Inférence de réseaux phylogénétiques pour la détection d'hybridations et d'introgressions

Charles-Elie Rabier

Vincent Berry, Jean-Christophe Glaszmann
Fabio Pardi et Céline Scornavacca

Genome Harvest / KIM Data & Life Sciences
ISEM, Institut des Sciences de l'Evolution de Montpellier
IMAG, Institut Montpelliérain Alexander Grothendieck
LIRMM, Laboratoire d'informatique, de Robotique et de Microélectronique
UMR AGAP, Amélioration Génétique et adaptation des plantes, CIRAD

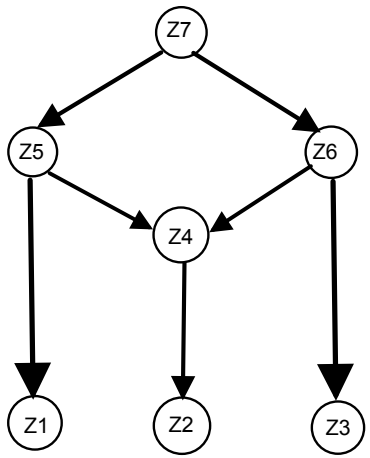
Plan

Plan

Réseaux phylogénétiques

Les réseaux phylogénétiques sont des DAG qui vont nous permettre de détecter des :

- hybridations (e.g. plantes)
- introgressions (e.g. plantes et animaux)
- transferts horizontaux (e.g. bactéries)



Quelques points importants :

- Longueur d'une arête = temps d'évolution
- Dépendance entre noeuds
- Les noeuds de réticulations ont 2 parents et représentent les évènements de réticulation
- On cherche à avoir une distribution de réseaux (incertitude sur des clades)
- Plus on collecte de données, plus on est en mesure d'inférer précisément le réseau

La domestication du riz, un sujet de grand intérêt

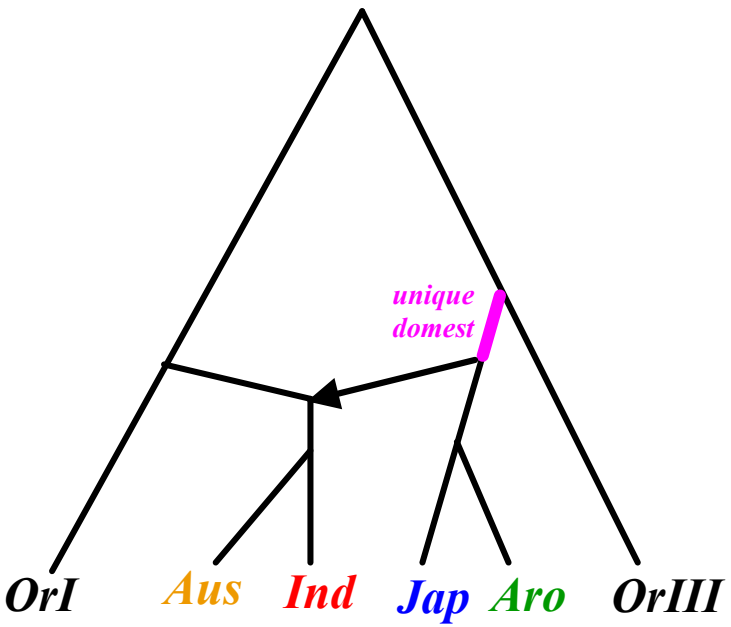
“ La domestication est un processus transformant une espèce sauvage en une espèce dépendant de l’homme ... Elle subira une évolution adaptative afin de satisfaire les besoins de l’homme.”

Choi et al. (MBE, 2017)

Un processus de domestication controversé et débattu

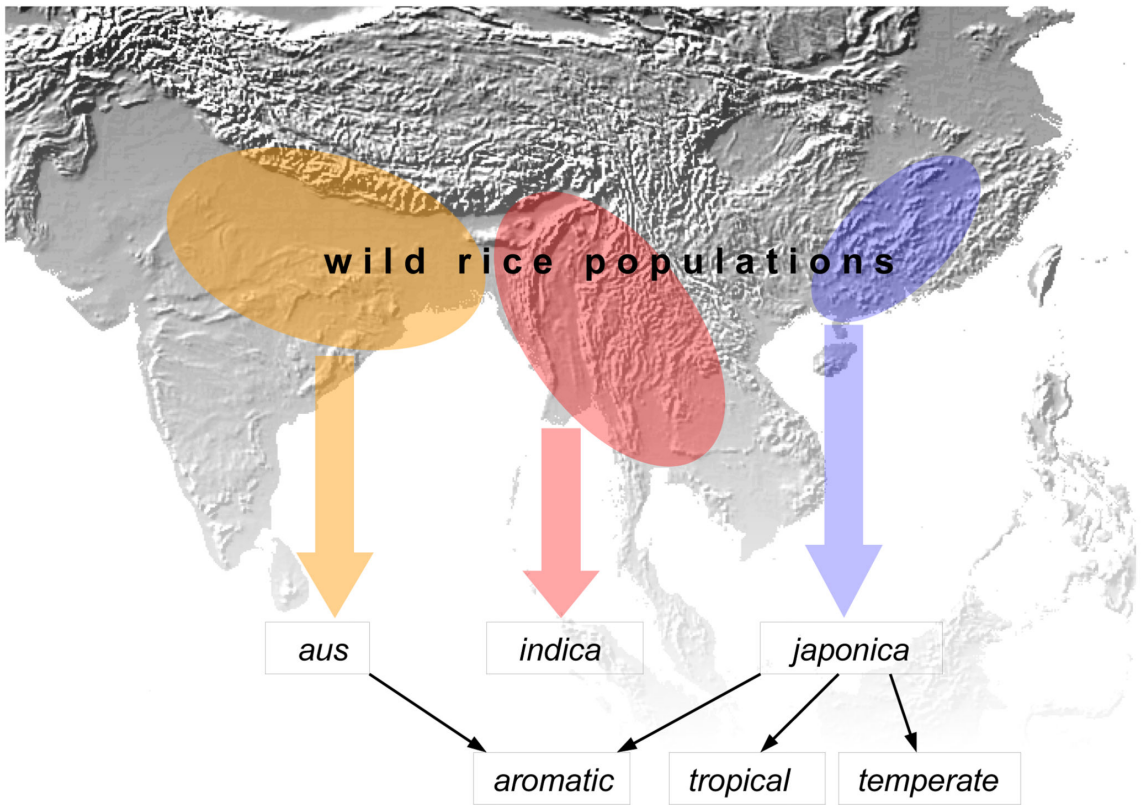
Quelques thèses sur la domestication du riz

- Huang et al. (Nature, 2012) : japonica domestiqué à partir d'un riz sauvage dans le sud de la Chine, puis croisé à un sauvage dans le sud est de l'Asie, générant indica



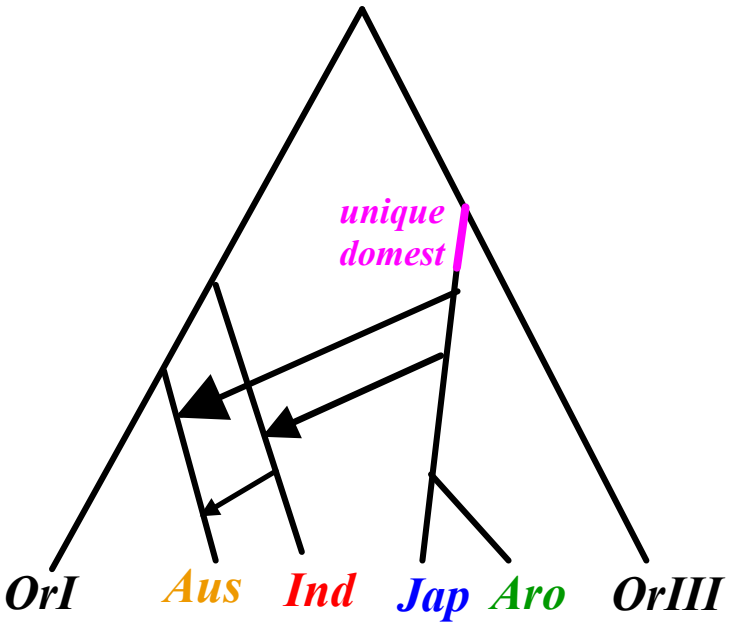
Quelques thèses sur la domestication du riz

- Civan et al. (Nature Plants, 2015) : indica, japonica et aus domestiqués séparément dans différentes parties d'Asie



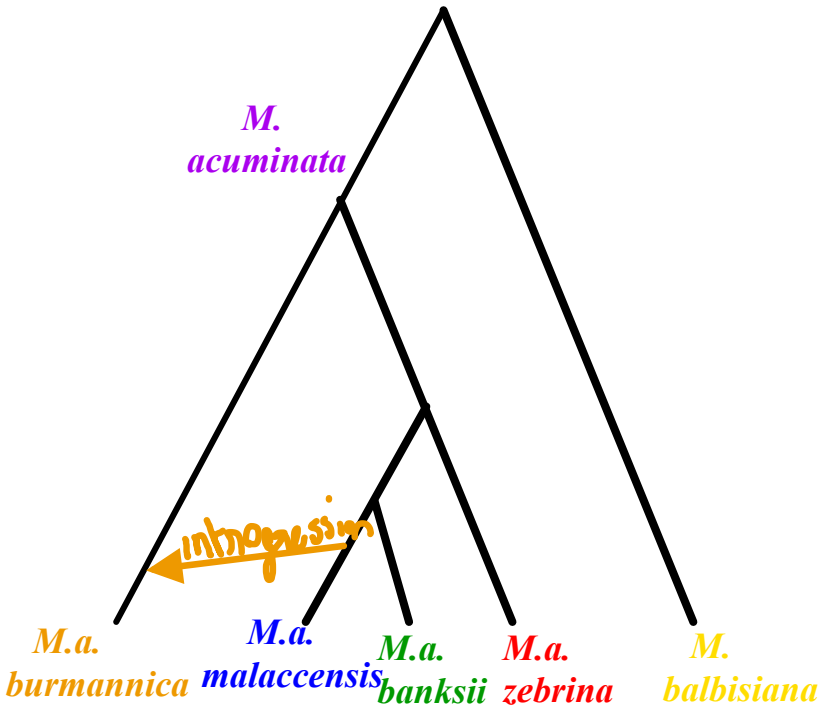
Quelques thèses sur la domestication du riz

- Choi et al. (MBE, 2017) soutiennent aussi un seul évènement de domestication (japonica). Introgression par hybridation de japonica et proto-indica et proto-aus, générant indica et aus



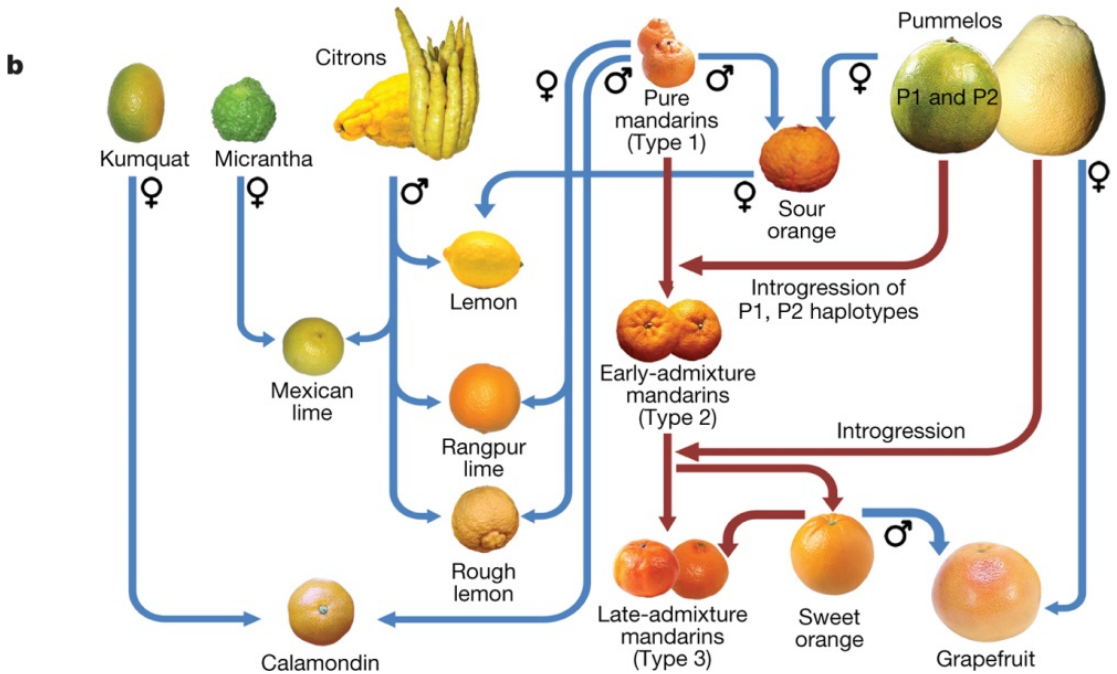
Introgressions chez les bananes cultivées

Bananes cultivées issues de Musa balbisiana et Musa acuminata.
Evènement d'introgression entre M.a.malaccensis et M.a.burmannica
(Rouard et al, GBE 2018)



Histoire évolutive chez citrus (Wu et al, Nature 2018)

Agrumes cultivés issus d'une évolution d'hybridation interspécifique impliquant cinq taxons ancestraux (C. fortunella, C. micrantha, C. medica, C. reticulata, C. maxima).



Méthodes utilisées chez le riz = méthodes assez génériques

- Huang et al. (Nature, 2012)
  - Données : 446 O.rufipogon (sauvages) + 1083 O.sativa (cultivés) + 8 Millions de SNPs
  - Reconstitution d'arbres phylogénétiques par Neighbour-Joining + Analyse en Composantes Principales
- Civan et al. (Nature Plants, 2015)
  - Même jeux de données que Huang + sélection de 31 régions de faible diversité (loci de domestication)
  - Reconstitution d'arbres phylogénétiques par Neighbour-Joining + Analyse en Composantes Principales
- Wang et al. (Genome Research, 2017)
  - Reconstitution d'arbres par TreeMix :maximum de vraisemblance basé sur un modèle Gaussien de changement des fréquences d'allèles au cours du temps. Modélisation des fréquences alléliques selon un graphe. Approximations...

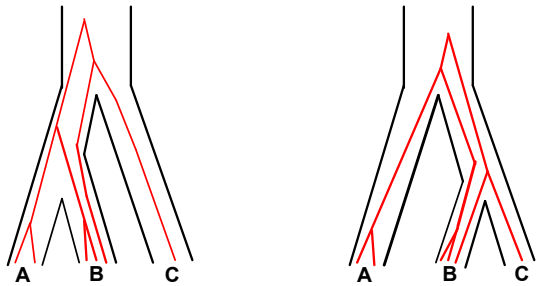
Notre approche méthodologique

On s'intéresse à un modèle qui, outre le tri de lignées, considère explicitement les mutations et hybridation. Modélisation Bayésienne plus fine.

Nos pistes :

1. Inférence d'arbres d'espèces + arbres résumés en réseaux phylogénétiques

SNAPP (Bryant et al. 2012, MBE) + SplitsTree



2. Inférence directe de réseaux

SNAPPNET = Extension de SNAPP aux réseaux

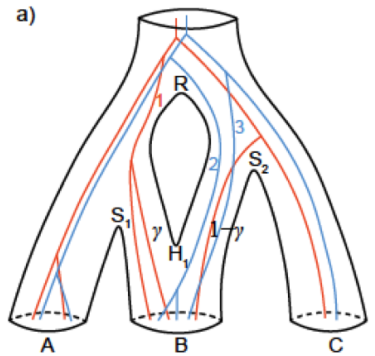


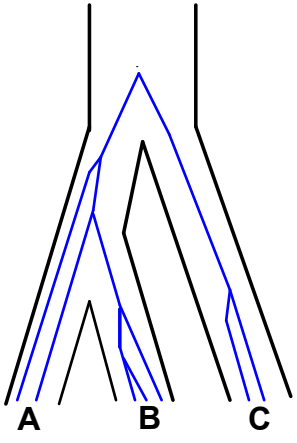
schéma tiré de Zhang et al. (2018)

Plan

Logiciel SNAPP pour l'inférence Bayésienne d'arbres (Bryant et al. 2012, MBE)

- Marqueurs biallélliques (SNPs) indépendants sachant l'arbre d'espèces

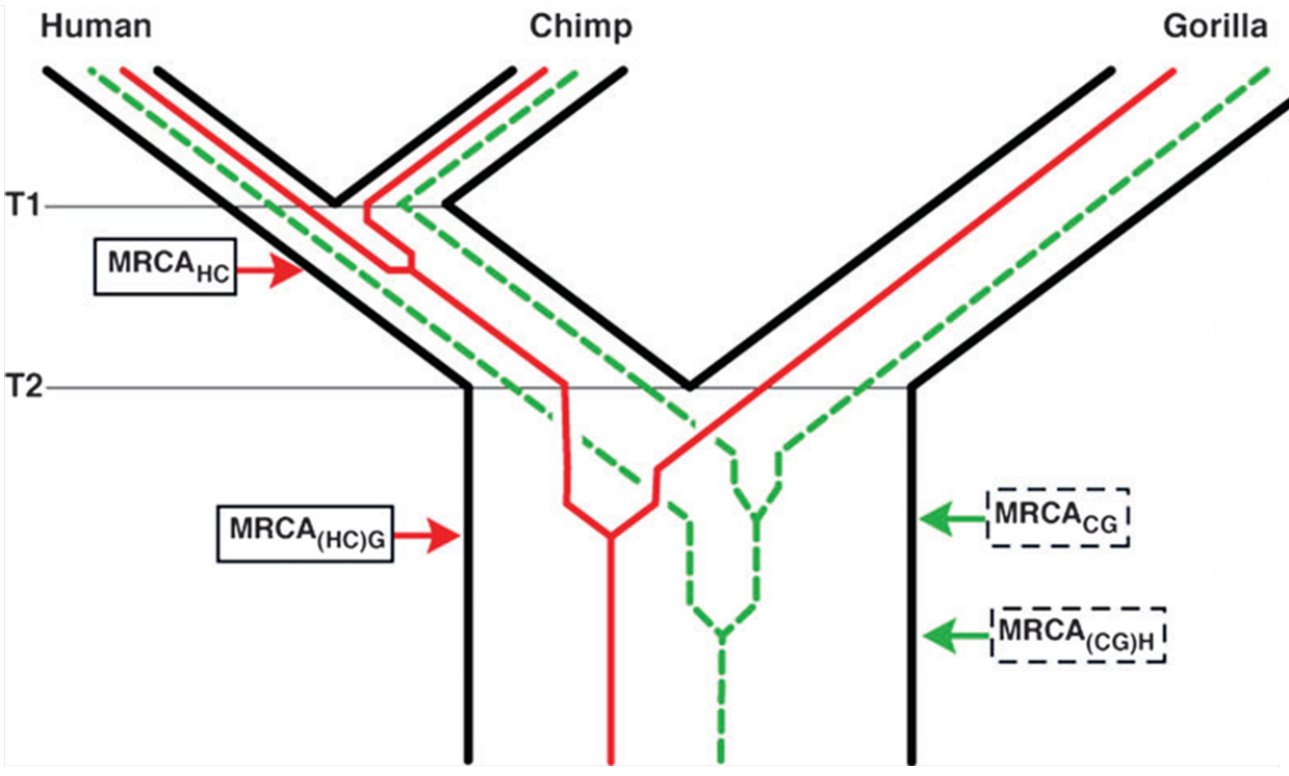
- Modélisation de l'arbre de locus (backward)
  - Processus de coalescence évoluant à l'intérieur d'un arbre d'espèces (MultiSpecies Coalescent)
  - Processus autorisant la discordance entre arbres de locus et arbres d'espèces (tri de lignées incomplet)



