Apport des approches phylogénétiques pour expliquer l'origine des génomes mosaïques, exemple chez le riz

Charles-Elie Rabier

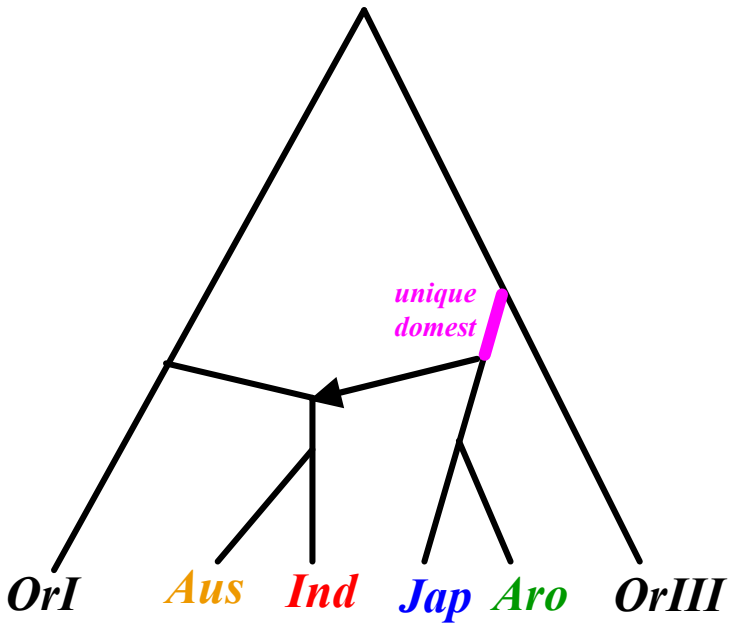
Vincent Berry, Fabio Pardi et Céline Scornavacca

ISEM, Institut des Sciences de l'Evolution de Montpellier
LIRMM, Laboratoire d'informatique, de Robotique et de Microélectronique
Genome Harvest

Collaborations : Jean-Christophe Glaszmann (CIRAD), Joao Santos (CIRAD)

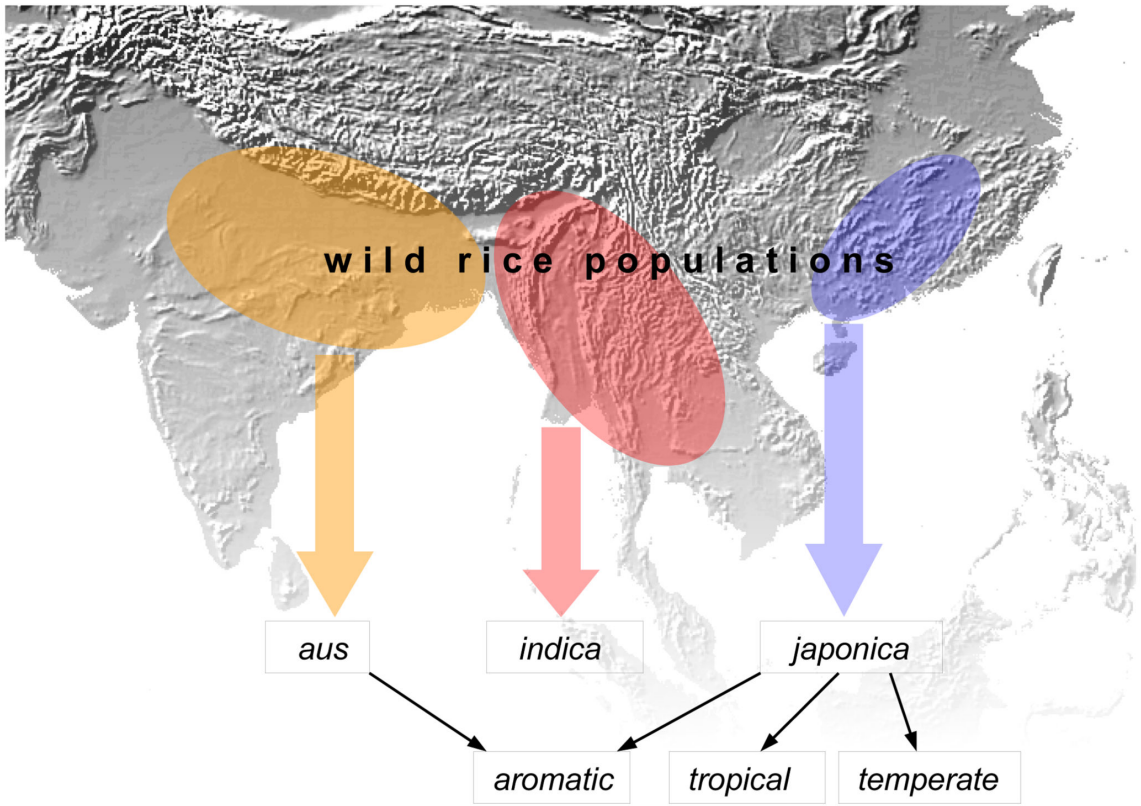
Quelques thèses sur la domestication

- 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

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



Réseaux phylogénétiques

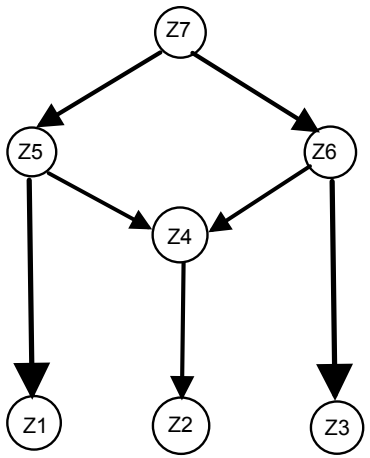
Les réseaux phylogénétiques sont des DAG qui vont nous permettre de détecter des :

- hybridations
- introgressions
- transferts horizontaux

Quelques points importants :

- Longueur d'une arête = temps d'évolution
- Dépendance entre noeuds (probabilités conditionnelles ?)
- 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

Un réseau Bayesien ? Non, mais ...



