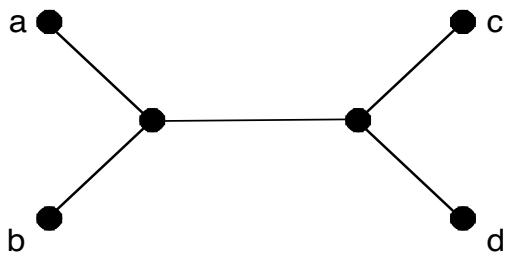
Quelques résultats sur le minimum d'évolution

Sylvain Guillemot
LIRMM, Montpellier

23 janvier 2006

Minimum d'évolution

Arbre T :



$\Rightarrow$

Matrice δ :

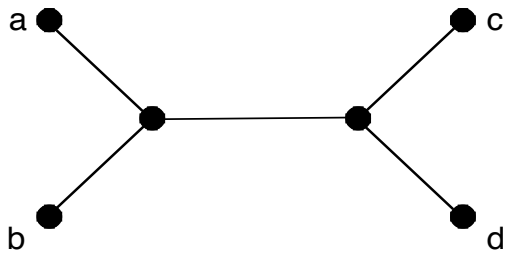
	a	b	c	d
a		2	3	3
b			2	3
c				2
d				

Minimum d'évolution : à partir de δ, on veut reconstruire T. On choisit T parmi un ensemble d'arbres candidats :

- pour chaque arbre candidat, on estime sa longueur par une formule $L^T(\delta)$
- on choisit un arbre candidat de longueur minimale.

Minimum d'évolution (2)

Arbre T :



$\Rightarrow$

Matrice δ :

	a	b	c	d
a		2	3	3
b			2	3
c				2
d				

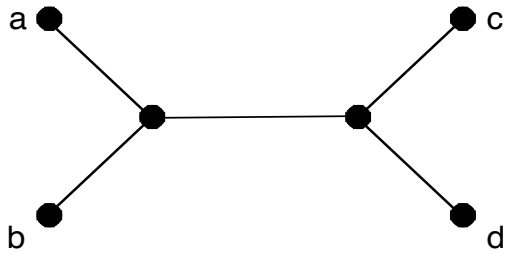
Comment définir $L^T(\delta)$?

Le plus naturel :

$$L^T(\delta) = \frac{1}{2}(\delta_{ab} + \delta_{cd}) + \frac{1}{4}(\delta_{ac} + \delta_{bd}) + \frac{1}{4}(\delta_{ad} + \delta_{bc})$$

Minimum d'évolution (3)

Arbre T :

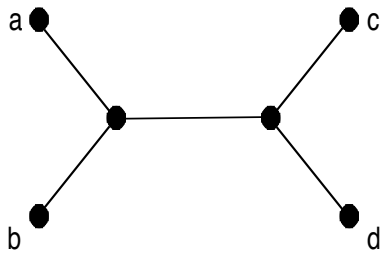


$\Rightarrow$

Matrice δ :

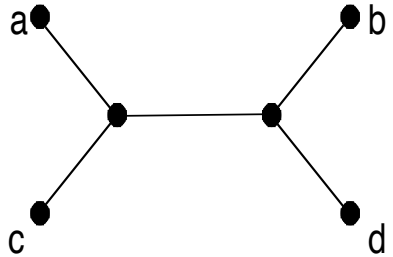
	a	b	c	d
a		2	3	3
b			2	3
c				2
d				

Arbre T_1 :



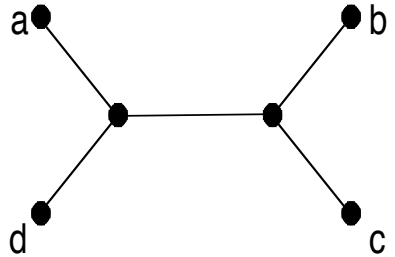
$L^{T_1}(\delta) = 5$

Arbre T_2 :



