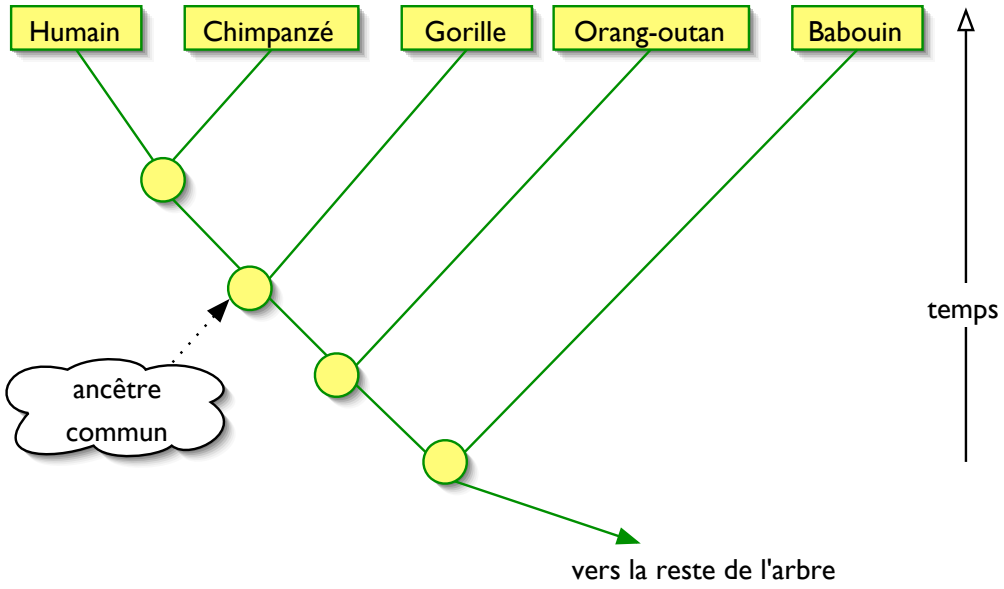
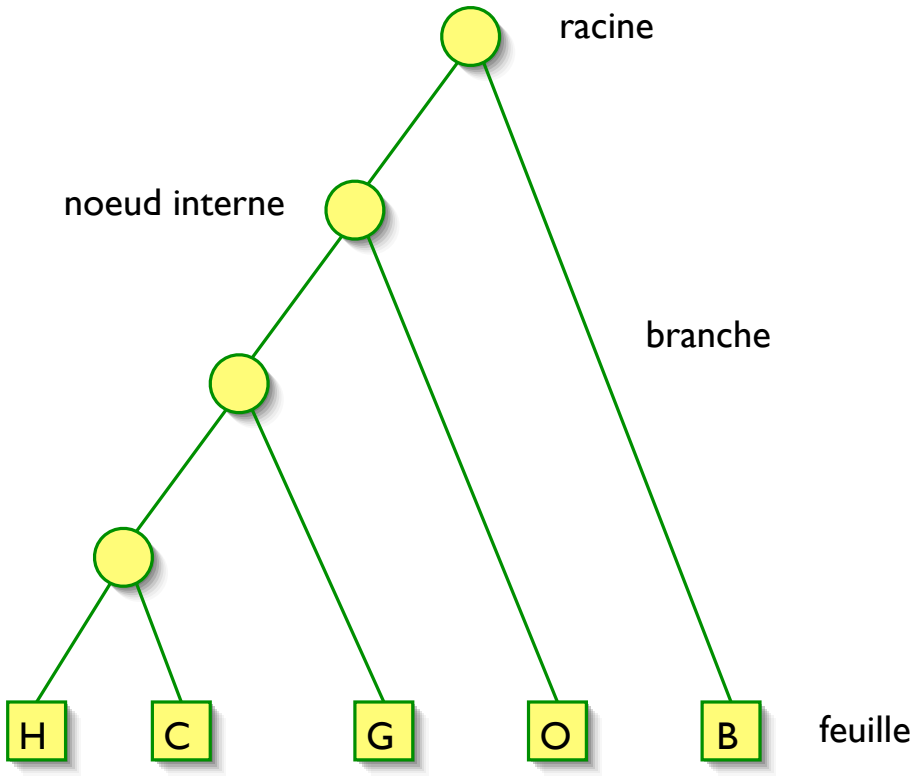
Phylogénies