- Noeud 4 : noeud de réticulation
- Noeud 1, 2 et 3 : feuilles du réseau
- Z_k : v.a. correspondant au noeud k
- Loi jointe :

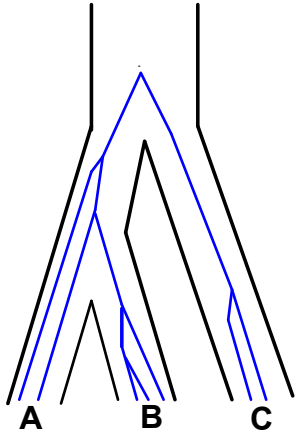
$$\begin{aligned}&\mathbb{P}(Z_1, Z_2, Z_3, Z_4, Z_5, Z_6, Z_7)\\ &= \mathbb{P}(Z_2 \mid Z_4)\,\mathbb{P}(Z_1, Z_3, Z_4 \mid Z_5, Z_6)\\ &\times \mathbb{P}(Z_5, Z_6 \mid Z_7)\,\mathbb{P}(Z_7)\end{aligned}$$

- Z_1, Z_2, Z_3 ne sont pas de même nature que Z_4, Z_5, Z_6, Z_7
- 2 types d'allèles (rouge/vert)
- N_k : v.a. pour le nombre de lignées au noeud k
- R_k : v.a. pour le nombre d'allèles rouge au noeud k
- $\forall k \in \{4,5,6,7\}$, $Z_k = (N_k, R_k)$
- $\forall k \in \{1,2,3\}$, $Z_k = R_k$
- N_1, N_2 et N_3 sont connus ! ! !
- Data=(Z_1, Z_2, Z_3)

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

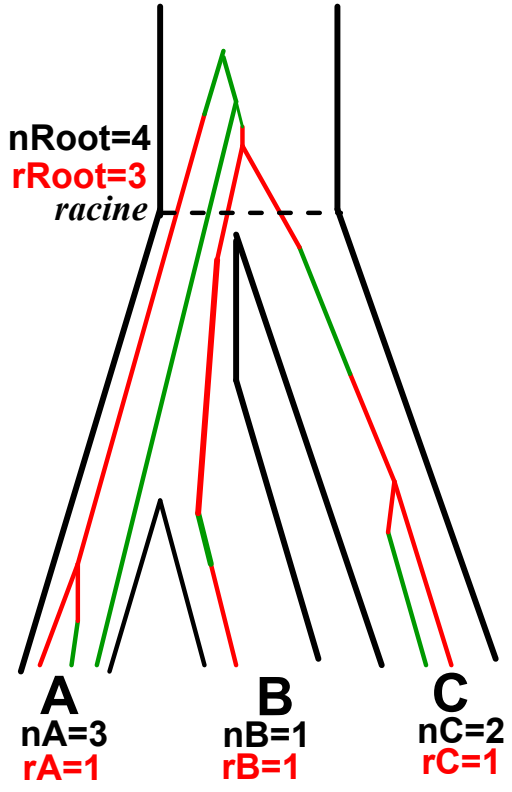
- Marqueurs biallélliques (SNPs) indépendants sachant l'arbre d'espèces

- 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)



Les mutations interviennent au cours du temps

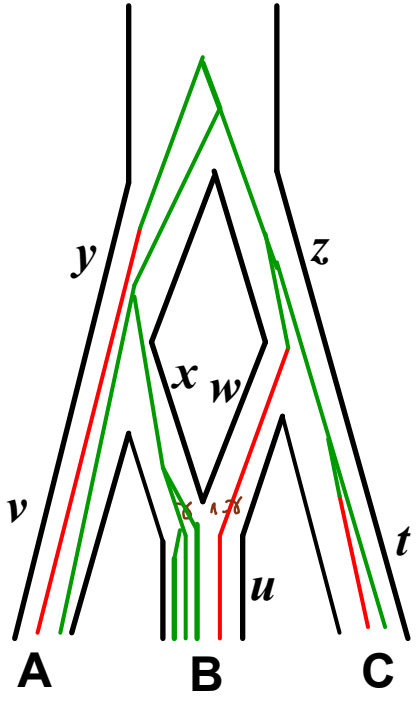
- Modélisation des données au SNP (forward)
  - mutation (rouge $\leftrightarrow$ vert) : modèle markovien évoluant le long des branches de l'arbre de locus
  - u : taux de mutation rouge $\rightarrow$ vert
  - v : taux de mutation vert $\rightarrow$ rouge



- 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

Cadre d'un réseau phylogénétique

- Modélisation de l'arbre de locus (backward) :
  - multispecies coalescent
  - modèle de Nakhleh au niveau du noeud de réticulation
- 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)
- X_i : données pour le SNP i
- G_i : arbre de locus pour le SNP i
- m SNPs

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

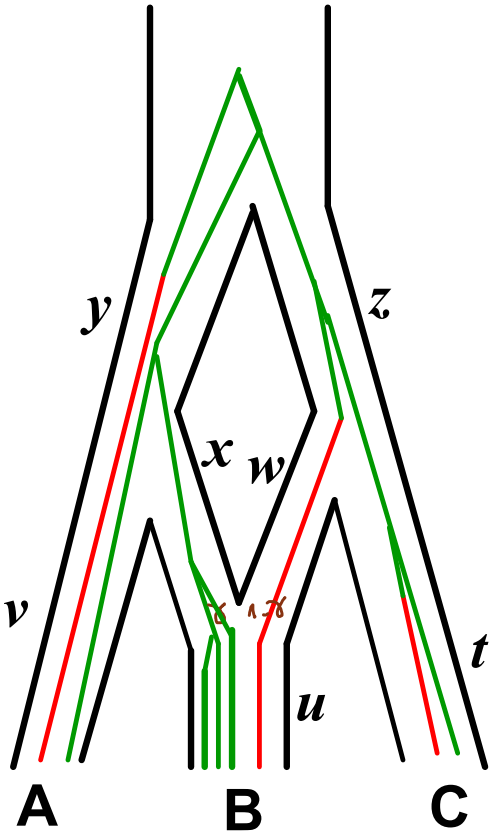
SNAPPNet intègre sur tous les arbres de locus (extension de SNAPP, Bryant et al. MBE 2012), à l'aide d'un nouvel algorithme de parcours du réseau

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

$\Rightarrow$ 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 ...

