschaft: Offenes sendschreiben an herrn Ernst Hackel*. Weimar, Germany: Hermann Bohlau.
- Schmidt, H., Strimmer, K., Vingron, M. & von Haeseler, A. (2002). TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, 18(3), 502–504.
- Schmidt, J. (1872). *Die Verwantschaftsverhaltnisse der indogermanischen Sprachen*. Weimar, Germany: Hermann Bohlau.
- Schwarz, G. & Gideon, E. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2), 461–464.
- Semple, C. (2007). Hybridization networks. In O. Gascuel & M. Steel (eds.). *Reconstructing Evolution: New Mathematical and Computational Advances*, pp. 277–314., Oxford, UK. Oxford University Press.
- Smith, N. & Bergin, T. G. (1984). *An old Provenal primer*. New York, NY, USA: Garland.
- Smouse, P. E. (2000). Reticulation inside the species boundary. *Journal of Classification*, 17(2), 165–173.

- Sneath, P. H. A., Sackin, M. J. & Ambler, R. P. (1975). Detecting Evolutionary Incompatibilities From Protein Sequences. *Systematic Biology*, 24(3), 311–332.
- Sneath, P. H. A. & Sokal, R. R. (1973). *Numerical Taxonomy. The Principles and Practice of Numerical Classification*. San Francisco, CA, USA: W. H. Freeman.
- Solís-Lemus, C. & Ané, C. (2016). Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS genetics*, 12(3), e1005896.
- Sonea, S. & Panisset, M. (1983). *A new bacteriology*. Burlington, MA, USA: Jones and Bartlett.
- Stace, C. A. (1991). *Plant Taxonomy and Biosystematics*. Cambridge, UK: Cambridge University Press.
- Stamatakis, A., Aberer, A. J., Goll, C., Smith, S. A., Berger, S. A. & Izquierdo-Carrasco, F. (2012). RAxML-Light: A Tool for computing TeraByte Phylogenies. <http://www.exelixis-lab.org/>.
- Stamatakis, A., Hoover, P. & Rougemont, J. (2008). A Rapid Bootstrap Algorithm for the RAxML Web Servers. *Systematic Biology*, 57(5), 758–771.
- Steiner, L., Stadler, P. F. & Cysouw, M. (2011). A pipeline for computational historical linguistics. *Language Dynamics and Change*, 1(1), 89–127.
- Stephens, J. C. (1985). Statistical methods of DNA sequence analysis: detection of intragenic recombination or gene conversion. *Molecular Biology and Evolution*, 2(6), 539–556.
- Strimmer, V. & von Haeseler, A. (1996). Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. *Molecular Biology and Evolution*, 13(7), 964–969.

- Studier, J. A. & Keppler, K. J. (1988). A note on the neighbor-joining algorithm of Saitou and Nei. *Molecular Biology and Evolution*, 5(6), 729–731.
- Swadesh, M. (1952). Lexico-Statistic Dating of Prehistoric Ethnic Contacts: With Special Reference to North American Indians and Eskimos. *Proceedings of the American Philosophical Society*, 96(4), 452–463.
- Swafford, D. (2002). PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods). Version 4.
- Thomason, S. G. & Kaufman, T. (1988). *Language contact, creolization, and genetic linguistics*. Berkeley and Los Angeles, CA, USA: University of California Press.
- Trask, R. L. (2000). *The dictionary of historical and comparative linguistics*. Edinburgh, UK: Edinburgh University Press.
- Van Der Aek, R., Menecier, P., Nerbonne, J. & Manni, F. (2007). Preliminary Identification of Language Groups and Loan Words in Central Asia. In P. Osenova (ed.). *conference Recent Advances in Natural Language Phonology*, pp. 13–20., Borovetz, Bulgaria.
- Van Iersel, L. & Kelk, S. (2011). Constructing the Simplest Possible Phylogenetic Network from Triplets. *Algorithmica*, 60(6097), 207–235.
- Van Iersel, L., Kelk, S., Rupp, R. & Huson, D. (2010). Phylogenetic networks do not need to be complex: using fewer reticulations to represent conflicting clusters. *Bioinformatics*, 26(12), i124–i131.
- Velupillai, V. (2015). *Pidgins, creoles and mixed languages: An introduction*. Amsterdam, Netherlands: John Benjamins Publishing Company.
- Vogt, T. & Geis, T. (2007). *Wort für Wort*. Beilefeld, Germany: Reise Know-How.

- Wang, W. S.-Y. & Minett, J. W. (2005). Vertical and horizontal transmission in language evolution. *Transactions of the Philological Society*, 103(2), 121–146.
- Wen, D., Yu, Y. & Nakhleh, L. (2016). Bayesian inference of reticulate phylogenies under the multispecies network coalescent. *PLoS genetics*, 12(5), e1006006.
- Wichmann, S., Walker, R., Rama, T. & Holman, E. W. (2011). Correlates of reticulation in linguistic phylogenies. *Language Dynamics and Change*, 1(2), 205–240.
- Willems, M., Lord, E., Laforest, L., Labelle, G., Lapointe, F.-J., Di Sciullo, A. M. & Makarenkov, V. (2016). Using hybridization networks to retrace the evolution of indo-european languages. *BMC Evolutionary Biology*, 16(1), 180.
- Willems, M., Tahiri, N. & Makarenkov, V. (2014). A new efficient algorithm for inferring explicit hybridization networks following the neighbor-joining principle. *Journal of Bioinformatics and Computational Biology*, 12(05), 1450024.
- Willson, S. J. (2013). Reconstruction of certain phylogenetic networks from their tree-average distances. *Bulletin of Mathematical Biology*, 75(10), 1840–1878.
- Woolley, S. M., Posada, D. & Crandall, K. A. (2008). A Comparison of Phylogenetic Network Methods Using Computer Simulation. *PLoS ONE*, 3(4), e1913.
- Wu, Y. (2010). Close lower and upper bounds for the minimum reticulate network of multiple phylogenetic trees. *Bioinformatics [ISMB]*, 26(12), 140–148.
- Yang, J., Grünewald, S. & Wan, X. (2013). Quartet-Net: A Quartet Based Method to Reconstruct Phylogenetic Networks. *Molecular Biology and Evolution*, 30(5), 1206–17.
- Zaretskii, K. A. (1965). Constructing a tree on the basis of a set of distances between hanging vertices. *Uspekhi Matematicheskikh Nauk*, 20(6), 90–92.

Zhaxybayeva, O., Lapierre, P. & Gogarten, J. P. (2004). Genome mosaicism and organismal lineages. *Trends in genetics : TIG*, 20(5), 254–260.