NOUVEAUX ALGORITHMES POUR L'INFÉRENCE DE RÉSEAUX PHYLOGÉNÉTIQUES

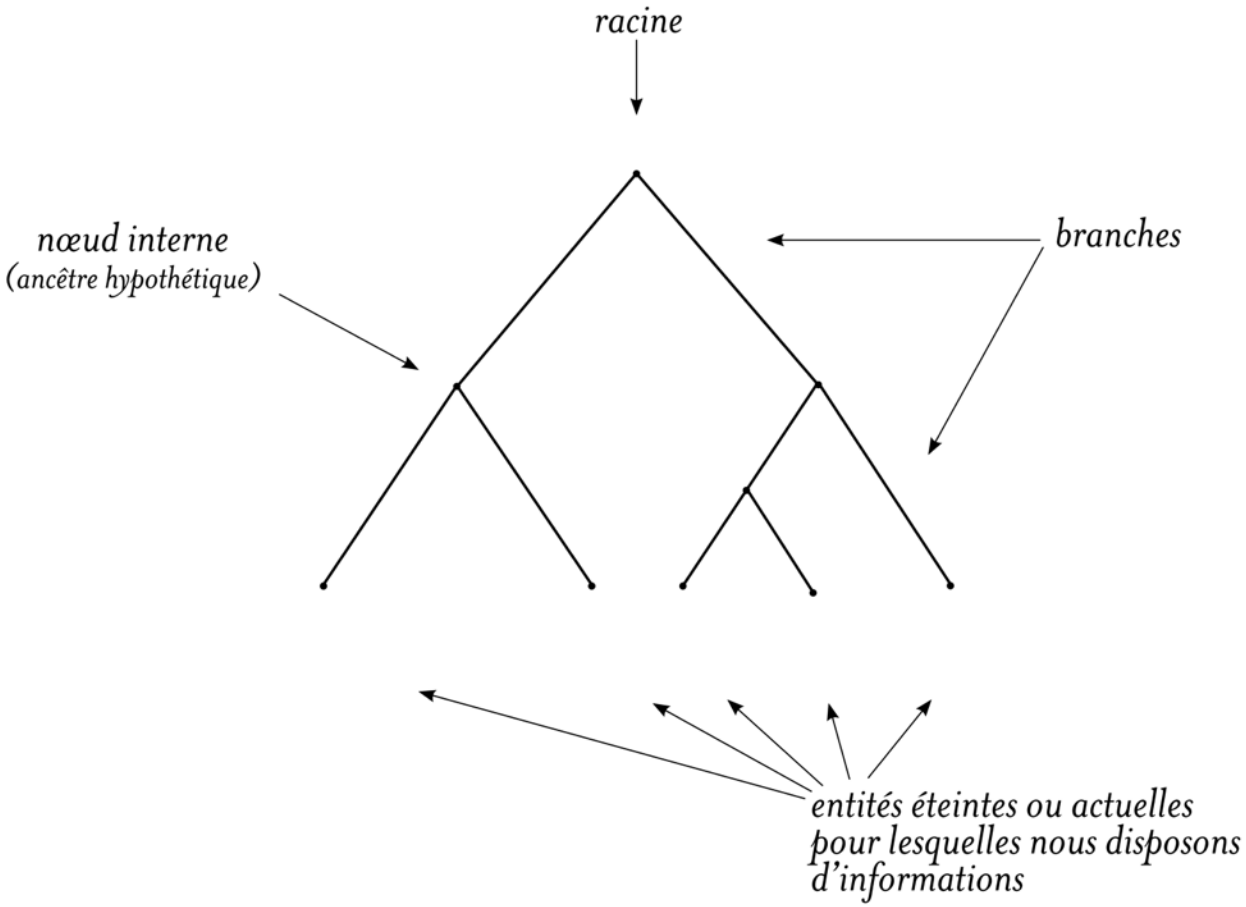
Matthieu Willems

PLAN

1. Introduction
    1. L'inférence phylogénétique
    2. L'évolution réticulée
    3. Les réseaux phylogénétiques
2. Un nouvel algorithme d'inférence de réseaux phylogénétique basé sur l'algorithme neighbor-joining
    1. Description de l'algorithme
    2. Résultats de différentes simulations
3. Application en biolinguistique
    1. Phylogénie et linguistique
    2. Réseaux biolinguistiques obtenus
4. Un nouvel algorithme d'inférence de réseaux phylogénétique basé sur les caractères
    1. Description de l'algorithme
    2. Résultats de différentes simulations
5. Perspectives

1.1. L'inférence phylogénétique

Objectif : reconstruire l'histoire évolutive d'un ensemble d'espèce à partir de données moléculaires



Entrées des algorithmes phylogénétiques

- Méthodes basées sur les caractères (séquences moléculaires) :

E1 : ATCGGATCGTATTCGA

E2 : ATGCCTAGGATATGGT

E3 : TTGGGAACGCATCCTA

- Matrices de distances :

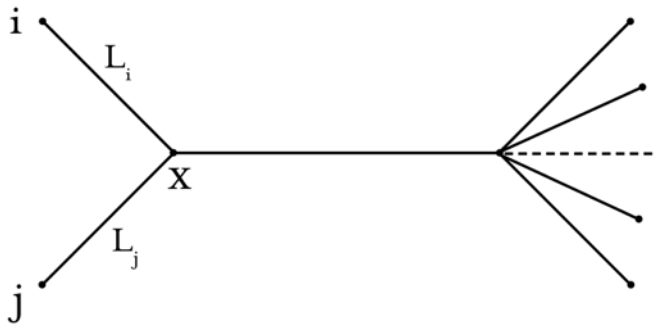
	E1	E2	E3
E1	0	11	6
E2	11	0	12
E3	6	12	0

Principales méthodes

1. Méthodes de distances (NJ, UPGMA).
2. Maximum de parcimonie.
3. Maximum de vraisemblance.
4. Méthodes bayésiennes.
5. Peu de nouveaux développements depuis 15 ans.

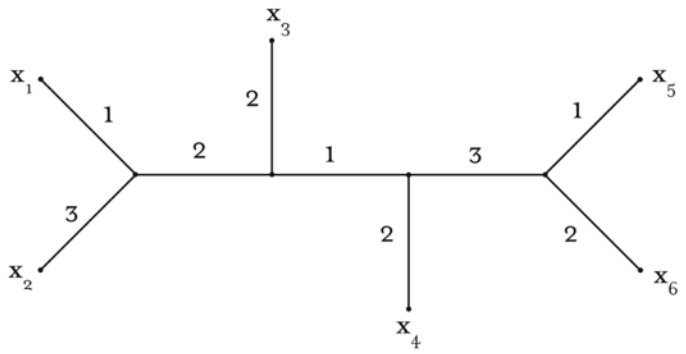
Neighbor-joining (NJ)

- Méthode développée par Saitou et Nei (1987).
- Méthode de distances la plus utilisée.
- Entrée : Matrice de distances sur n espèces.
- Sortie : Arbre phylogénétique non enraciné dont les feuilles sont en correspondance avec les n espèces considérées.
- Complexité : $O(n^3)$.
- Algorithme itératif.



Arbres additifs

	X_1	X_2	X_3	X_4	X_5	X_6
X_1		4	5	6	8	9
X_2			7	8	10	11
X_3				5	7	8
X_4					6	7
X_5						3
X_6						



Condition des 4 points :

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

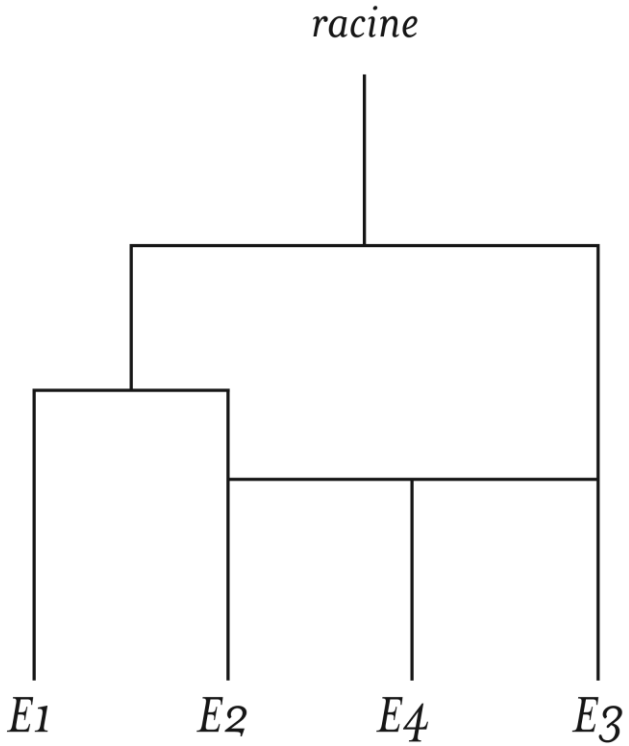
$\leftrightarrow \text{Min}\{d(x;y)+d(z;w)-d(x;z)-d(y;w) ; d(x;y)+d(z;w)-d(x;w)-d(y;z)\} \leq 0.$