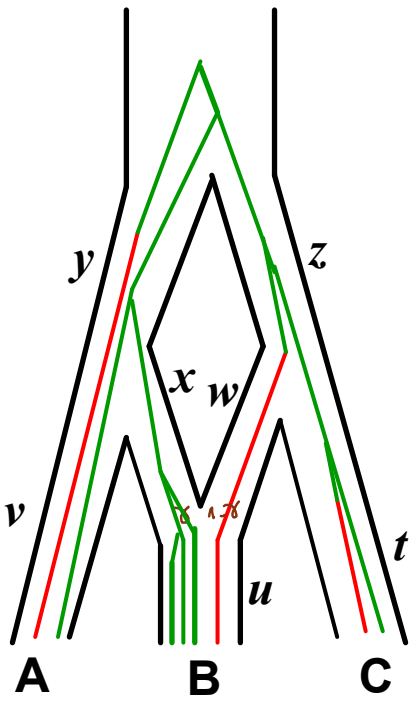
$Data_{zT}$ et $Data_{yT}$ comprennent les allèles rouges et verts de l'espèce hybride

Calcul de la vraisemblance dans un réseau

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

- $\mathbb{P}(r_{root} = j \mid n_{root} = i)$ 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 :
    $\mathbb{P}(r_{root} = j \mid n_{root} = i) = C_i^j B(j + \theta, i - j + \theta)/B(\theta, \theta)$
- $\mathbb{P}(\mathit{Data} \mid n_{root} = i, r_{root} = j)\, \mathbb{P}(n_{root} = i)$ 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) $\mathbb{P}\left(Data_{yB}\, Data_{wT} \mid n_{yB}, r_{yB}, n_{wT}, r_{wT}\right)$

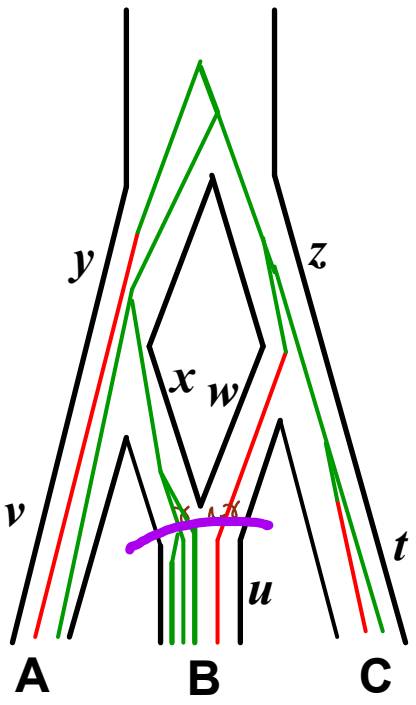
(8) $\mathbb{P}\left(Data_{yB}\, Data_{zB} \mid n_{yB}, r_{yB}, n_{zB}, r_{zB}\right)$

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

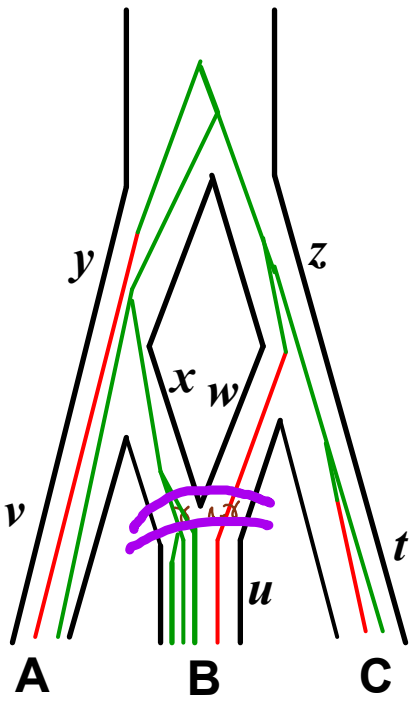
(8) $\mathbb{P}\left(Data_{yB}\, Data_{zB} \mid n_{yB}, r_{yB}, n_{zB}, r_{zB}\right)$

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

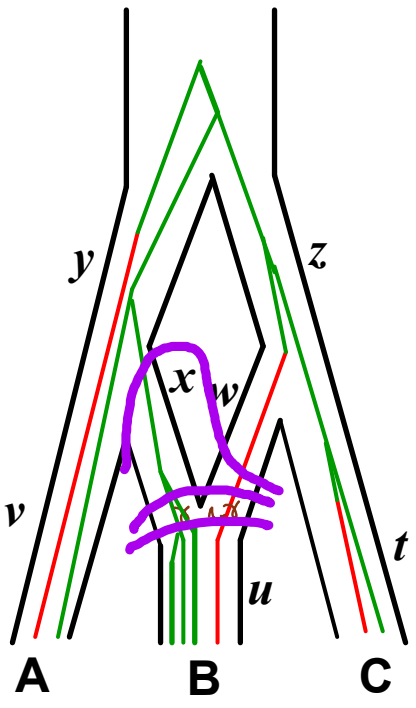
(8) $\mathbb{P}\left(Data_{yB}\, Data_{zB} \mid n_{yB}, r_{yB}, n_{zB}, r_{zB}\right)$