Histoires évolutives discordantes

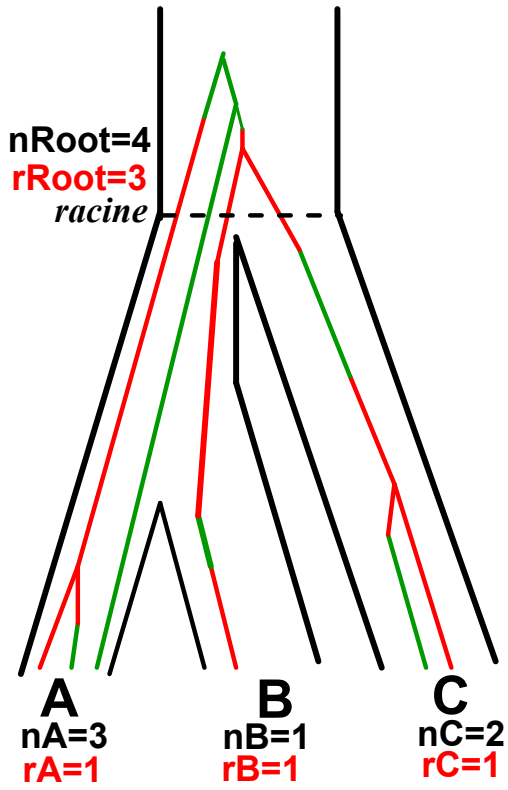
Ebersberger *et al.* (2007)

Tri de lignées incomplet (ILS) :
processus biologique causant une discordance avec l'arbre de locus



Les mutations interviennent au cours du temps

- Modélisation des données au SNP (forward)
  - mutation (rouge $\leftrightarrow$ vert) : modèle markovien évoluant le long des branches de l'arbre de locus
  - u : taux de mutation rouge $\rightarrow$ vert
  - v : taux de mutation vert $\rightarrow$ rouge



- V.a. : rRoot, nRoot, rIntNode, nIntNode, rA, rB, rC
- pas d'aléa dans nA, nB, nC
- Data=(rA, rB, rC)
- Vraisemblance : $\mathbb{P}(Data \mid S)$ avec S arbre d'espèce

Calcul de vraisemblance dans un arbre

- $\mathbb{P}(n_{root} = i \mid Count)$ calculé récursivement en remontant dans le temps (postorder)

  Tavaré (Theor Pop Biol, 1984), Watterson (Theor Pop Biol, 1984), Takahata and Nei (Genetics, 1985) ...

- $\mathbb{P}(Data \mid Count, n_{root} = i, r_{root} = j)$ calculé récursivement en remontant dans le temps (postorder)

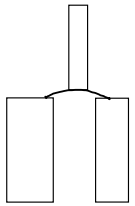
  Slatkin (Genetics, 1996) vs. Griffiths and Tavaré (Springer, 1997)

- $\mathbb{P}(r_{root} = j \mid n_{root} = i)$ calculé par
  - la loi Binomiale : $\mathbb{P}(r_{root} = j \mid n_{root} = i) = C_i^j p^j (1-p)^{i-j}$
  - la loi $\beta(\theta, \theta)$ sur le paramètre p de la Binomiale : $\mathbb{P}(r_{root} = j \mid n_{root} = i) = C_i^j B(j+\theta, i-j+\theta)/B(\theta, \theta)$

- Astuces afin de raccourcir les calculs : Vraisemblances partielles...

La statistique Bayésienne dans SNAPP

- S : arbre d'espèces (topologie, longueurs de branches, tailles de populations)
- X_i : données pour le locus i
- G_i : arbre de locus pour le locus i
- m loci



Par la théorème de Bayes

$$\mathbb{P}(S|X_1,\ldots,X_m) \propto \left(\prod_{i=1}^{m}\int_{\psi}\mathbb{P}(X_i|G_i)\mathbb{P}(G_i|S)dG_i\right)P(S)$$

$$\propto \mathbb{P}\left(Data \mid S\right)P(S)$$

SNAPP intègre sur tous les arbres de locus

Calcul de la *prior* P(S) par le processus de naissances

$\Rightarrow$ Markov Chain Monte Carlo (MCMC) afin d'estimer la distribution à posteriori de $\mathbb{P}(S|X_1,\ldots,X_m)$

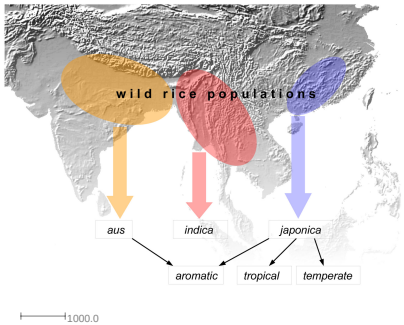
Implémenté dans BEAST

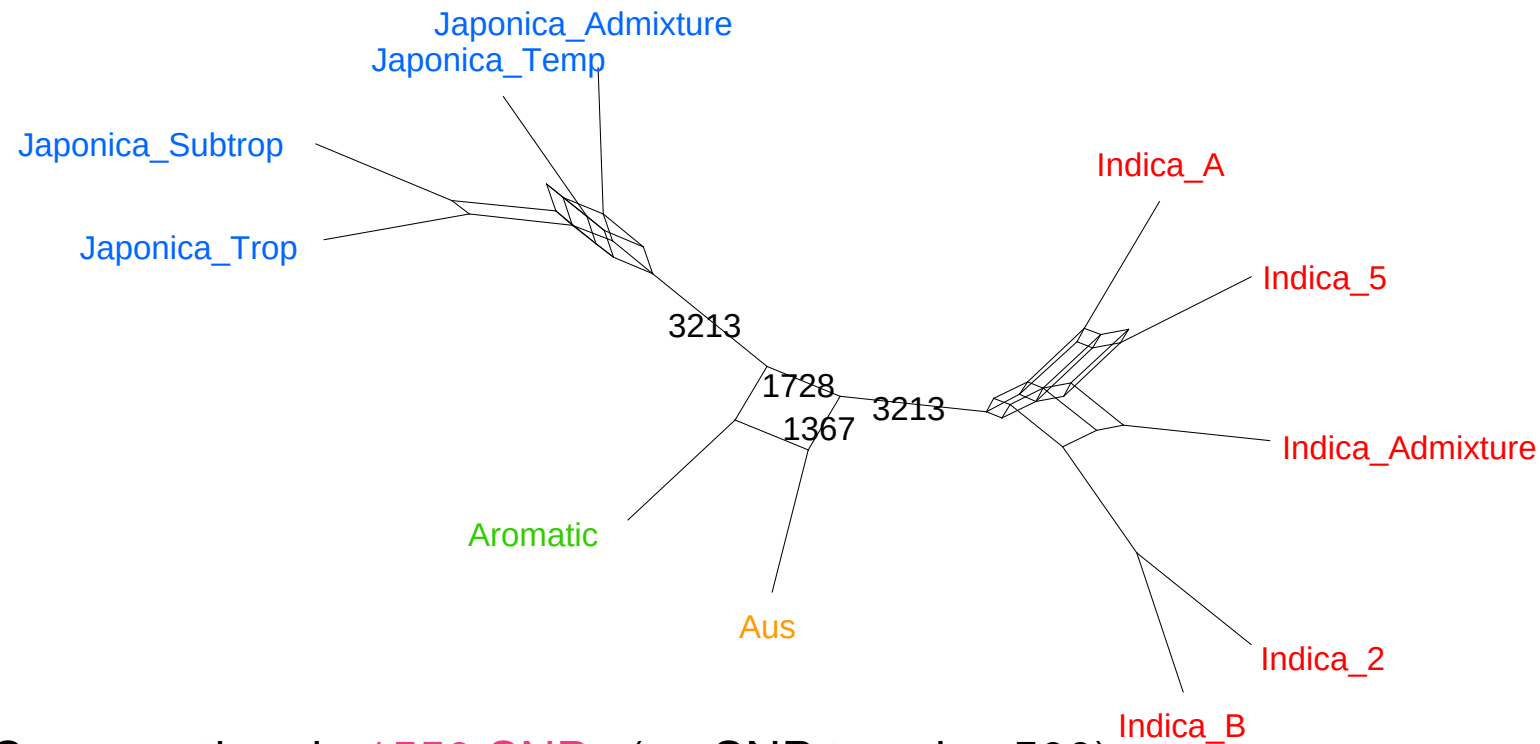
Plan

Les données de riz

1. Données tirées de Wang et al. (Nature 2018), 3000 variétés de riz cultivés
2. Prétraitement par Joao Santos (données manquantes ...)
3. 895 977 marqueurs disponibles sur le chromosome 6
4. 1 jeu de données proposés par JC Glaszmann (core collections)
   - 44 variétés (7 Aromatic, 7 Aus, 13 Indica, 17 Japonica)

Chromosome 6 (données J. Santos, J-C. Glaszmann)





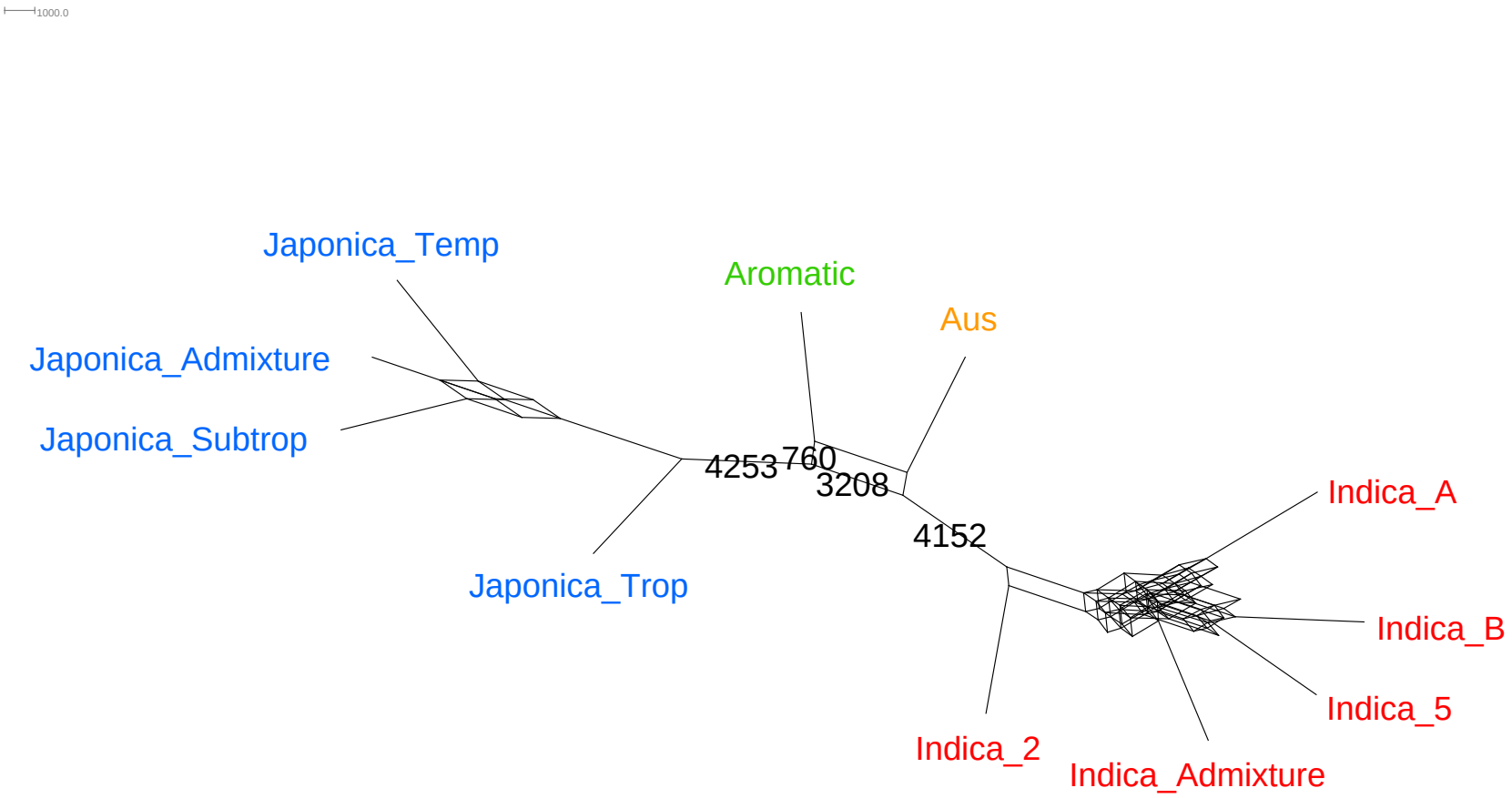
Conservation de 1550 SNPs (un SNP tous les 500)

- JDD1 (1er SNP= 1er SNP du chromosome 6)

Chromosome 10 (données J. Santos, J-C. Glaszmann)

Conservation de 1089 SNPs (un SNP tous les 500)

- JDD2 (1er SNP= 50ème SNP du chromosome 10)



Plan

Simulateur basé sur un réseau

SNAPPSimNET construit sur la base du simulateur SNAPPSim de Bryant et al. (2012)

- Génération d'arbres de locus évoluant à l'intérieur d'un réseau selon un processus de coalescence (Multispecies Network Coalescent)

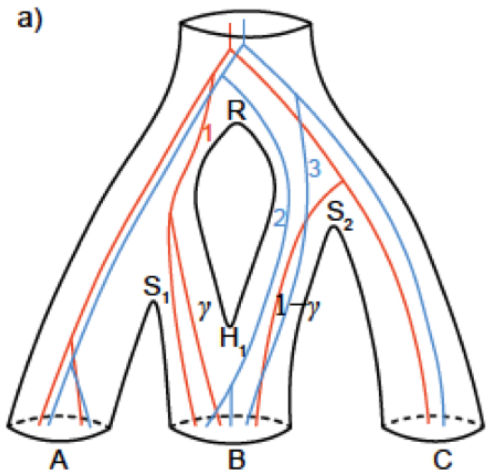


schéma tiré de Zhang et al. (2018)

- Modèle markovien évoluant le long des branches de l'arbre de locus
- Chaque arbre de locus locus donne un SNP
- SNAPP est fortement attiré par un scénario sous-jacent au réseau

cf. Solis-Lemus et al., Syst Biol 2016

Plan

A propos de l'inférence directe de réseaux

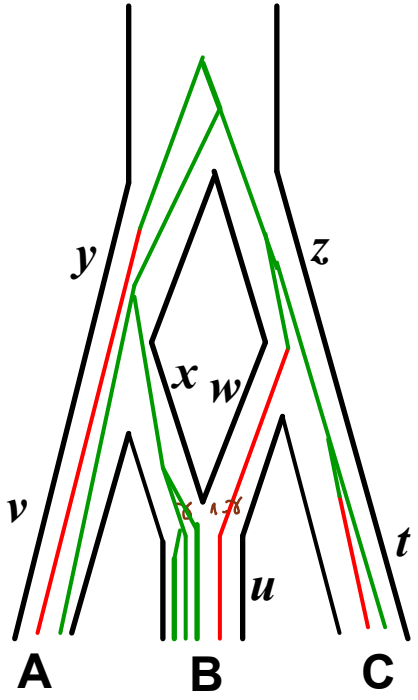
Méthodes explicites pour la reconstruction de réseaux

- Approches combinatoires qui tiennent compte de l'hybridation mais ne gèrent pas l'ILS (e.g. Gambette et al., J. Bioinform Comput Biol 2012)
- Approches basées sur un modèle stochastique avec de nombreux paramètres
  - Méthodes de pseudo-vraisemblance : la vraisemblance du réseau est calculée pour des sous parties de sa topologie. Ces valeurs sont combinées par la suite (e.g.Solis-Lemus et Ané, Plos Genetics 2016, Zhu et Nakhkeh, Bioinformatics 2018, Zhu et al, Bioinformatics 2019 ...)
  - Méthodes maximum de vraisemblance (e.g. Yu et al. Plos Genetics 2012, Yu et al. PNAS 2014 ...)
  - Méthodes Bayésiennes (e.g. Wen et al., Plos Genetics 2016, Zhang et al, MBE 2018, Zhu et al, Plos Comp Biol 2018 ...)

Plan

Cadre d'un réseau phylogénétique

- Modélisation de l'arbre de locus (backward) :
  - processus de coalescence
  - modèle de Nakhleh au niveau du noeud de réticulation
    $\Rightarrow$ Multispecies Network Coalescent
- Modélisation des données au SNP (forward)



- V.a. : rRoot, nRoot, rIntNode, nIntNode, rA, rB, rC
- pas d'aléa dans nA, nB, nC
- Data=(rA, rB, rC)
- Vraisemblance : $\mathbb{P}(Data \mid N)$ avec N réseau

Une méthode Bayésienne d'inférence de réseaux

- N : réseau phylogénétique (topologie, longueurs de branches, tailles de populations, proba d'hybridation)
- X_i : données pour le locus i
- G_i : arbre de locus pour le locus i
- m loci

$$\mathbb{P}(N|X_1,\ldots,X_m) \propto \left(\prod_{i=1}^{m}\int_{\psi}\mathbb{P}(X_i|G_i)\mathbb{P}(G_i|S)dG_i\right)P(N)$$
$$\propto \mathbb{P}\left(Data \mid N\right)P(N)$$

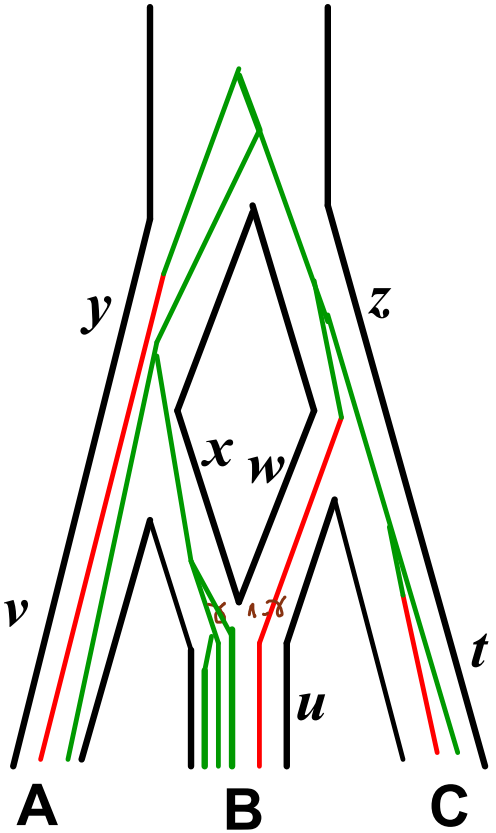
SNAPPNET intègre sur tous les arbres de locus (extension de SNAPP, Bryant et al. MBE 2012), à l'aide d'un nouvel algorithme de parcours du réseau

Calcul de la *prior* P(N) par le processus de naissances hybridation de Zhang et al. (MBE 2018)

$\Rightarrow$ Markov Chain Monte Carlo (MCMC) afin d'estimer la distribution à posteriori de $\mathbb{P}(N|X_1,\ldots,X_m)$

Implémenté dans BEAST

Problème sous-jacent aux réseaux phylogénétiques



$Data_z$: proportion de rouge/vert dans les espèces sous la branche z

$Data_y$: proportion de rouge/vert dans les espèces sous la branche y

$Data_{zT}$ et $Data_{yT}$ ne sont pas indépendantes ...