1.2. Évolution réticulée (non arborescente)

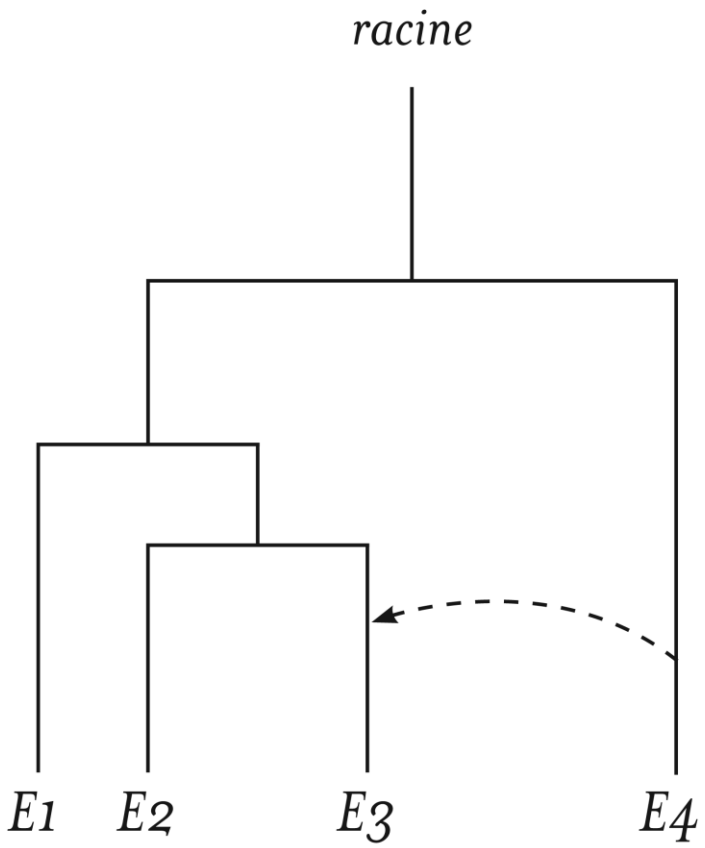
- Hybridation
- Transferts horizontaux de gènes
- Homoplasie
- Autres phénomènes

Comment prendre en compte ces phénomènes dans les algorithmes d'inférences d'arbres phylogénétiques ?

Hybridation



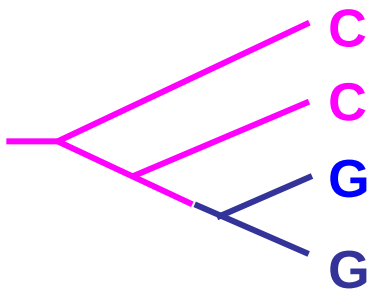
Transfert horizontal



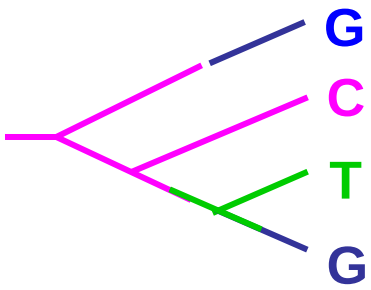
Homoplasie

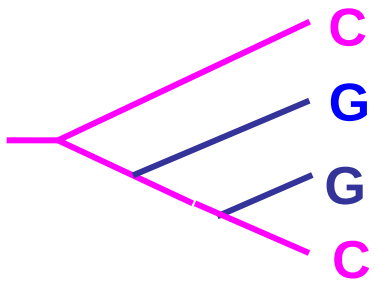
La similarité entre deux entités

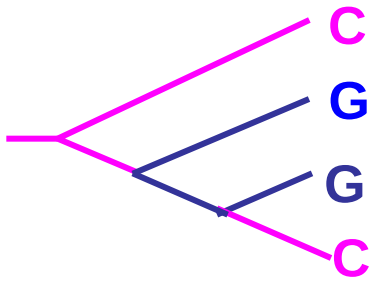
Relation évolutionnaire :
Caractères ancestraux partagés (‘plésiomorphies’)
Caractères dérivés partagés (‘’synapomorphie’)



Homoplasie (évolution indépendante du même caractère) :
Évènements convergents, évènements parallèles, évènements inverses.







Adapté de C-B Stewart Lecture (2000)

Autres phénomènes

- Duplications
- Pertes de gènes
- Recombinaisons génétiques

1.3. Réseaux phylogénétiques

- Définition : Graphe utilisé pour représenter des relations d'évolution entre un ensemble de taxons qui sont associés à certains des nœuds du graphe (généralement les feuilles).

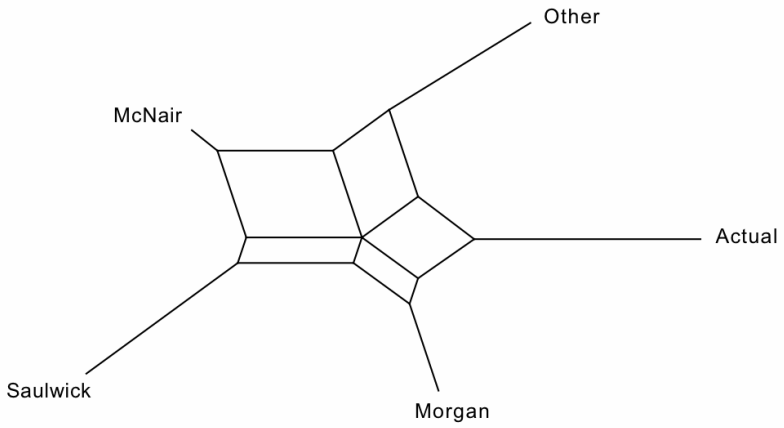
- Réseau explicite : arbre auquel on rajoute des réticulations qui représentent explicitement certains phénomènes évolutifs comme l'hybridation.

- Réseau abstrait : réseau qui permet seulement de visualiser certaines incompatibilités dans les données sans expliciter des phénomènes biologiques particuliers.

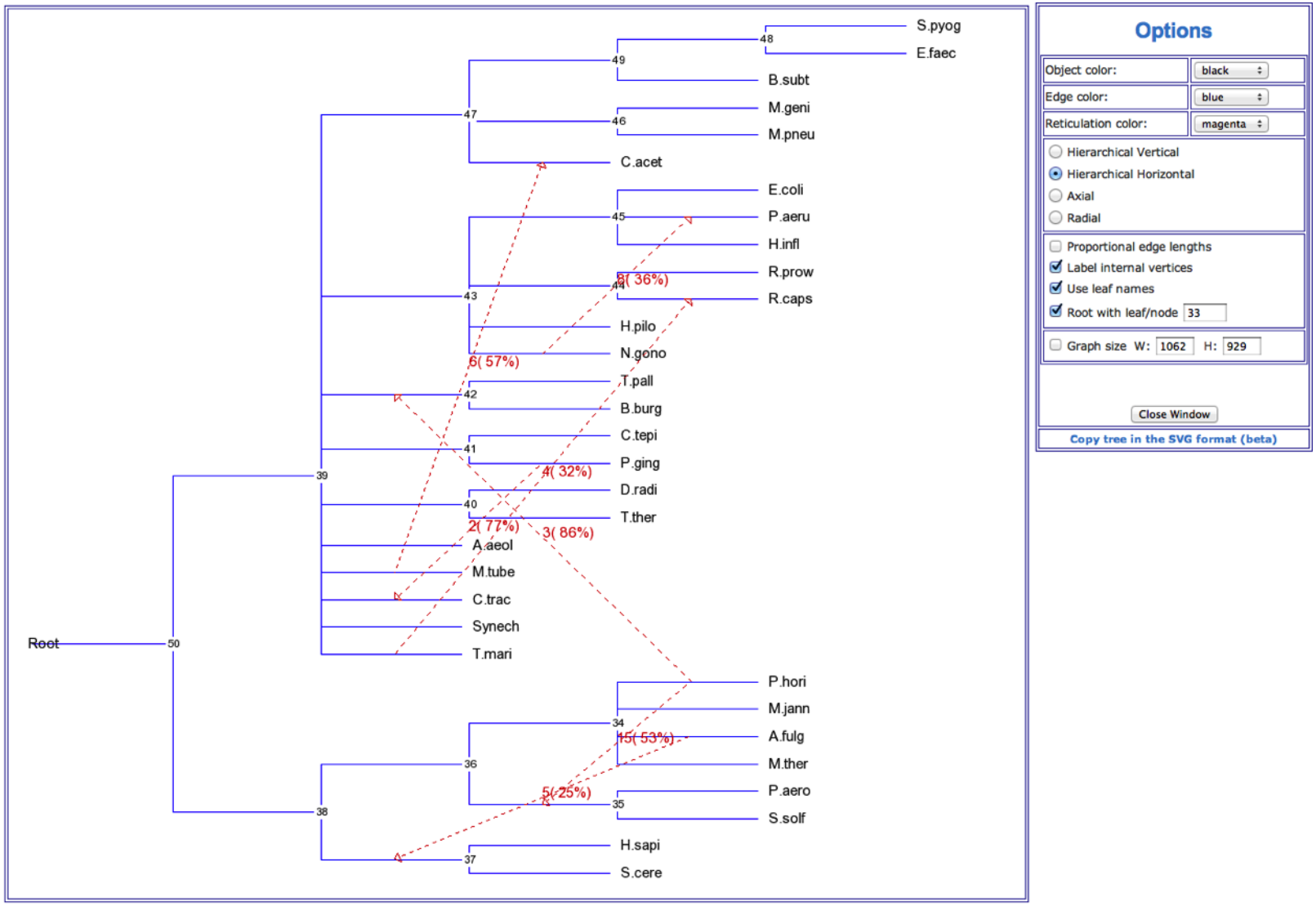
- De nombreuses méthodes d'inférence de tels réseaux ont été développées depuis une quinzaine d'années.