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

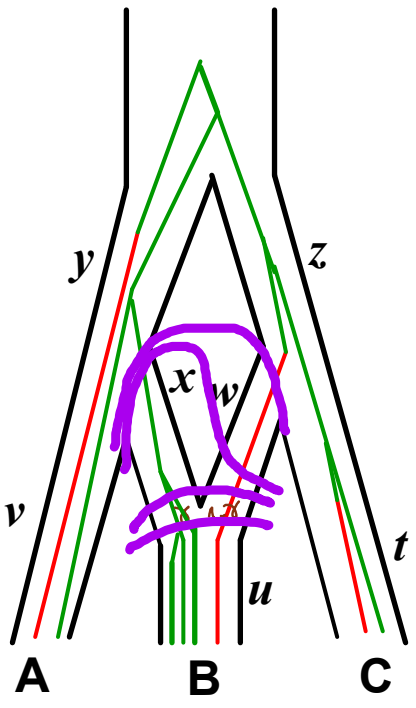
(8) $\mathbb{P}\left(Data_{yB}\, Data_{zB} \mid n_{yB}, r_{yB}, n_{zB}, r_{zB}\right)$

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

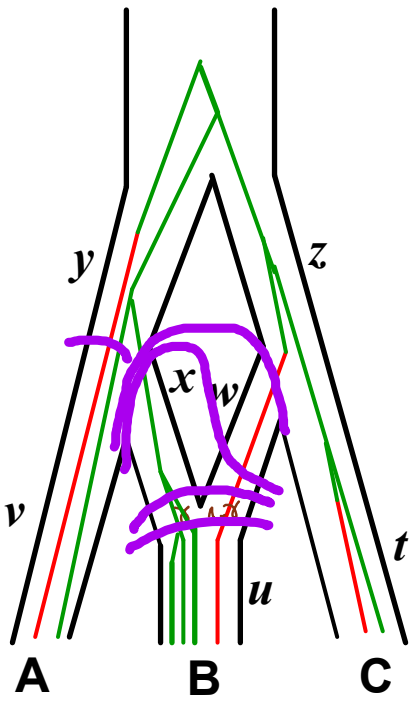
(8) $\mathbb{P}\left(Data_{yB}\, Data_{zB} \mid n_{yB}, r_{yB}, n_{zB}, r_{zB}\right)$

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

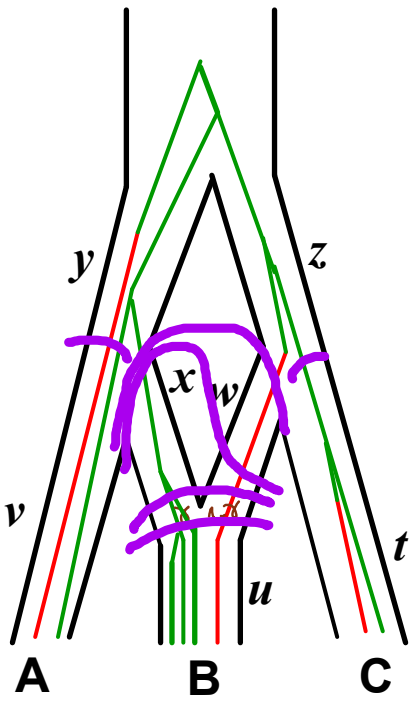
(8) $\mathbb{P}\left(Data_{yB}\, Data_{zB} \mid n_{yB}, r_{yB}, n_{zB}, r_{zB}\right)$

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

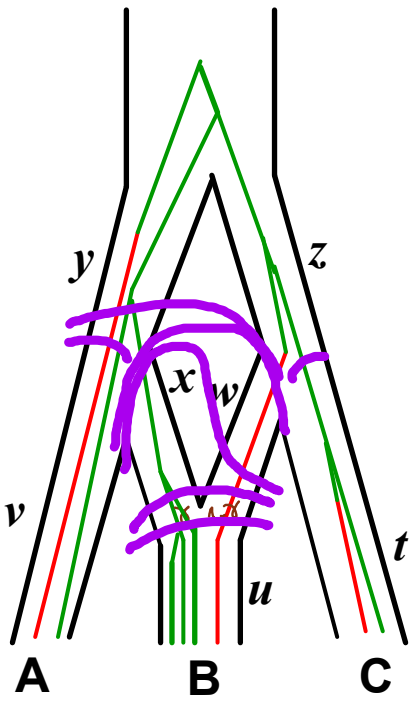
(8) $\mathbb{P}\left(Data_{yB}\, Data_{zB} \mid n_{yB}, r_{yB}, n_{zB}, r_{zB}\right)$

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

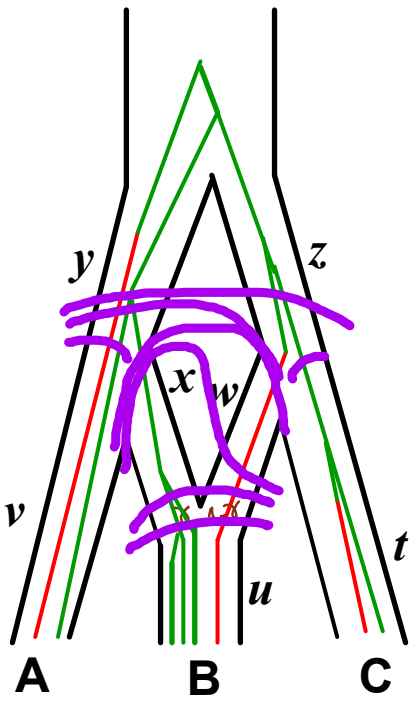
(8) $\mathbb{P}\left(Data_{yB}\, Data_{zB} \mid n_{yB}, r_{yB}, n_{zB}, r_{zB}\right)$