$L^{T_2}(\delta) = 5 + \frac{1}{2}$

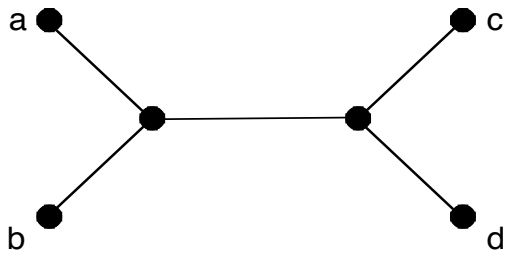
Arbre T_3 :



$L^{T_3}(\delta) = 5 + \frac{1}{2}$

Minimum d'évolution (4)

Arbre T :



$\Rightarrow$

Matrice δ :

	a	b	c	d
a		2	3	3
b			2	3
c				2
d				

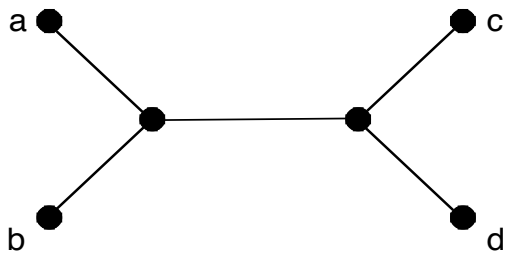
Mais d'autres formules sont possibles.

Ainsi si λ est un réel quelconque :

$$L^T(\delta) = \frac{1}{2}(\delta_{ab} + \delta_{cd}) + \lambda(\delta_{ac} + \delta_{bd}) + (\frac{1}{2} - \lambda)(\delta_{ad} + \delta_{bc})$$

Minimum d'évolution (5)

Arbre T :

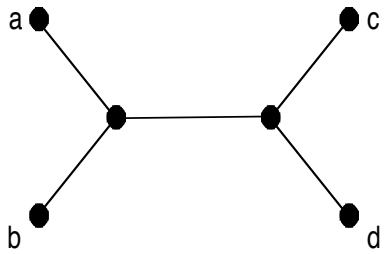


$\Rightarrow$

Matrice δ :

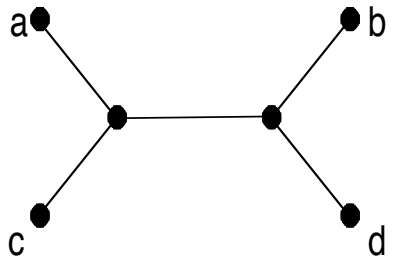
	a	b	c	d
a		2	3	3
b			2	3
c				2
d				

Arbre T_1 :



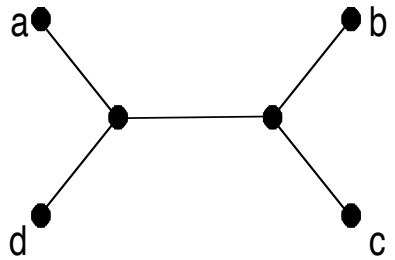
$L^{T_1}(\delta) = 5$

Arbre T_2 :



$L^{T_2}(\delta) = 6 - 2\lambda$

Arbre T_3 :



$L^{T_3}(\delta) = 6 - 2\lambda$

Question : quelles formules utiliser pour $L^T(\delta)$?

Méthode OLS : on cherche une distance d associée à T qui minimise l'écart quadratique

$$\sum_{i,j \in X} (\delta_{ij} - d_{ij})^2$$

on pose $L^T(\delta)$ égale à la somme des longueurs de branches obtenues.

$\Rightarrow$ simple, mais donne la même importance à tous les coefficients δ_{ij} indépendamment de leur variance.

Méthodes OLS, WLS et BLS (2)

Question : quelles formules utiliser pour $L^T(\delta)$?