Principales méthodes d'inférence de réseaux phylogénétiques

- Réseaux de bipartitions (Bandelt et Dress, 1992).
- NeigborNet (Bryant et Moulton, 2004) : Méthode d'inférence de split-graphes (représentations graphiques de réseaux de bipartition) à partir de matrices de distances.
- SplitsTree (Huson et Bryant, 2006) : Logiciel d'inférence de split-graphes à partir de différents types de données en utilisant la méthode NeighborNet.

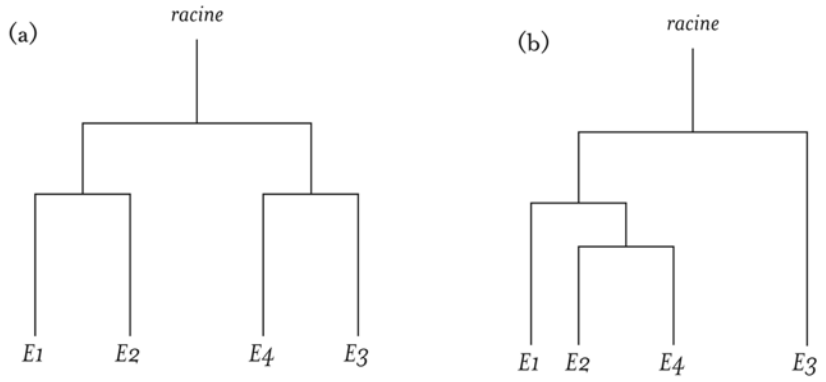


- T-Rex (Boc *et al*, 2012) : Identification de transferts horizontaux de gènes à partir d'un ensemble d'arbres contradictoires.

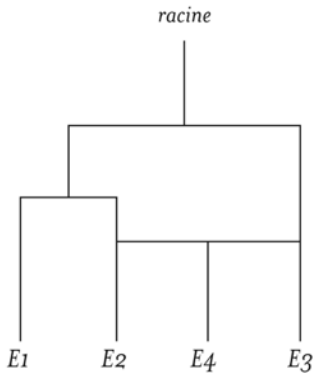


- Inférence de réseaux d'hybridation à partir d'un ensemble d'arbres contradictoires : (Albrecht *et al*, 2012), (Chen et Wang, 2012).

2 arbres contradictoires



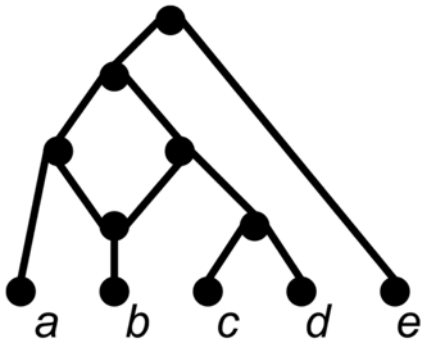
Réseau d'hybridation correspondant



- Dendroscope (Huson et Scornavacca, 2012) : Inférence de réseaux de clusters.

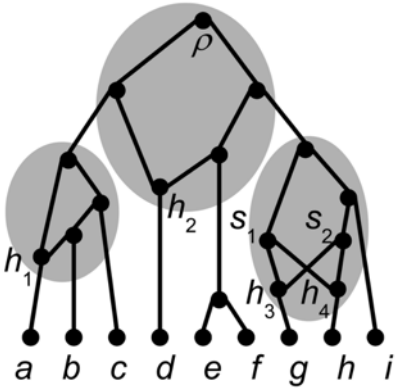
- Réseaux de niveau k (Van Iersel *et al*, 2010) : réseaux où le nombre maximum de réticulations contenues entièrement dans une composante biconnexe de N est égal à k.

Niveau 1



(a)

Niveau 2



(b)

2.1. Description d'un nouvel algorithme basé sur neighbor-joining

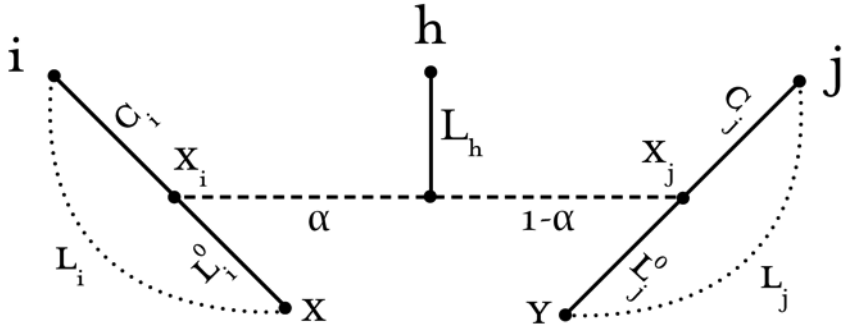
Deux principes fondamentaux :

1. Regroupement hiérarchique (algorithme neigbor-joining).
2. Principe des moindres carrés.

Willems, M., Tahiri, N. et Makarenkov, V. (2014). A new efficient algorithm for inferring explicit hybridization networks following the Neighbor-Joining principle, *Journal of Bioinformatics and Computational Biology*, **12**(5), DOI: 10.1142/S0219720014500243

Définition des distances (dans un réseau additif)

- $D[i][h]=L_h+L_i- \alpha L^0_i+(1- \alpha)(L^0_j+d(Y;X))$, **(1)**
- $D[j][h]=L_h+L_j- (1-\alpha)L^0_j+\alpha(L^0_i+d(Y;X))$, **(2)**
- $D[k][h]=L_h+\alpha(L^0_i+d(X;k))+(1- \alpha)(L^0_j+d(Y;k))$. **(3)**



Propriété fondamentale

Dans un réseau additif, si h est l'hybride de i et j :

$$\text{Min}\{ D[i][j]+D[k][h]-D[i][h]-D[k][j] \; ; \; D[i][j]+D[k][h]-D[j][h]-D[k][i] \} > 0,$$

ce minimum étant pris sur toutes les espèces k différentes de i, j et h.

On note MIN_{ijh} ce minimum.

Utilisation du principe des moindres carrés

Pour tout triplet (i, j, h), on peut calculer :

1. Un degré d'hybridation : α_{ijh},

2. Un score d'hybridation L_{ijh} (Plus ce score est petit, plus la probabilité que h soit l'hybride de i et j est élevée).

Principes du nouvel algorithme

- Entrée : Une matrice de distances sur n espèces, un seuil d'hybridation minimal α_{Min} et maximal α_{Max} (entre 0 et 1).
- Sortie : Un réseau d'hybridation explicite dont les n feuilles sont en correspondance avec les n espèces en entrée.
- Complexité : $O(n^3)$.
- Algorithme itératif.