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

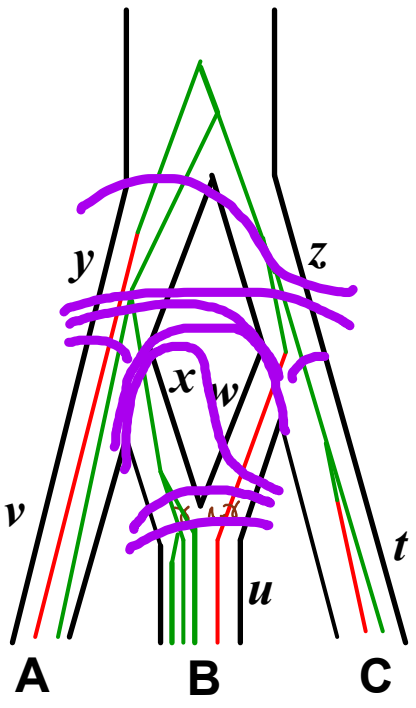
(8) $\mathbb{P}\left(Data_{yB}\, Data_{zB} \mid n_{yB}, r_{yB}, n_{zB}, r_{zB}\right)$

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

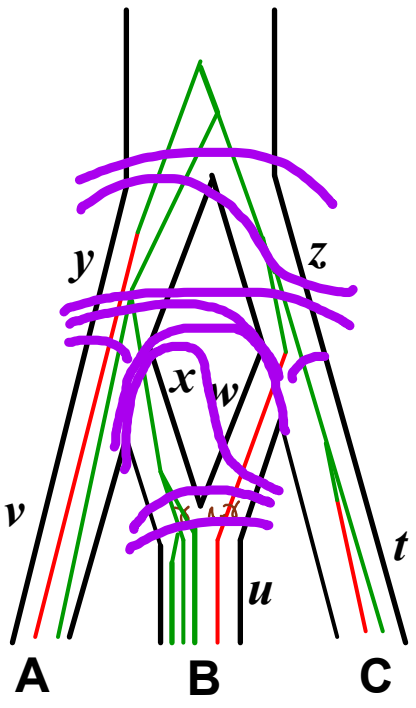
(8) $\mathbb{P}\left(Data_{yB}\, Data_{zB} \mid n_{yB}, r_{yB}, n_{zB}, r_{zB}\right)$

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

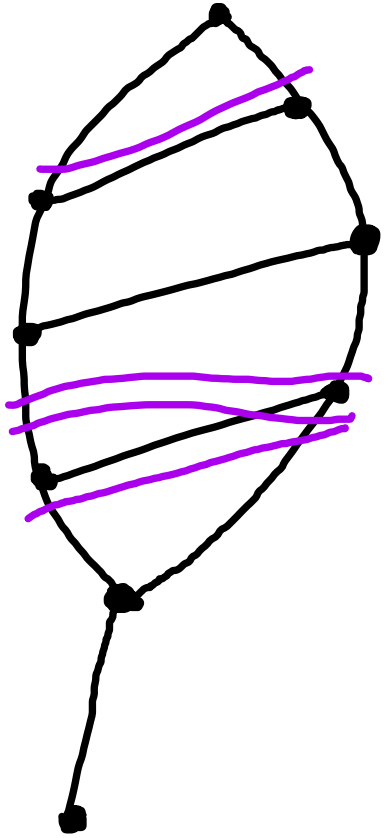
(9) $\mathbb{P}\left(Data_{yT}\, Data_{zB} \mid n_{yT}, r_{yT}, n_{zB}, r_{zB}\right)$

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

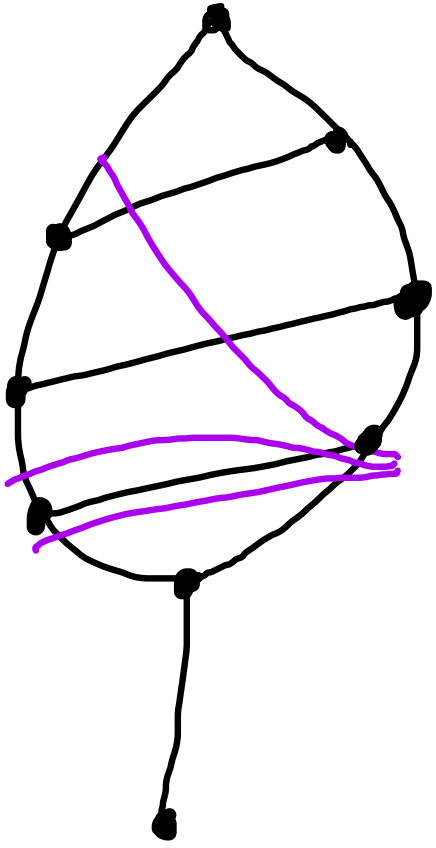
(11) $\mathbb{P}\left(Data \mid n_{root}, r_{root}\right)$

