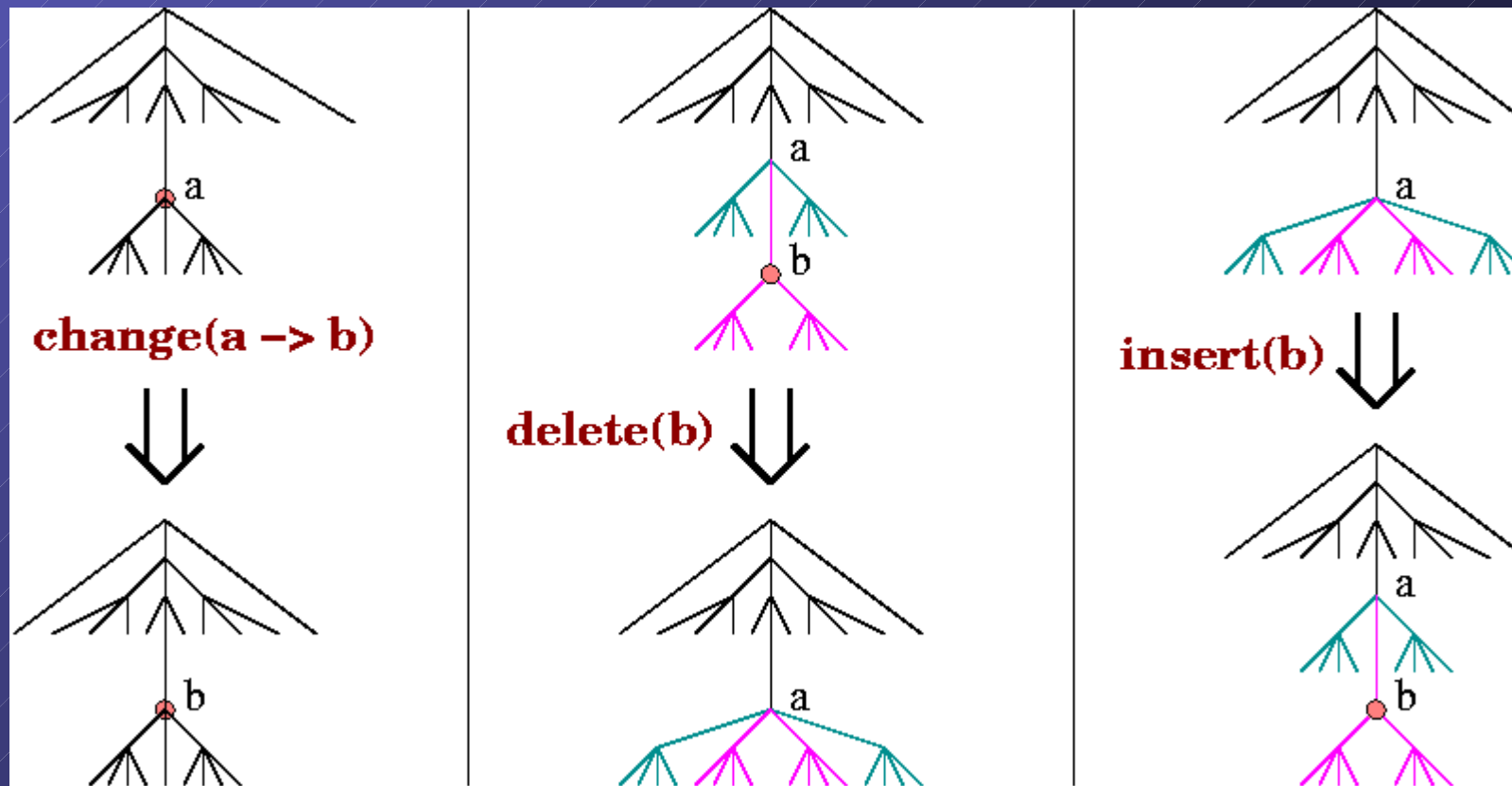


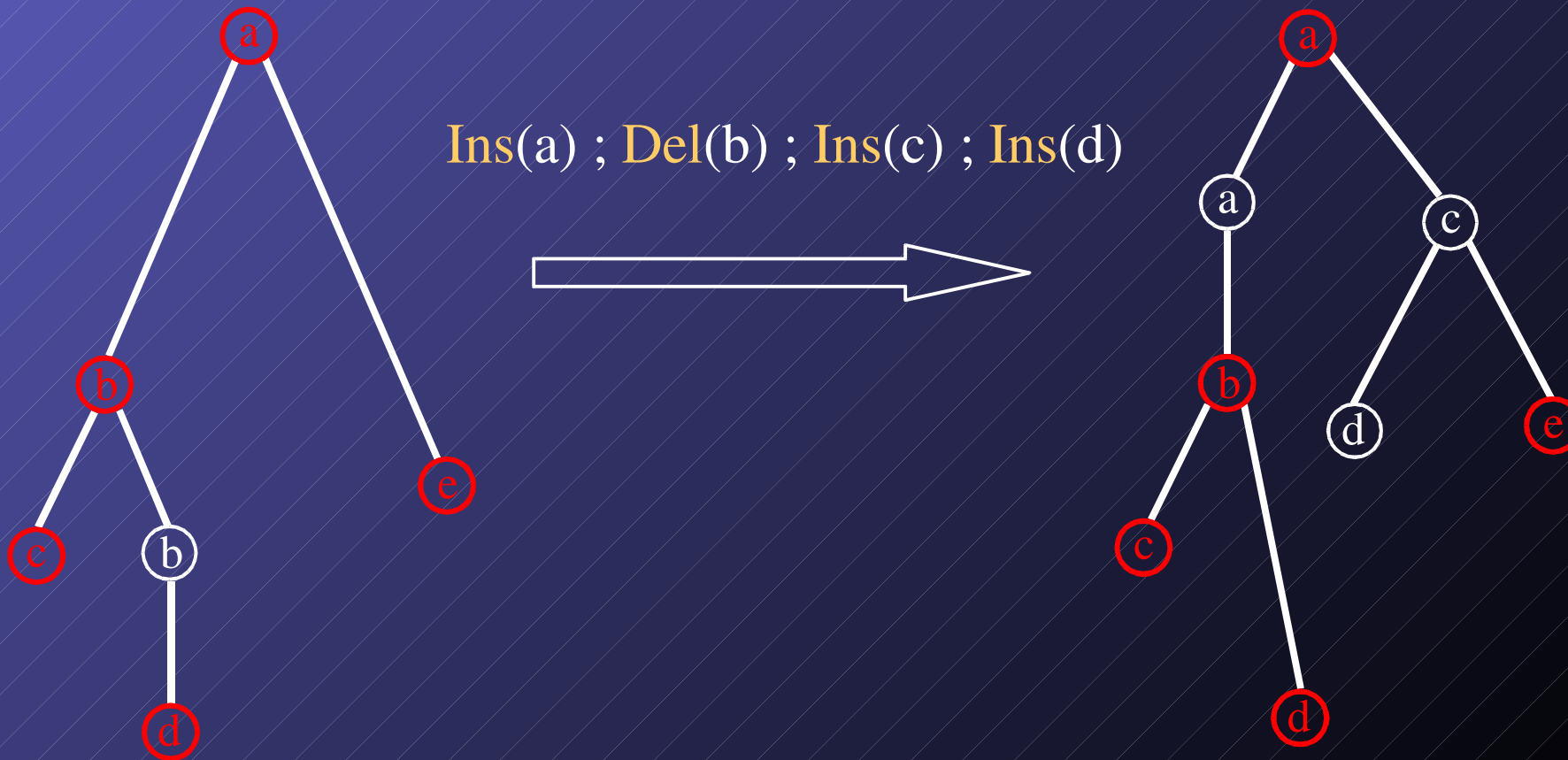
Comparaison d'arborescences ordonnées étiquetées

Comparaison de deux arborescences

Opérations de base



« *Transformation* » d'un arbre utilisant ces opérations de base



Matrices de substitution

Sur l'alphabet de symboles suivants (étiquettes des nœuds des arbres) :

- A, C, G, U,
- AU, UA, CG, GC, GU, UG, AG, GA, AC, CA, ...
- H, B5', B3', S, I, M
- + mots représentant les bases composant l'élément de structure
- + symboles d'éléments de structure tertiaire

⇒ détermination de scores de substitution à partir de structures secondaires d'ARN alignées par les biologistes

Comment organiser les calculs ?

Recours à la
"*programmation dynamique*"
utilisée dans le cas de
l'alignement de séquences

$$Tscore \left[\begin{array}{c} \bullet \\ \diagup \quad \diagdown \\ \triangle \quad \triangle \quad \triangle \end{array}, \begin{array}{c} \bullet \\ \diagup \quad \diagdown \\ \triangle \quad \triangle \end{array} \right]$$

$$= \max \left\{ \begin{array}{l} Fscore \left[\triangle \triangle \triangle, \triangle \triangle \right] + Subst \left[\bullet, \bullet \right] \\ Fscore \left[\triangle \triangle \triangle, \begin{array}{c} \bullet \\ \diagup \quad \diagdown \\ \triangle \quad \triangle \end{array} \right] + Del \left[\bullet \right] \\ Fscore \left[\begin{array}{c} \bullet \\ \diagup \quad \diagdown \\ \triangle \quad \triangle \quad \triangle \end{array}, \triangle \triangle \right] + Ins \left[\bullet \right] \end{array} \right\}$$