Principes de chaque itération

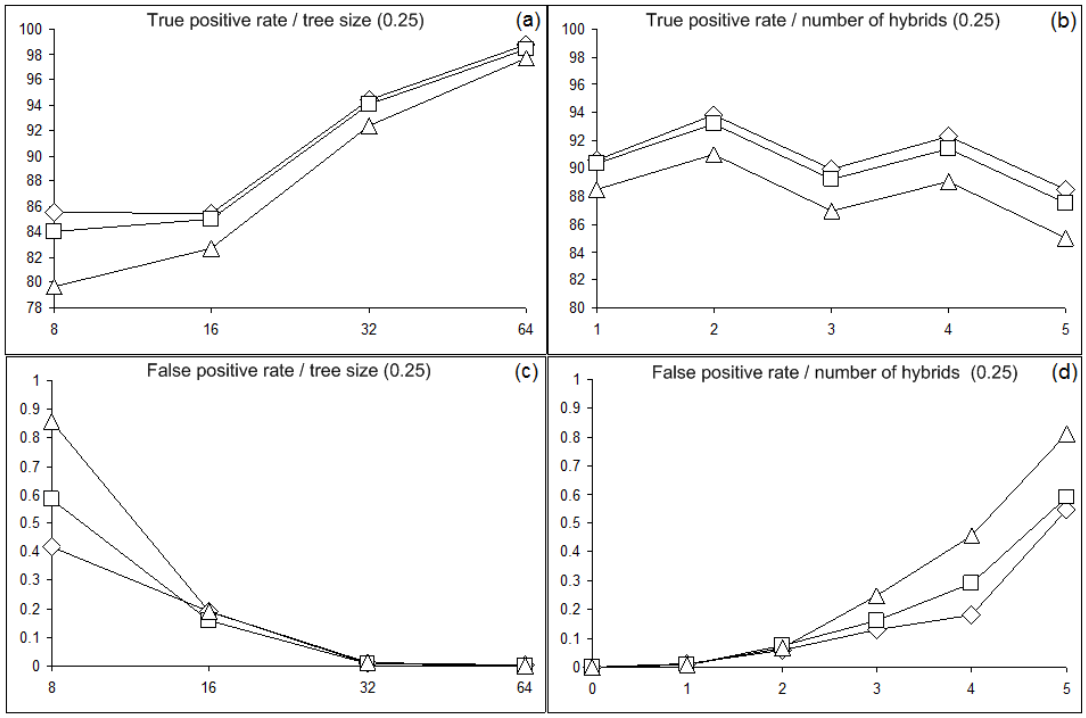
- On détermine deux espèces voisines ii et jj par NJ.
- On cherche le meilleur hybride h (de i et j) ayant ii ou jj pour parent : le triplet i, j, h doit vérifier $\alpha_{Min} \leq \alpha_{ijh} \leq \alpha_{Max}$, $MIN_{ijh} > 0$ et avoir le score L_{ijh} le plus petit possible.
- Si L_{ijh} est plus petit qu'un seuil défini à chaque itération, on élimine l'espèce h qui est considérée comme l'hybride de i et j.
- Sinon, on remplace ii et jj par leur ancêtre commun, comme dans l'algorithme NJ classique.

2.2. Résultats de différentes simulations

Données additives :

- Génération de 1000 matrices de distances d'arbres additifs de tailles 8, 16, 32, 64 (avec T-Rex).
- Ajout de 0 à 5 hybrides dans ces matrices en utilisant les formules 1 à 3.
- Degrés d'hybridation : 0,3 ; 0,4 ; 0,5.
- Au total : 64 000 matrices de distances.

Résultats des simulations (données additives)

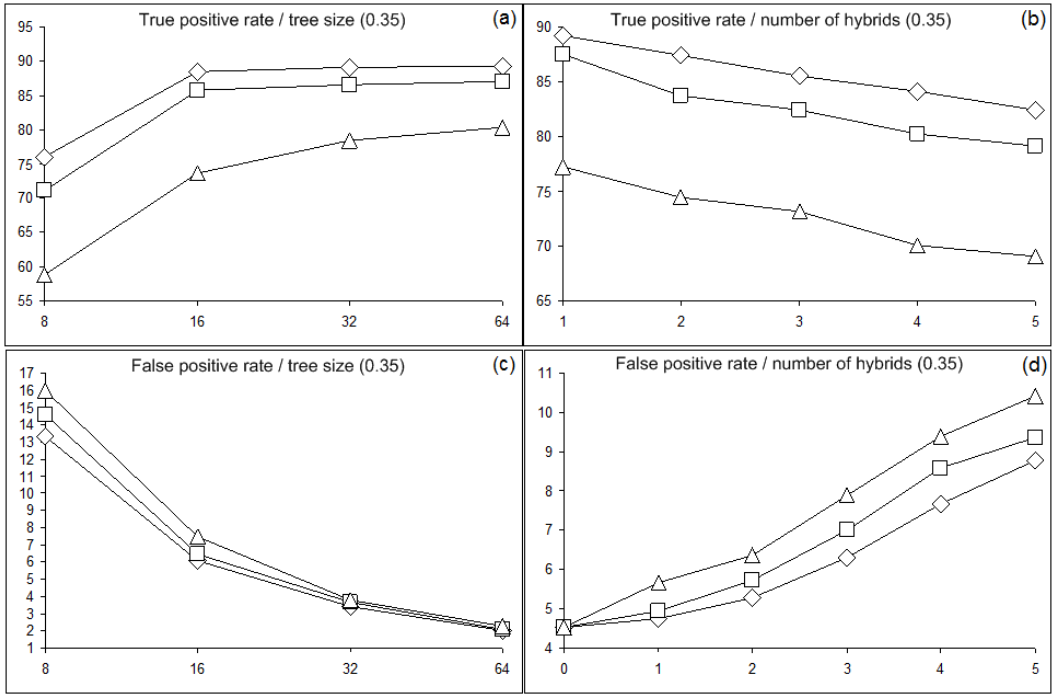


Δ : α=0,3 ; ☐ : α=0,4 ; ◊ : α=0,5

Données non additives :

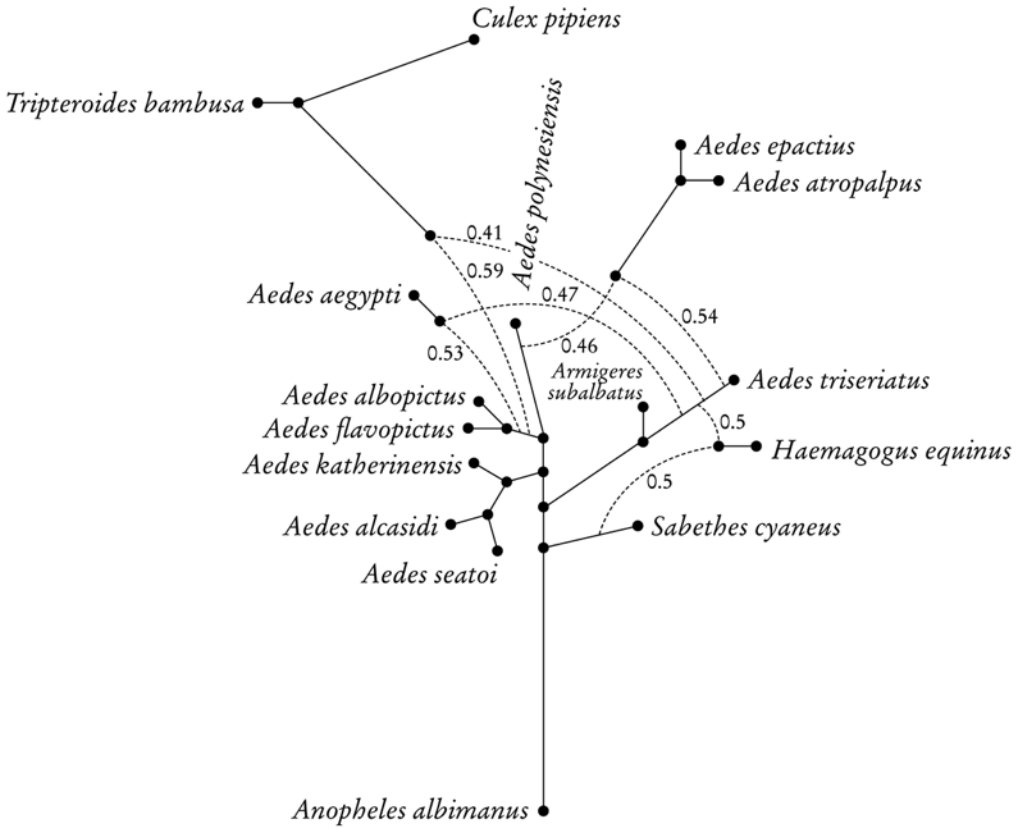
- Génération de 1000 arbres additifs de tailles 8, 16, 32, 64 (avec T-Rex).
- Génération de séquences pour les feuilles de chacun de ces arbres (avec SeqGen) en utilisant le modèle d'évolution K2P.
- Ajout de 0 à 5 séquences hybrides.
- Degrés d'hybridation : 0,3 ; 0,4 ; 0,5.
- Calcul des matrices de distances à partir de ces séquences et du modèle K2P.
- Au total : 64 000 matrices de distances.