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

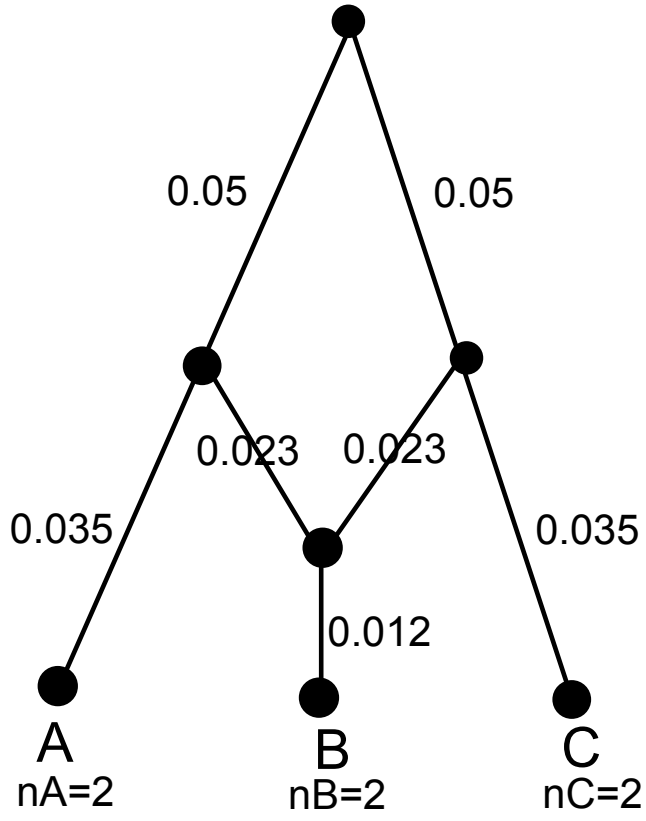
Un parcours intéressant



Un parcours à éviter



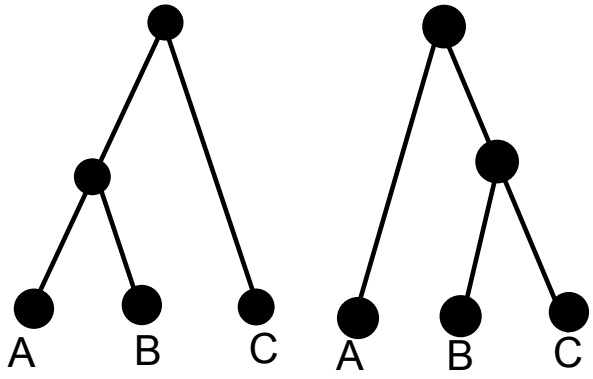
Un exemple de réseau étudié par simulation



- 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 $\mathbb{E}(T) = 0.005/2 = 0.025$

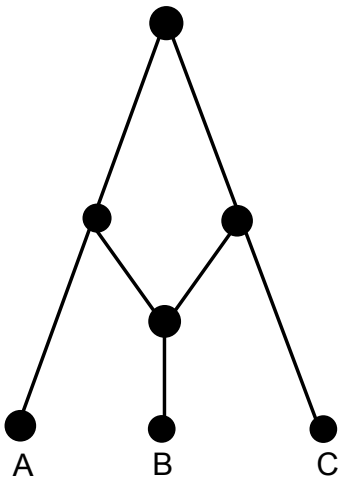
Réseaux échantillonnés par MCMC

- 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, $\theta = 0.05$

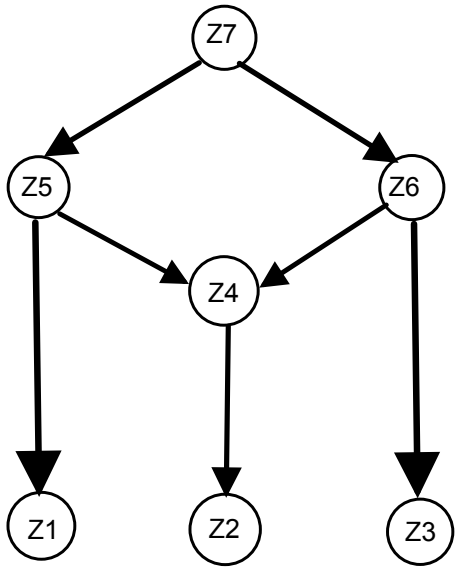


100%

Conclusion

- L'inférence de réseau est un sujet compétitif : Tanja Stadler (ETH Zurich), Luay Nakhleh (Rice University, USA). Notre approche devrait être plus performante sur des réseaux aux nombreuses hybridations
- Jusqu'alors travail sur le riz $\rightarrow$ autre plante d'intérêt ? Genome Harvest : Banane, Citrus, Caféier, Riz, Tomate, Canne à sucre ...
- SNAPP disponible sur http ://snapp.otago.ac.nz
- Nous testons SNAPPNet sur données simulées et réelles ...

Autre représentation



Au final, en (11), on a calculé :

$$\mathbb{P}(Z_1, Z_2, Z_3 \mid Z_7)$$
$$= \sum_{Z_5} \sum_{Z_6} \sum_{Z_4} \mathbb{P}(Z_2 \mid Z_4)\, \mathbb{P}(Z_1, Z_3, Z_4 \mid Z_5, Z_6)$$
$$\times\, \mathbb{P}(Z_5, Z_6 \mid Z_7)$$

Chromosome Painting

Zhang et al. (MBE, 2017) : Equipe de Tanja Stadler (ETH Zurich)

Statistique Bayésienne dans le cadre d'un réseau

- N : réseau phylogénétique
- X_i : données pour le SNP i
- G_i : arbre de locus pour le SNP i
- m SNPs

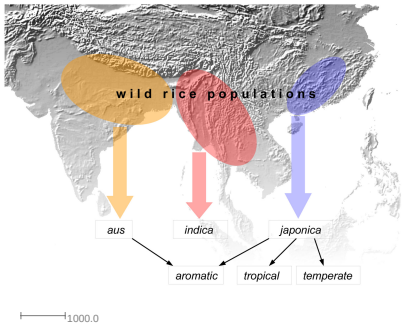
$$\mathbb{P}(N, G_1, \ldots, G_m | X_1, \ldots, X_m) \propto \left(\prod_{i=1}^{m} \mathbb{P}(X_i | G_i) \mathbb{P}(G_i | N) \right) \mathbb{P}(N)$$

