Modèles et outils à l'interface entre la génétique des populations et de la phylogénétique pour la phylogénomique du 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

Plan

1. Introduction
2. Inférence d'arbres d'espèces + arbres résumés en réseaux
   - Données réelles de riz
   - Données simulées
3. Inférence directe de réseaux
   - Nouvelle méthode (en cours)
   - Nos concurrents
4. Conclusion

Plan

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

- 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 la chine, générant indica
- Civan et al. (Nature Plants, 2015) : indica, japonica et aus domestiqués séparément dans différentes parties d'Asie
- Choi et al. (MBE, 2017) : Différentes sous espèces de riz ont différentes origines, mais une seule domestication (japonica), et introgression par hybridation de japonica et proto-indica et proto-aus, générant indica et aus

Méthodes utilisées = 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 apport méthodologique

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

Nos pistes :

1. Arbre de gènes par blocs (Parcimonie) + Phylonet (Wen,Yu and Nakhleh 2016, Plos Genetics)
2. Inférence d'arbres d'espèces + arbres résumés en réseaux
   - SNAPP (Bryant et al. 2012, MBE) + SplitsTree
3. Inférence directe de réseaux
   - Extension de SNAPP au réseau

Plan

Les données de riz

1. Données disponibles à l'état brut sur le site de l'IRRI
2. Prétraitement de Joao (données manquantes ...)
3. 3023 variétés avec 895 977 marqueurs disponibles sur le chromosome 6 (Merci Joao !)
4. 2 jeux de données proposés par JC Glazmann (core collections)
   - 20 variétés
   - 50 variétés (7 aromatic, 7 aus, 13 indica, 17 japonica, 4 indéterminés)

Hypothèses propres à SNAPP (Bryant et al. 2012, MBE)

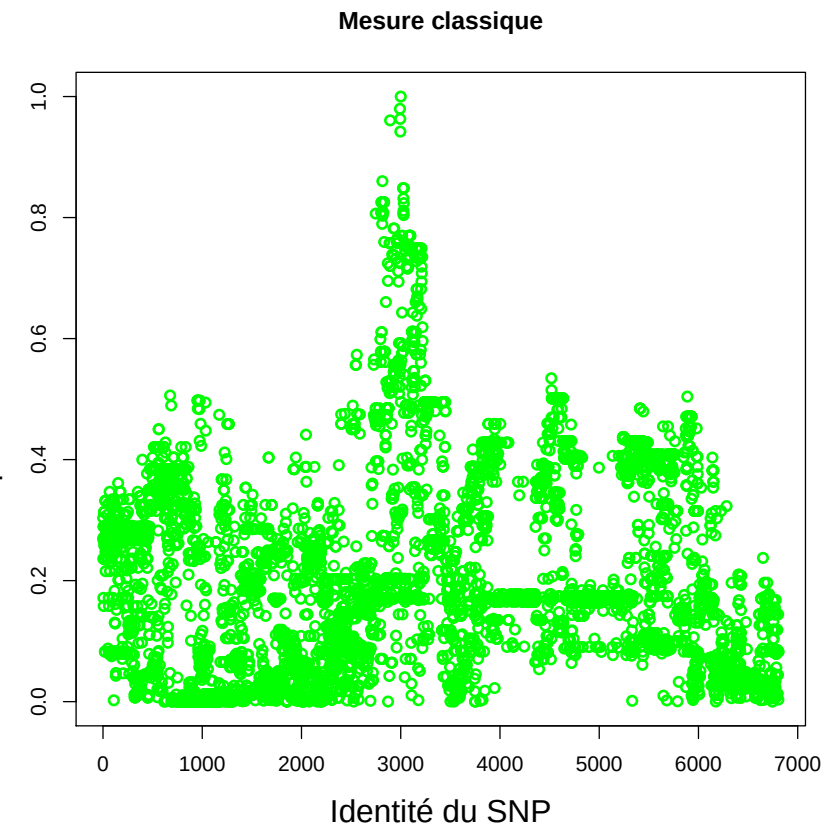
- Marqueurs bialléliques
- Généalogie de chaque marqueur indépendante sachant l'arbre d'espèces (i.e. marqueurs indépendants ou très peu de déséquilibre de liaison)

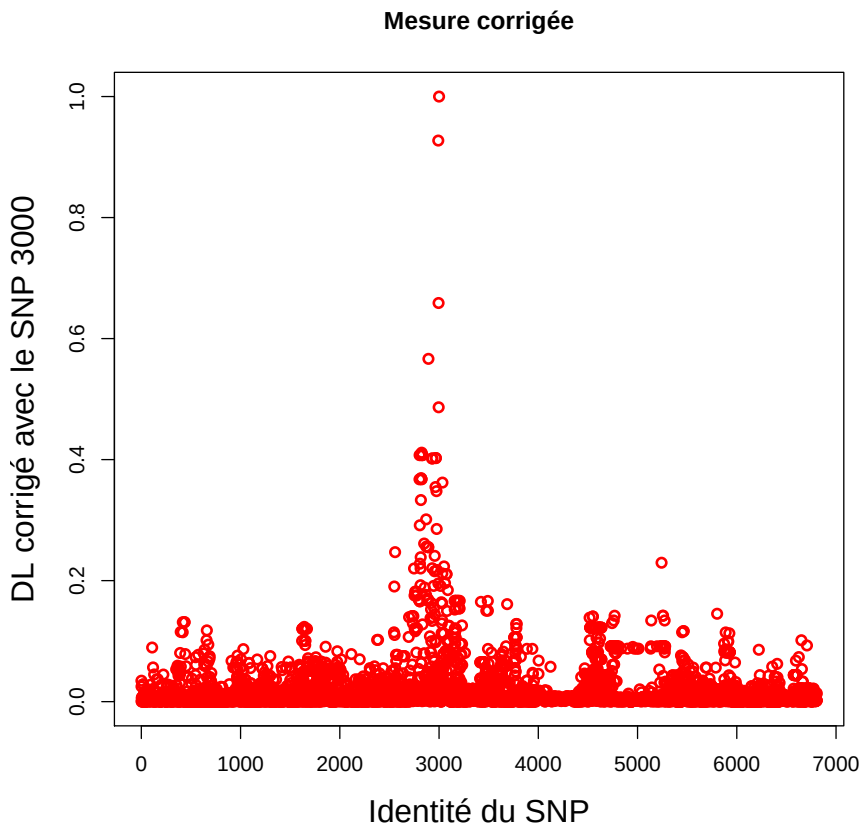
Il nous faut des SNPs suffisamment distants
le long du génome ...

Déséquilibre de liaison (tranche 100000-200000 SNPs du chromosome 6 du riz)

50 variétés, 6806 SNPs après filtrage 100000 SNPs ...