$$Fscore \left[\begin{array}{c} \triangle \quad \triangle \quad \triangle \end{array}, \begin{array}{c} \triangle \quad \triangle \end{array} \right]$$

$$= \max \left\{ \begin{array}{l} Fscore \left[\begin{array}{c} \triangle \quad \triangle \quad \triangle \end{array}, \triangle \right] + Tscore \left[\begin{array}{c} \triangle \quad \triangle \quad \triangle \end{array}, \begin{array}{c} \triangle \quad \triangle \end{array} \right] \\ Fscore \left[\begin{array}{c} \triangle \quad \triangle \quad \triangle \end{array}, \begin{array}{c} \triangle \quad \triangle \quad \triangle \end{array} \right] + Del \left[\begin{array}{c} \bullet \end{array} \right] \\ Fscore \left[\begin{array}{c} \triangle \quad \triangle \quad \triangle \end{array}, \begin{array}{c} \triangle \quad \triangle \quad \triangle \end{array} \right] + Ins \left[\begin{array}{c} \bullet \end{array} \right] \end{array} \right\}$$

Édition d'arborescences ordonnées étiquetées

Plusieurs algorithmes de comparaison d'arborescences basés sur la notion d'édition implémentés :

- algorithme d'édition d'arborescences (Zhang –Shasha 1989)
- algorithme d'alignement d'arborescences (Jiang-Wang-Zhang 1994)
- versions « locales » de ces algorithmes
- versions minimisant le nombre de composantes connexes

En pratique

Comparaison d'arborescences de 400 sommets

⇒ Structures secondaires de 600 nucléotides codées très finement
(modèle 2)

⇒ Structures secondaires de 4000 nucléotides codées grossièrement
(modèle 6)

Un exemple d'édition de structures secondaires

Comparaison de C. Trachomatis et de H. Cutirubrum

408 nucléotides

375 nucléotides

Modèle grossier (6)

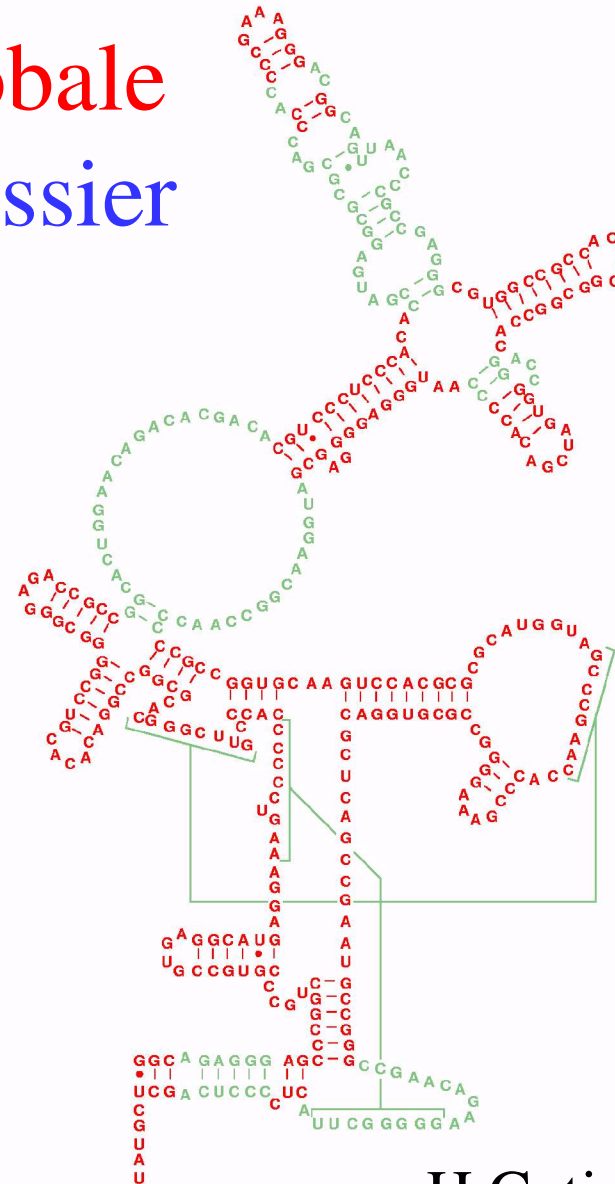
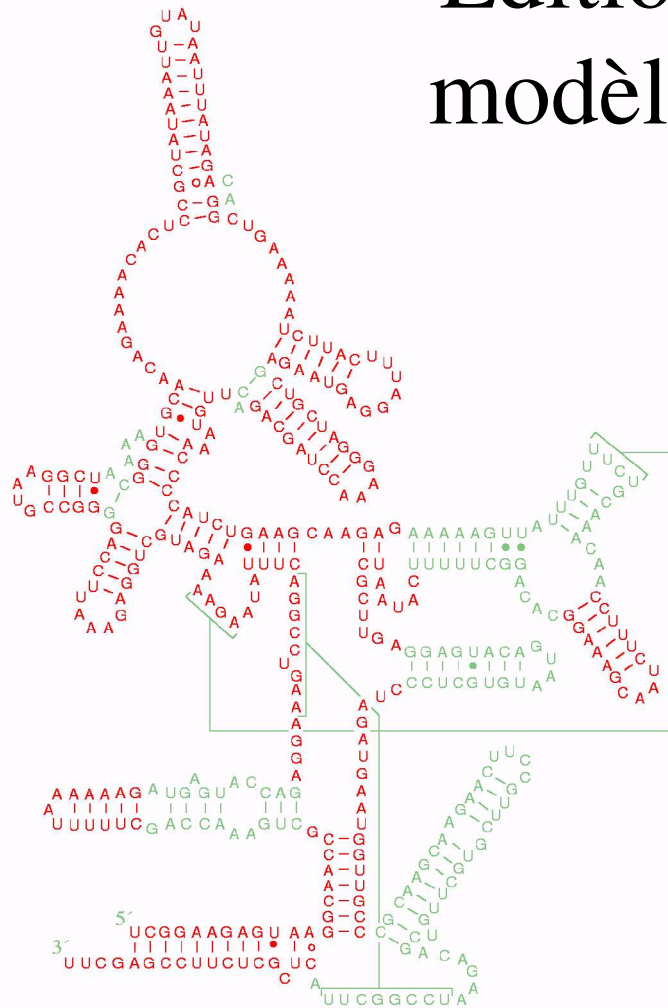
VS