Méthode WLS : on cherche une distance d associée à T qui minimise l'écart quadratique

$$\sum_{i,j \in X} w_{ij}(\delta_{ij} - d_{ij})^2$$

on pose $L^T(\delta)$ égale à la somme des longueurs de branches obtenues.

$\Rightarrow$ plus général, mais n'a pas de bonnes propriétés mathématiques (inconsistant).

Méthodes OLS, WLS et BLS (3)

Question : quelles formules utiliser pour $L^T(\delta)$?

Méthode BLS : (Pauplin)
Formule directe :

$$L^T(\delta) = \sum_{i,j \in X} 2^{-p_{ij}^T} \delta_{ij}$$

(Remarque : la méthode BLS peut aussi se voir comme un cas particulier de WLS).

$\Rightarrow$ bien adapté aux données biologiques, et bonnes propriétés mathématiques.

Consistance (1)

Définition : une méthode de reconstruction est consistante si la probabilité d'inférer l'arbre réel tend vers 1 à mesure que la longueur des séquences augmente.

Pour les méthodes de distance : c'est équivalent à demander que pour tout arbre T : si la méthode est exécutée a partir d'une distance d^T associée à T, alors elle infère correctement l'arbre T.

Pour ME ?

Consistance (2)

Pour ME : si d est associée à un arbre T de longueur totale L,

- estime correctement la longueur de T : $L^T(d) = L$
- tout arbre $W \neq T$ est strictement plus long que T : $L^W(d) > L$

Resultats :

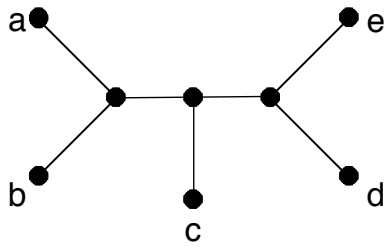
- ME+OLS est consistant [RN93]
- ME+WLS est inconsistant [GBD01]
- ME+BLS est consistant [DG04]

Consistance (3)

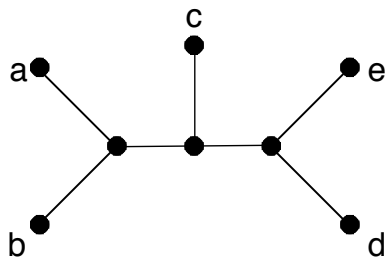
Exemple : consistance de ME+BLS

- plusieurs preuves ([DG04], ...)
- une preuve élémentaire utilisant un résultat de [SS04].

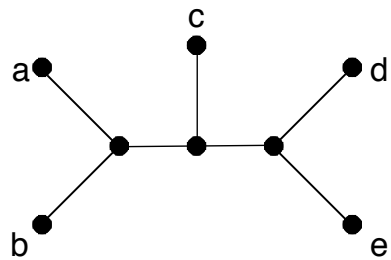
Etant donné un arbre T, plusieurs façons de le dessiner dans le plan (*arbre planaire*). A chaque façon correspond un *ordre circulaire* sur les taxons.



Ordre (*abcde*)



Ordre (*abdec*)



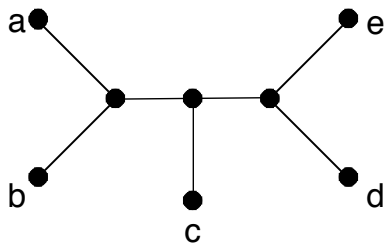
Ordre (*abedc*)

Notation $o(T)$: ensemble des ordres circulaires associés à T.