Mangin et al. (Heredity, 2012)



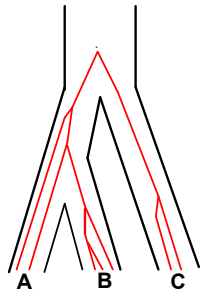


Constitutions de 4 jeux de données de manière empirique

- Conservation de 1550 SNPs (un SNP tous les 500)
  - JDD1 (1er SNP= 1er SNP du chromosome 6)
  - JDD2 (1er SNP= 50e SNP du chromosome 6)
- Conservation de 7749 SNPs (un SNP tous les 100)
  - JDD3 (1er SNP= 1er SNP du chromosome 6)
  - JDD4 (1er SNP= 50e SNP du chromosome 6)

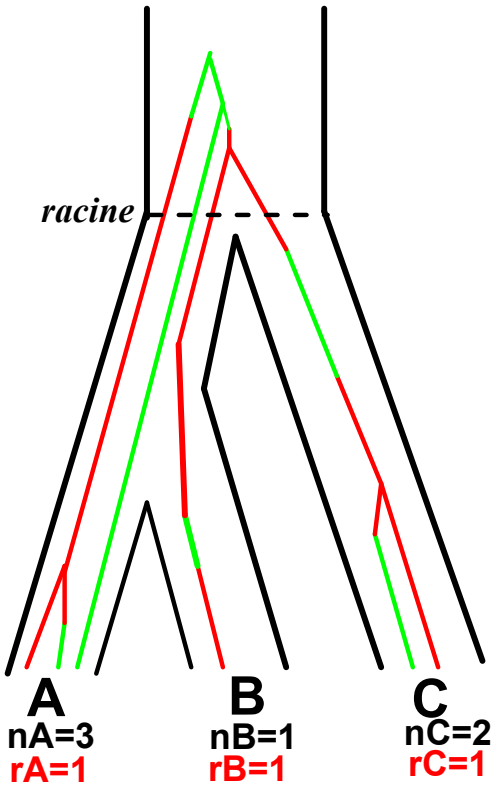
Modèle propre à SNAPP

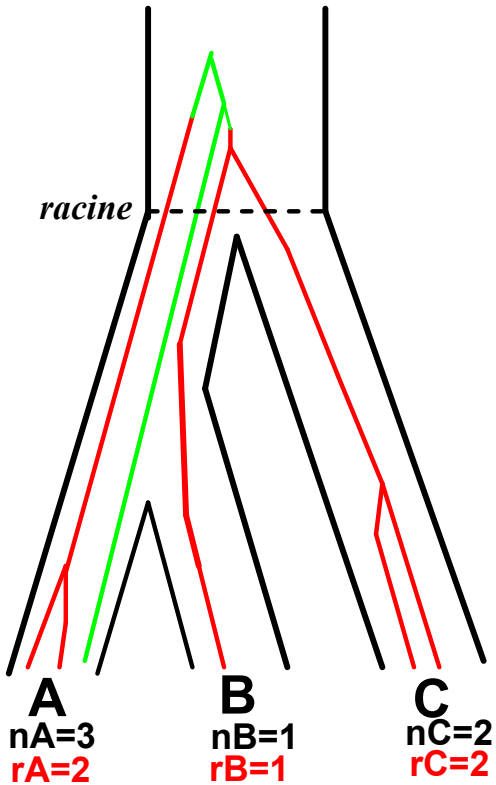
- Modélisation de l'arbre de gènes (backward)
  - Processus de coalescence évoluant à l'intérieur d'un arbre d'espèces
  - Processus autorisant la discordance entre arbres de gènes et arbres d'espèces (tri de lignées incomplet)



- Modélisation des séquences (forward)
  - Modèle markovien évoluant le long des branches de l'arbre de gènes
  - u : taux de mutation rouge $\rightarrow$ vert
  - v : taux de mutation vert $\rightarrow$ rouge
  - probabilité stationnaire à la racine (estimée sur les données)
    - rouge avec probabilité $u/u+v$
    - vert avec probabilité $v/u+v$

Mutations au cours du temps (Bryant et al.) vs Mutations au dessus de la racine (Roychoudhury et al., Genetics 2008)

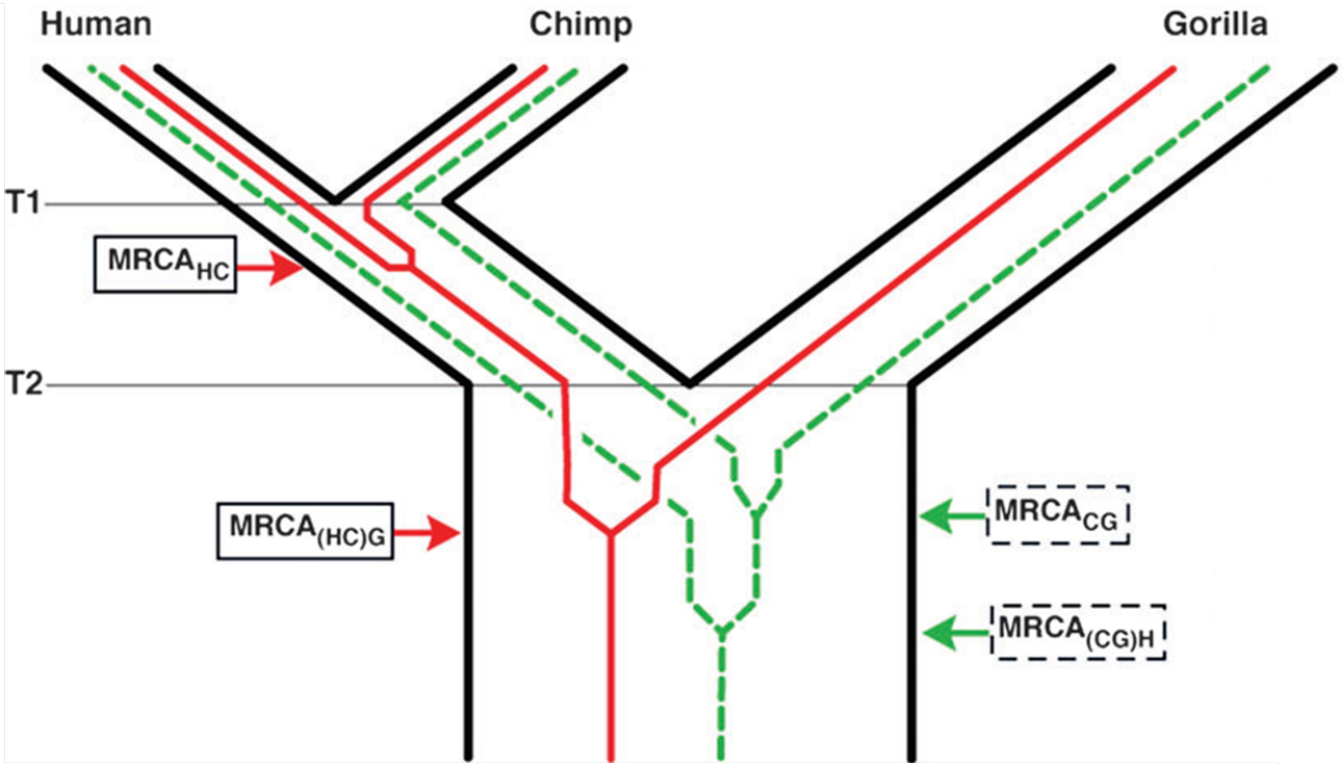




Histoires évolutives discordantes

Ebersberger *et al.* (2007)

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



Calcul de vraisemblance dans SNAPP (1)

Où se situe l'aléatoire dans le modèle ?