Codage (modèle 2)

31/51 sommets

31/41 sommets

Edition **globale** modèle **grossier**

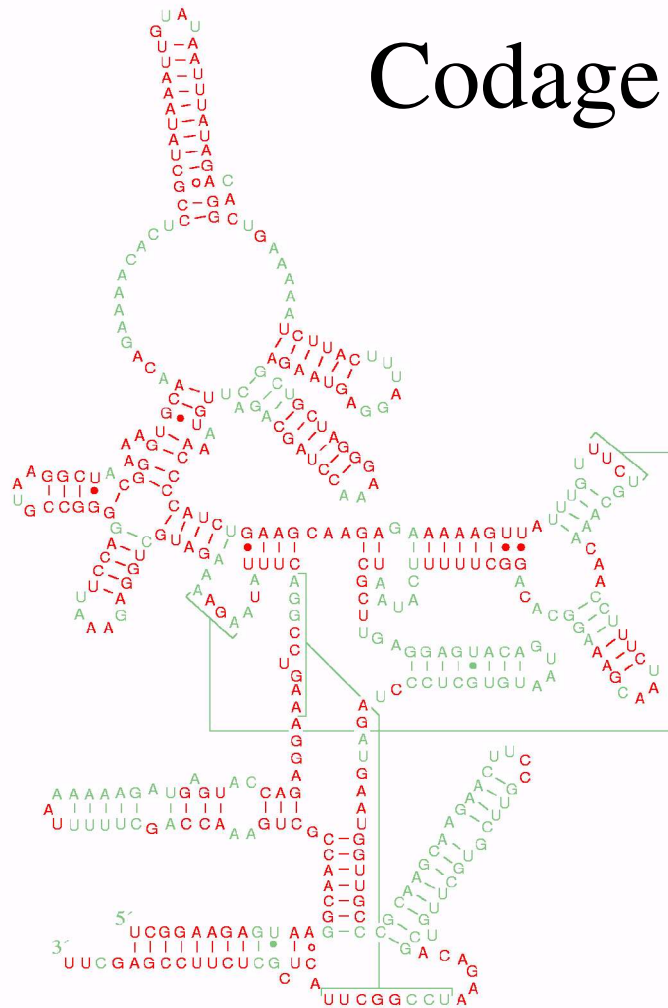


148/306 sommets

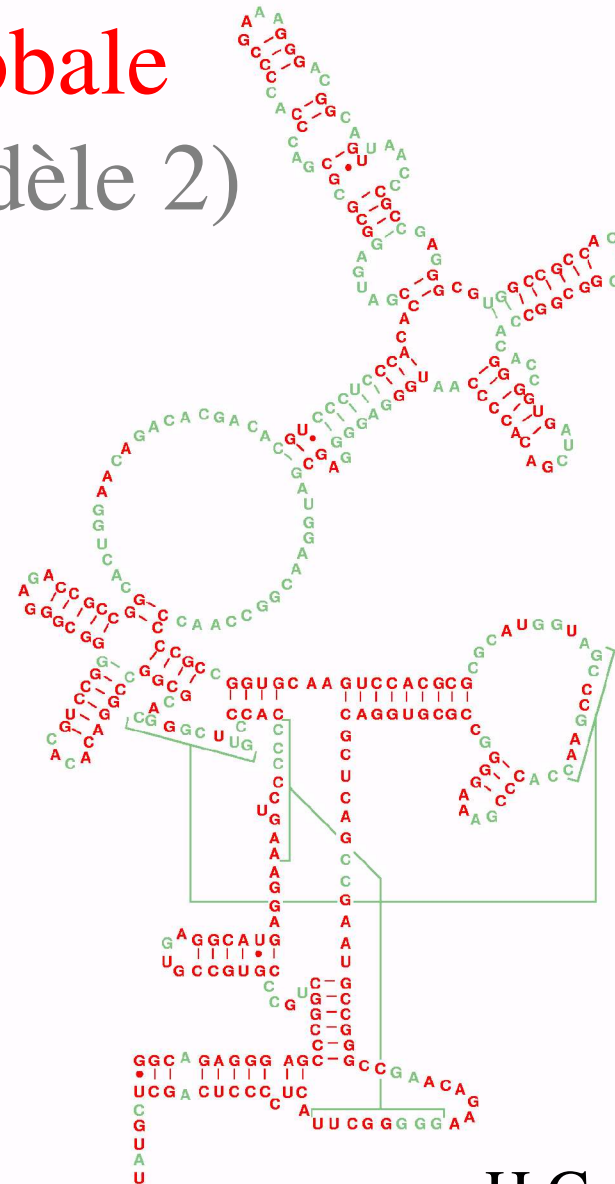
148/300 sommets

Edition globale

Codage (modèle 2)



C. Trachomatis



H. Cutirubrum

Modèle Grossier vs Codage

Modèle grossier :

- rapide, peu gourmand en espace mémoire (+)
- alignement relativement dense (+)
- comparaison peu précise (−)

Codage :

- résultat plus précis (+)
- alignement dispersé (−)
- résultat souvent non concluant (−)