Résultats des simulations (données non-additives)



Δ : α=0,3 ; ☐ : α=0,4 ; ◊ : α=0,5

Réseau d'hybridation pour 16 espèces de moustiques



Sites de restrictions du rDNA de 16 espèces de moustiques

Aedes albopictus	11110101010100010101010010
Aedes aegypti	11110101000100010101000010
Aedes seatoi	11110101010100010101010000
Aedes flavopictus	11110101010100010101010010
Aedes alcasidi	11110101010100010101010000
Aedes katherinensis	11110101010100010101010000
Aedes polynesiensis	11110101000100010101010010
Aedes triseriatus	10110101000110010101000000
Aedes atropalpus	10110101000100010111000010
Aedes epactius	10110101000100010111000010
Haemagogus equinus	10110101000110010101010000
Armigeres subalbatus	10110101000100010101000000
Culex pipiens	11110111000100011101001011
Tripteroides bambusa	11110111000100010101000010
Sabethes cyaneus	11110101001100010101010000
Anopheles albimanus	11011101100101110101110100

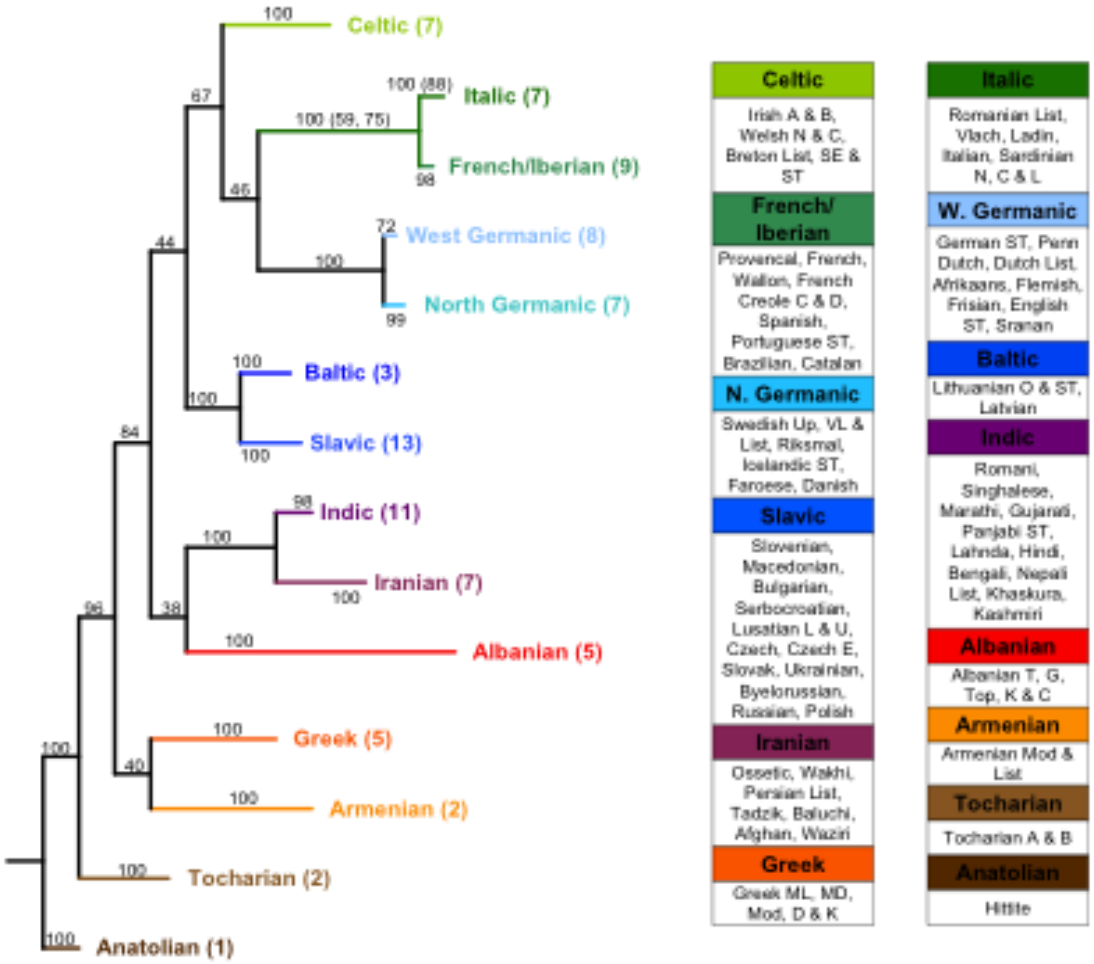
3.1. Phylogénie et linguistique

- Utilisation du nouvel algorithme dans le cadre de l'histoire évolutive des langue indo-européennes (IE).

- Les méthodes d'analyse phylogénétique sont de plus en plus utilisées pour inférer l'histoire des langues.

Willems, M., Lord, E., Laforest, L., Labelle, G., Lapointe, FJ., Di Sciullo, AM. et Makarenkov, V. (2016), Using hybridization networks to retrace the evolution of Indo-European languages, *BMC Evolutionary Biology*, 2016, 16:180

Histoire des langues IE



Arbre obtenu par Gray et Atkinson (2003).

Base de données

- Tirée de (Boc *et al* 2010).
- Basée sur la liste Swadesh (200 mots).
- 87 langues considérées.
- Regroupement des 200 mots en 1315 cognats.
- Cognat : groupe de mots apparentés ayant une racine commune.
- Pour chaque langue, on obtient une séquence binaire de taille 1315.

Distances entre langues

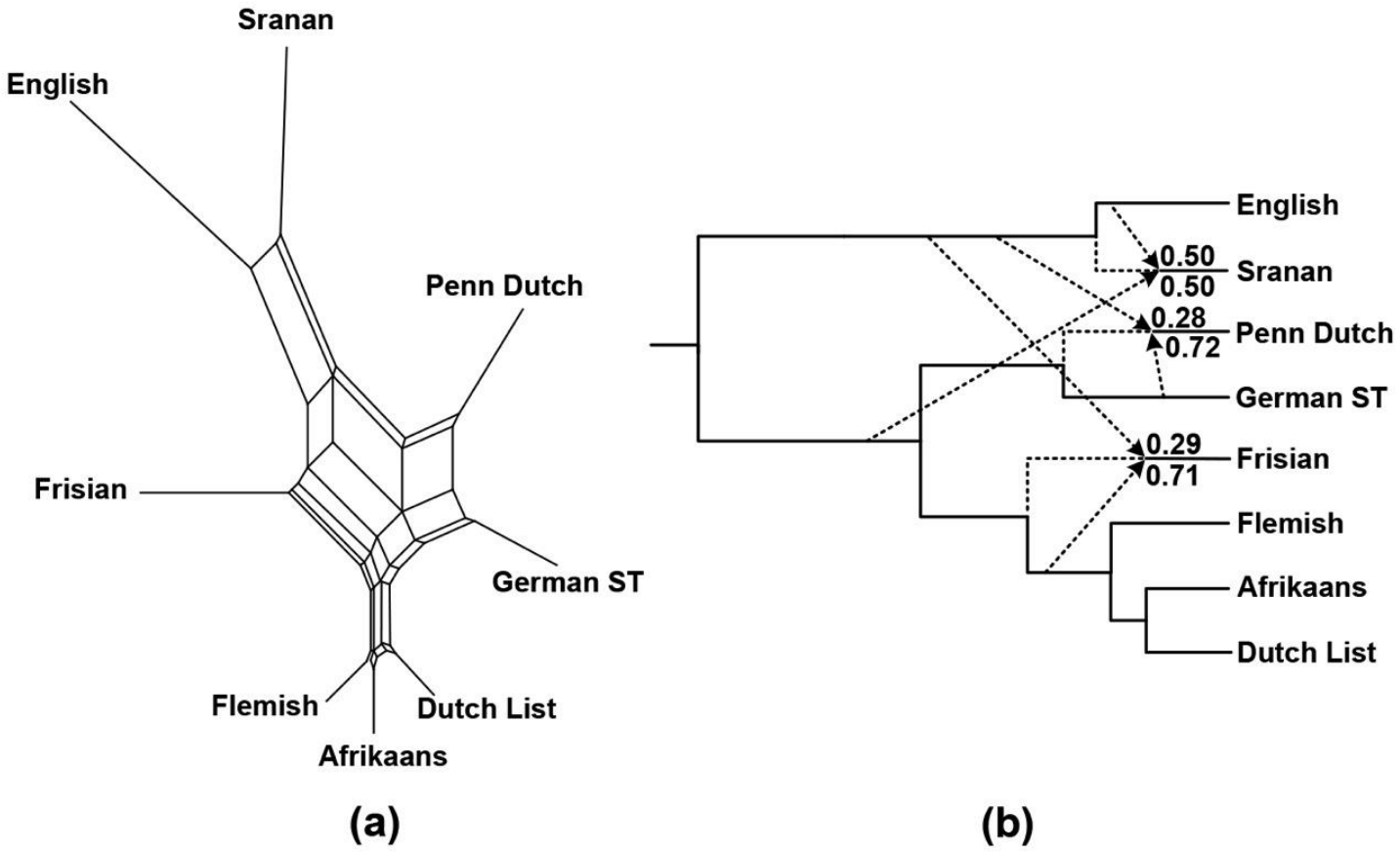
- Distance de Hamming : Nombre de cognats contenant seulement une des deux langues considérées.
- Distance de Levenshtein : distance d'édition.
- Distance de Levenshtein modifiée en tenant compte de la proximité de certaines lettres (Boc *et al* 2010).