- Modèle de coalescence $\rightarrow$ pas d'aléatoire dans le nombre de lignées (Count) dans chaque espèce !
- L'aléatoire réside dans la répartition d'allèles rouges et verts dans chaque espèce

$$\mathbb{P}(Data)$$

$$= \sum_i \sum_j \mathbb{P}(Data \mid Count, n_{root} = i, r_{root} = j)\, \mathbb{P}(n_{root} = i, r_{root} = j \mid Count)$$

$$= \sum_i \sum_j \mathbb{P}(Data \mid Count, n_{root} = i, r_{root} = j)\, \mathbb{P}(r_{root} = j \mid n_{root} = i)$$

$$\times\, \mathbb{P}(n_{root} = i \mid Count)$$

Calcul de vraisemblance dans SNAPP (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...

La statistique Bayésienne dans SNAPP

- S : arbre d'espèces
- X_i : alignements pour le SNP i
- G_i : arbre de gènes pour le SNP i
- m SNPs

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

SNAPP intègre sur tous les arbres de gènes (algorithme + probabilités inspirées de Griffiths et Tavaré, 1997)

Calcul de l'a priori P(S) par le processus de Yule (processus de naissances)

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

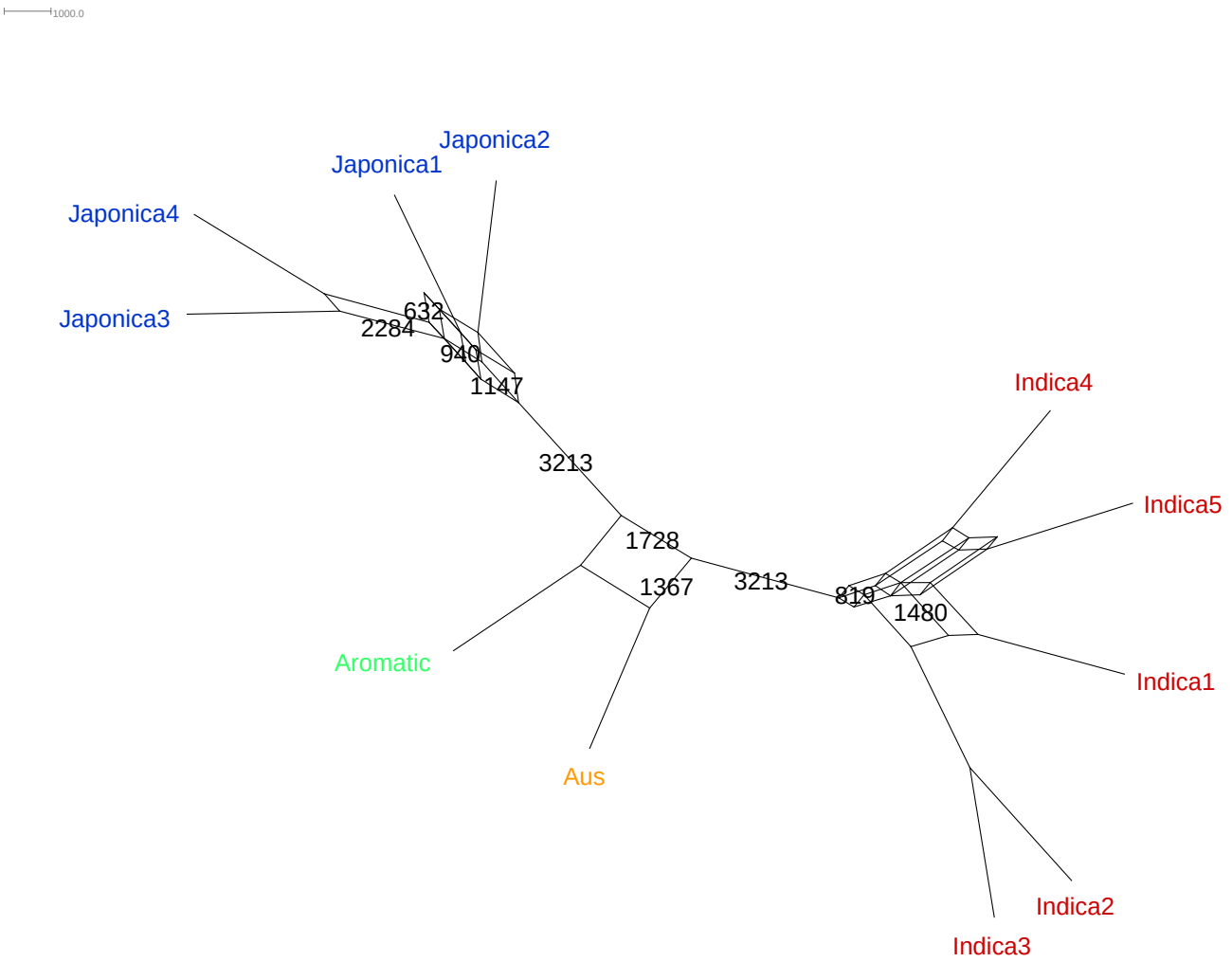
Implémenté dans BEAST

Retour à nos données de riz ...

A l'aide de SNAPP+SplitsTree

Inférence d'un réseau à partir de la distribution a posteriori de l'arbre d'espèces sachant les séquences

JDD1 (1550 SNPs), CHR6, données Joao+Jean-Christophe



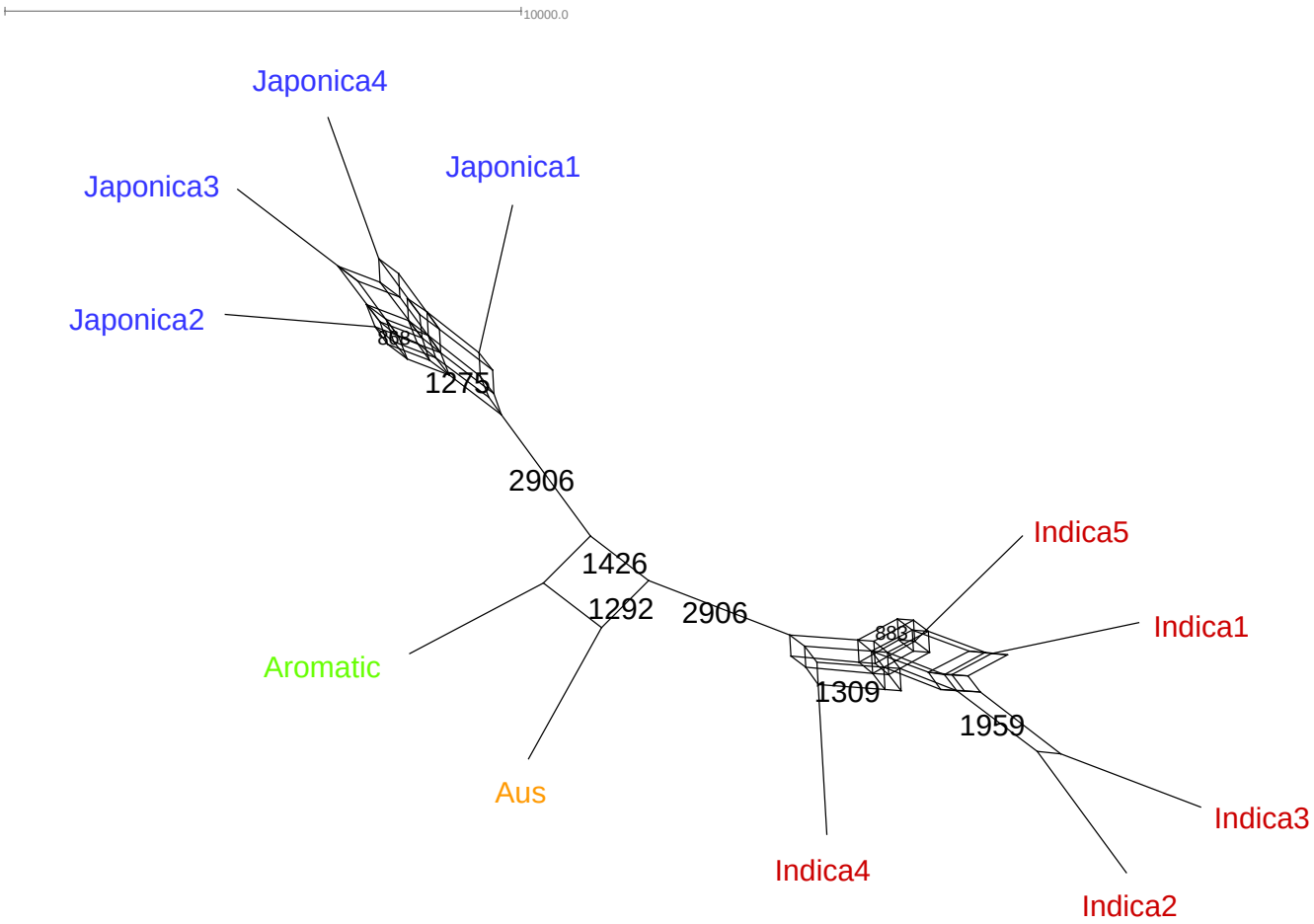
Japonica :