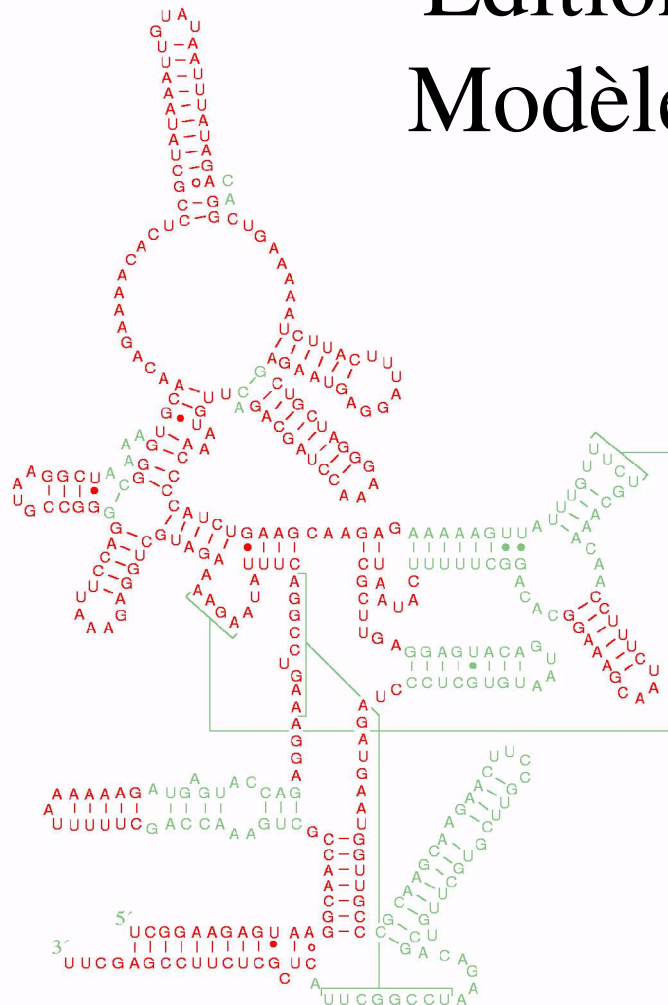
Edition Globale

VS

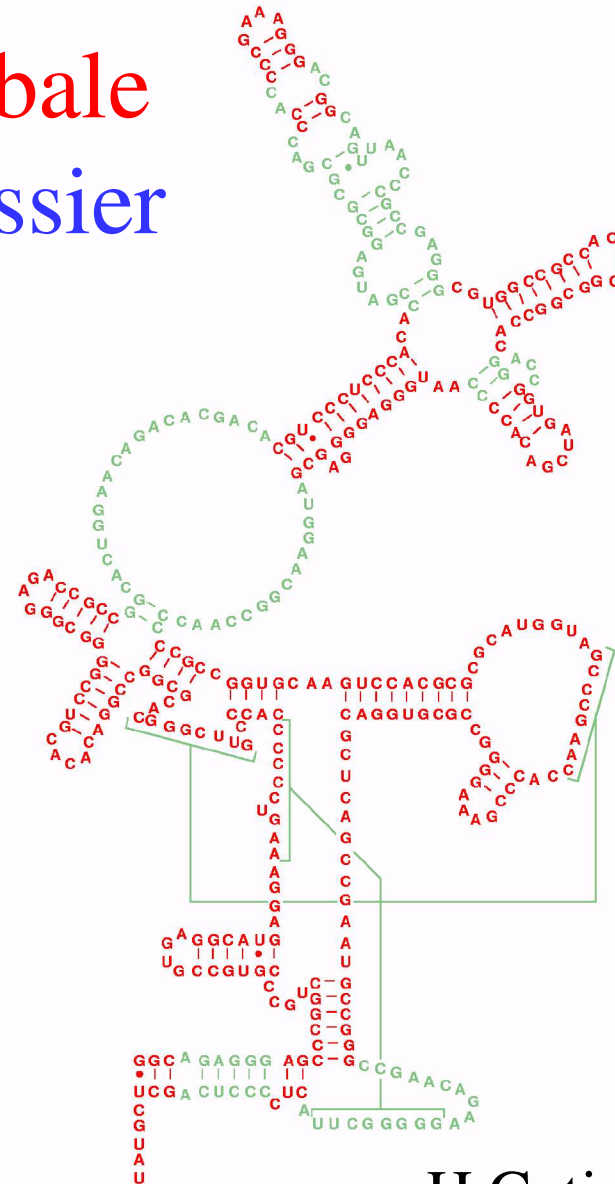
Edition Locale

31/41 sommets

Edition **globale**
Modèle **grossier**



C.Trachomatis



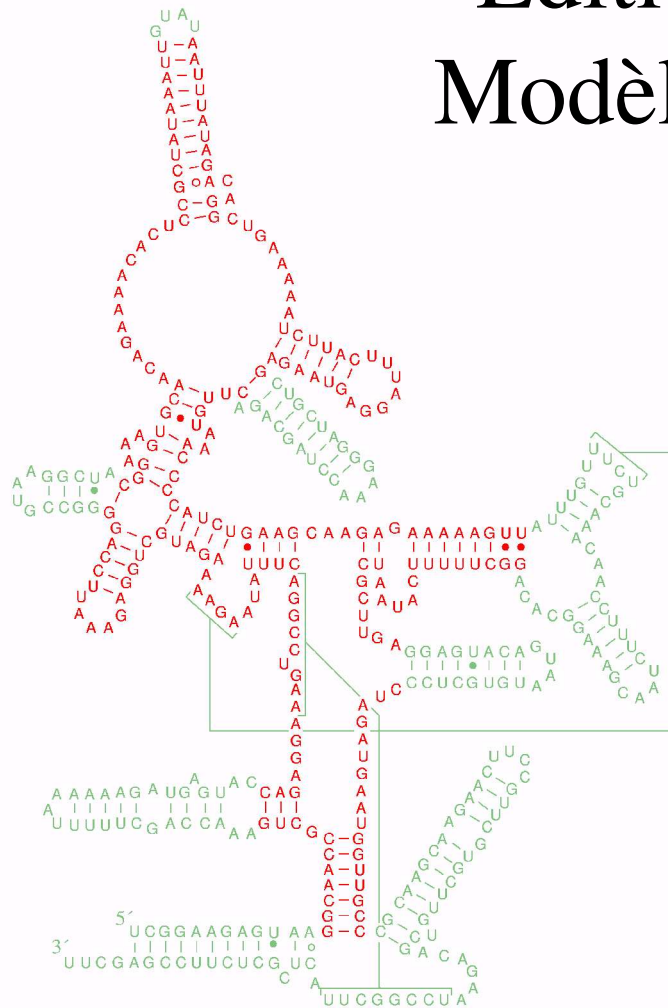
H.Cutirubrum

25/51 sommets

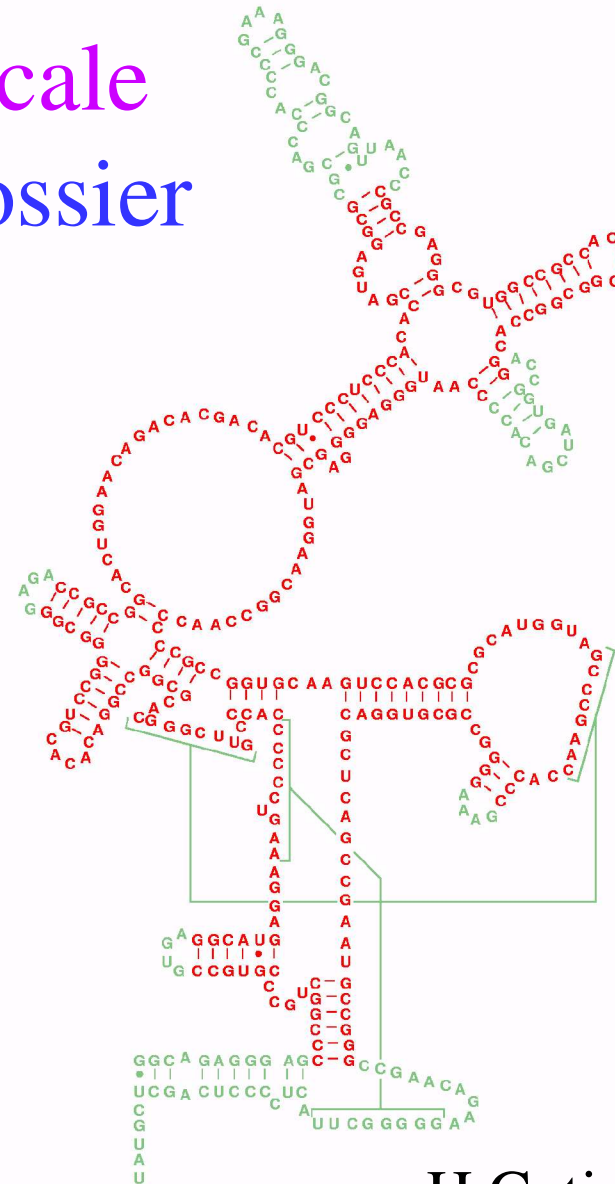
25/41 sommets

Edition locale

Modèle grossier



C. Trachomatis



H. Cutirubrum

Edition Globale vs Edition Locale

L'**édition locale** met en évidence une partie commune plus réduite (en nombre de nucléotides) que l'édition globale.

Cependant, les nucléotides alignés sont **moins dispersés**, ce qui permet de mettre en évidence un motif précis.

Edition Globale

 V_S

Edition Globale avec Densification

(minimisation du nombre de composantes
connexes)