$Data_{zT}$ et $Data_{yT}$ comprennent les allèles rouges et verts de l'espèce hybride

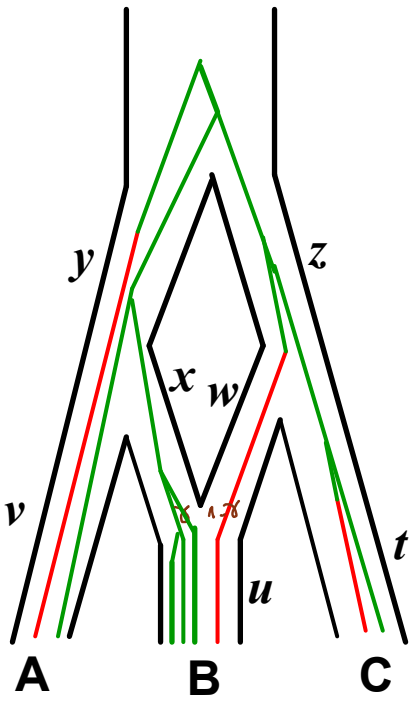
Plan

Calcul de la vraisemblance dans un réseau

$$
\begin{aligned}
&\mathbb{P}(Data) \\
=&\sum_i \sum_j \mathbb{P}(Data \mid n_{root} = i, r_{root} = j)\, \mathbb{P}(r_{root} = j \mid n_{root} = i) \\
&\mathbb{P}(n_{root} = i) \\
=&\sum_i \sum_j \sum_{i'} \sum_{j'} \mathbb{P}(Data_{zT}\, Data_{yT} \mid n_{yT} = i', n_{zT} = i - i', r_{yT} = j', \\
&\quad r_{zT} = j - j')\, \mathbb{P}(r_{yT} = j',\ r_{zT} = j - j' \mid n_{yT} = i', n_{zT} = i - i', r_{root} = j) \\
&\quad \mathbb{P}(n_{yT} = i', n_{zT} = i - i' \mid n_{root} = i)\, \mathbb{P}(r_{root} = j \mid n_{root} = i)\, \mathbb{P}(n_{root} = i)
\end{aligned}
$$

- $\mathbb{P}(r_{root} = j \mid n_{root} = i)$ calculé par
  - la loi Binomiale : $\mathbb{P}(r_{root} = j \mid n_{root} = i) = C_i^j p^j (1-p)^{i-j}$
  - la loi $\beta(\theta, \theta)$ sur le paramètre p de la Binomiale :
    $\mathbb{P}(r_{root} = j \mid n_{root} = i) = C_i^j B(j+\theta, i-j+\theta)/B(\theta, \theta)$
- $\mathbb{P}(Data \mid n_{root} = i, r_{root} = j)\, \mathbb{P}(n_{root} = i)$ calculé par un nouvel algorithme

Notre algorithme : calcul des lois jointes



Quantités calculées successivement

(1) $\mathbb{P}\left(Data_{uT} \mid n_{uT}, r_{uT}\right)$

(2) $\mathbb{P}\left(Data_{xB}\, Data_{wB} \mid n_{xB}, r_{xB}, n_{wB}, r_{wB}\right)$

(3) $\mathbb{P}\left(Data_{xT}\, Data_{wB} \mid n_{xT}, r_{xT}, n_{wB}, r_{wB}\right)$

(4) $\mathbb{P}\left(Data_{xT}\, Data_{wT} \mid n_{xT}, r_{xT}, n_{wT}, r_{wT}\right)$

(5) $\mathbb{P}\left(Data_{vT} \mid n_{vT}, r_{vT}\right)$

(6) $\mathbb{P}\left(Data_{tT} \mid n_{tT}, r_{tT}\right)$

(7) $\mathbb{P}\left(Data_{yB}\, Data_{wT} \mid n_{yB}, r_{yB}, n_{wT}, r_{wT}\right)$

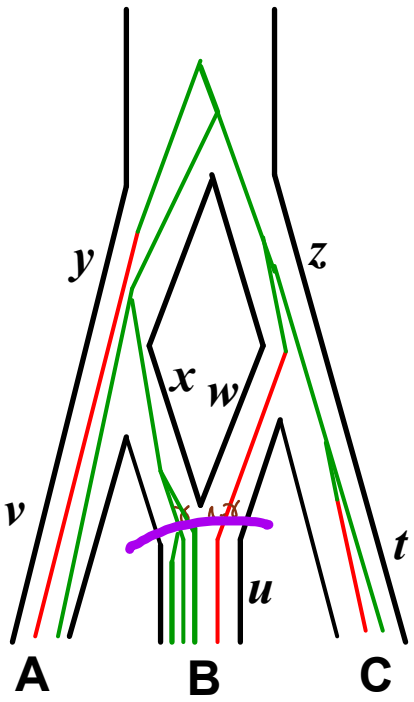
(8) $\mathbb{P}\left(Data_{yB}\, Data_{zB} \mid n_{yB}, r_{yB}, n_{zB}, r_{zB}\right)$

(9) $\mathbb{P}\left(Data_{yT}\, Data_{zB} \mid n_{yT}, r_{yT}, n_{zB}, r_{zB}\right)$

(10) $\mathbb{P}\left(Data_{yT}\, Data_{zT} \mid n_{yT}, r_{yT}, n_{zT}, r_{zT}\right)$

(11) $\mathbb{P}\left(Data \mid n_{root}, r_{root}\right)$

Notre algorithme : calcul des lois jointes



Quantités calculées successivement

(1) $\mathbb{P}\left(Data_{uT} \mid n_{uT}, r_{uT}\right)$

(2) $\mathbb{P}\left(Data_{xB}\, Data_{wB} \mid n_{xB}, r_{xB}, n_{wB}, r_{wB}\right)$

(3) $\mathbb{P}\left(Data_{xT}\, Data_{wB} \mid n_{xT}, r_{xT}, n_{wB}, r_{wB}\right)$

(4) $\mathbb{P}\left(Data_{xT}\, Data_{wT} \mid n_{xT}, r_{xT}, n_{wT}, r_{wT}\right)$

(5) $\mathbb{P}\left(Data_{vT} \mid n_{vT}, r_{vT}\right)$

(6) $\mathbb{P}\left(Data_{tT} \mid n_{tT}, r_{tT}\right)$

(7) $\mathbb{P}\left(Data_{yB}\, Data_{wT} \mid n_{yB}, r_{yB}, n_{wT}, r_{wT}\right)$

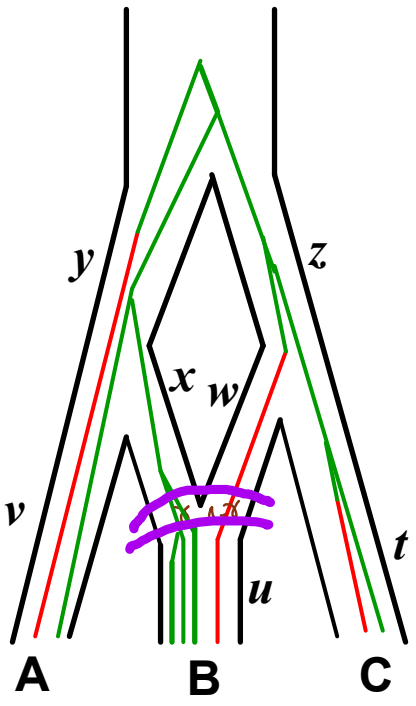
(8) $\mathbb{P}\left(Data_{yB}\, Data_{zB} \mid n_{yB}, r_{yB}, n_{zB}, r_{zB}\right)$

(9) $\mathbb{P}\left(Data_{yT}\, Data_{zB} \mid n_{yT}, r_{yT}, n_{zB}, r_{zB}\right)$

(10) $\mathbb{P}\left(Data_{yT}\, Data_{zT} \mid n_{yT}, r_{yT}, n_{zT}, r_{zT}\right)$

(11) $\mathbb{P}\left(Data \mid n_{root}, r_{root}\right)$

Notre algorithme : calcul des lois jointes



Quantités calculées successivement

(1) $\mathbb{P}\left(Data_{uT} \mid n_{uT}, r_{uT}\right)$

(2) $\mathbb{P}\left(Data_{xB}\, Data_{wB} \mid n_{xB}, r_{xB}, n_{wB}, r_{wB}\right)$

(3) $\mathbb{P}\left(Data_{xT}\, Data_{wB} \mid n_{xT}, r_{xT}, n_{wB}, r_{wB}\right)$

(4) $\mathbb{P}\left(Data_{xT}\, Data_{wT} \mid n_{xT}, r_{xT}, n_{wT}, r_{wT}\right)$

(5) $\mathbb{P}\left(Data_{vT} \mid n_{vT}, r_{vT}\right)$

(6) $\mathbb{P}\left(Data_{tT} \mid n_{tT}, r_{tT}\right)$

(7) $\mathbb{P}\left(Data_{yB}\, Data_{wT} \mid n_{yB}, r_{yB}, n_{wT}, r_{wT}\right)$

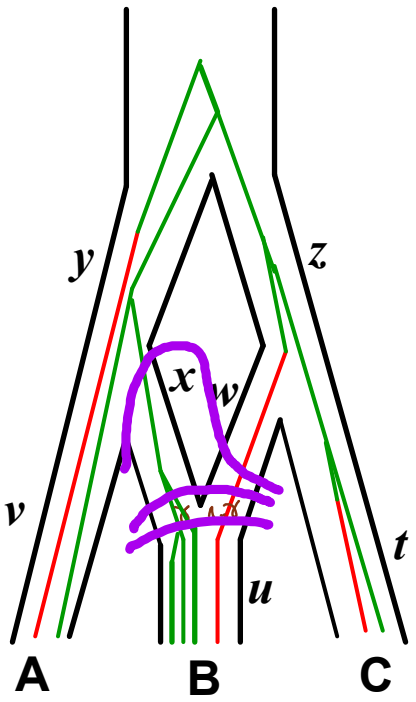
(8) $\mathbb{P}\left(Data_{yB}\, Data_{zB} \mid n_{yB}, r_{yB}, n_{zB}, r_{zB}\right)$

(9) $\mathbb{P}\left(Data_{yT}\, Data_{zB} \mid n_{yT}, r_{yT}, n_{zB}, r_{zB}\right)$

(10) $\mathbb{P}\left(Data_{yT}\, Data_{zT} \mid n_{yT}, r_{yT}, n_{zT}, r_{zT}\right)$

(11) $\mathbb{P}\left(Data \mid n_{root}, r_{root}\right)$

Notre algorithme : calcul des lois jointes



Quantités calculées successivement

(1) $\mathbb{P}\left(Data_{uT} \mid n_{uT}, r_{uT}\right)$

(2) $\mathbb{P}\left(Data_{xB}\, Data_{wB} \mid n_{xB}, r_{xB}, n_{wB}, r_{wB}\right)$

(3) $\mathbb{P}\left(Data_{xT}\, Data_{wB} \mid n_{xT}, r_{xT}, n_{wB}, r_{wB}\right)$

(4) $\mathbb{P}\left(Data_{xT}\, Data_{wT} \mid n_{xT}, r_{xT}, n_{wT}, r_{wT}\right)$

(5) $\mathbb{P}\left(Data_{vT} \mid n_{vT}, r_{vT}\right)$

(6) $\mathbb{P}\left(Data_{tT} \mid n_{tT}, r_{tT}\right)$

(7) $\mathbb{P}\left(Data_{yB}\, Data_{wT} \mid n_{yB}, r_{yB}, n_{wT}, r_{wT}\right)$

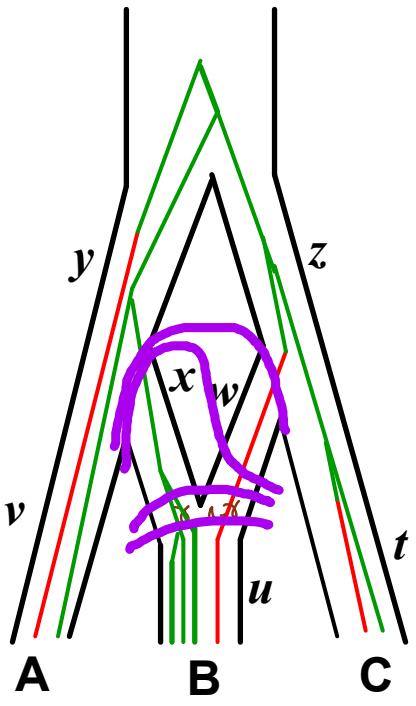
(8) $\mathbb{P}\left(Data_{yB}\, Data_{zB} \mid n_{yB}, r_{yB}, n_{zB}, r_{zB}\right)$

(9) $\mathbb{P}\left(Data_{yT}\, Data_{zB} \mid n_{yT}, r_{yT}, n_{zB}, r_{zB}\right)$

(10) $\mathbb{P}\left(Data_{yT}\, Data_{zT} \mid n_{yT}, r_{yT}, n_{zT}, r_{zT}\right)$

(11) $\mathbb{P}\left(Data \mid n_{root}, r_{root}\right)$

Notre algorithme : calcul des lois jointes



Quantités calculées successivement

(1) $\mathbb{P}\left(Data_{uT} \mid n_{uT}, r_{uT}\right)$

(2) $\mathbb{P}\left(Data_{xB}\, Data_{wB} \mid n_{xB}, r_{xB}, n_{wB}, r_{wB}\right)$

(3) $\mathbb{P}\left(Data_{xT}\, Data_{wB} \mid n_{xT}, r_{xT}, n_{wB}, r_{wB}\right)$

(4) $\mathbb{P}\left(Data_{xT}\, Data_{wT} \mid n_{xT}, r_{xT}, n_{wT}, r_{wT}\right)$

(5) $\mathbb{P}\left(Data_{vT} \mid n_{vT}, r_{vT}\right)$

(6) $\mathbb{P}\left(Data_{tT} \mid n_{tT}, r_{tT}\right)$

(7) $\mathbb{P}\left(Data_{yB}\, Data_{wT} \mid n_{yB}, r_{yB}, n_{wT}, r_{wT}\right)$

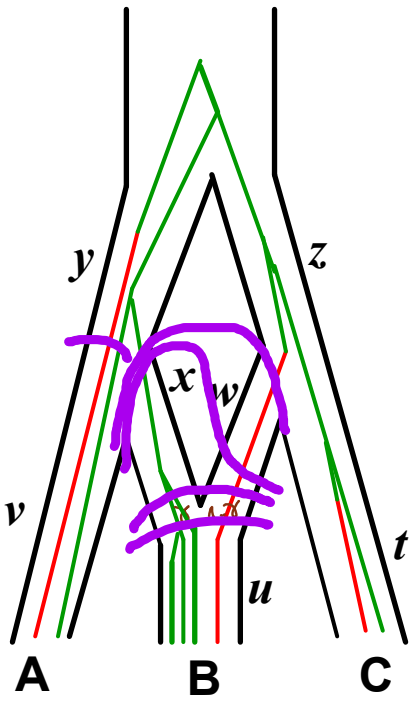
(8) $\mathbb{P}\left(Data_{yB}\, Data_{zB} \mid n_{yB}, r_{yB}, n_{zB}, r_{zB}\right)$

(9) $\mathbb{P}\left(Data_{yT}\, Data_{zB} \mid n_{yT}, r_{yT}, n_{zB}, r_{zB}\right)$

(10) $\mathbb{P}\left(Data_{yT}\, Data_{zT} \mid n_{yT}, r_{yT}, n_{zT}, r_{zT}\right)$

(11) $\mathbb{P}\left(Data \mid n_{root}, r_{root}\right)$

Notre algorithme : calcul des lois jointes



Quantités calculées successivement

(1) $\mathbb{P}\left(Data_{uT} \mid n_{uT}, r_{uT}\right)$

(2) $\mathbb{P}\left(Data_{xB}\, Data_{wB} \mid n_{xB}, r_{xB}, n_{wB}, r_{wB}\right)$

(3) $\mathbb{P}\left(Data_{xT}\, Data_{wB} \mid n_{xT}, r_{xT}, n_{wB}, r_{wB}\right)$

(4) $\mathbb{P}\left(Data_{xT}\, Data_{wT} \mid n_{xT}, r_{xT}, n_{wT}, r_{wT}\right)$

(5) $\mathbb{P}\left(Data_{vT} \mid n_{vT}, r_{vT}\right)$

(6) $\mathbb{P}\left(Data_{tT} \mid n_{tT}, r_{tT}\right)$

(7) $\mathbb{P}\left(Data_{yB}\, Data_{wT} \mid n_{yB}, r_{yB}, n_{wT}, r_{wT}\right)$

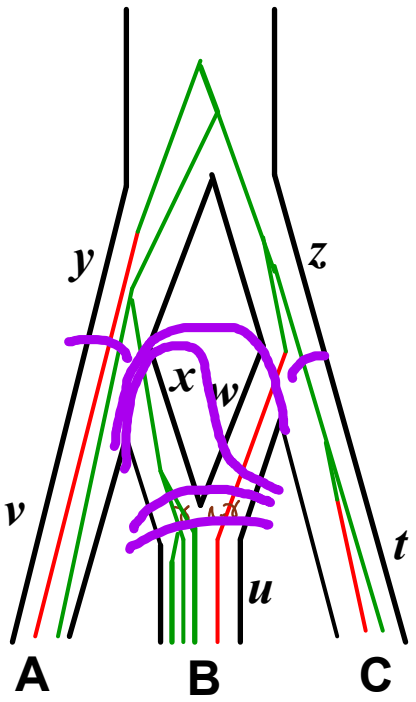
(8) $\mathbb{P}\left(Data_{yB}\, Data_{zB} \mid n_{yB}, r_{yB}, n_{zB}, r_{zB}\right)$

(9) $\mathbb{P}\left(Data_{yT}\, Data_{zB} \mid n_{yT}, r_{yT}, n_{zB}, r_{zB}\right)$

(10) $\mathbb{P}\left(Data_{yT}\, Data_{zT} \mid n_{yT}, r_{yT}, n_{zT}, r_{zT}\right)$

(11) $\mathbb{P}\left(Data \mid n_{root}, r_{root}\right)$

Notre algorithme : calcul des lois jointes



Quantités calculées successivement

(1) $\mathbb{P}\left(Data_{uT} \mid n_{uT}, r_{uT}\right)$

(2) $\mathbb{P}\left(Data_{xB}\, Data_{wB} \mid n_{xB}, r_{xB}, n_{wB}, r_{wB}\right)$

(3) $\mathbb{P}\left(Data_{xT}\, Data_{wB} \mid n_{xT}, r_{xT}, n_{wB}, r_{wB}\right)$

(4) $\mathbb{P}\left(Data_{xT}\, Data_{wT} \mid n_{xT}, r_{xT}, n_{wT}, r_{wT}\right)$