- 1=tempéré
- 2=admixture
- 3=tropicaux
- 4=subtrop

Indica :

- 1=admixture
- 2
- 3=B
- 4=A
- 5

JDD2 (1550 SNPs), CHR6, données Joao+Jean-Christophe



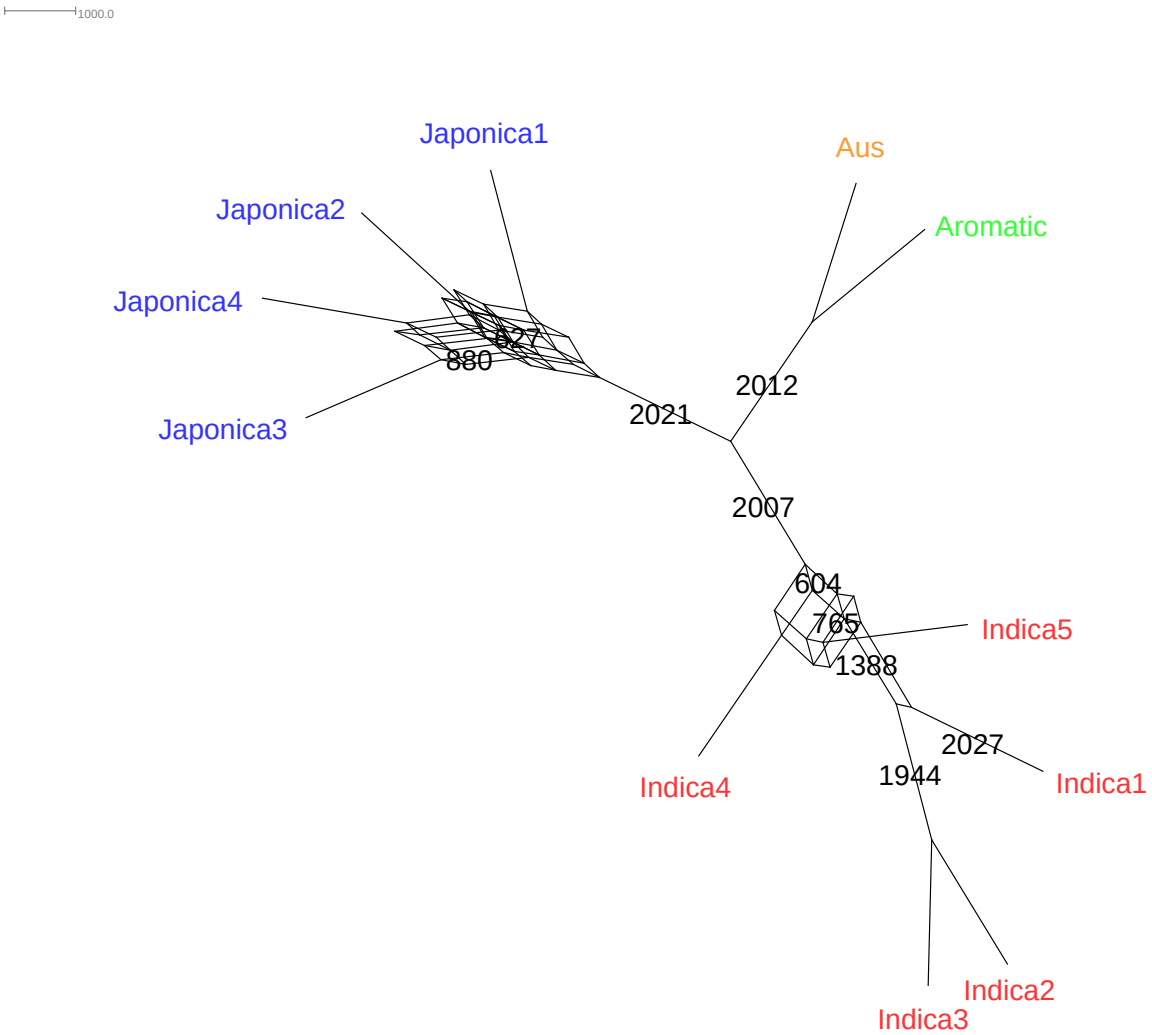
Japonica :

- 1=tempéré
- 2=admixture
- 3=tropicaux
- 4=subtrop

Indica :