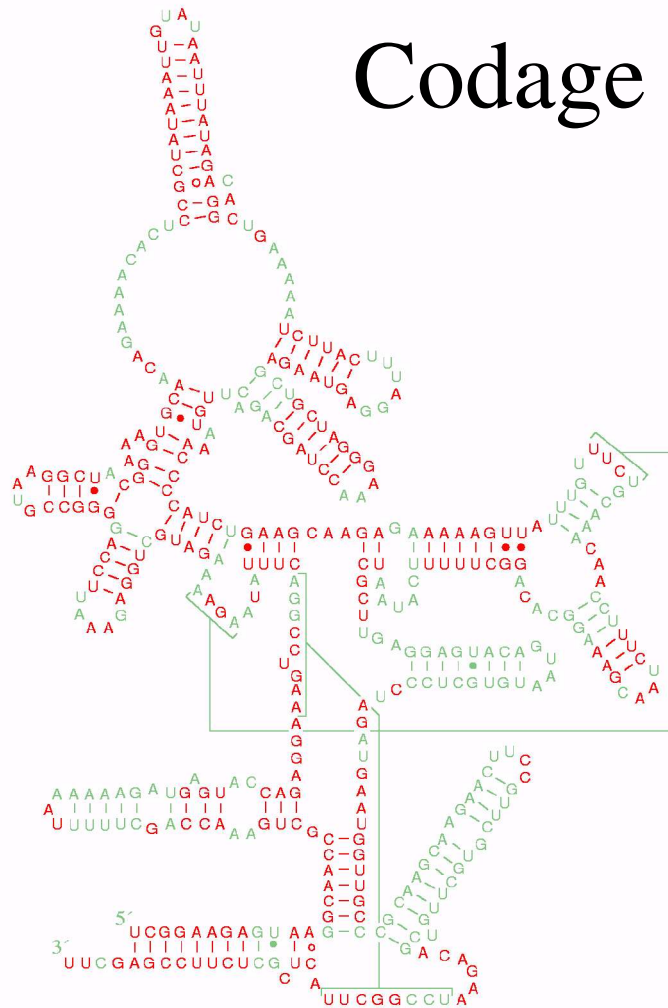


148/306 sommets

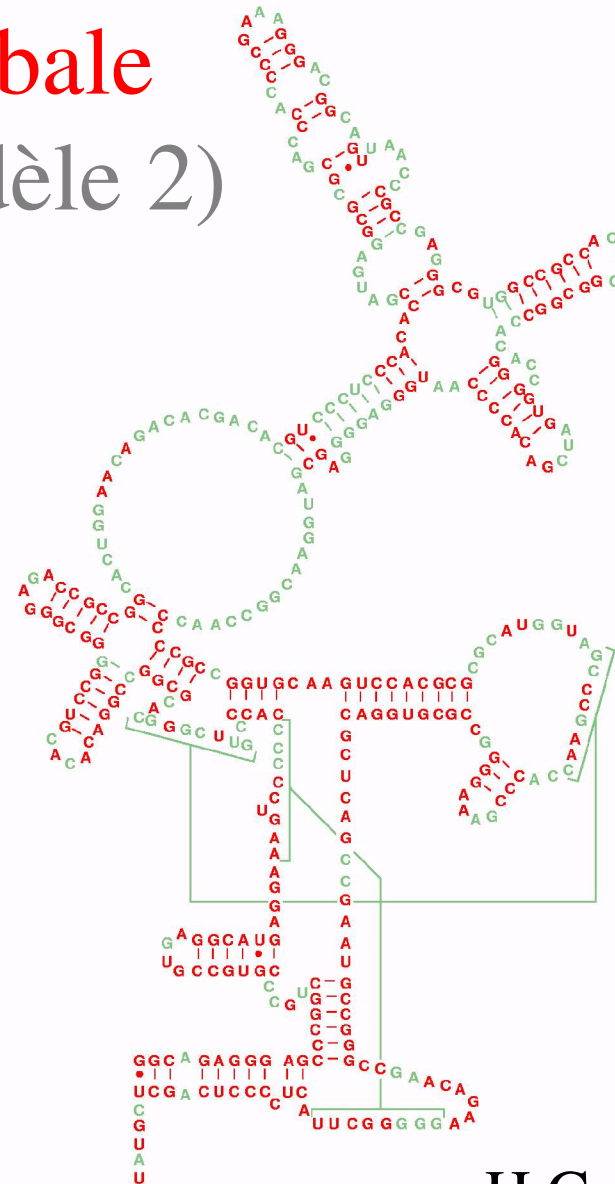
148/300 sommets