(5) $\mathbb{P}\left(Data_{vT} \mid n_{vT}, r_{vT}\right)$

(6) $\mathbb{P}\left(Data_{tT} \mid n_{tT}, r_{tT}\right)$

(7) $\mathbb{P}\left(Data_{yB}\, Data_{wT} \mid n_{yB}, r_{yB}, n_{wT}, r_{wT}\right)$

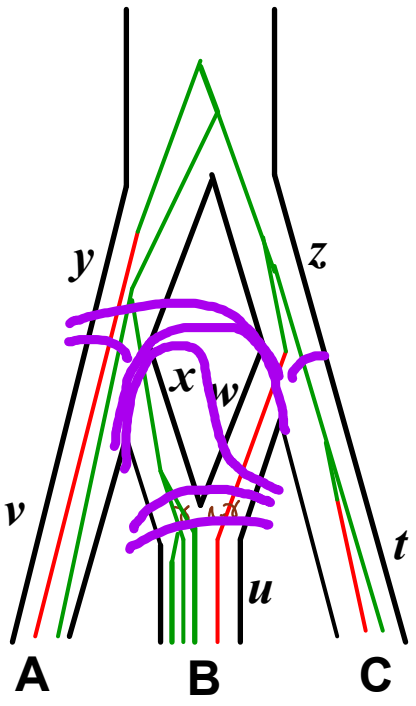
(8) $\mathbb{P}\left(Data_{yB}\, Data_{zB} \mid n_{yB}, r_{yB}, n_{zB}, r_{zB}\right)$

(9) $\mathbb{P}\left(Data_{yT}\, Data_{zB} \mid n_{yT}, r_{yT}, n_{zB}, r_{zB}\right)$

(10) $\mathbb{P}\left(Data_{yT}\, Data_{zT} \mid n_{yT}, r_{yT}, n_{zT}, r_{zT}\right)$

(11) $\mathbb{P}\left(Data \mid n_{root}, r_{root}\right)$

Notre algorithme : calcul des lois jointes



Quantités calculées successivement

(1) $\mathbb{P}\left(Data_{uT} \mid n_{uT}, r_{uT}\right)$

(2) $\mathbb{P}\left(Data_{xB}\, Data_{wB} \mid n_{xB}, r_{xB}, n_{wB}, r_{wB}\right)$

(3) $\mathbb{P}\left(Data_{xT}\, Data_{wB} \mid n_{xT}, r_{xT}, n_{wB}, r_{wB}\right)$

(4) $\mathbb{P}\left(Data_{xT}\, Data_{wT} \mid n_{xT}, r_{xT}, n_{wT}, r_{wT}\right)$

(5) $\mathbb{P}\left(Data_{vT} \mid n_{vT}, r_{vT}\right)$

(6) $\mathbb{P}\left(Data_{tT} \mid n_{tT}, r_{tT}\right)$

(7) $\mathbb{P}\left(Data_{yB}\, Data_{wT} \mid n_{yB}, r_{yB}, n_{wT}, r_{wT}\right)$

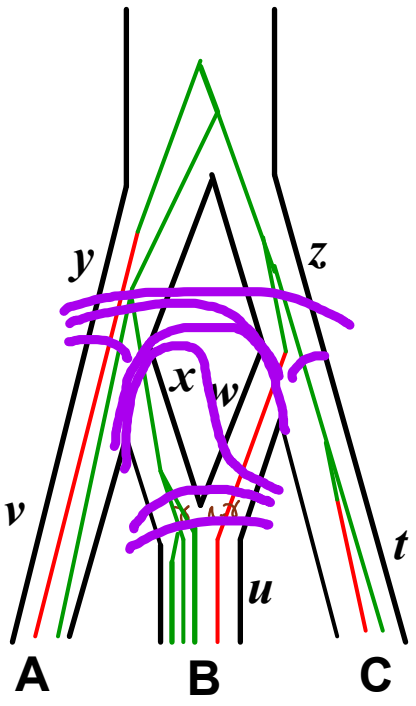
(8) $\mathbb{P}\left(Data_{yB}\, Data_{zB} \mid n_{yB}, r_{yB}, n_{zB}, r_{zB}\right)$

(9) $\mathbb{P}\left(Data_{yT}\, Data_{zB} \mid n_{yT}, r_{yT}, n_{zB}, r_{zB}\right)$

(10) $\mathbb{P}\left(Data_{yT}\, Data_{zT} \mid n_{yT}, r_{yT}, n_{zT}, r_{zT}\right)$

(11) $\mathbb{P}\left(Data \mid n_{root}, r_{root}\right)$

Notre algorithme : calcul des lois jointes



Quantités calculées successivement

(1) $\mathbb{P}\left(Data_{uT} \mid n_{uT}, r_{uT}\right)$

(2) $\mathbb{P}\left(Data_{xB}\, Data_{wB} \mid n_{xB}, r_{xB}, n_{wB}, r_{wB}\right)$

(3) $\mathbb{P}\left(Data_{xT}\, Data_{wB} \mid n_{xT}, r_{xT}, n_{wB}, r_{wB}\right)$

(4) $\mathbb{P}\left(Data_{xT}\, Data_{wT} \mid n_{xT}, r_{xT}, n_{wT}, r_{wT}\right)$

(5) $\mathbb{P}\left(Data_{vT} \mid n_{vT}, r_{vT}\right)$

(6) $\mathbb{P}\left(Data_{tT} \mid n_{tT}, r_{tT}\right)$

(7) $\mathbb{P}\left(Data_{yB}\, Data_{wT} \mid n_{yB}, r_{yB}, n_{wT}, r_{wT}\right)$

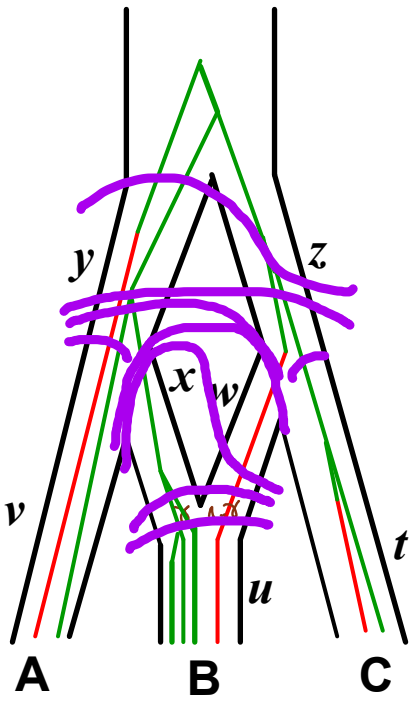
(8) $\mathbb{P}\left(Data_{yB}\, Data_{zB} \mid n_{yB}, r_{yB}, n_{zB}, r_{zB}\right)$

(9) $\mathbb{P}\left(Data_{yT}\, Data_{zB} \mid n_{yT}, r_{yT}, n_{zB}, r_{zB}\right)$

(10) $\mathbb{P}\left(Data_{yT}\, Data_{zT} \mid n_{yT}, r_{yT}, n_{zT}, r_{zT}\right)$

(11) $\mathbb{P}\left(Data \mid n_{root}, r_{root}\right)$

Notre algorithme : calcul des lois jointes



Quantités calculées successivement

(1) $\mathbb{P}\left(Data_{uT} \mid n_{uT}, r_{uT}\right)$

(2) $\mathbb{P}\left(Data_{xB}\, Data_{wB} \mid n_{xB}, r_{xB}, n_{wB}, r_{wB}\right)$

(3) $\mathbb{P}\left(Data_{xT}\, Data_{wB} \mid n_{xT}, r_{xT}, n_{wB}, r_{wB}\right)$

(4) $\mathbb{P}\left(Data_{xT}\, Data_{wT} \mid n_{xT}, r_{xT}, n_{wT}, r_{wT}\right)$

(5) $\mathbb{P}\left(Data_{vT} \mid n_{vT}, r_{vT}\right)$

(6) $\mathbb{P}\left(Data_{tT} \mid n_{tT}, r_{tT}\right)$

(7) $\mathbb{P}\left(Data_{yB}\, Data_{wT} \mid n_{yB}, r_{yB}, n_{wT}, r_{wT}\right)$

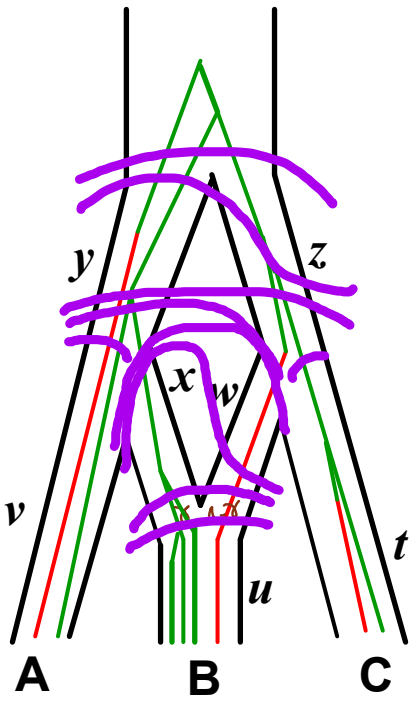
(8) $\mathbb{P}\left(Data_{yB}\, Data_{zB} \mid n_{yB}, r_{yB}, n_{zB}, r_{zB}\right)$

(9) $\mathbb{P}\left(Data_{yT}\, Data_{zB} \mid n_{yT}, r_{yT}, n_{zB}, r_{zB}\right)$

(10) $\mathbb{P}\left(Data_{yT}\, Data_{zT} \mid n_{yT}, r_{yT}, n_{zT}, r_{zT}\right)$

(11) $\mathbb{P}\left(Data \mid n_{root}, r_{root}\right)$

Notre algorithme : calcul des lois jointes



Quantités calculées successivement

(1) $\mathbb{P}\left(Data_{uT} \mid n_{uT}, r_{uT}\right)$

(2) $\mathbb{P}\left(Data_{xB}\, Data_{wB} \mid n_{xB}, r_{xB}, n_{wB}, r_{wB}\right)$

(3) $\mathbb{P}\left(Data_{xT}\, Data_{wB} \mid n_{xT}, r_{xT}, n_{wB}, r_{wB}\right)$

(4) $\mathbb{P}\left(Data_{xT}\, Data_{wT} \mid n_{xT}, r_{xT}, n_{wT}, r_{wT}\right)$

(5) $\mathbb{P}\left(Data_{vT} \mid n_{vT}, r_{vT}\right)$

(6) $\mathbb{P}\left(Data_{tT} \mid n_{tT}, r_{tT}\right)$

(7) $\mathbb{P}\left(Data_{yB}\, Data_{wT} \mid n_{yB}, r_{yB}, n_{wT}, r_{wT}\right)$

(8) $\mathbb{P}\left(Data_{yB}\, Data_{zB} \mid n_{yB}, r_{yB}, n_{zB}, r_{zB}\right)$

(9) $\mathbb{P}\left(Data_{yT}\, Data_{zB} \mid n_{yT}, r_{yT}, n_{zB}, r_{zB}\right)$

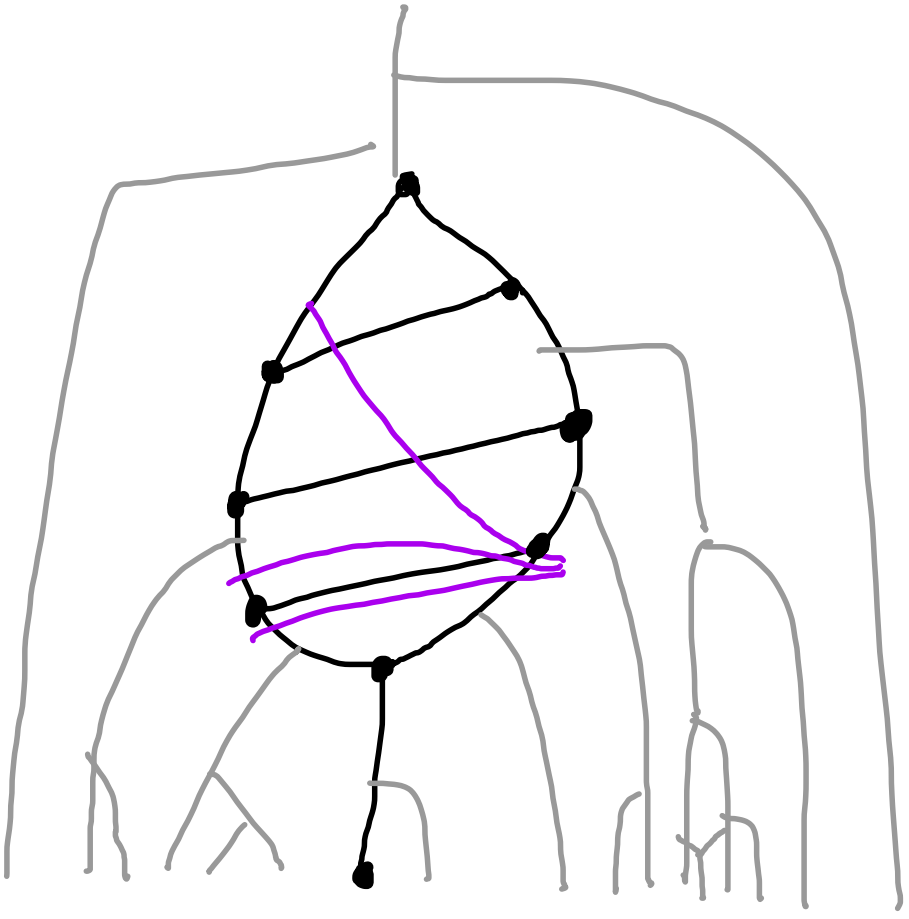
(10) $\mathbb{P}\left(Data_{yT}\, Data_{zT} \mid n_{yT}, r_{yT}, n_{zT}, r_{zT}\right)$

(11) $\mathbb{P}\left(Data \mid n_{root}, r_{root}\right)$

Nous cherchons à minimiser le nombre d'arêtes à considérer simultanément dans nos calculs de probabilités

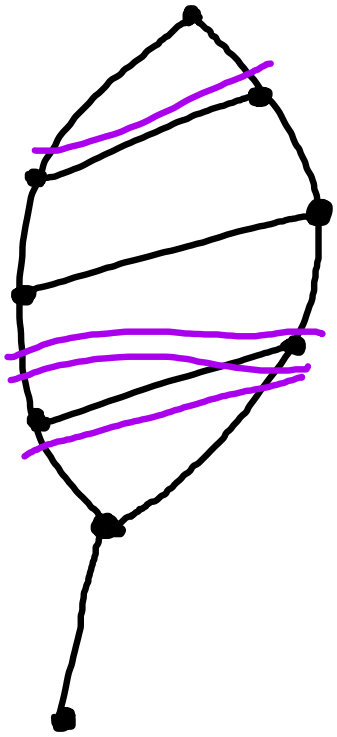
Un parcours à éviter

Un maximum de 5 arêtes



Un parcours intéressant

Un maximum de 3 arêtes



Plan

A propos du prior sur le réseau (Zhang et al., MBE 2018)

Rappel au niveau du posterior :

$$\mathbb{P}(N|X_1, \ldots, X_m) \propto \mathbb{P}\left(Data \mid N\right) P(N)$$

Le prior est $P(N)$.
Le processus de naissance hybridation dépend de paramètres ν et λ.

Des lois sont imposées sur ν et λ : on parle d'hyper prior.

$$\mathbb{P}(N, X, Y) = \mathbb{P}(N \mid X, Y)P(X)P(Y)$$

- X : v.a. pour la valeur $\lambda - \nu$
- la loi choisie pour X est une loi exponentielle
- Y : v.a. pour la valeur $\frac{\nu}{\lambda}$
- la loi choisie pour Y est une loi Beta

Plan

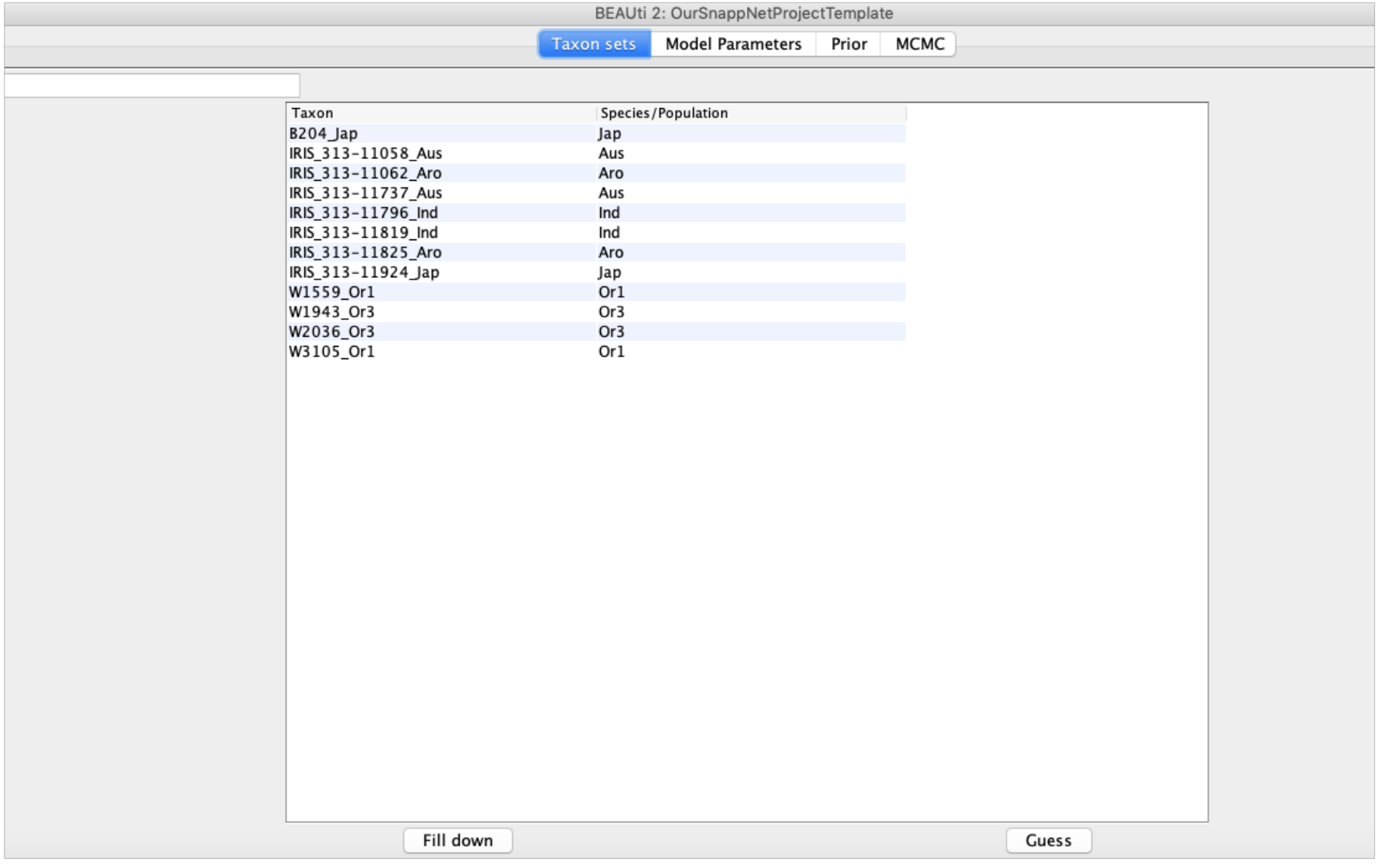
A propos du xml de SNAPPNET (Add On pour Beast)