- 1=admixture
- 2
- 3=B
- 4=A
- 5

JDD3 (7749 SNPs), CHR6, données Joao+Jean-Christophe



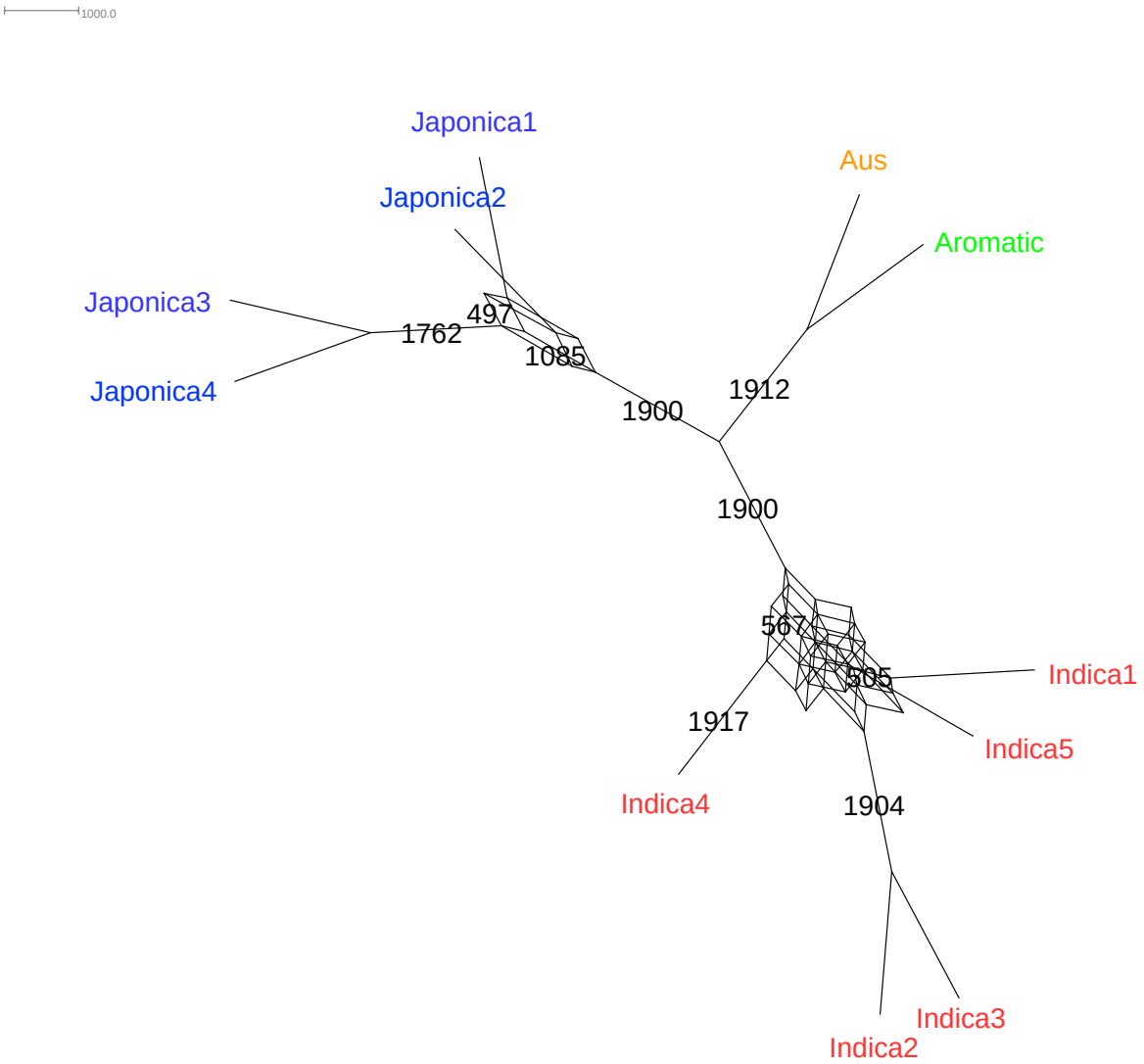
Japonica :

- 1=tempéré
- 2=admixture
- 3=tropicaux
- 4=subtrop

Indica :

- 1=admixture
- 2
- 3=B
- 4=A
- 5

JDD4 (7749 SNPs), CHR6, données Joao+Jean-Christophe



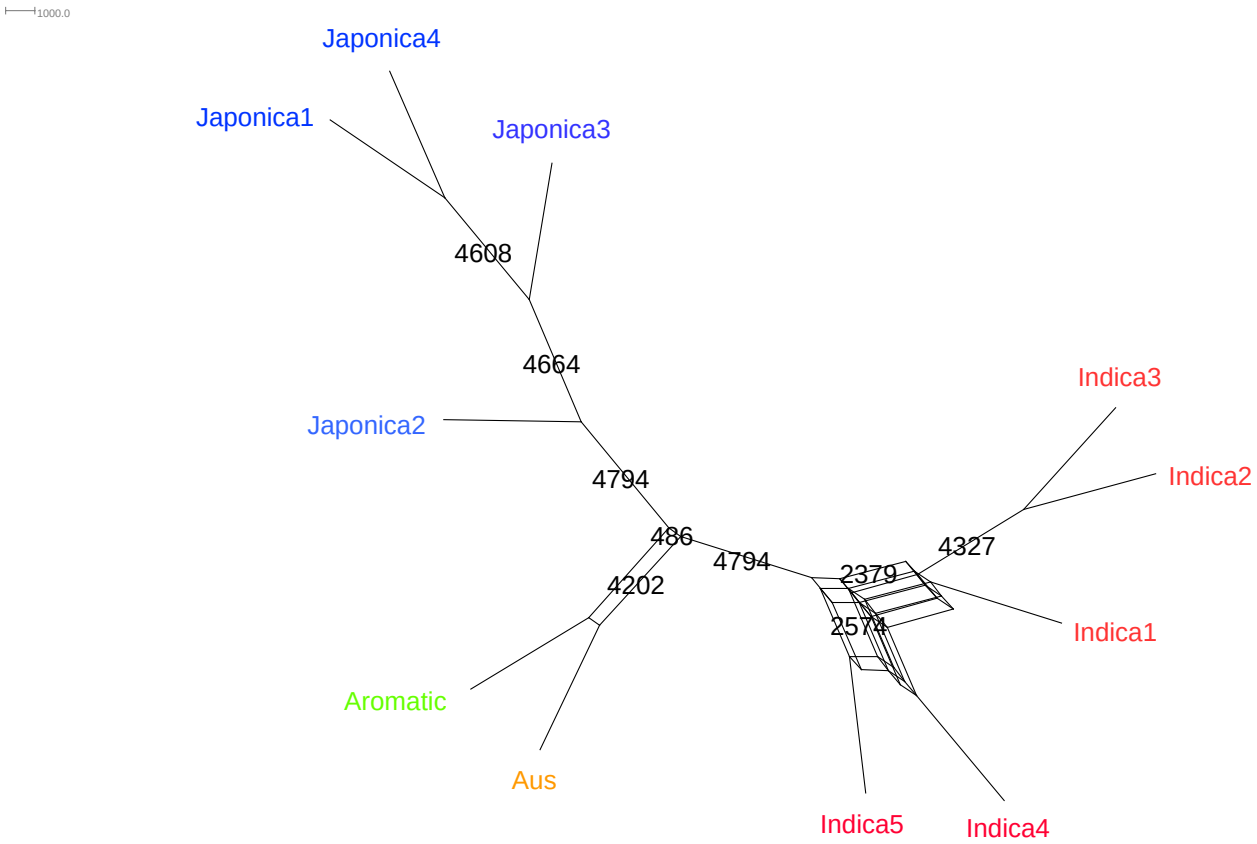
Japonica :

- 1=tempéré
- 2=admixture
- 3=tropicaux
- 4=subtrop

Indica :