Consistance (4)

A partir d'un ordre circulaire $\sigma \in o(T)$, avec $\sigma = (x_1 ... x_n)$, on définit :

$$l_\sigma(\delta) = \frac{1}{2} \sum_{i=1}^{n} \delta(x_i, x_{i+1})$$



$$\Rightarrow l_\sigma(\delta) = \tfrac{1}{2}(\delta_{ab} + \delta_{bc} + \delta_{cd} + \delta_{de} + \delta_{ea})$$

$$\sigma = (abcde)$$

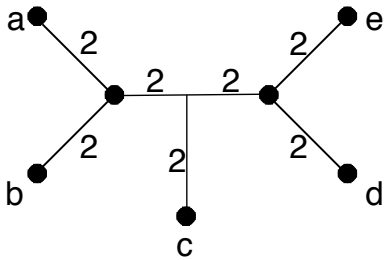
Résultat (Semple & Steel) :

$$L^T(\delta) = \frac{1}{|o(T)|} \sum_{\sigma \in o(T)} l_\sigma(\delta)$$

Consistance (5)

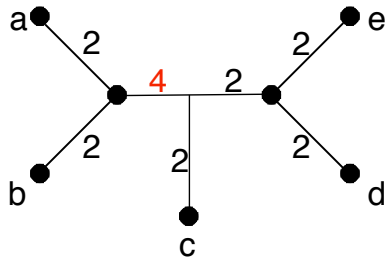
Conséquence : d distance associée à T, L longueur totale, alors

si $\sigma \in o(T)$ alors $l_\sigma(d) = L$



$\sigma = (abcde)$

si $\sigma \notin o(T)$ alors $l_\sigma(d) > L$



$\sigma = (adebc)$

Si W est un arbre quelconque

- si $W = T$ alors $L^T(d) = L$
- si $W \neq T$, alors il existe $\sigma \in o(W) \backslash o(T)$, et donc $L^W(d) > L$.

Rayon de sécurité

Définition : Une méthode de reconstruction a un rayon de securité ρ si : à partir d'une estimation δ d'une distance d^T associée à T, dès que

$$\frac{||\delta - d^T||_\infty}{w_{min}} < \rho$$

alors la méthode infère l'arbre T à partir de δ.

Propriété [Atteson97] : $\rho \leq \frac{1}{2}$.

Résultats pour ME :

- ME+OLS : $\rho \to 0$ quand $n \to \infty$ (new)
  ([Wil05] : $\rho \leq 1/4$)

- ME+BLS : $\rho \geq 1/4$ (new)

ME versus autres méthodes (1)

Minimum d'évolution :

- ME+WLS : inconsistant
- ME+OLS : consistant, rayon 0 (new)
- ME+BLS : consistant, rayon $\geq 1/4$ (new)

Autres méthodes :

- moindres carrés : consistant, NP-difficile [Day86]
- parcimonie : inconsistant, NP-difficile [DJS86]
- maximum de vraisemblance : consistant, NP-difficile [CT05]

ME versus autres méthodes (2)

Algorithmes :

- algorithme GME [DG02] : deux versions
  - GME+OLS ("FastME") : consistant, rayon 0 (new) complexité en temps : $O(n^2)$
  - GME+BLS ("BME") : consistant, rayon 1/2 (new) complexité en temps : $O(n^3)$
- algorithme NJ [SN81] : consistant, rayon 1/2 [Atteson97] complexité en temps : $O(n^3)$

Remarque : Ces deux algorithmes peuvent être vus comme des heuristiques gloutonnes pour ME.


R. Desper and O. Gascuel.
Theoretical foundation of the balanced minimum evolution ...
Molecular Biology and Evolution, 21 :587–598, 2004.


O. Gascuel, D. Bryant, and F. Denis.
Strengths and limitations of the minimum-evolution principle.
Systematic Biology, 50(5) :621–627, 2001.


A. Rzhetsky and M. Nei.
Theoretical foundation of the minimum evolution method.
Molecular Biology and Evolution, 10 :1073–1095, 1993.


C. Semple and M. Steel.
Cyclic permutations and evolutionary trees.
Advances in Applied Mathematics, 32 :669–680, 2004.


S.J. Willson.
Minimum evolution using ordinary least squares is less robust than neighbor-joining.
Bulletin of Mathematical Biology, 67 :261–279, 2005.