<distribution id="networkPrior"
spec="speciesnetwork.BirthHybridizationModel"
network="@network :species" netDiversification="@netDivRate :species"
turnOver="@turnOverRate :species"/>
<prior id="netDivPrior" name="distribution" x="@netDivRate :species">
<Exponential id="exponential.01" name="distr" mean="10.0"/>
</prior>
<prior id="turnOverPrior" name="distribution" x="@turnOverRate :species">
<Beta id="betadistr.01" name="distr" alpha="1.0" beta="1.0"/>
</prior>

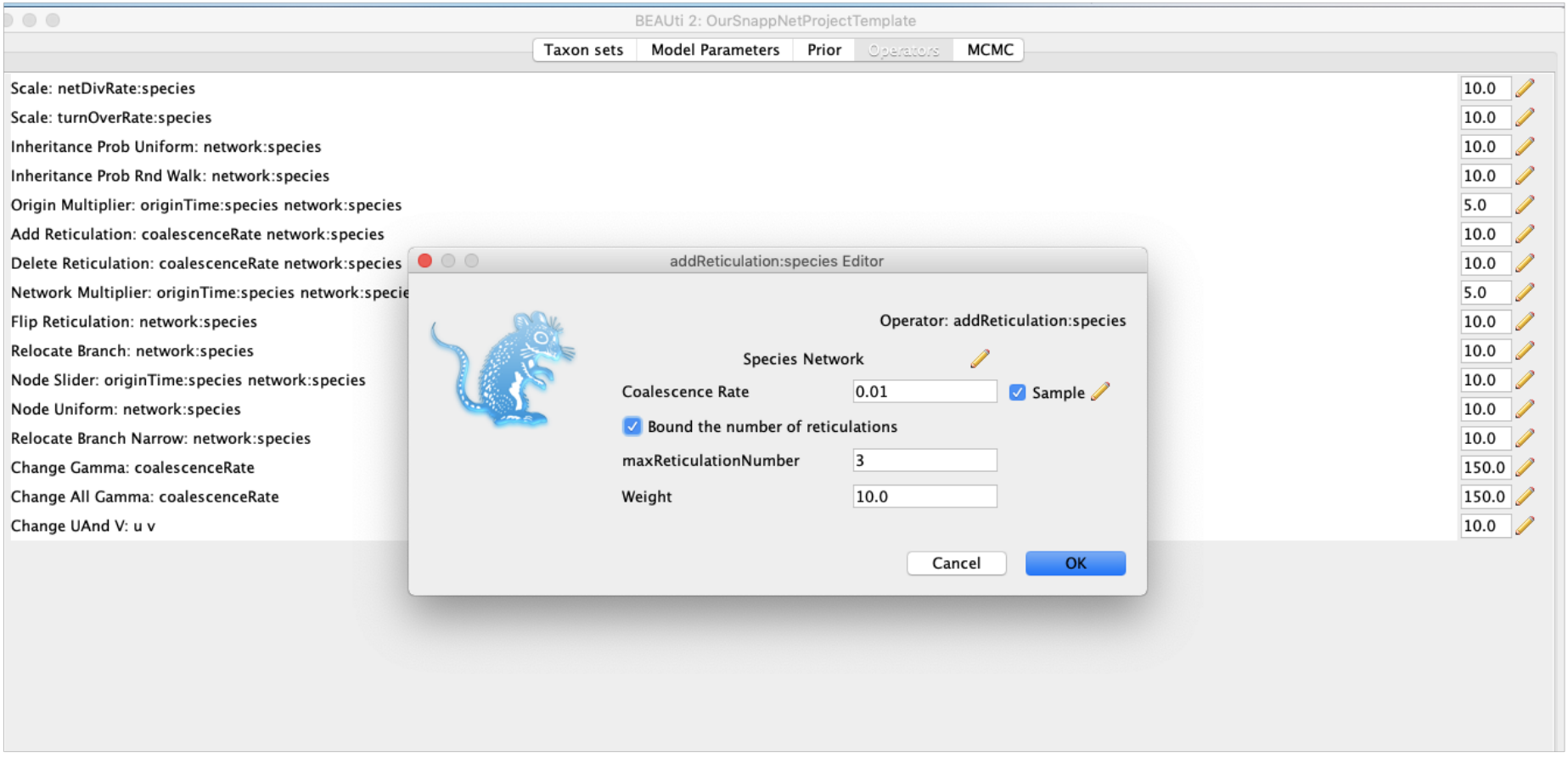
Obtention du xml de SNAPPNET :

BEAUti — Bayesian Evolutionary Analysis Utility.
This program is used to import data, design the analysis, and generate the BEAST control file.

BEAUti : correspondance entre variétés et espèces



BEAUti : choix du nombre maximum de réticulations



Plan

A propos des opérateurs implémentés

16 opérateurs pour l'échantillonnage par MCMC (Bryant et al, MBE 2012 ; Zhang et al., MBE 2017)

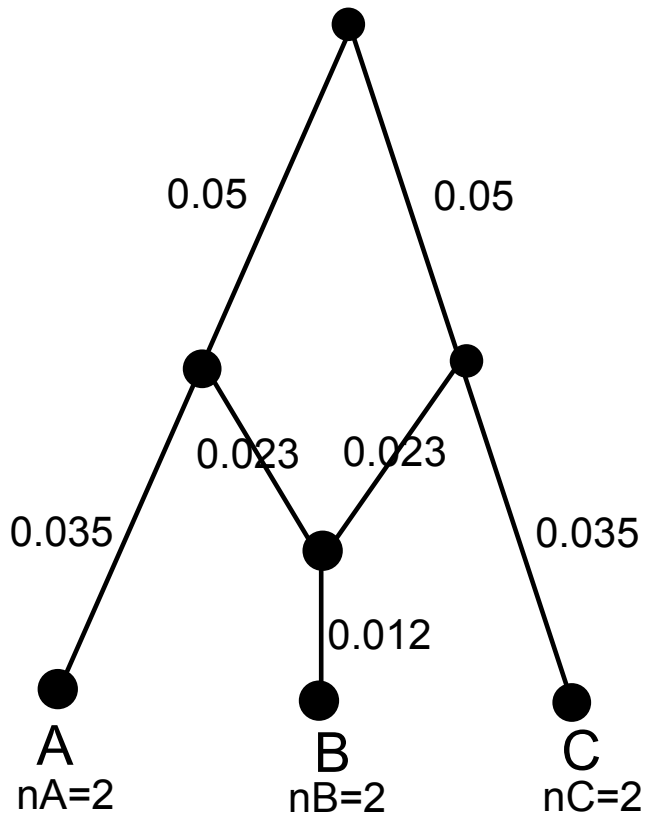
Opérateurs de changement topologique

- addReticulation : ajout d'un noeud de réticulation
- deleteReticulation : suppression d'un noeud de réticulation
- flipReticulation : inversion de l'orientation d'une branche de réticulation
- relocateBranch

Autres opérateurs, comme par exemple

- MutationMover : changement des valeurs des taux de mutation u (rouge $\rightarrow$ vert) et v (vert $\rightarrow$ rouge), sous la contrainte $\frac{2uv}{u+v} = 1$
- ChangeTheta : changement de la taille de population θ liée à une branche
- ChangeAllTheta : changement de toutes les tailles de population θ
- turnOverScale : changement de la valeur du paramètre $\frac{\nu}{\lambda}$ lié au processus de naissance hybridation
  (ν taux d'hybridation, λ taux de spéciation)
- divrRateScale : changement de valeur du paramètre λ - ν lié au processus de naissance hybridation

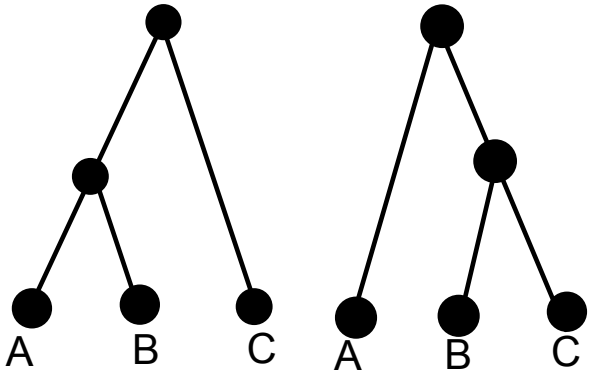
Un exemple de réseau étudié par simulation



- Longueurs de branches en nombre de mutations par site
- nA=2, nB=2, nC=2
- 1 000 sites ou 10 000 sites
- Tailles de population θ égales à 0.005 ou 0.05
- T : temps de coalescence entre 2 lignées (en mutations par site)
  - si $\theta = 0.005$, alors $\mathbb{E}(T) = 0.005/2 = 0.0025$
  - si $\theta = 0.05$, alors $\mathbb{E}(T) = 0.005/2 = 0.025$

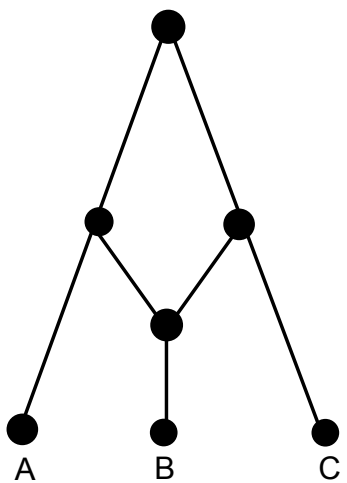
Réseaux échantillonnés par MCMC

- 1 000 sites, $\theta = 0.005$



51.94% 48.06%

- 10 000 sites, $\theta = 0.005$
- 1 000 sites, $\theta = 0.05$
- 10 000 sites, $\theta = 0.05$



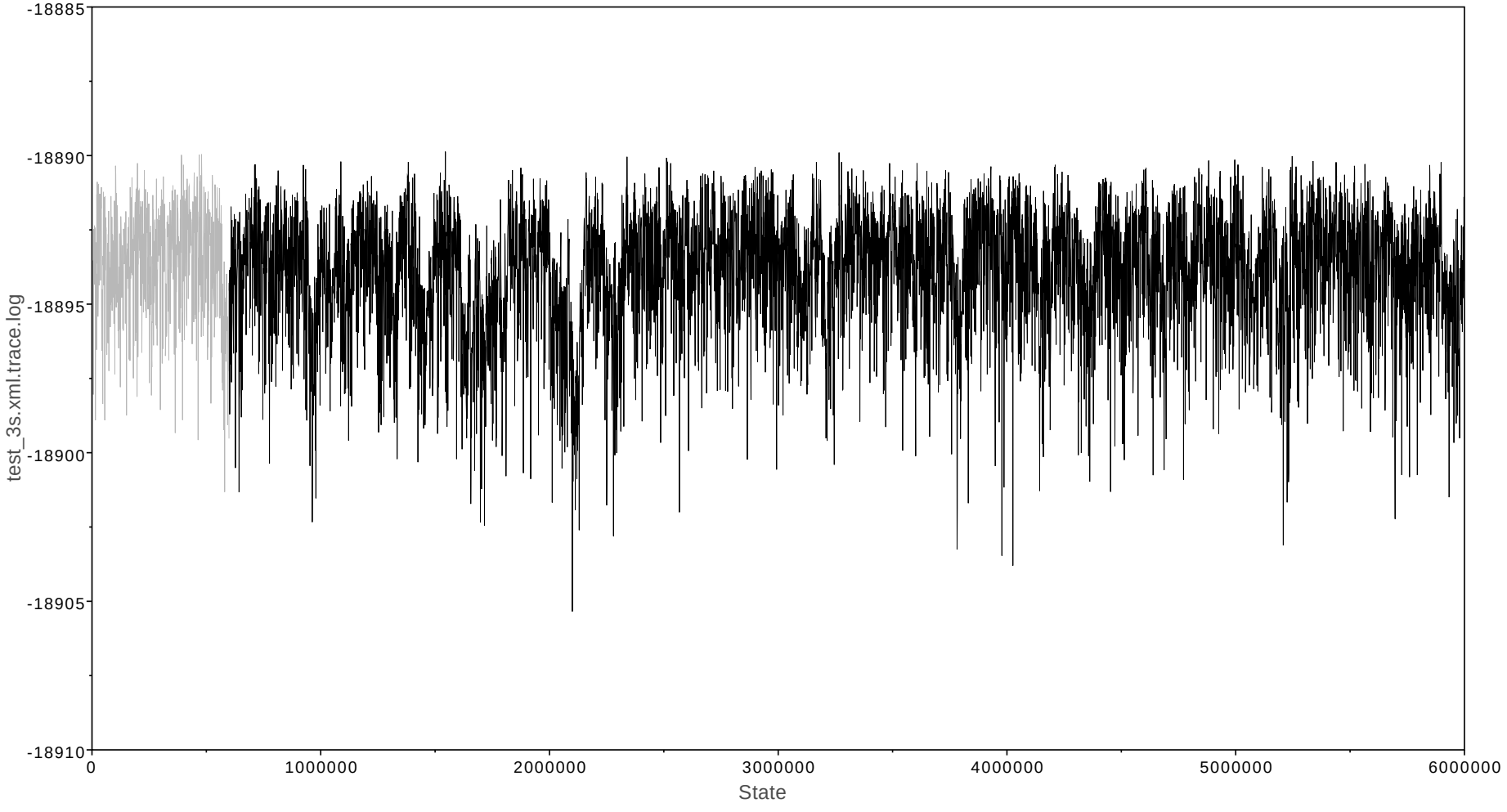
100%

Avec une taille de population plus importante,
on a besoin de moins de sites pour retrouver le réseau ! ! !

Trace plot
Cas 10 000 sites, $\theta = 0.05$

Analyse avec le logiciel Tracer

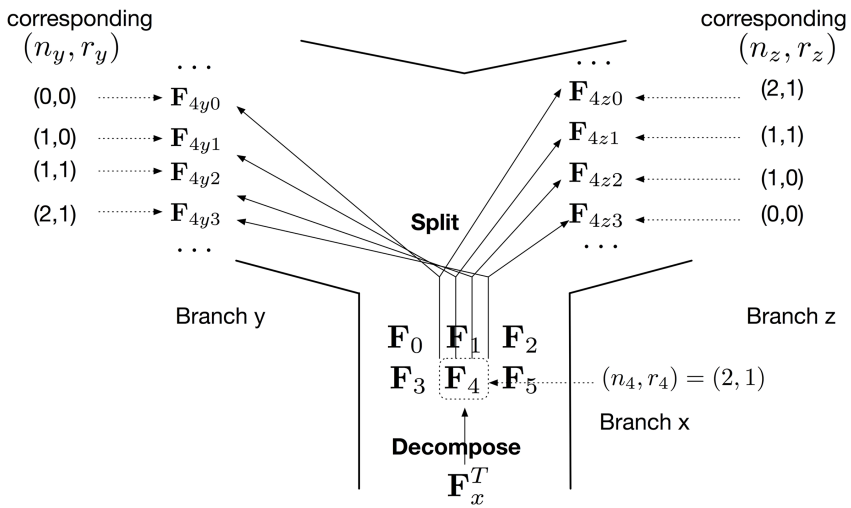
Effective Sample Size (ESS) = 413



Plan

La méthode MCMCBiMarkers de Zhu et al (Plos Comput Biol, 2018)

- Enumération de toutes les partitions possibles des lignées au sein du réseau



MCMCBiMarkers implémenté dans PHYLONET

Complexité pour le calcul de vraisemblance de l'ordre : MCMCBiMarkers vs SNAPPNET

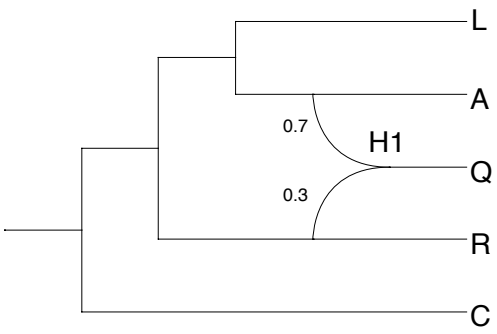
Notations

- m : nombre de noeuds
- n : nombre maximal de varietés dans une espèce

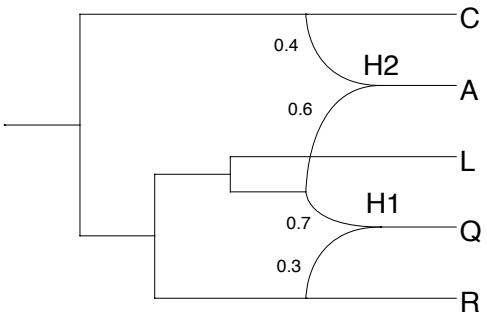
Complexités

- MCMCBiMarkers : $O(mn^{4k+4})$,
  - k : niveau du réseau
- SNAPPNET : $O(mn^{2k'+2})$
  - k' : nombre maximum de branches simultanément considérées dans les vraisemblances partielles

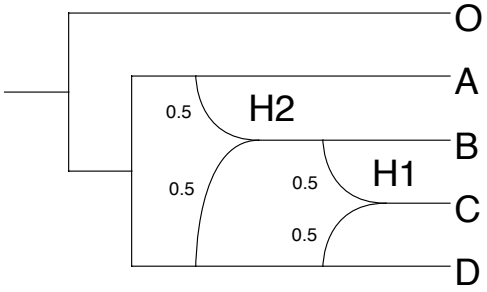
3 Réseaux étudiés par simulation



Réseau A



Réseau B



Réseau C

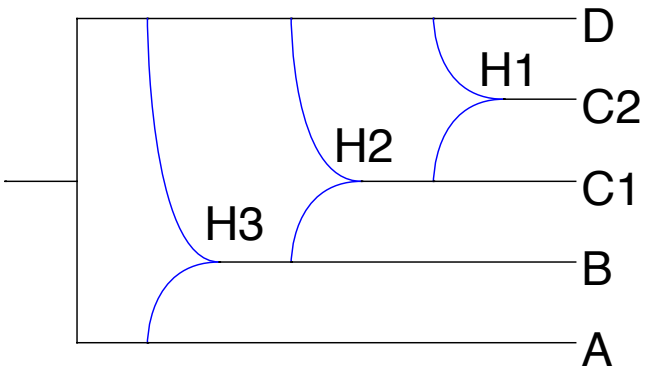
Réseaux A et B tirés de Zhu et al (Plos Comput Biol, 2018)

MCMCBiMarkers vs SNAPPNET

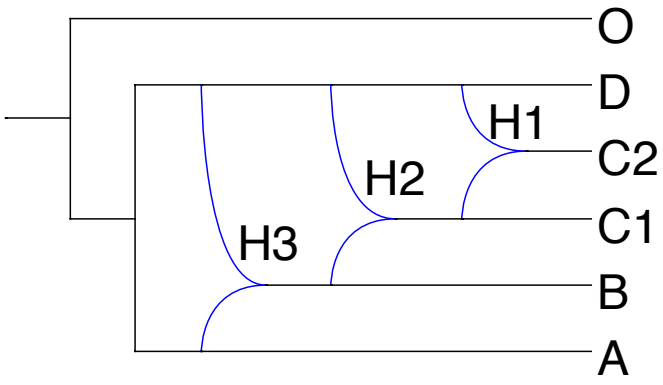
- Réseau A : $k = 1$ $O(n^8)$, $k' = 2$ $O(n^6)$
- Réseaux B et C : $k = 2$ $O(n^{12})$, $k' = 3$ $O(n^8)$