phylogénie ou arbre évolutif

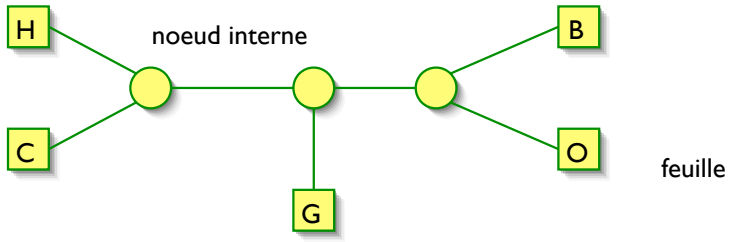
(aussi : a. phylogénétique, a. évolutionnaire, a. de l'Évolution, a. d'évolution)



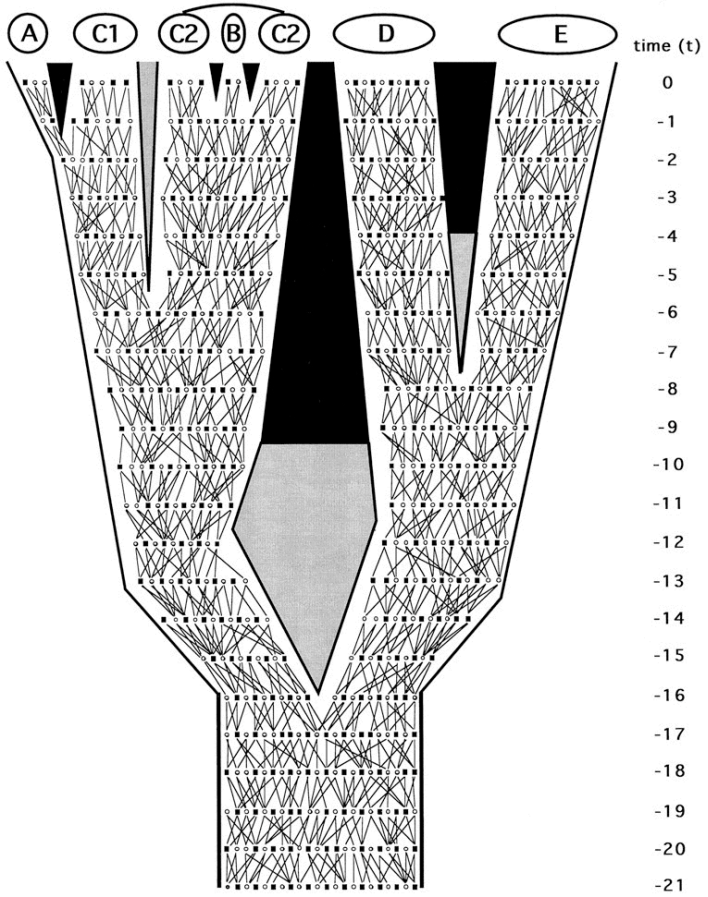
Arbre raciné



Arbre non-raciné



D’où vient l’arbre ?



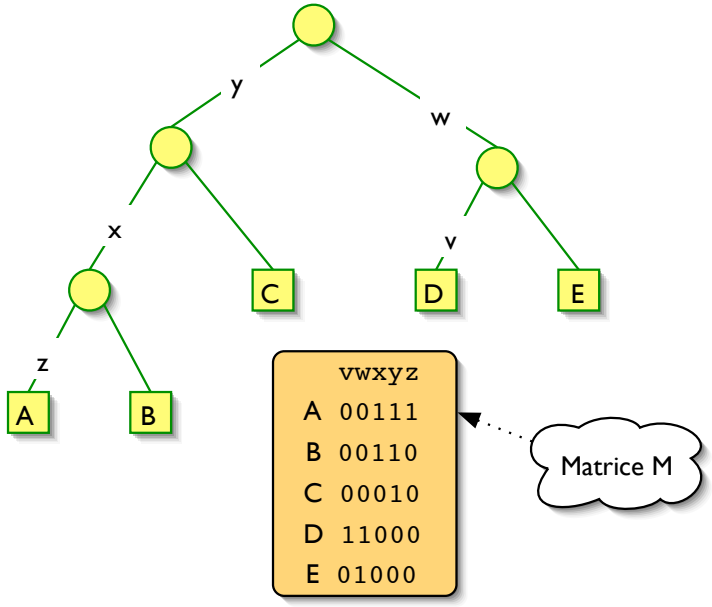
Phylogénie parfaite

«caractères» binaires :
- possède des ailes, présence de colonne vertébrale

qqch qui apparaît à un point dans l'Évolution et tous les descendants le possèdent

chaque caractère est associée à une branche de l'arbre : le caractère y change $0 \rightarrow 1$