Matrices de distances utilisées

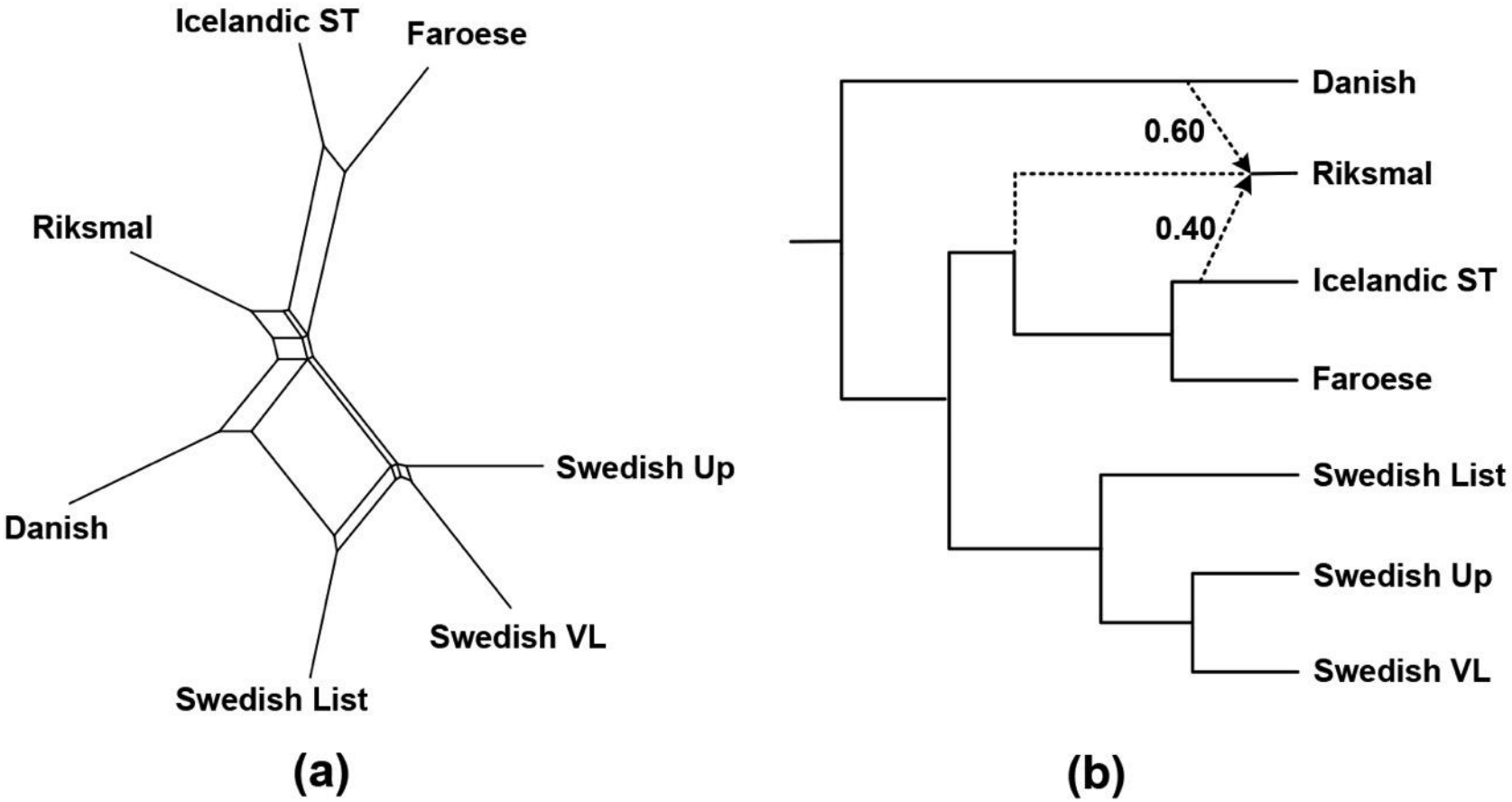
- Distance de Hamming pour le nouvel algorithme (réseau d'hybridation) : une seule matrice de distances.

- Distances de Levenshtein modifiées pour les split-graphes (arbres de mots) : 200 matrices de distances.

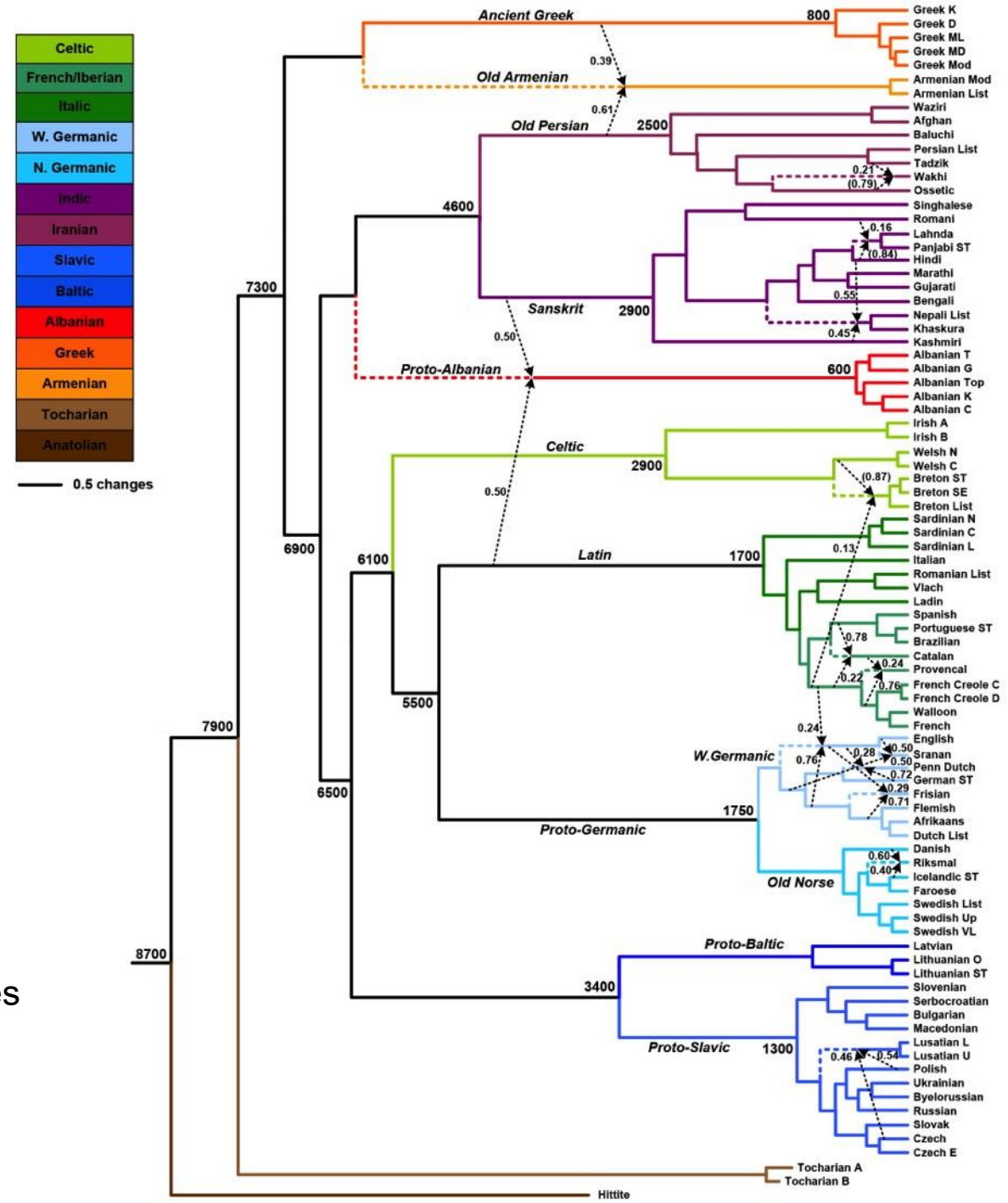
3.2. Réseaux biolinguistiques obtenus



Split-graphe (a) et réseau d'hybridation (b) pour huit langues ouest-germaniques



Split-graphe (a) et réseau d'hybridation (b) pour sept langues nord-germaniques



Réseau d'hybridation pour 87 langues IE

4.1. Description d'un nouvel algorithme basé sur les caractères

Maximum de vraisemblance :

- $\Pr(t, N_1, N_2)$: probabilité que le caractère N_1 (ADN, acide aminé ou binaire) évolue vers le caractère N_2 pendant le temps évolutionnaire t.
- Il existe plusieurs modèles d'évolution basés sur des processus markoviens.
- Vraisemblance d'un arbre : $L(T) = \prod_{l=1}^{L} L_l(T)$ où $L_l(T)$ est la somme des probabilités de tous les scénarios d'évolution possibles à la position l.
- Objectif : Trouver l'arbre ayant la vraisemblance la plus grande possible : NP-difficile.
- Il existe plusieurs heuristiques.