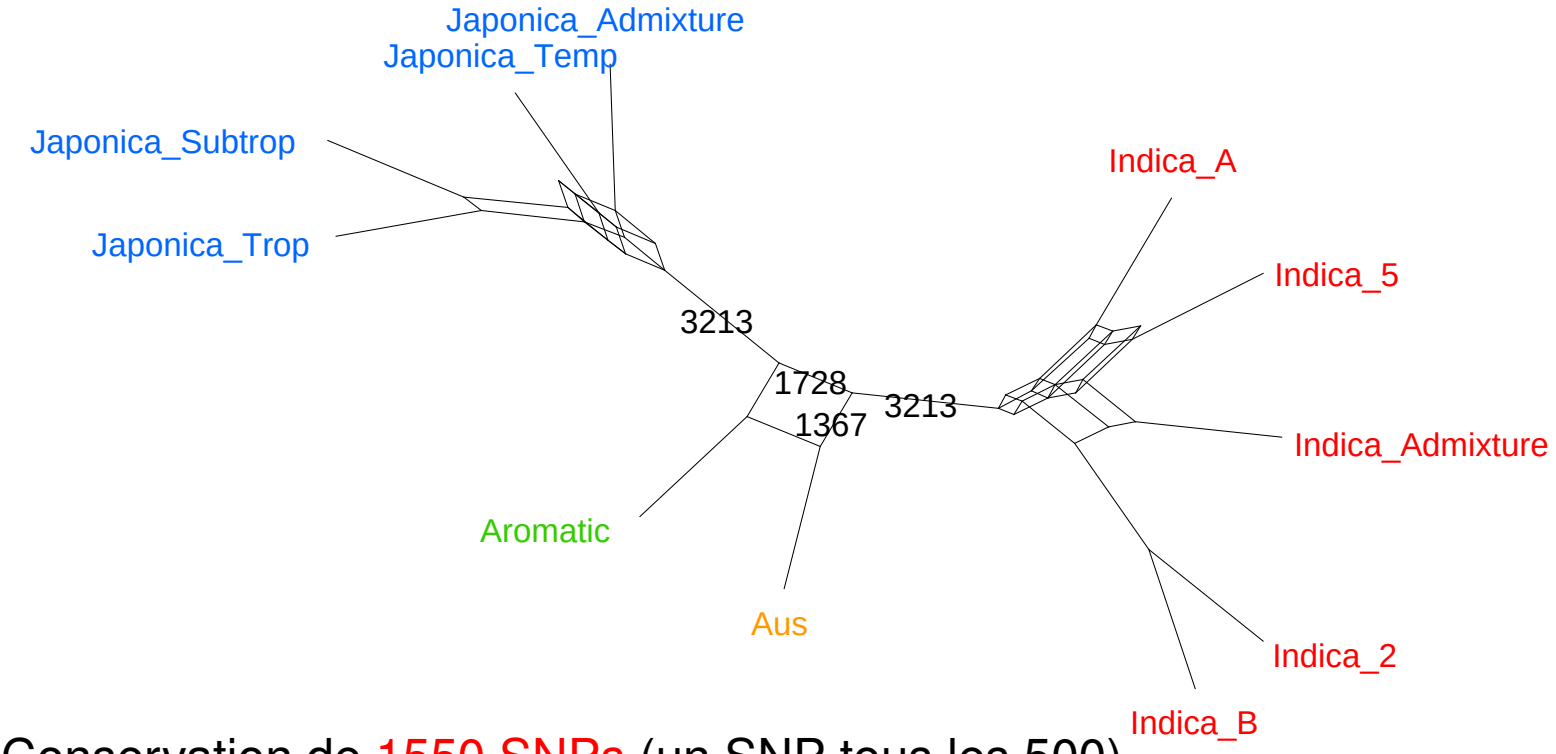
Calcul de l'a priori P(N) par un processus de naissance/hybridation

$\Rightarrow$ Markov Chain Monte Carlo afin d'estimer la distribution à posteriori de $\mathbb{P}(N, G_1, \ldots, G_m | X_1, \ldots, X_m)$.

Echantillonnage de réseaux et d'arbres de locus $\rightarrow$ chromosome painting