- 1=admixture
- 2
- 3=B
- 4=A
- 5

JDD1 (1550 SNPs), mutations au dessus de la racine



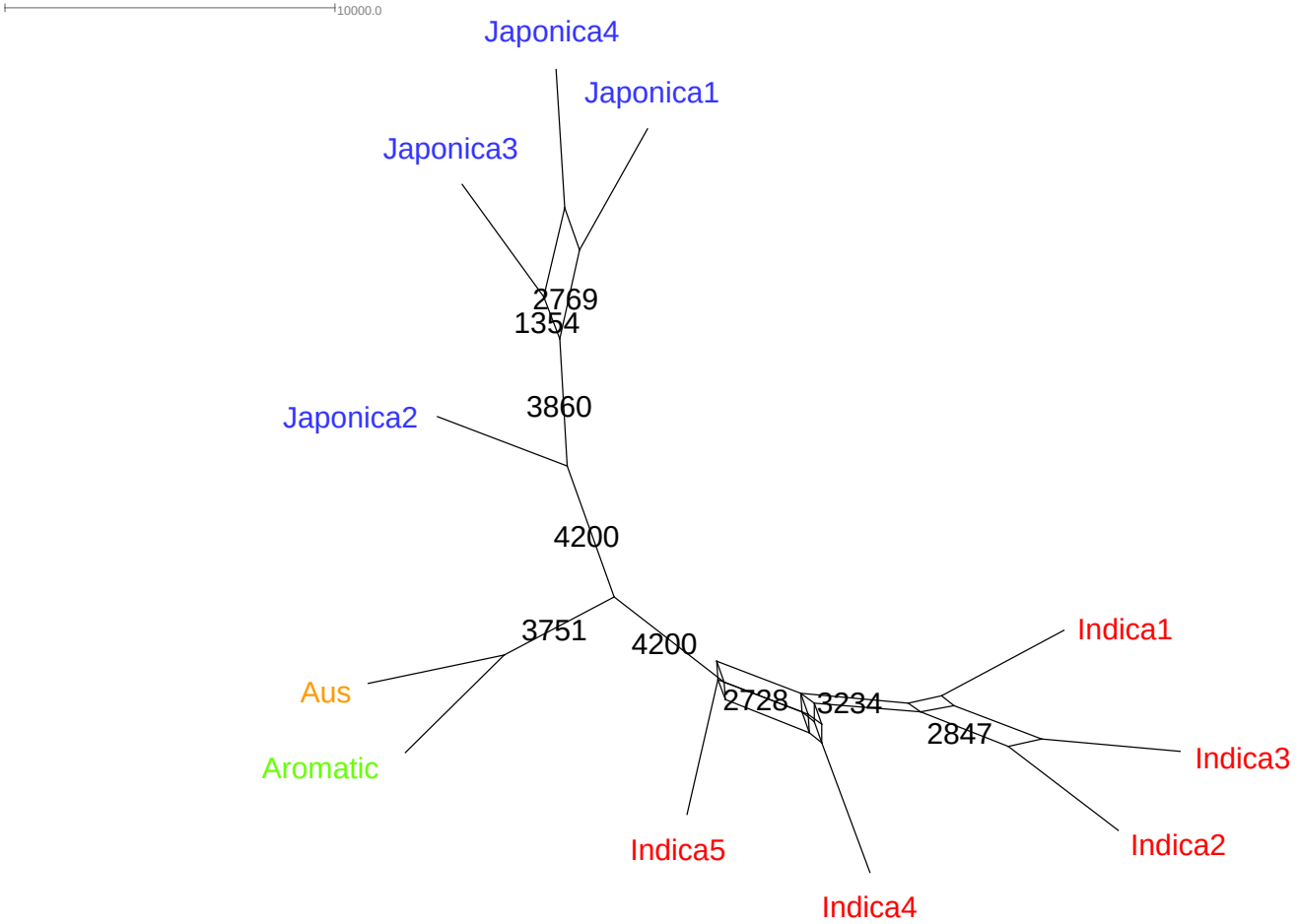
Japonica :

- 1=tempéré
- 2=admixture
- 3=tropicaux
- 4=subtrop

Indica :

- 1=admixture
- 2
- 3=B
- 4=A
- 5

JDD2(1550 SNPs), mutations au dessus de la racine



Japonica :

- 1=tempéré
- 2=admixture
- 3=tropicaux
- 4=subtrop

Indica :

- 1=admixture
- 2
- 3=B
- 4=A
- 5

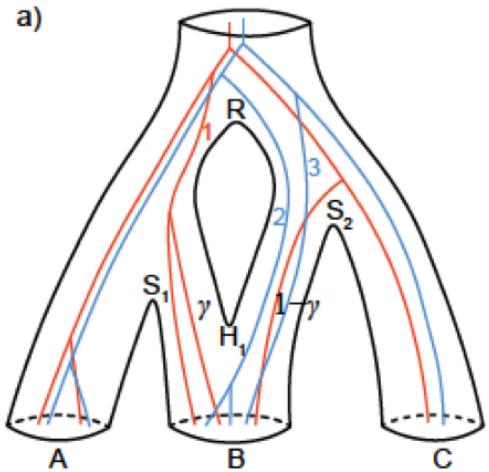
Plan

Simulateur basé sur un réseau (Génome Harvest)

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

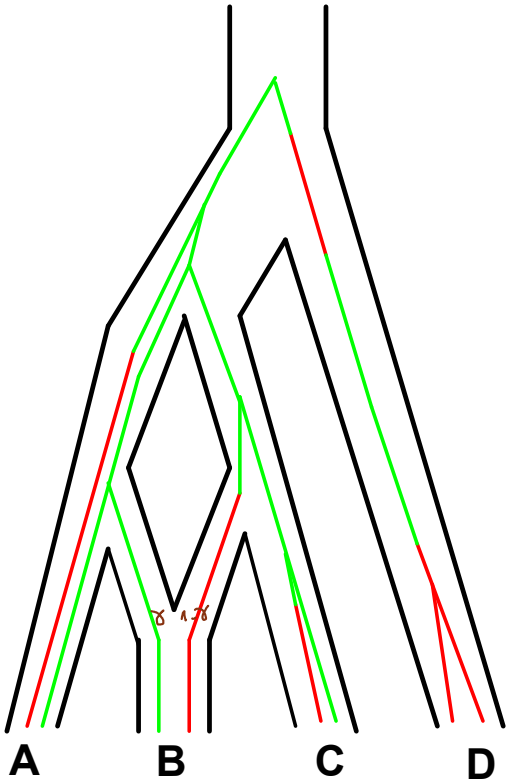
- Génération d'arbres de gènes évoluant à l'intérieur d'un réseau selon un processus de coalescence + modèle de Nakhleh au niveau du noeud de réticulation

Zhang et al. (2017)



- Modèle markovien évoluant le long des branches de l'arbre de gènes
- Chaque arbre de gènes généré donne un SNP

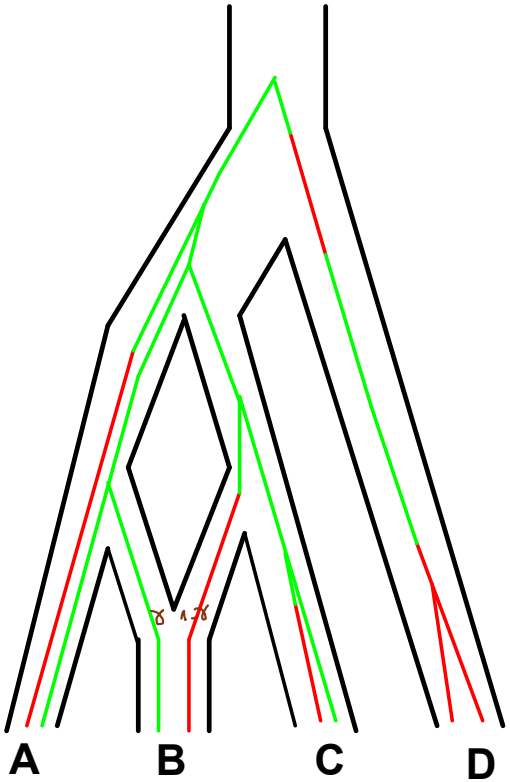
Un exemple illustrant les performances de SNAPP+SplitsTree dans l'inférence de réseaux