Phylogénie parfaite - 2



Phylogénie parfaite - 3

ensemble de feuilles : $\mathcal{L}$ avec $|\mathcal{L}| = n$
chaque feuille est identifiée par ℓ caractères binaires

$\Rightarrow$ chaque feuille $k \in \mathcal{L}$ est identifiée par une séquence $S_k \in \{0, 1\}^\ell$.

Matrice M de taille $n \times \ell$: $\mathrm{M}[k, j] = S_k[j]$.

Problème : trouver la p.p. T à partir de la matrice M.

Phylogénie parfaite - 4

Pour chaque $j = 1, \ldots, \ell$, soit O_j l'ensemble de feuilles pour lesquelles $S_k[j] = 1$

$$O_j = \{k \in \mathcal{L} \colon S_k[j] = 1\}.$$

Déf. Deux positions $i, j = 1, \ldots, n$ sont compatibles ssi (1) $O_i \cap O_j = \emptyset$, (2) $O_i \subseteq O_j$, ou (3) $O_j \subseteq O_i$.

Thm. La matrice M correspond à une p.p. ssi toutes les paires de positions sont compatibles.

Phylogénie parfaite - 5

Preuve.

$\rightarrow$ s'il existe une p.p., alors toutes les paires sont compatibles [. . .]

$\leftarrow$ si toutes les paires sont compatibles, alors il existe une p.p.

[1] les colonnes de M sont triées de $\mathtt{11\ldots1} \rightarrow \mathtt{00..0}$.

[2] soit $k, k' \in \mathcal{L}$ et soit j la dernière position avec $k, k' \in O_j$. Pour toute $i < j$ avec $k \in O_i$, on a $k' \in O_i$ (car $O_j \subseteq O_i$).

[3] $\mathrm{M}[k, i] = \mathrm{M}[k', i]$ si $i \leq j$ et $\mathrm{M}[k, i'] = [k', i']$ pour $i' > j$ seulement si $\mathrm{M}[k, i'] = 0$.

[4] construire l'arbre défini par préfixes communs des S

Arbre de mots-clés

Déf. L'ensemble $\mathcal{S}$ de mots est un ensemble préfixe-libre s'il n'existe aucune paire $u \neq v \in \mathcal{S}$ telle que u est le préfixe de v.

Thm. Si l'ensemble $\mathcal{S}$ est préfixe-libre, alors il existe un arbre raciné T avec les propriétés suivantes.

– chaque branche e de T est étiquetée par un caractère
– pour chaque feuille de T, les caractères concatenés sur le chemin de la racine forment un mot dans $\mathcal{S}$
– chaque mot de $\mathcal{S}$ est associé avec une feuille

Preuve. (construction de l'arbre)

Phylo parfaite - 6

Retour à la question de phylogénie parfaite

associer un mot avec chaque rangée : les caractères sont les indices de colonnes où on a 1, $ à la fin

C'est un ensemble préfixe-libre $\Rightarrow$ construction de l'arbre

Phylogénie parfaite - 7

Et si pas de phylo parfaite ?

Sous-ensemble de feuilles avec p.p ? – problème NP-complet.

Alphabet quelconque ? — ditto.

Maximum de parcimonie (MP)

minimiser le nombre total de changements de caractères sur les branches (valeur de parcimonie)

calculer la valuer de parcimonie pour un arbre donné : PD

MP - 2

MP pondéré : PD

trouver l'arbre : NP-difficile

Bipartitions

Soit $\mathcal{L}$ un ensemble de feuilles.

Déf. Une bipartition («split») de $\mathcal{L}$ est une paire d'ensembles $(L, \mathcal{L} - L)$.