Réseaux de la famille C

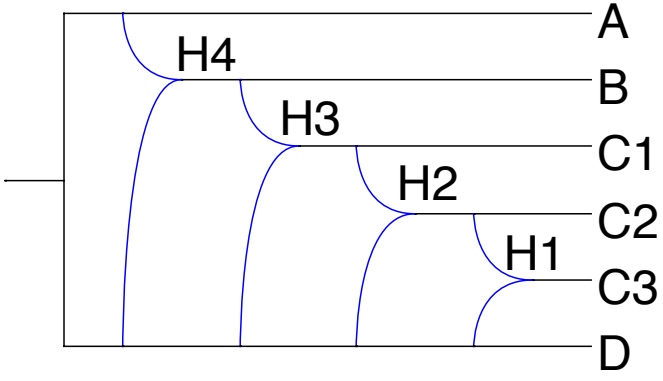
Réseaux avec 3 ou 4 noeuds de réticulations, et avec ou sans Outgroup O



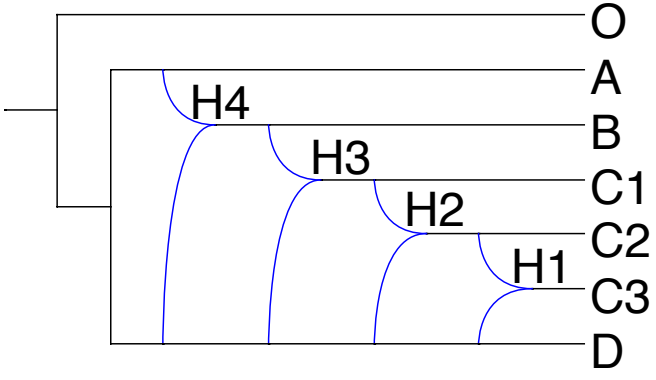
Réseau C(3) sans O



Réseau C(3) avec O



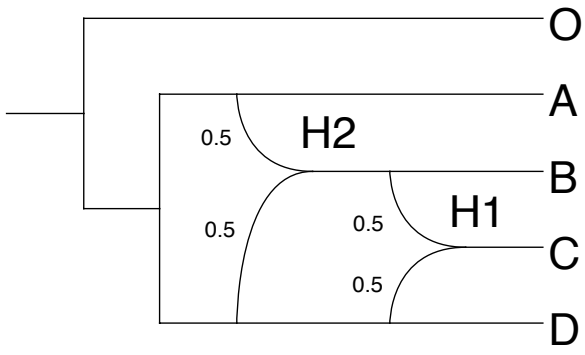
Réseau C(4) sans O



Réseau C(4) avec O

Comparaison des temps de calcul de la vraisemblance $\mathbb{P}(Data \mid N)$

Dataset ID	CPU time	
	SNAPPNET (en minutes)	MCMCBiMarker (en heures)
1	5.559	35.9354
2	5.6763	34.2433
3	5.7351	32.6519
4	5.446	34.2011
5	5.5996	33.2354



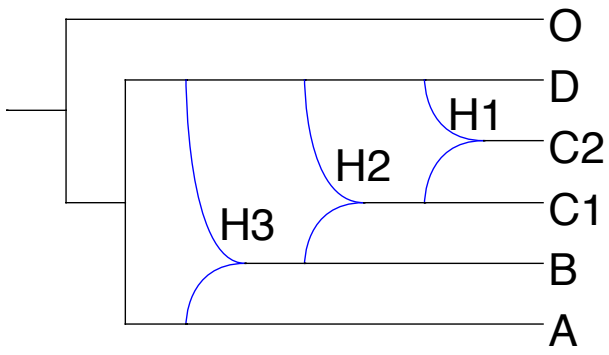
Réseau C de niveau 2

SNAPPNET vs MCMCBiMarker
(Zhu et al., Plos Comp Biol 2018)

$O(n^8)$ vs $O(n^{12})$

Comparaison des temps de calcul de la vraisemblance $\mathbb{P}(Data \mid N)$

Dataset ID	CPU time	
	SNAPPNET (en minutes)	MCMCBiMarker (en heures)
1	23.75	> 336
2	25.14	> 336
3	25.31	> 336
4	24.90	> 336
5	25.15	> 336



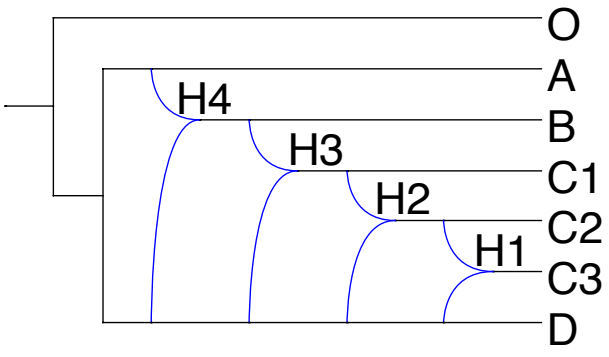
Réseau C(3)

SNAPPNET vs MCMCBiMarker
(Zhu et al., Plos Comp Biol 2018)

$O(n^8)$ vs $O(n^{16})$

Comparaison des temps de calcul de la vraisemblance $\mathbb{P}(Data \mid N)$

Dataset ID	CPU time	
	SNAPPNET (en minutes)	MCMCBiMarker (en heures)
1	69.72	> 336
2	66.91	> 336
3	70.13	> 336
4	69.36	> 336
5	69.38	> 336



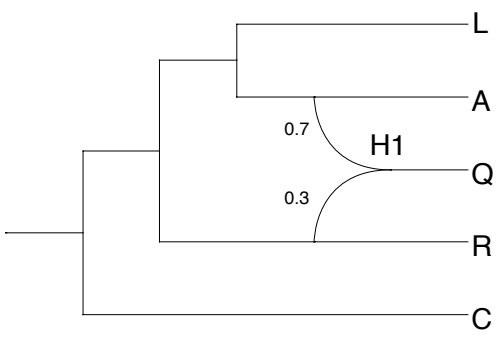
Réseau C(4)

SNAPPNET vs MCMCBiMarker
(Zhu et al., Plos Comp Biol 2018)

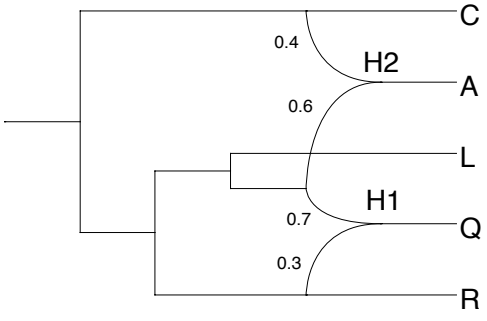
$O(n^8)$ vs $O(n^{20})$

Capacité de SNAPPNET à retrouver la topologie des réseaux A et B

Nombre de sites	1,000	10,000	100,000
Réseau A	0%	100%	100%
Réseau B	0%	81.25%	100%



Réseau A



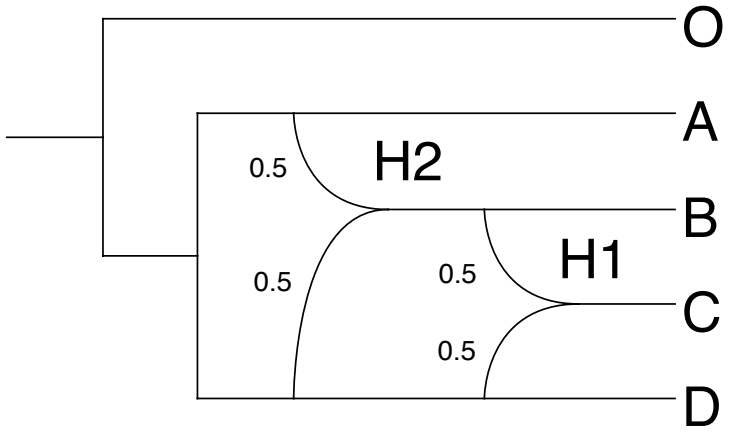
Réseau B

- 1000 sites : MCMCBiMarker $>$ SNAPPNET
- $\geq$ 10000 sites : SNAPPNET $\approx$ MCMCBiMarker
- 10 000 sites requis pour inférer ces réseaux

Capacité à retrouver la topologie du réseau C

SNAPPNET vs MCMCBiMarker

	Nombre de sites		
Nb lignées pour B et pour C	**1,000**	**10,000**	**100,000**
1	0% 0%	7.87% 4.84%	54.90% 0%
4	0% 0%	50.00% 0%	49.60% 0%

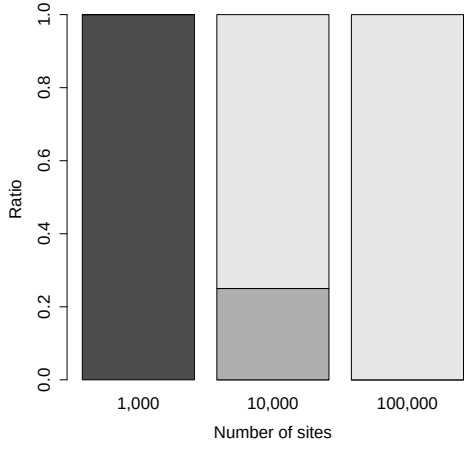


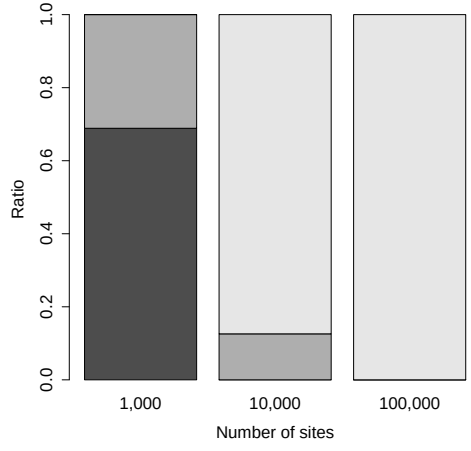
Réseau C

Réseaux échantillonnés lors du MCMC

Ratio arbres / réseau à 1-reticulation / réseau à 2-reticulations

1 lignée pour B et C

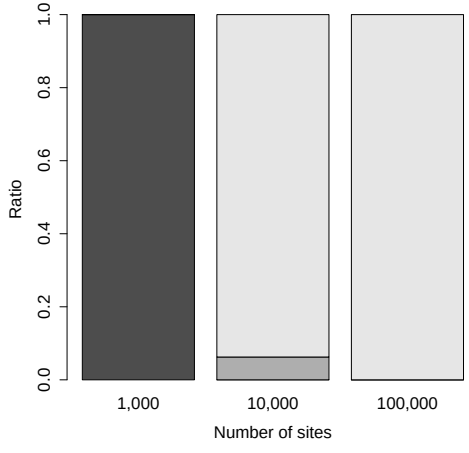


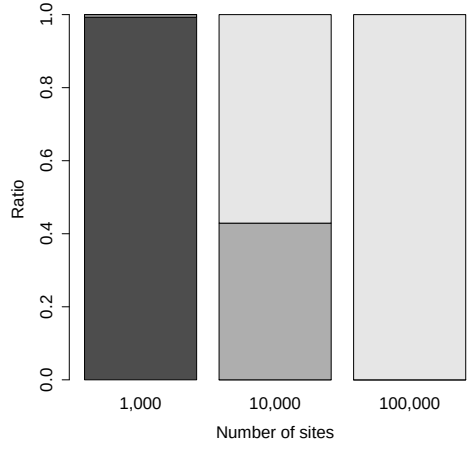


SNAPPNET

MCMCBiMarkers

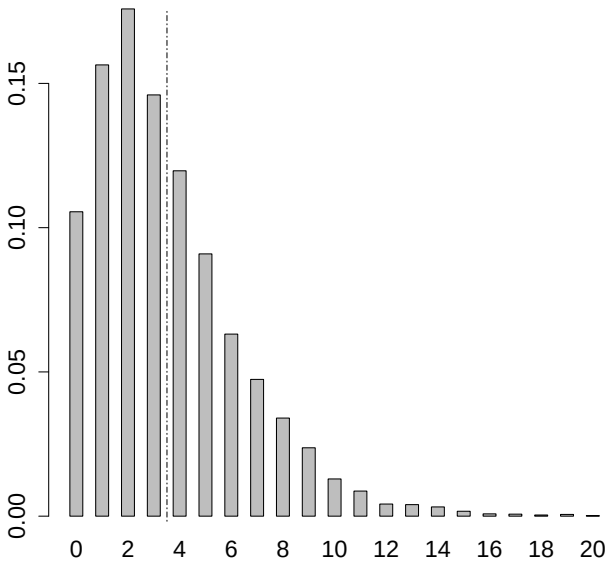
4 lignées pour B et C



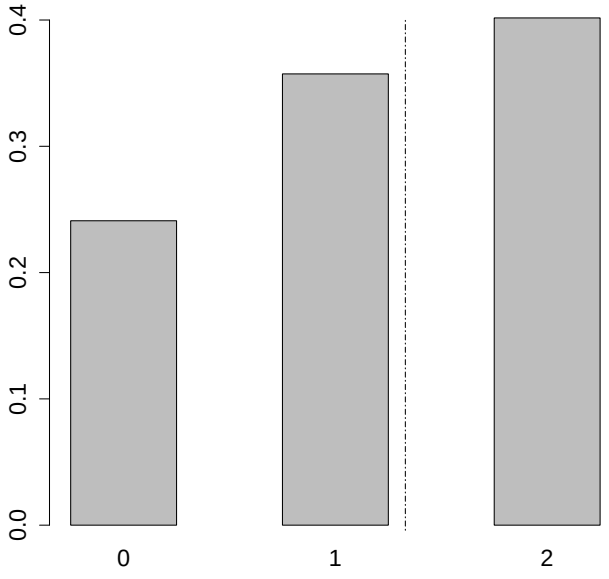


A propos du processus de naissance hybridation dans SNAPPNET (Zhang et al., MBE 2018)

Distribution du nombre de réticulations



Sans contrainte



en bornant à 2 réticulations

Afin de limiter l'espace d'exploration $\Rightarrow$ un maximum de 2 réticulations

dans MCMCBiMarker : a priori basé sur graphes de recombinaison ancestrale

Capacité de SNAPPNET à retrouver la topologie en fonction de l'a priori sur les tailles de population

Vraie valeur $\theta = 0.005$

$\theta \sim \Gamma(1, 200)$
$\mathbb{E}(\theta) = 0.005$

Nombre de sites	1,000	10,000	100,000
Réseau A	0%	100%	100%
Réseau B	0%	81.25%	100%

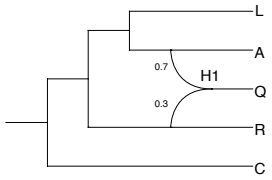
$\theta \sim \Gamma(1, 1000)$
$\mathbb{E}(\theta) = 0.001$

Nombre de sites	1,000	10,000	100,000
Réseau A	0%	94.73%	91.30%
Réseau B	0%	80%	95.65%

$\theta \sim \Gamma(1, 2000)$
$\mathbb{E}(\theta) = 0.0005$

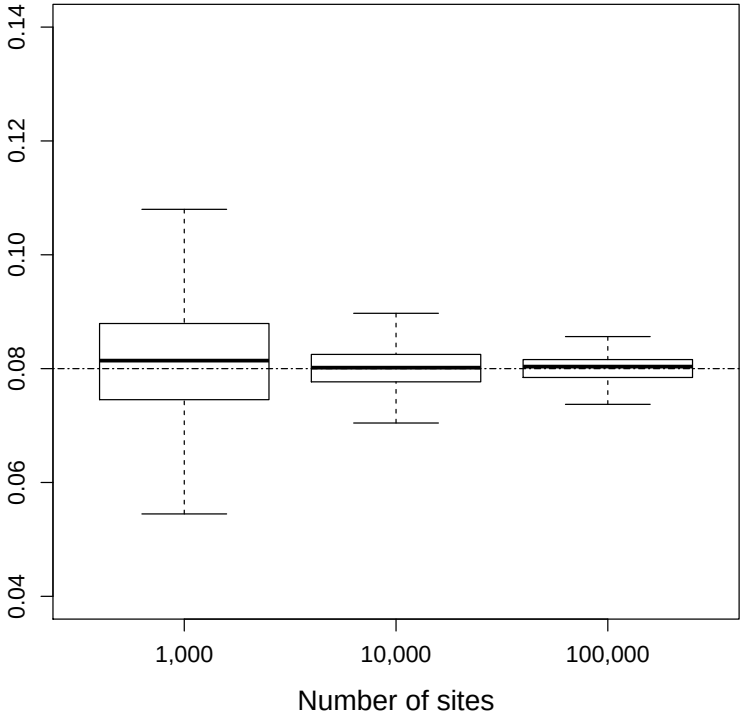
Nombre de sites	1,000	10,000	100,000
Réseau A	0%	100%	80%
Réseau B	0%	85%	85.71%