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



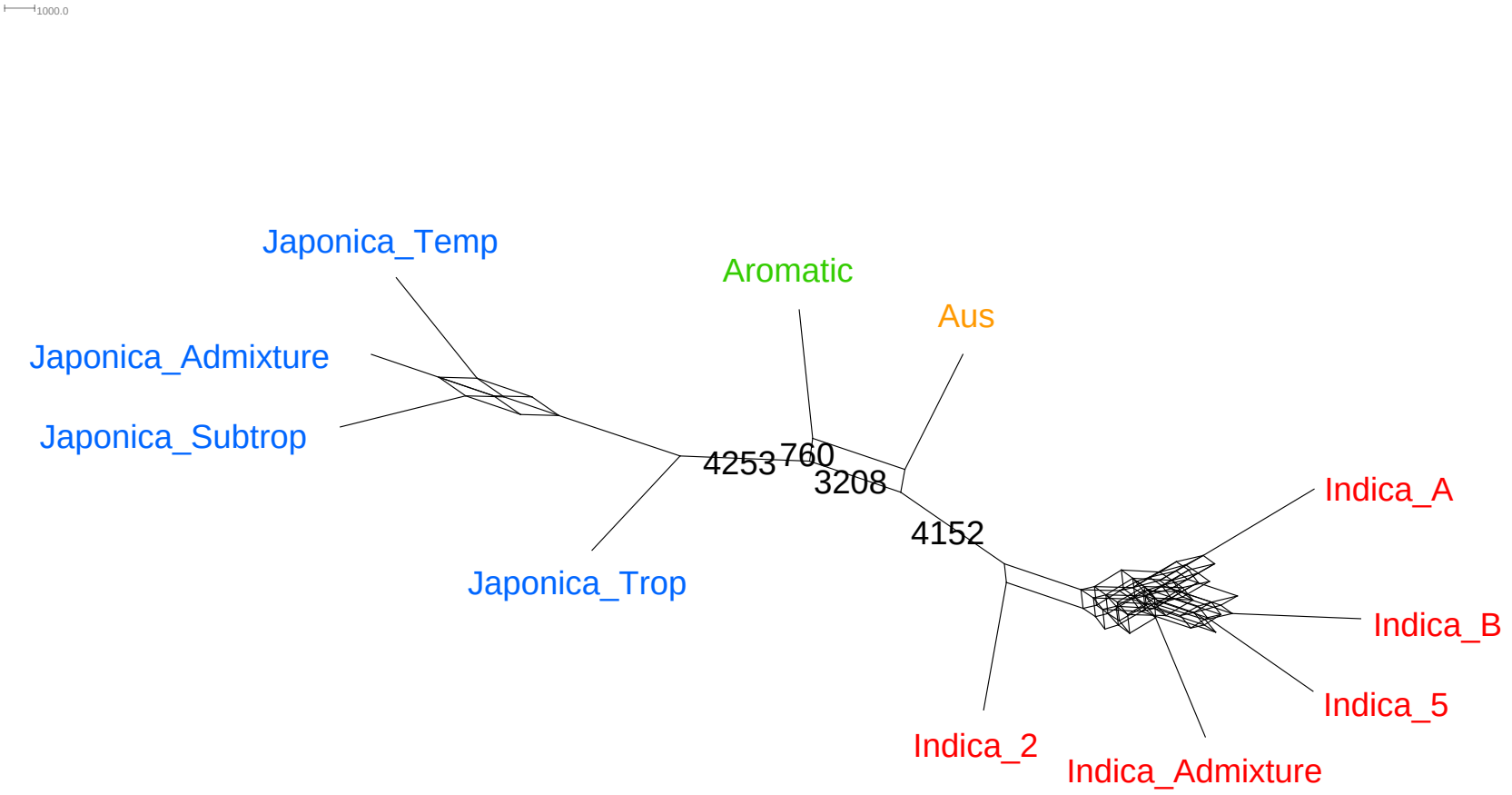


Conservation de 1550 SNPs (un SNP tous les 500)

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)



La statistique Bayésienne dans SNAPP

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

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

SNAPP intègre sur tous les arbres de locus

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

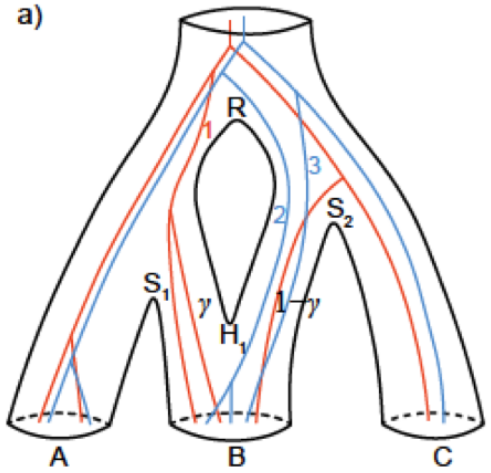
$\Rightarrow$ Markov Chain Monte Carlo (MCMC) afin d'estimer la distribution à posteriori de $\mathbb{P}(S|X_1,\ldots,X_m)$

Implémenté dans BEAST

Simulateur basé sur un réseau (Genome Harvest)

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



- Snapp est fortement attiré par un scénario sous-jacent au réseau

Calcul de vraisemblance dans un arbre (1)

$$
\begin{aligned}
& \mathbb{P}(\mathit{Data}) \\
&= \sum_i \sum_j \mathbb{P}(\mathit{Data} \mid \mathit{Count}, n_{root} = i, r_{root} = j)\, \mathbb{P}(n_{root} = i, r_{root} = j \mid \mathit{Count}) \\
&= \sum_i \sum_j \mathbb{P}(\mathit{Data} \mid \mathit{Count}, n_{root} = i, r_{root} = j)\, \mathbb{P}(r_{root} = j \mid n_{root} = i) \\
&\qquad \times \mathbb{P}(n_{root} = i \mid \mathit{Count})
\end{aligned}
$$

Calcul de vraisemblance dans un arbre (2)

- $\mathbb{P}(n_{root} = i \mid Count)$ 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...

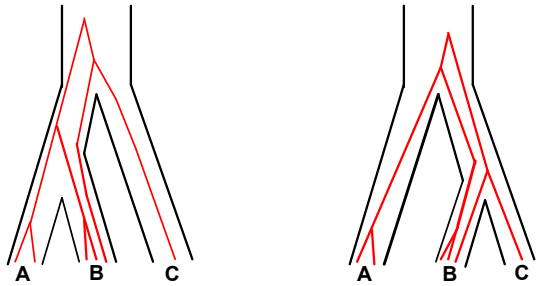
Notre approche méthodologique

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

Nos pistes :

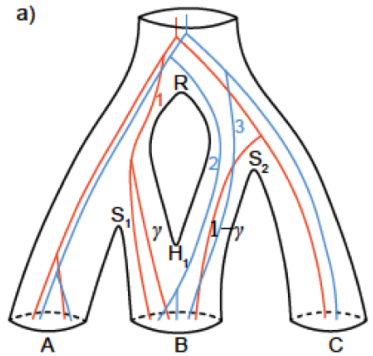
1. Inférence d'arbres d'espèces + arbres résumés en réseaux phylogénétiques

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



2. Inférence directe de réseaux

Extension de SNAPP aux réseaux



Quelques thèses sur la domestication

- 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