Soit T un arbre (non-raciné) avec feuilles $\mathcal{L}$. Chaque arête e induit une bipartition $(L_e, \bar{L}_e)$: en suppressant cet arête, T est coupé en deux arbres T_1 et T_2, L_e et $\bar{L}_e$ correspondent aux deux ensembles de feuilles.

Bipartitions - 2

Soit $\mathcal{S}$ un ensemble de bipartitions sur un ensemble $\mathcal{L}$.

Thm. (compatibilité de bipartitions) $\mathcal{S}$ est l'ensemble de bipartitions induites par les arêtes d'un arbre T ssi il possède la propriété suivante. Pour deux bipartitions $(L_1, \bar{L}_1), (L_2, \bar{L}_2) \in \mathcal{S}$, exactement un des ensembles $L_1 \cap L_2$, $L_1 \cap \bar{L}_2$, $\bar{L}_1 \cap L_2$, $\bar{L}_1 \cap \bar{L}_2$ est vide.

Preuve.
$\leftarrow$ prenons les deux arêtes e_1, e_2 qui induisent les bipartitions. . .

Bipartitions - 3

Preuve cont.

$\rightarrow$ Réduction au problème de phylogénie parfaite. Séquences binaire pour encoder les bipartitions, 1 pour l'ensemble plus petit entre $(L, \bar{L}) \Rightarrow$ compatibilité

QED

Donc, l'ensemble de ses bipartitions définit un arbre.

Distance sur un arbre

Déf. $D : \mathcal{L} \times \mathcal{L} \mapsto [0, \infty)$ est une distance sur $\mathcal{L}$ ssi

- $d(i, j) = 0$ ssi $i = j$
- $d(i, j) = d(j, i)$
- $(d(i, j) + d(j, k) \geq d(i, k)$

Notre but : construire un arbre qui «représente» les distances d — on veut trouver un ensemble de bipartitions compatibles

arbre $\rightarrow$ distance : facile

1. nombre d'arêtes sur le chemin entre i et j
2. somme des poids (positifs) des arêtes sur le chemin entre i et j

Distance - 2

Soit T un arbre avec feuilles $\mathcal{L}$ et soit $\mathcal{S}$ l'ensemble de bipartitions pour T. Considerons la pondération pour chaque bipartition $S = (L, \bar{L}) \in \mathcal{S}$

$$\mu_S = \frac{1}{2} \min\Big\{ d(i,k) + d(j,l) - d(i,j) - d(k,l) : i,j \in L, k,l \in \bar{L} \Big\}.$$

(μ_S est positif si $d(\cdot)$ vient d'un arbre...)

Thm. Si $S_1 = (L_1, \bar{L}_1)$ et $S_2 = (L_2, \bar{L}_2)$ sont deux bipartitions avec $\mu_{S_1} > 0$ et $\mu_{S_2} > 0$, alors exactement un des ensembles $L_1 \cap L_2$, $L_1 \cap \bar{L}_2$, $\bar{L}_1 \cap L_2$, $\bar{L}_1 \cap \bar{L}_2$ est vide.

Distance - 3

Preuve. Par contradiction : prenons $i \in L_1 \cap L_2$, $j \in L_1 \cap \bar{L}_2$, $k \in \bar{L}_1 \cap L_2$, $l \in \bar{L}_1 \cap \bar{L}_2$. Or,

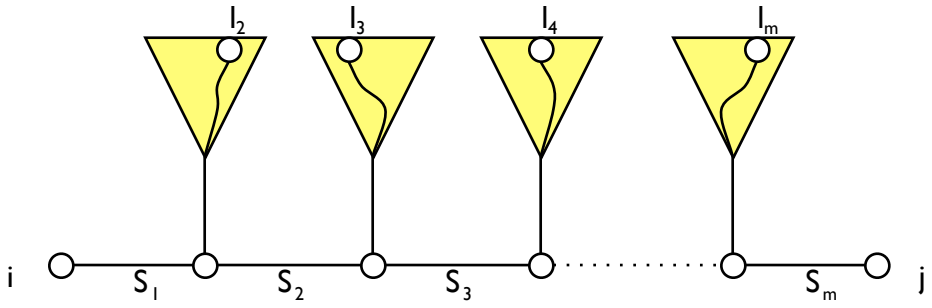
$$d(i,k) + d(j,l) - d(i,j) - d(k,l) \geq 2\mu_{S_1} > 0$$
$$d(i,k) + d(j,l) - d(i,j) - d(k,l) \leq -2\mu_{S_2} < 0,$$

contradiction.