Estimation des paramètres continus par SNAPPNET

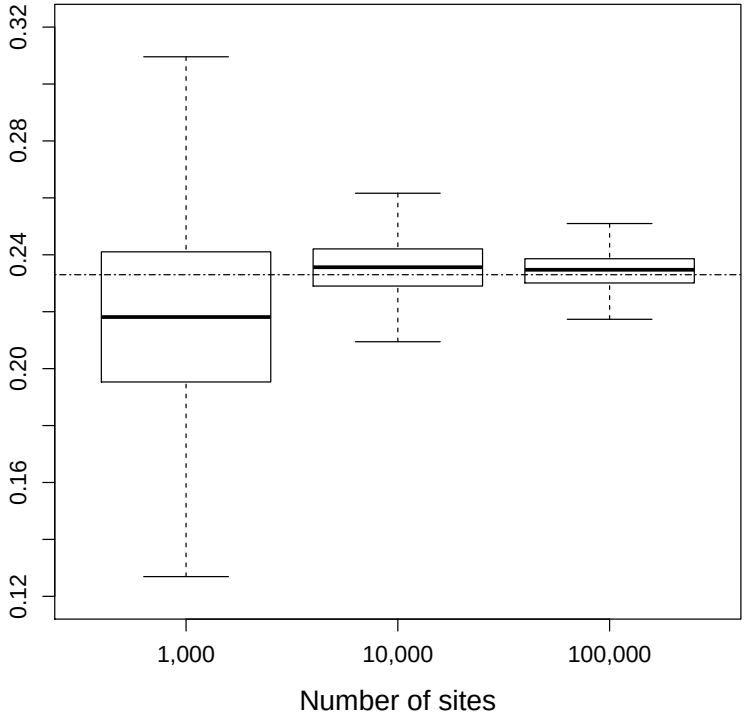


Réseau A

2 varietés par espèces

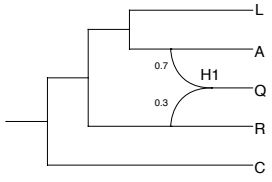


Hauteur du réseau



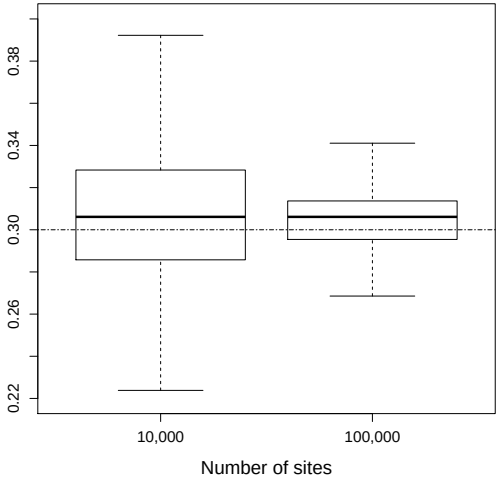
Longueur du réseau

Estimation des paramètres continus par SNAPPNET

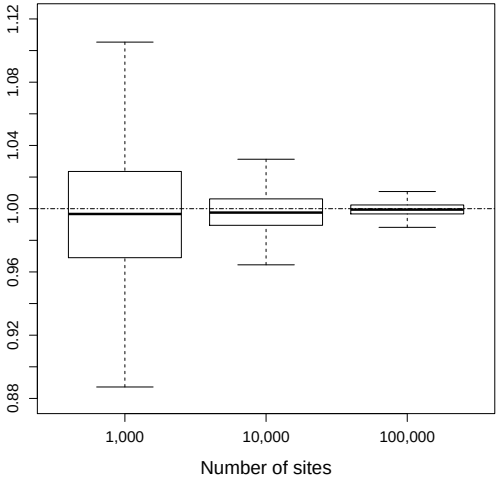


Réseau A

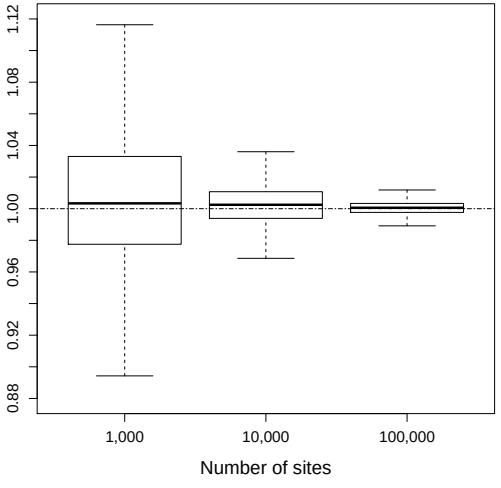
2 varietés par espèces



Probabilité d'hybridation γ

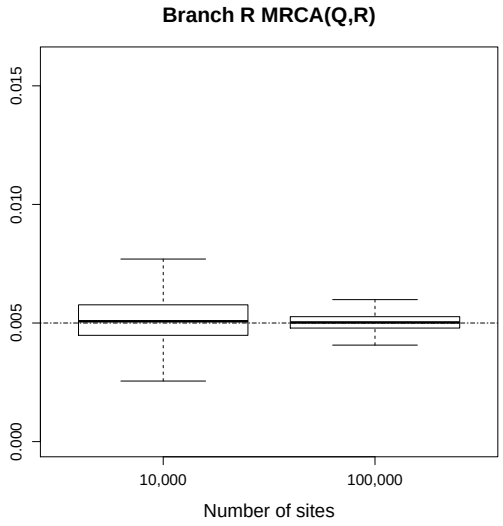


Taux instantané *u*

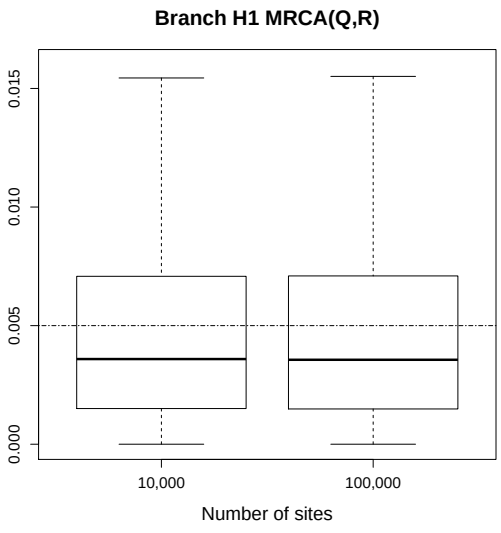


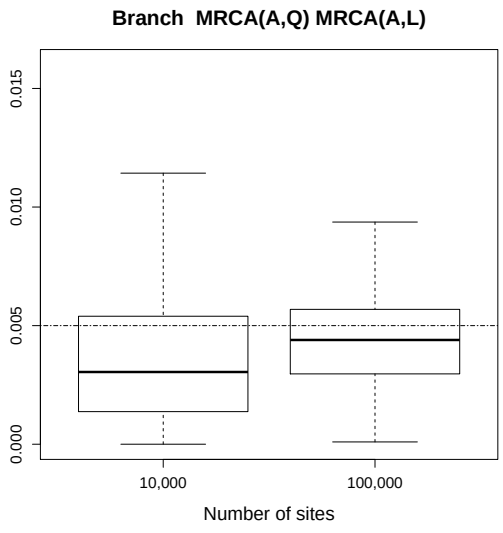
Taux instantané *v*

Les tailles de population récentes sont plus facilement estimables



- 2 branches externes





- 2 branches internes

A propos des sites polymorphes

Les sites polymorphes sont utiles pour la reconstruction de réseau (Zhu et al, 2018)

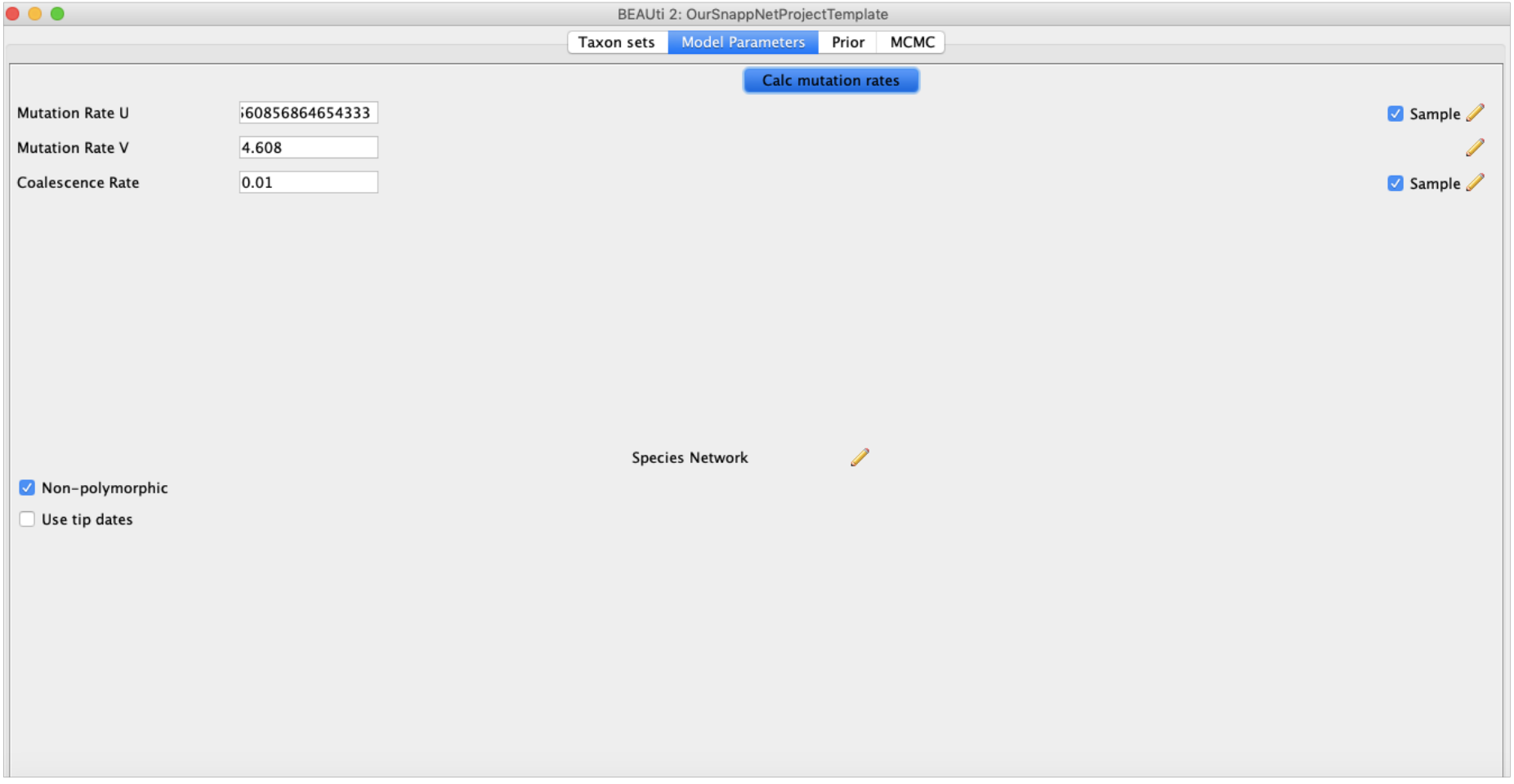
Inference de la topologie du reseau A par SNAPPNET

- 1000 sites polymorphes
  - Le taux de reconstruction est désormais de 94.45%
  - Si on oublie de conditionner, le taux de reconstruction chute à 23.81%
- $\geq$ 10000 sites : le taux de reconstruction est proche de 100% avec ou sans conditionnement

Paramètres continus :

- Estimations semblables à celles obtenues dans l'experience incluant les sites invariants
- Si on oublie de conditionner : longueur et hauteur biaisés, taux u et v non biaisés

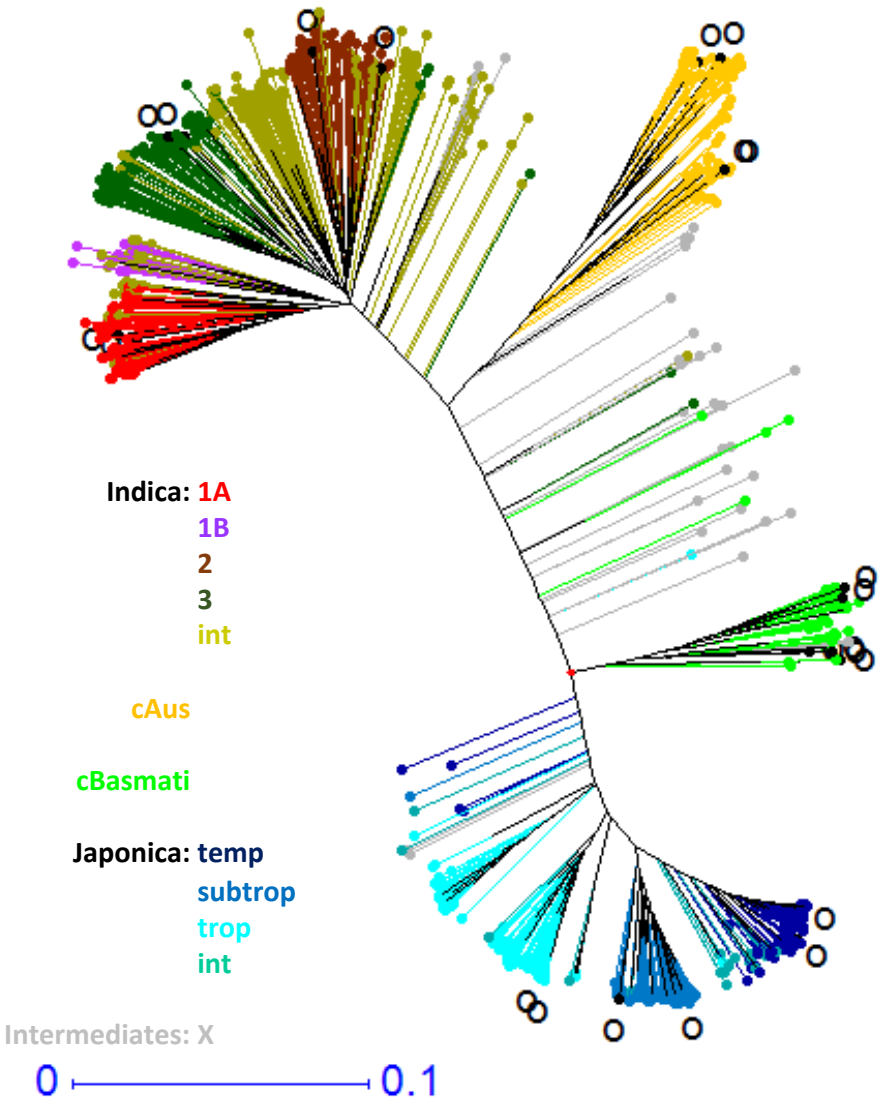
Indiquez à SNAPPNET si vous avez conservé ou non les sites polymorphes



Plan

24 variétés selectionnées par J.C. Glaszmann

Arbre neighbour joining basé sur Wang et al. (Nature, 2018), 3000 variétés de riz (4.8 millions de SNPs)



24 variétés selectionnées par J.C. Glaszmann

Sub-populations	Dataset	Variety ID	Country	Variety name
circum Aus	1	IRIS-313-11058	Bangladesh	AUS 329
		IRIS 313-11737	India	CHUNDI
	2	IRIS-313-10852	India	ARC 7336
		IRIS-313-11027	Pakistan	JHONA 101
circum Basmati	1	IRIS-313-11062	Bangladesh	BEGUNBICHI 33
		IRIS-313-11825	India	HANSRAJ
		IRIS-313-8326	India	JC1
	2	IRIS-313-11258	India	ARC 13502
		IRIS-313-10851	India	ARC 7296
		IRIS-313-12094	Bangladesh	ARC KASHA
Indica	1	IRIS-313-11819	Myanmar	PADINTHUMA
		IRIS-313-11796	China	DU GEN CHUAN
		IRIS-313-11089	Cambodia	SRAU THMOR
	2	IS-313-11646	India	NCS771 A
		CX270	Taiwan	TAICHUNGNATIVE1
		IRIS-313-11741	SriLanka	HERATH BANDA
Japonica	1	B204	China	LONGHUAMAOHU
		IRIS-313-11924	Thailand	NAM JAM
		IRIS-313-10577	Philippines	IFUGAO RICE
	2	IRIS-313-11691	Bhutan	SHANGYIPA
		IRIS-313-7883	Indonesia	GANIGI
		B269	China	YUEFU

Dataset 1 : de l'Inde aux Philippines, Dataset 2 : du Pakistan à l'Indonésie

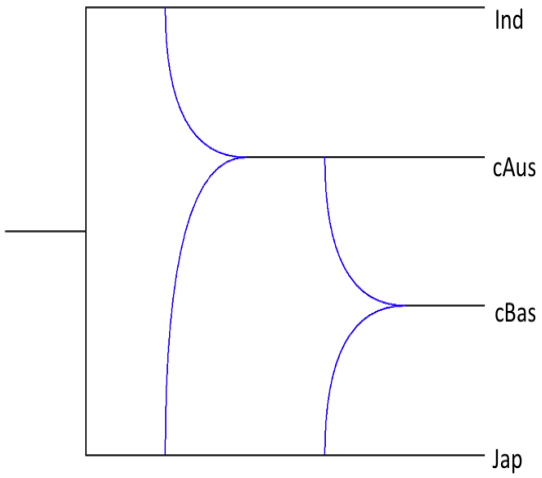
24 variétés selectionnées par J.C. Glaszmann

Dataset 1 : de l'Inde aux Philippines, Dataset 2 : du Pakistan à l'Indonésie

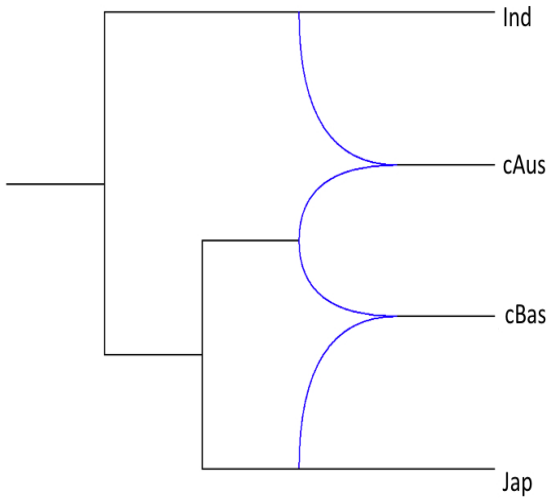
source http ://axl.cefan.ulaval.ca/

Les dix réseaux riz étudiés

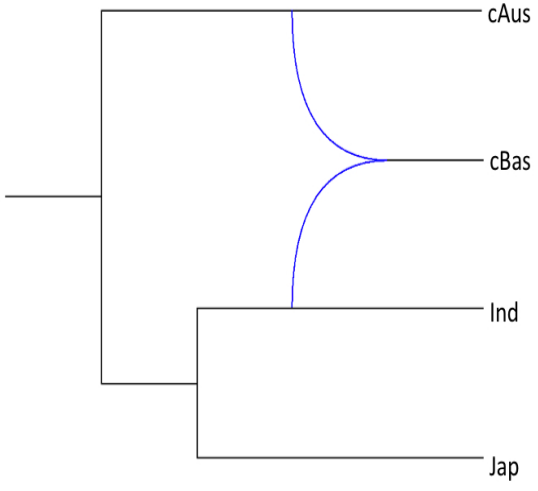
Réseau 1



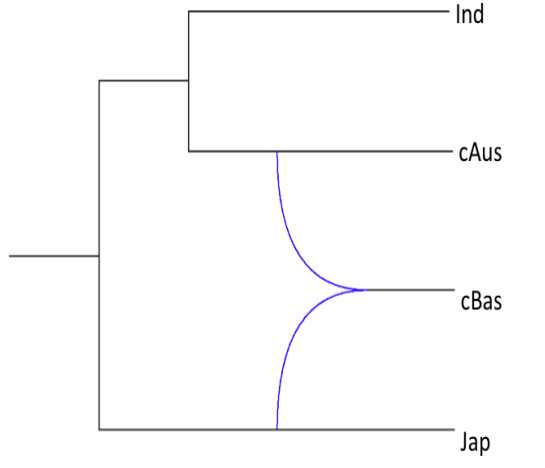
Réseau 2



Réseau 3 (peu conforme aux attentes)

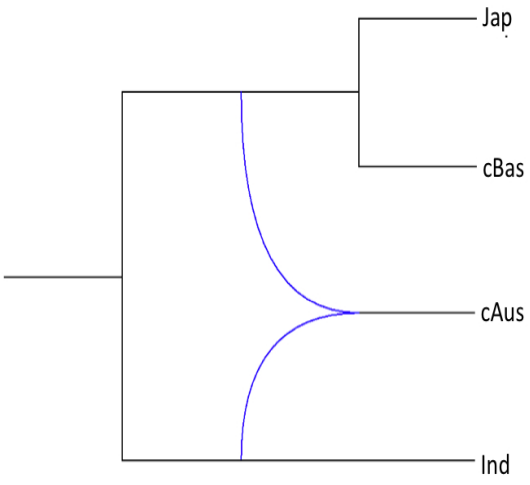


Réseau 4 (le plus conforme aux attentes)