F81 model (Felsenstein, 1981)

- Le plus utilisé pour les données binaires.
- π_0 et π_1 : proportion de 0 (respectivement, 1) dans les données d'entrée.
- $\beta = \frac{1}{1-\pi_0^2-\pi_1^2}$
- $\Pr(t,i,j) = e^{-\beta t} + \pi_j\left(1 - e^{-\beta t}\right)$ si $i = j$
- $\Pr(t,i,j) = \pi_j\left(1 - e^{-\beta t}\right)$ si $i \neq j$

Vecteurs de probabilité

Pour chaque espèce i, on considère sa séquence binaire comme un vecteur de probabilité de dimension L (probabilité d'avoir 1 à la position l) :

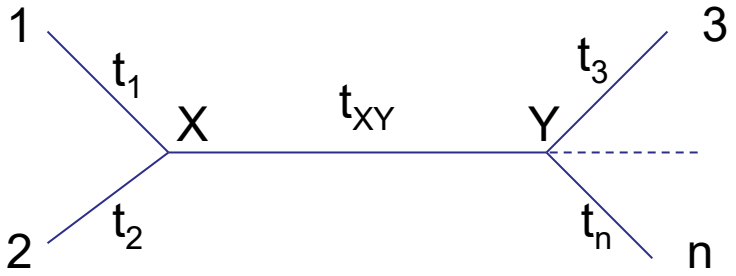
- $P(i,l) = 0$ si le $l$ème caractère de la séquence i est égal à 0.
- $P(i,l) = 1$ si le $l$ème caractère de la séquence i est égal à 1.

Vraisemblance d'un arbre de NJ

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

où :

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

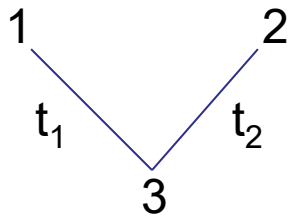


Vraisemblance d'un hybride

$$L^{H}{}_{3,1,2} = \prod_{l=1}^{L} \left(\sum_{i \in \{1,2\}} \sum_{(\varepsilon_i, \varepsilon_3) \in \{0;1\}^2} P_i P_3 \Pr(t_i, \varepsilon_i, \varepsilon_3) \right),$$

où :

- $P_i = (1 - \varepsilon_i)(1 - \mathrm{P}(i, l)) + \varepsilon_i \mathrm{P}(i, l)$
- $P_3 = (1 - \varepsilon_3)(1 - \mathrm{P}(3, l)) + \varepsilon_i \mathrm{P}(3, l)$



L'espèce 3 est l'hybride des espèces 1 et 2 dans cette configuration

Algorithme

Entrée : n séquences binaires de taille L.

1. $n_A = n$
2. Tant que ($n_A > 3$)

- On détermine les meilleurs voisins (i_T, j_T) selon NJ.
- On calcule la vraisemblance $L^T{}_{i_T, j_T}$ de l'arbre de NJ dans lequel i_T et j_T sont voisins.
- On détermine l'hybride le plus vraisemblable (h, i, j).
  - Si $L^H{}_{h,i,j} < L^T{}_{i_T, j_T}$) : on considère h comme l'hybride de i et j, et on le retire des données.
  - Sinon, on considère i_T et j_T comme voisins, et on les remplace par leur ancêtre commun direct déterminé par NJ.

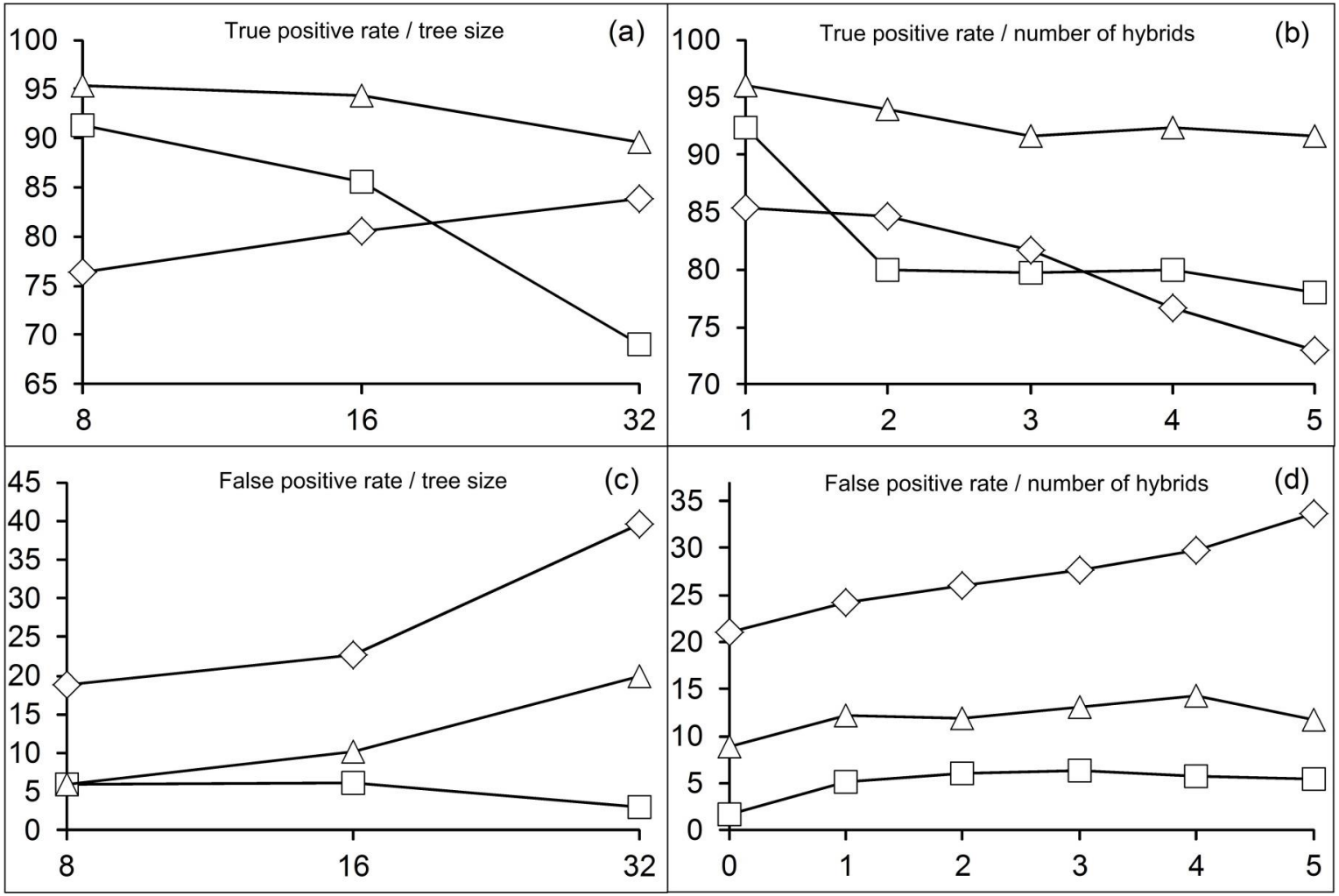
3. On joint les trois dernières espèces.

Sortie : Un réseau d'hybridation **explicite** dont les nœuds terminaux sont en correspondance avec les espèces initiales.