Donc $\{S : \mu_S > 0\}$ définit un arbre T_d. Chaque arête e de cet arbre est pondéré par un μ_S (S est induit par e). Définissons la distance $\Delta : \mathcal{L} \times \mathcal{L} \mapsto [0, \infty)$: $\Delta(i,j)$ est la sommme des poids sur les arêtes du chemin entre i et j.

Distance - 4

Thm. $\triangle \leq d$.



Preuve (Waterman) Idée pour feuilles i, j : séquence de bipartitions $S_1, \ldots, S_m$ sur le chemin entre eux. Prenons les vertices arbitraires $l_2, \ldots, l_m$ dans les sous-arbres, sommation de μ_{S_i}+inégalités.

Distance - 5

$$\mu_{S_1} \leq \frac{1}{2}\Big(d(i,j) + d(i,l_2) - d(i,i) - d(j,l_2)\Big) \qquad i, i \leftrightarrow l_2, j$$

$$\mu_{S_2} \leq \frac{1}{2}\Big(d(i,l_3) + d(j,l_2) - d(i,l_2) - d(j,l_3)\Big) \qquad i, l_2 \leftrightarrow l_3, j$$

$$\mu_{S_3} \leq \frac{1}{2}\Big(d(i,l_4) + d(j,l_3) - d(i,l_3) - d(j,l_4)\Big) \qquad i, l_2 \leftrightarrow l_4, j$$

$$\vdots$$

$$\mu_{S_m} \leq \frac{1}{2}\Big(d(i,j) + d(j,l_m) - d(i,l_m) - d(j,j)\Big) \qquad i, l_m \leftrightarrow j, j$$

Sommation (tous les $d(i,l_k)$ et $d(l_k,j)$ s'éliminent) :

$$\Delta(i,j) = \sum_{k=2}^{m} \mu_{S_k} \leq \frac{1}{2}\Big(2d(i,j) - d(i,i) - d(j,j)\Big) = d(i,j). \quad \square$$

Condition des quatre points

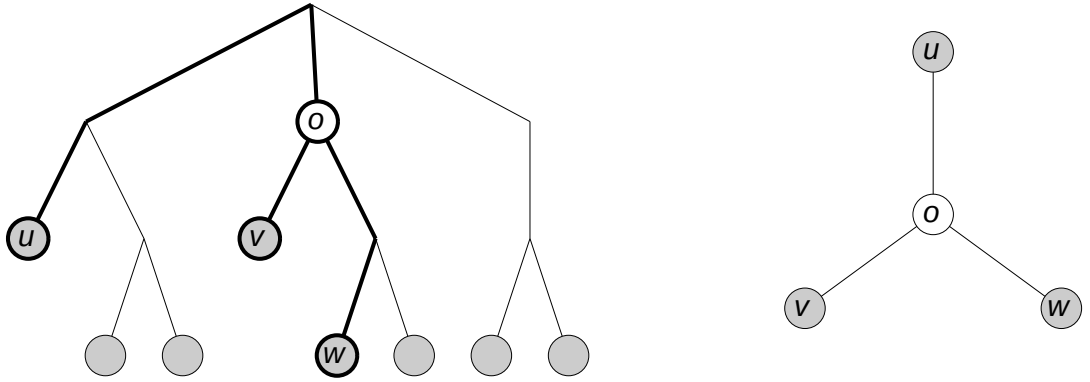
$$d(i,j) + d(k,l) \leq d(i,k) + d(j,l) = d(i,l) + d(j,k)$$

Thm. $\triangle = d$ ssi la condition des quatre points est satisfaite pour chaque quadruplet $i, j, k, l \in \mathcal{L}$.