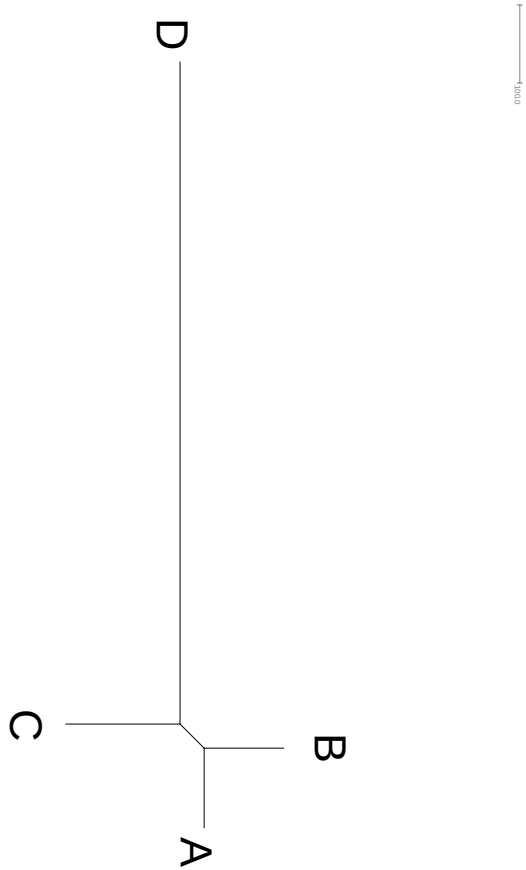
Réseau étudié

- 2 lignées dans chaque espèces
- $10,000$ sites polymorphes
- $\gamma = 0.5$
- 3 millions d'itérations par MCMC