Remarques

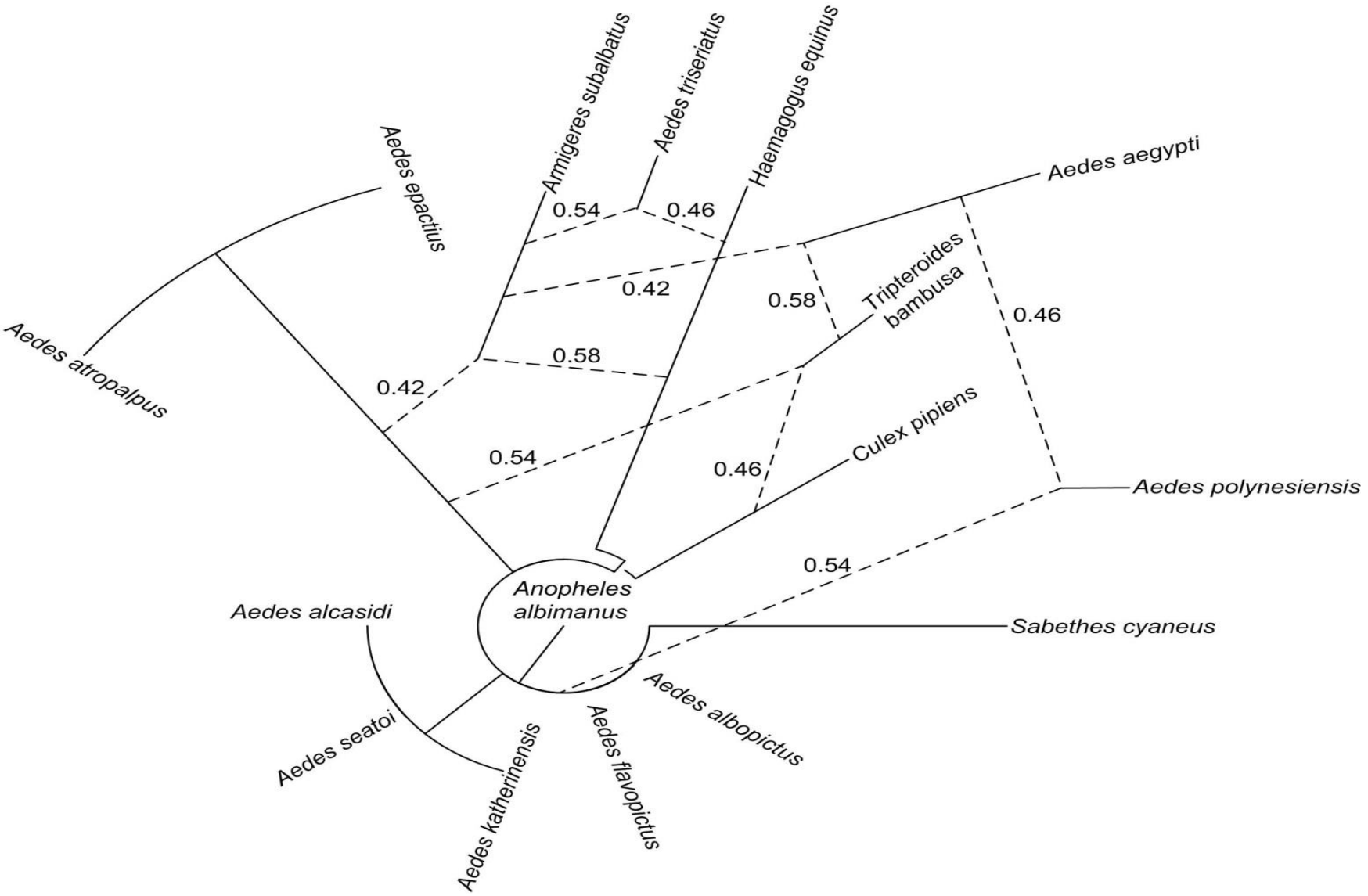
- Quand on remplace i et j par leur ancêtre commun direct X, la $l$ème composante du vecteur de probabilité de X est calculée en divisant $L(X, l, 1)$ (la vraisemblance d'avoir 1 à la position l) par $L(X, l, 1) + L(X, l, 0)$.
- Degré d'hybridation : pour chaque position, on calcule la probabilité que le caractère hybride provienne de chacun de ses deux parents.
- Les longueurs de branches sont optimisées avec la méthode de Newton-Raphson.

4.2. Résultats de différentes simulations



Résultats obtenus lors de simulations avec 0 à 5 hybrides pour des arbres avec 8, 16 et 32 feuilles, avec notre méthode de distances (◊), avec notre méthode baseé sur les caractères (□) et avec cette dernière modifiée à l'aide d'un critère d'information bayésien (BIC) (Δ)

Réseau d'hybridation pour 16 espèces de moustiques



Remarques sur l'exemple des moustiques

- On a trouvé 4 hybrides avec notre première méthode et 5 hybrides avec notre méthode basée sur les caractères.

- Huson et Klöpper (2007) ont identifié 4 hybrides à l'aide d'un réseau de recombinaisons.

- Dans tous les cas, les hybrides identifiés sont très similaires (on observe essentiellement des permutations entre les hybrides et leurs parents).

5. Perspectives

- Méthode basée sur les caractères :
  - Application en biolinguistique.
  - Adaptation pour des données moléculaires.
  - Comparaison des arbres obtenus avec d'autres méthodes de maximum de vraisemblance.
  - Reconstruction des séquences ancestrales.
  - Optimisation du critère statistique pour déterminer s'il y a un hybride à chaque pas de l'algorithme.
- Ajout du nombre d'hybrides souhaité comme paramètre d'entrée.
- Tests de nos algorithmes sur des données biologiques.

Références

- **Adachi, J. et Hasegawa, M.** (1992). *Computer Science Monographs, No. 27. MOLPHY : Programs for Molecular Phylogenetics, I. – PROTML : Maximum Likelihood Inference of Protein Phylogeny*. Tokyo, Japon : Institute of Statistical Mathematics.
- **Albrecht, B., Scornavacca, C., Cenci, A. et Huson, D. H.** (2012). Fast computation of minimum hybridization networks. *Bioinformatics*, **28(2)**, 191–197.
- **Boc, A., Di Sciullo, A. M. et Makarenkov, V.** (2010). Classification of the Indo-European languages using a phylogenetic network approach. In H. Locarek-Junge et C. Weihs (dir.), *Classification as a Tool for Research* 647–655. Berlin Heidelberg, Allemagne : Springer.
- **Bandelt, H.-J. et Dress, A. W. M.** (1992). Split decomposition : A new and useful approach to phylogenetic analysis of distance data. *Molecular Phylogenetics and Evolution*, **1(3)**, 242–252.
- **Bryant, D., Filimon, F. et Gray, R.** (2005). Untangling our past : Languages, trees, splits and networks. In R. Mace, S. Holden, et S. Shennan (dir.), *The Evolution of Cultural Diversity : A Phylogenetic Approach*. Walnut Creek, CA, États-Unis : Left Coast Press.
- **Bryant, D. et Moulton, V.** (2004). Neighbor-net : an agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution*, **21(2)**, 255–265.
- **Chen, Z.-Z. et Wang, L.** (2012). Algorithms for Reticulate Networks of Multiple Phylogenetic Trees. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **9(2)**, 372–384.
- **Felsenstein, J.** (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution,* **17(6)**, 368–376.
- **Gray, R. D. et Atkinson, Q. D.** (2003). Language-tree Divergence Times Support the Anatolian Theory of Indo-European Origin. *Nature*, **426(696)**, 435–439.
- **Jukes, T. H. et Cantor, C. R.** (1969). *Evolution of Protein Molecules*. New York, NY, États-Unis : Academy Press.
- **Huson, D. H. et Bryant, D.** (2006). Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, **23(2)**, 254–267.
- **Huson, D. H. et Klöpper, T. H.** (2007). Beyond galled trees – decomposition and computation of galled networks. In T. P. Speed et H. Huang (dir.), *Research in Computational Molecular Biology* 211–225. Berlin Heidelberg, Allemagne : Springer.
- **Huson, D. H. et Scornavacca, C.** (2012). Dendroscope 3 : An Interactive Tool for Rooted Phylogenetic Trees and Networks. *Systematic Biology*, **61(6)**, 1061–1067.
- **Kumar, A., Black, W. C. et 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)**, 367–373.
- **Saitou, N. et Nei, M.** (1987). The neighbor-joining method : a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, **4(4)**, 406–425.
- **Schwarz, G. et Gideon, E.** (1978). Estimating the dimension of a model. *Annals of statistics* **6(2)**, 461–464.
- **Willems, M., Tahiri, N. et Makarenkov, V.** (2014). A new efficient algorithm for inferring explicit hybridization networks following the Neighbor-Joining principle. *Journal of Bioinformatics and Computational Biology*, **12(5)**, 1450024.
- **Willems, M., Lord, E., Laforest, L., Labelle, G., Lapointe, F.-J., Di Sciullo, A.-M. et Makarenkov, V.** (2016). Using hybridization networks to retrace the evolution of Indo-European languages, *BMC Evolutionary Biology*, **16(1)**, 180.