Preuve. (en une direction seulement)

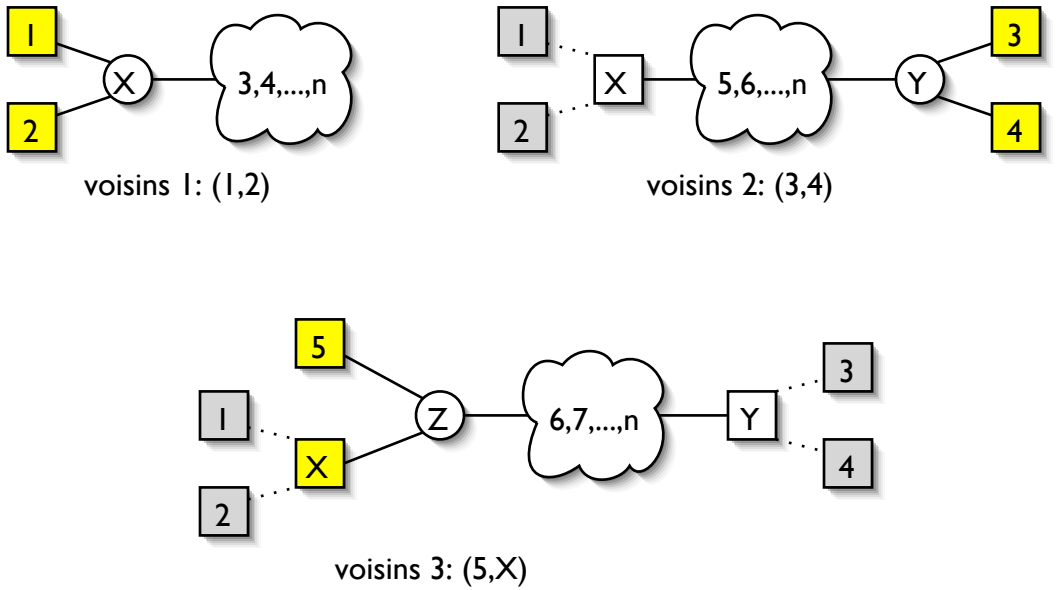
Distance - 6

Algorithme pour la construction d'un arbre à partir de Δ : triples

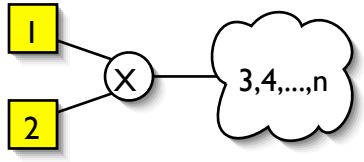


$$\Delta(u,o) = \Big(\Delta(u,v) + \Delta(u,w) - \Delta(v,w)\Big)/2.$$

Neighbor joining



Neighbor joining - 2



calcul de longueurs de branches : prenons la moyenne

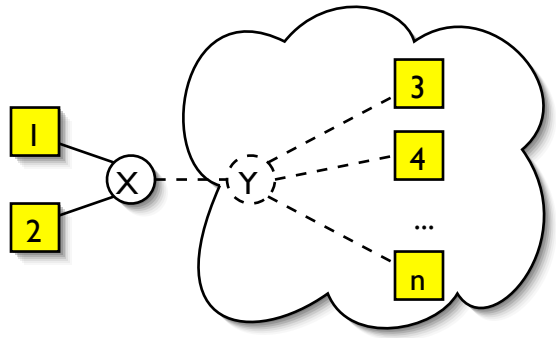
$$L_{1X} = \frac{1}{n-2} \sum_{i=3}^{n} (d_{1i} + d_{12} - d_{2i})/2.$$

L_{2X} même chose

$L_{iX} = (d_{1i} + d_{2i} - d_{12})/2.$

Neighbor joining - 3

choix des voisins : minimiser le somme total de longueurs de branches



$$S_{12} = L_{1X} + L_{2X} + L_{XY} + \sum_{i=3}^{n} L_{Yi}.$$

où $L_{XY} = (\sum_{i \geq 3} L_{Xi} - \sum_{i \geq 3} L_{Yi})/(n-2)$.

Neighbor joining - 4

On a $\sum_{i=3}^{n} L_{Yi} = \frac{1}{n-3} \sum_{3 \leq i < j \leq n} d_{ij}$. Après un peu d'arithmétique :

$$S_{12} = \frac{1}{2} d_{12} + \frac{1}{n-2} \sum_{i<j} d_{ij} - \frac{1}{n-2} \left(\frac{\sum_{i=1}^{n} d_{1i} + \sum_{i=1}^{n} d_{2i}}{2} \right)$$

Calculer en avance : $R_k = \sum_{i=1}^{n} d_{ki}$ pour chaque feuille k,
choisir la paire qui minimise $(n-2)d_{kk'} - R_k - R_{k'}$: elle minimise $S_{kk'}$ aussi

Neighbor joining - 5

algorithme en $O(n^3)$ (pas itératif) :
(1) trouver 1,2 en minimisant S_{12} (temps $O(n^2)$) ;
(2) calculer L_{1X}, L_{2X}, et d_{iX} pour $i \geq 3$ (temps $O(n)$) ;
(3) remplacer $1, 2$ par X dans la matrice de distances.