Comparaison vrai réseau vs réseau inféré

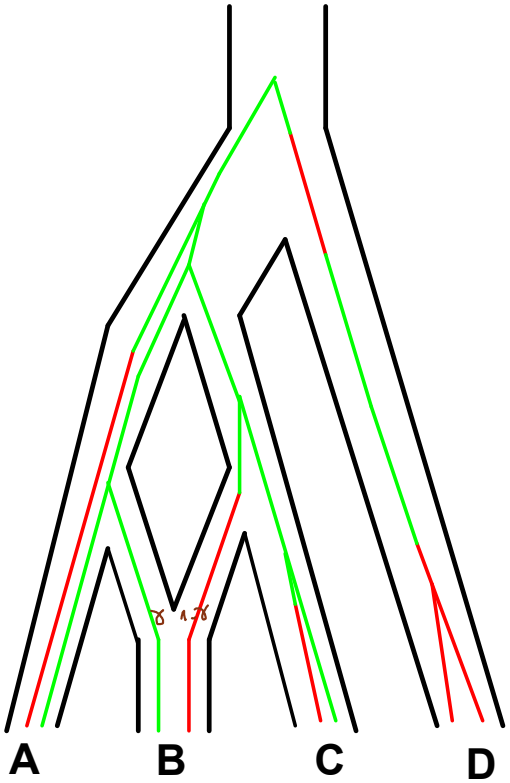


Vrai Réseau

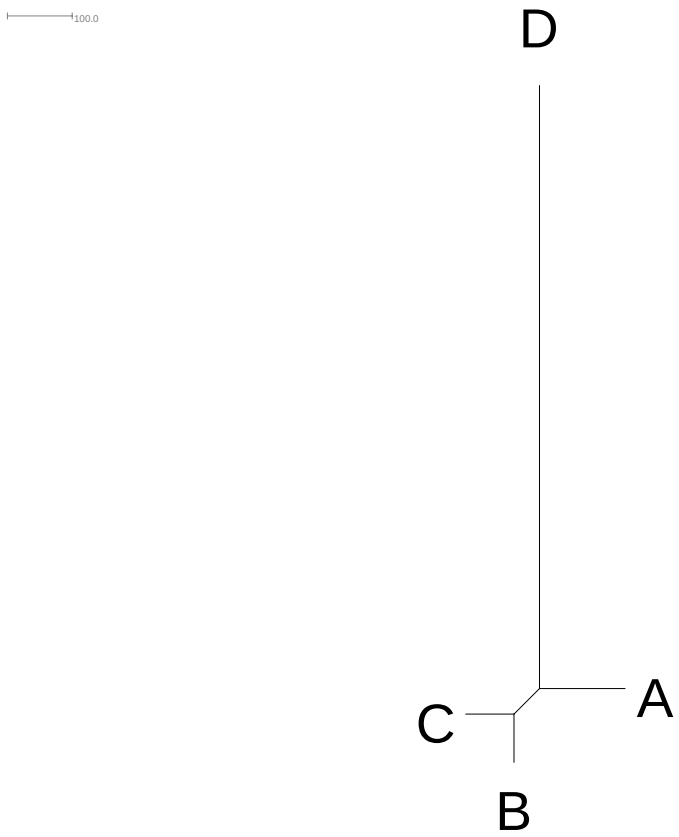


Réseau inféré par SNAPP+SplitsTree

En multipliant les taux de mutation *u* et *v* par 4

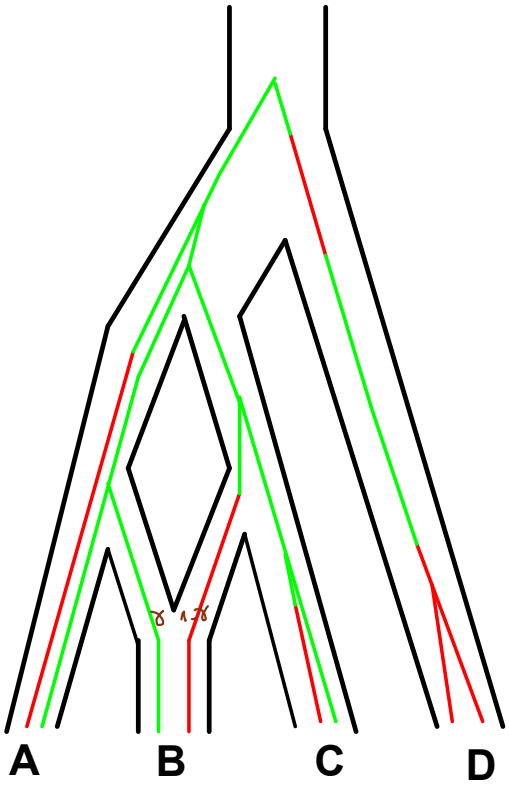


Vrai Réseau

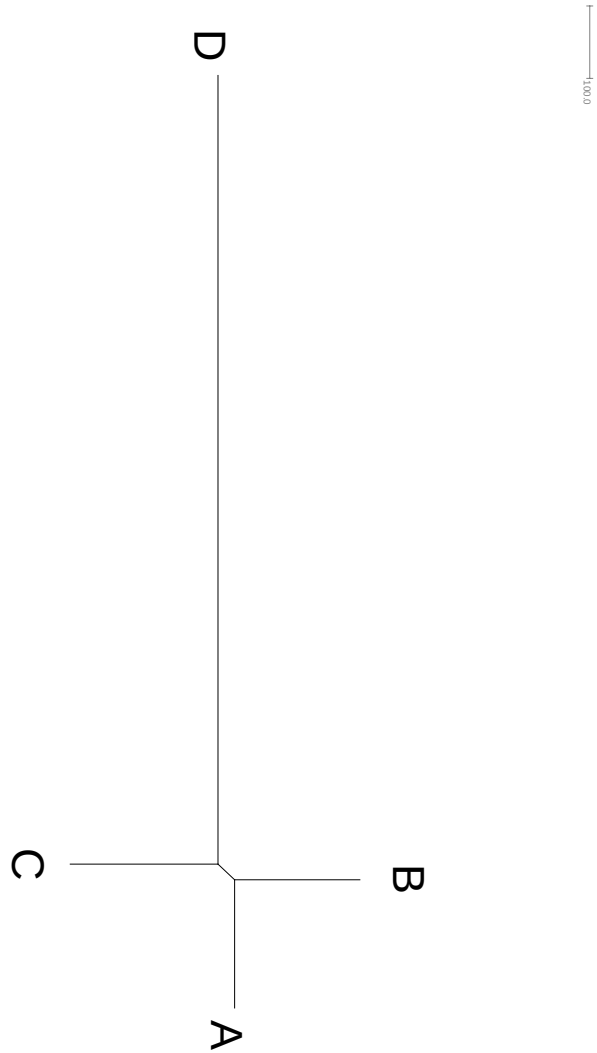


Réseau inféré par SNAPP+SplitsTree

En multipliant les tailles de population θ par 10



Vrai Réseau



Réseau inféré par SNAPP+SplitsTree

Bilan sur l'inférence de réseaux en 2 étapes successives

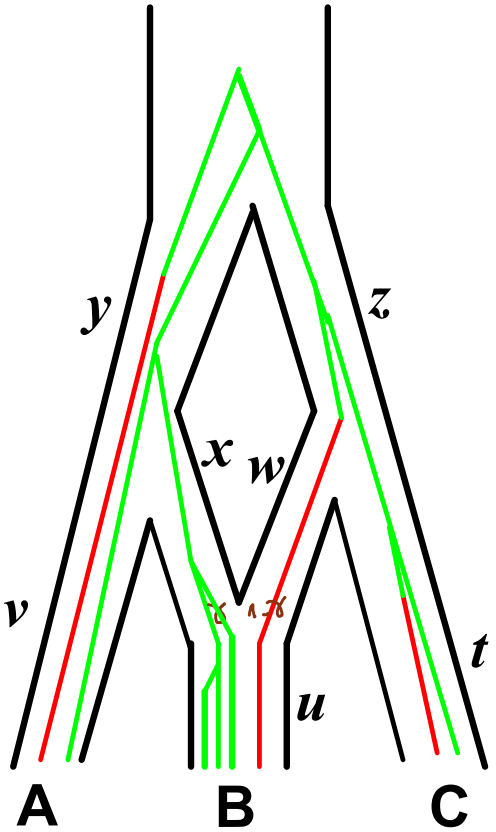
Quelques questions émergentes :

- Les réseaux inférés sur les données de Joao+Jean-Christophe sont-ils fiables ?
- Pourquoi avait-on réussi à obtenir des arbres différents (i.e. boites dans SplitTree) sur les données réelles ?
  - en raison du plus grand nombre d'espèces ?
  - en raison du plus grand nombre d'hybridations ?

Plan

Une nouvelle méthode d'inférence directe de réseaux (en cours)

Extension de SNAPP aux réseaux



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

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

Vraisemblance

Pour raccourcir, *Count* est implicite dans les probabilités ...

$$\mathbb{P}(Data)$$

$$= \sum_i \sum_j \mathbb{P}(Data \mid n_{root} = i, r_{root} = j)\, \mathbb{P}(r_{root} = j \mid n_{root} = i)$$

$$\mathbb{P}(n_{root} = i)$$

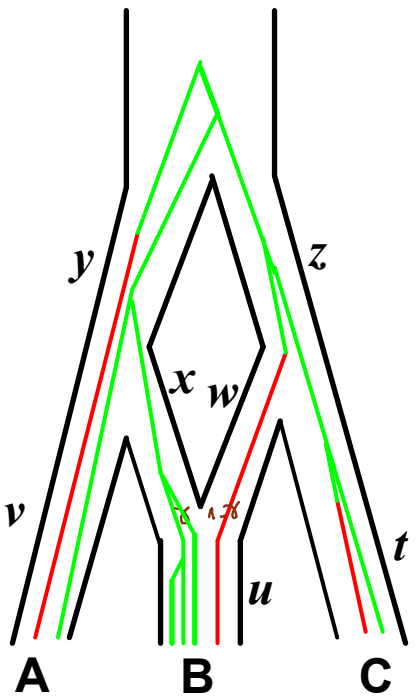
$$= \sum_i \sum_j \sum_{i'} \sum_{j'} \mathbb{P}\left(Data_{z^T} Data_{y^T} \mid n_{y^T} = i', n_{z^T} = i - i', r_{y^T} = j',\right.$$

$$\left. r_{z^T} = j - j'\right)\, \mathbb{P}(r_{root} = j \mid n_{root} = i)\, \mathbb{P}(n_{root} = i)$$

$Data_{z^T}$ et $Data_{y^T}$ ne sont plus indépendantes...
$Data_{z^T}$ et $Data_{y^T}$ comprennent les allèles rouges et verts de l'espèce hybride B !!!

On ne peut plus effectuer le produit des probabilités $\mathbb{P}(Data_{z^T}) \times \mathbb{P}(Data_{y^T})$!!!

Notre algorithme : calcul des lois jointes



Quantités calculées successivement

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

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

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

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

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

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

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

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

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

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

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

Plan

Equipe de Tanja Stadler (ETH Zurich)

Statistique Bayésienne dans le cadre d'un réseau

- N : réseau phylogénétique
- X_i : alignements pour le SNP i
- G_i : arbre de gènes 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 $\mathbb{P}(N, G_1, \ldots, G_m | X_1, \ldots, X_m)$.

Ils n'intègrent pas sur tous les arbres de gènes

Zhang et al (bioRxiv, Avril 2017)

Equipe de Luay Nakhleh (Université de Rice, USA)

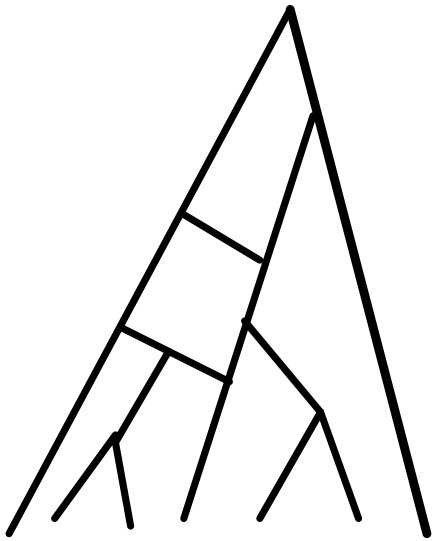
- Wen et al. (Plos Genetics, 2016)
  - Arbres de gènes inférés lors d'un étape préliminaire
  - Données = arbres de gènes !!!

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

- Zhu et al. (bioRxiv, 29 Mai 2017)
  - Données = alignements
  - Intégration sur tous les arbres de gènes ...
  - Algorithme inspiré de SNAPP mais très couteux en temps de calcul

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

Un réseau sur lequel on devrait etre plus performant



Plan

Conclusion

- Jusqu'alors travail sur le riz $\rightarrow$ autre plante d'intérêt ?
  Genome Harvest : Banane, Citrus, Caféier, Riz, Tomate, Canne à sucre ...
- Afin de comprendre l'histoire des riz cultivés, nécessité de disposer de riz sauvages, à l'instar de Choi et al. (MBE, 2017), Wang et al (Genome Research, 2017) ...
- SNAPP disponible sur http ://snapp.otago.ac.nz