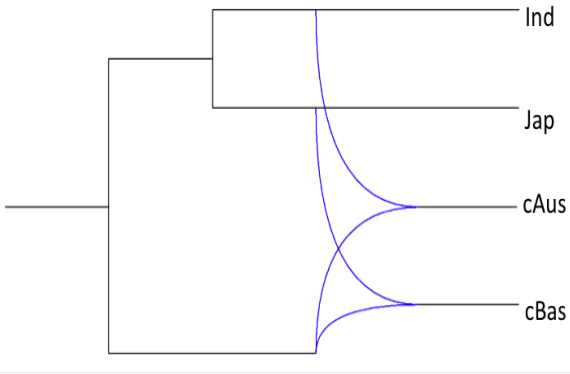


Les dix réseaux riz étudiés

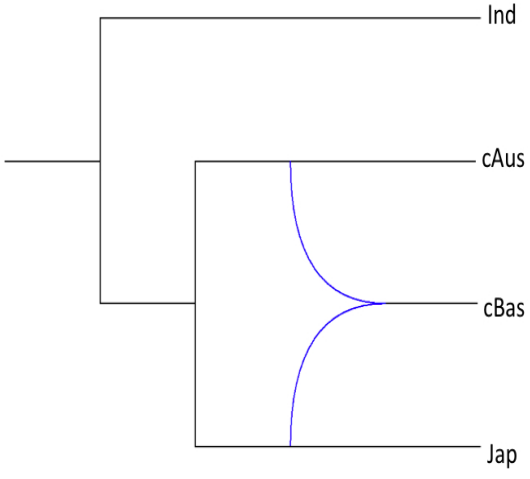
Réseau 5



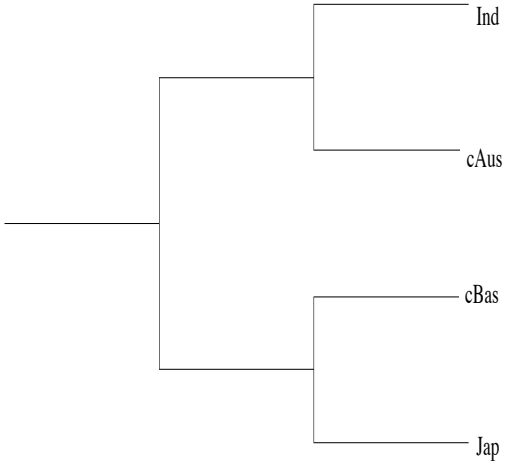
Réseau 6



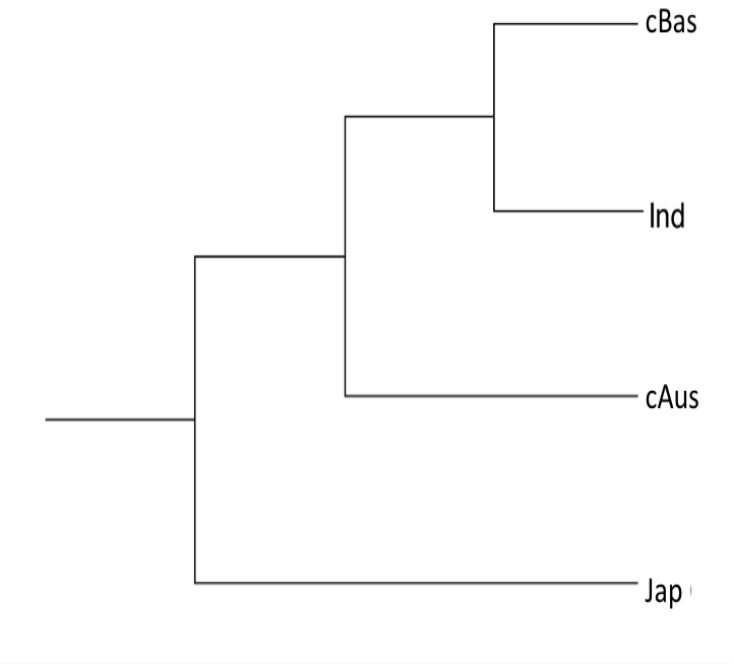
Réseau 7



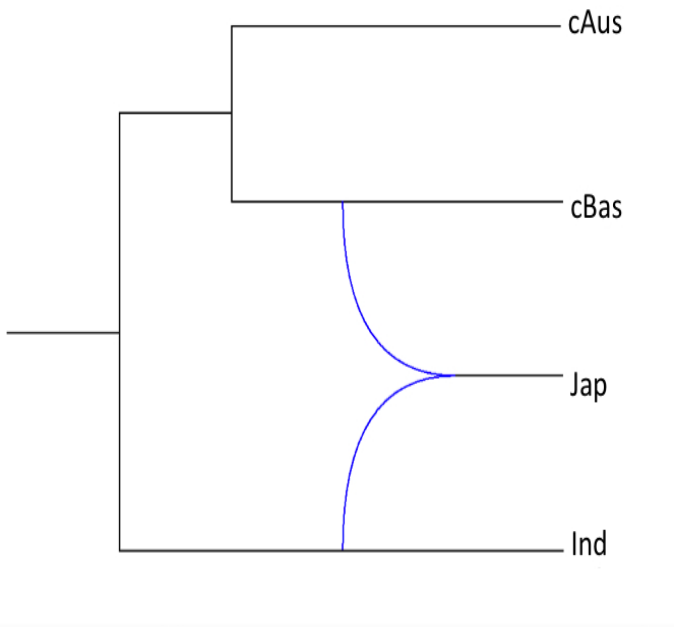
Réseau 8 (conforme aux attentes)



Les dix réseaux riz étudiés



Réseau 9



Réseau 10 (peu conforme aux attentes)

Calcul de vraisemblance pénalisée (AIC ou BIC) sur nos 10 reseaux

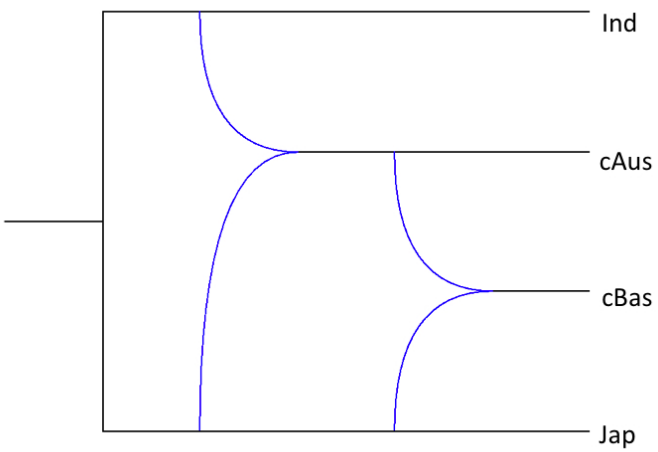
		Data set 1			Data set 2		
Network	Nb Parameters	Log Likelihood	AIC	BIC	Log Likelihood	AIC	BIC
1	**29**	-41699.09	83456.18 (2)	83670.57 (3)	-38860.03	77778.07 (1)	77992.45 (3)
7	**22**	-41724.92	83493.84 (3)	83656.48 (2)	-38878.57	77801.14 (4)	77963.78 (2)
4	**22**	-41669.73	83383.46 (1)	83546.10 (1)	-39036.92	78117.84 (6)	78280.48 (6)
2	**29**	-41755.40	83568.81 (4)	83783.19 (4)	-38867.12	77792.24 (2)	78006.63 (4)
10	**22**	-41936.12	83916.24 (7)	84078.88 (7)	-38878.48	77800.96 (3)	77963.60 (1)
5	**22**	-41921.53	83887.06 (6)	84049.70 (6)	-38964.73	77973.46 (5)	78136.10 (5)
6	**29**	-41874.35	83806.70 (5)	84021.09 (5)	-39040.20	78138.40 (7)	78352.79 (7)
3	**22**	-44264.72	88573.44 (8)	88736.08 (8)	-41197.22	82438.44 (9)	82601.08 (9)
8	**15**	-45586.44	91202.87 (10)	91313.76 (10)	-39209.63	78449.26 (8)	78560.15 (8)
9	**15**	-45041.92	90113.84 (9)	90224.73 (9)	-41498.45	83026.9 (10)	83137.79 (10)

Calcul de vraisemblance pénalisée (AIC ou BIC) sur nos 10 reseaux

		Data set 1			Data set 2		
Network	Nb Parameters	Log Likelihood	AIC	BIC	Log Likelihood	AIC	BIC
1	29	-41347.93	82753.86 (1)	82968.25 (1)	-38493.86	77045.72 (1)	77260.11 (1)
7	22	-41387.08	82818.16 (2)	82980.80 (2)	-38531.24	77106.48 (2)	77269.12 (2)
2	29	-41464.70	82987.39 (3)	83201.78 (3)	-38543.21	77144.40 (4)	77358.79 (4)
10	22	-41625.53	83295.06 (6)	83457.70 (6)	-38546.30	77136.60 (3)	77299.24 (3)
4	22	-41597.64	83239.28 (5)	83401.92 (5)	-38686.75	77417.50 (6)	77580.14 (6)
6	29	-41542.08	83142.16 (4)	83356.55 (4)	-38690.65	77439.30 (7)	77653.69 (7)
5	22	-41669.71	83383.42 (7)	83546.06 (7)	-38658.04	77360.08 (5)	77522.72 (5)
8	15	-42353.95	84737.90 (8)	84848.79 (8)	-38940.40	77910.80 (8)	78021.69 (8)
3	22	-44009.53	88063.06 (9)	88225.70 (9)	-40851.45	81746.90 (9)	81909.54 (9)
9	15	-44790.69	89611.38 (10)	89722.27 (10)	-41141.28	82312.56 (10)	82423.45 (10)

Rice evolution scenario inferred by SNAPPNET

- Average ranking : Net 1 > Net 7 > Net 2 > Net 4 > Net 10 > Net 5 > Net 6 > Net 8 > Net 3 > Net 9
- The proportion of Japonica present in cBasmati genome is approximately 73% (agreement across analyses on networks 1, 7, 2 and 4)
- Network 1 is the rice evolution scenario inferred by SNAPPNET

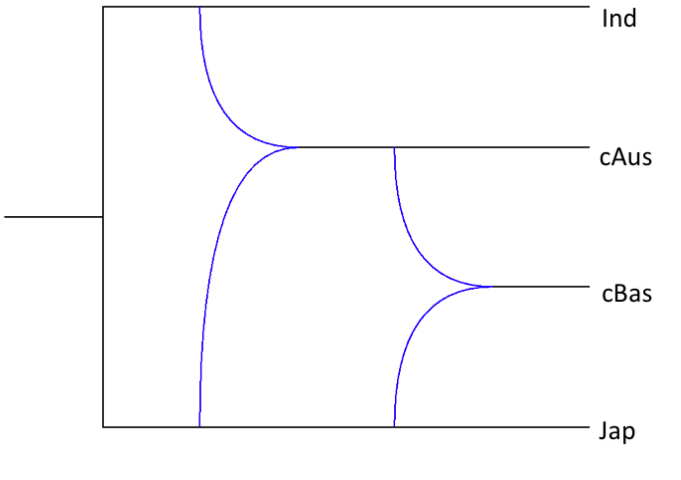


$_{c}$*Aus* : result of an early(old) combination between the *Ind* and the *Jap* lineages

$_{c}$*Bas* : a later combination between the $_{c}$*Aus* and the *Jap* lineages

New scheme that appears compatible within the simpler schemes that have been proposed so far (e.g. *Choi et al, Genome Biology 2020*)

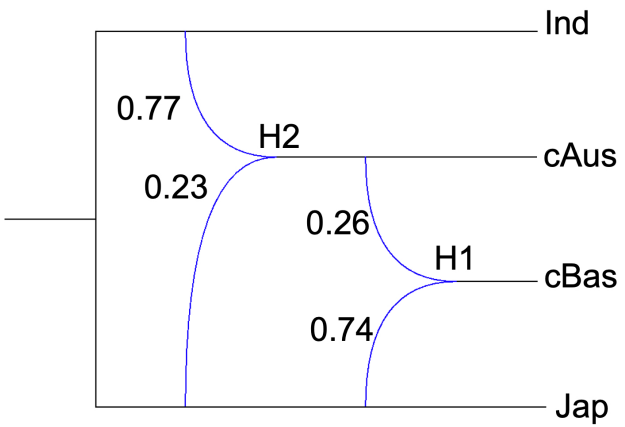
Extra informations on the oldest reticulation



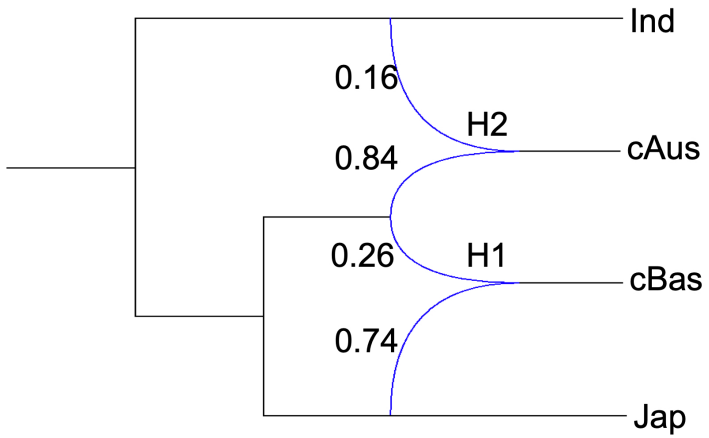
Network 1

- Contributions to cAus : 77% from Indica and 23% from Japonica
- Compatible with the geographical distribution of the three groups :
  - cAus in the Northeastern part of the Indian subcontinent, contacts between South Asia and East Asia because of the southern silk road (Chakraborty et al., 2019)
- Possible explanation : wide cross compatibility of the cAus varieties which tends to make them fertile in crosses with both Indica and Japonica (Morinaga et al., 1955, 1958).

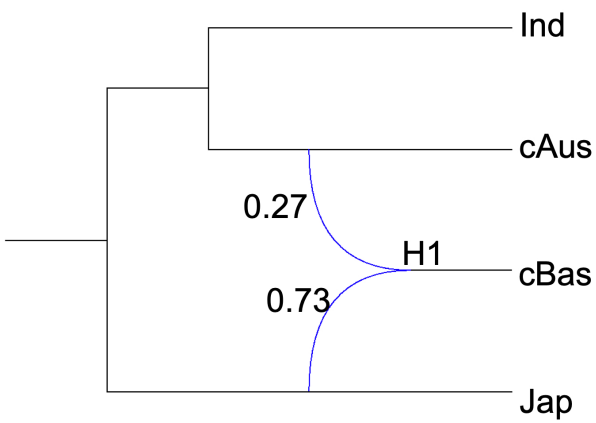
Probabilités d'hybridation moyennées sur les 4 jeux de données



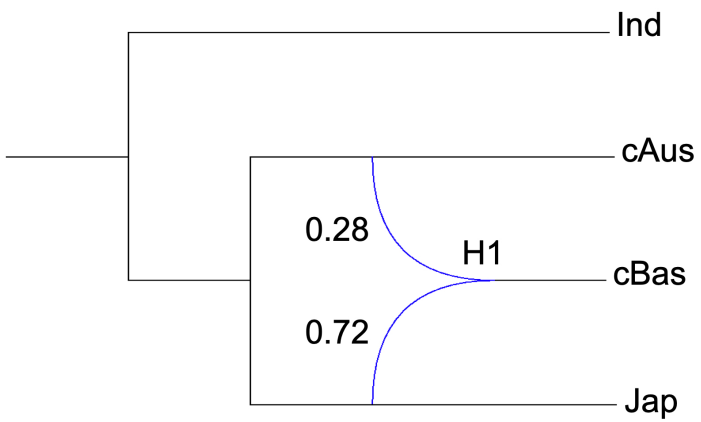
Réseau 1



Réseau 2



Réseau 4

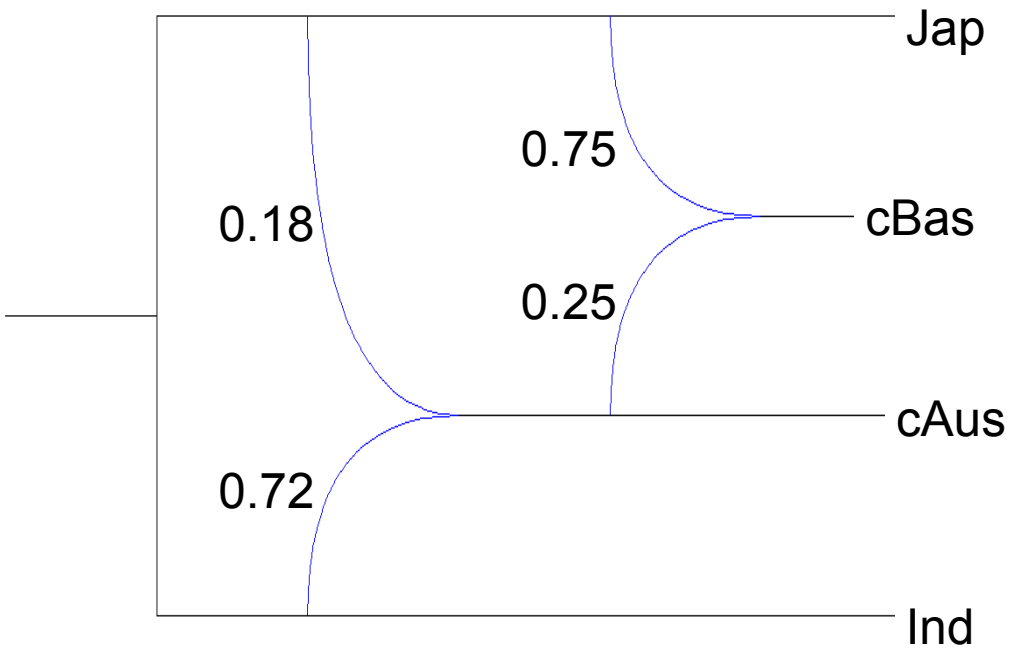


Réseau 7

SNAPPNET en version MCMC

Dataset 1 :

- 1 174 000 itérations effectuees
- Effective Sample Size = 130
- SNAPPNET bloqué sur le Réseau 1



Dernier réseau échantillonné

mais une autre chaine est sur un autre réseau

Plan

Conclusion

- Jusqu'alors travail sur le riz $\rightarrow$ autre plante d'intérêt ? Vigne, Sorgho, Banane, Citrus, Caféier, Canne à sucre ...
- SNAPPNET disponible sur https ://github.com/rabier/MySnappNet
- Papier disponible sur biorxiv : "On the inference of complex phylogenetic networks by Markov Chain Monte-Carlo"
- Le tri de lignées incomplet (ILS) aide à la reconstruction de reseau (sans ILS, incapacité à retrouver le réseau C)
- Les gains en vitesse de calcul permettent de considérer des scénarios évolutifs plus complexes mais convergence du MCMC assez lente
- Package BEAST coupled MCMC (Mueller and Bouckaert, 2020), afin de sortir des optimums locaux grace à des chaines chaudes et froides
- Le Multispecies Network Coalescent suppose l'indépendance entre les sites $\Rightarrow$ étude de reseaux à l'intérieur de réseaux pour modéliser la recombinaison proprement (cf. Degnan, 2018)
- Etudier plus en détail le réseau 1 avec les 2 hybridations l'une sous l'autre
- Considerer plus de variétes par des approches ABC random forest (Pudlo et al, 2015)

Remerciements

Céline Scornavacca

Vincent Berry
Fabio Pardi

Jean-Christophe Glaszmann
João D. Santos

isem
Institut des Sciences de l'Evolution-Montpellier

LIRMM

agap

agropolis fondation
flagship project

Jean-Michel Marin

Angélique D'Hont
Manuel Ruiz

bioRχiv
THE PREPRINT SERVER FOR BIOLOGY