Thm. Ça donne l'arbre correct pour des distances additives.

Preuve. Idée seulement : le calcul des longueurs de branches est correct si 12 sont les voisins. Il faut prouver que 12 sont les voisins quand on minimise S_{12}.

Distance - 7

D’où viennent les distances ?

Modèle d’évolution de séquences sur un arbre :
relai d’un message de la racine vers les feuilles
transmission avec du «bruit» sur chaque branche
bruit = mutations

modèle probabiliste pour les mutations
les séquences correspond à un «échantillon» aléatoire

Cavender-Farris

modèle de Cavender-Farris : séquences binaires, mutations symmetriques

Exemple : noeud x est le parent de y, transmission d'un bit $X \to Y$.
Probabilité de mutation sur la branche : p_{xy}.
Alors $\mathbb{P}\{X \neq Y\} = p_{xy}$.

Similarité de deux noeuds : $S(x, y) = \mathbb{P}\{X = Y\} - \mathbb{P}\{X \neq Y\}$.

Thm. $S(x, y)$ se multiplie sur chaque chemin.

Donc $D(x, y) = -\log S(x, y)$ est une distance additive sur l'arbre.

Taux de mutations

Modèle de Markov : probabilités de mutations [pour un temps d'unité] données par une matrice de transitions $\mathbf{M}$

Matrice de transition pour distance t (entier) est $\mathbf{P}(t) = \mathbf{M}^t$. Qu'est-ce qu'on fait quand t n'est pas entier ?

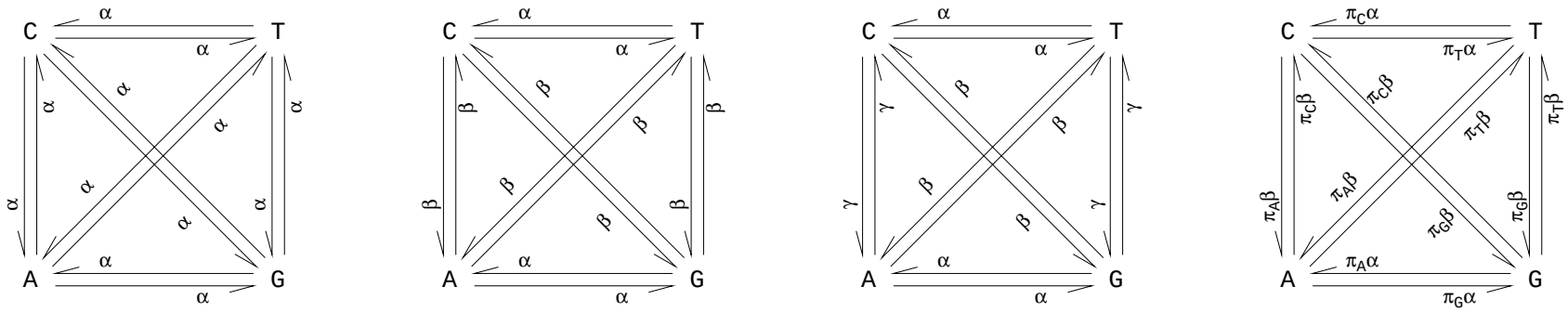
Derivation : $\mathbf{P}'(t) = \lim_{h \to +0} \frac{\mathbf{P}(t+h) - \mathbf{P}(t)}{h}$

Chapman-Kolmogorov : $\mathbf{P}(t+s) = \mathbf{P}(t)\mathbf{P}(s)$

$\mathbf{P}'(t) = \mathbf{P}(t)\mathbf{Q}$ avec $\mathbf{Q} = \lim_{h \to +0} \frac{\mathbf{P}(t) - \mathbf{I}}{h}$

Distance - 9

Taux de mutations pour $e^{\mathbf{Q}t}$



distance paralinéaire et LogDet