Edition globale

Codage (modèle 2)



C.Trachomatis

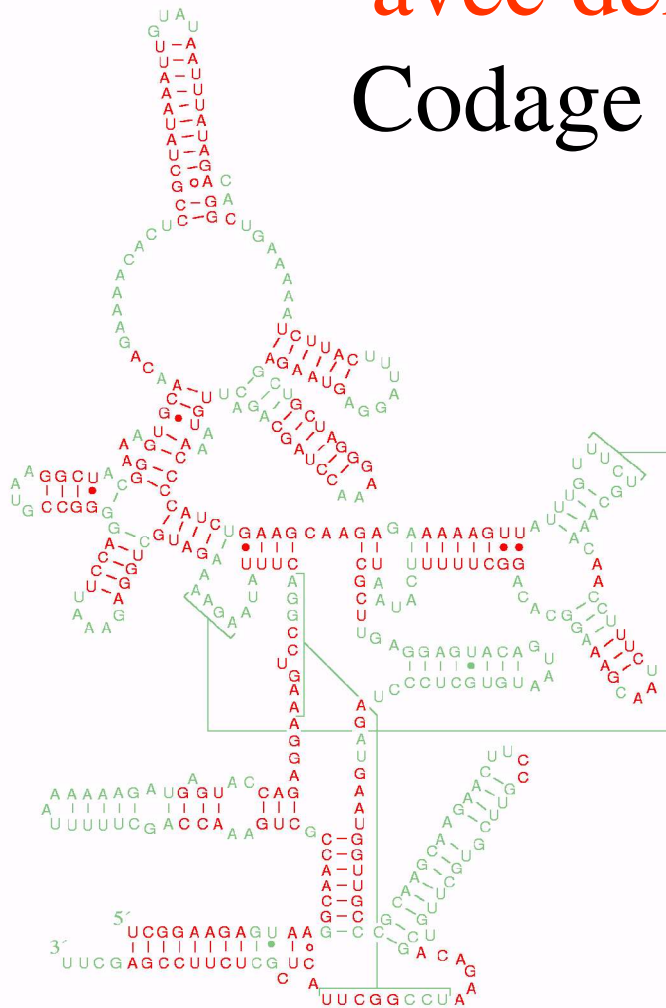


H.Cutirubrum

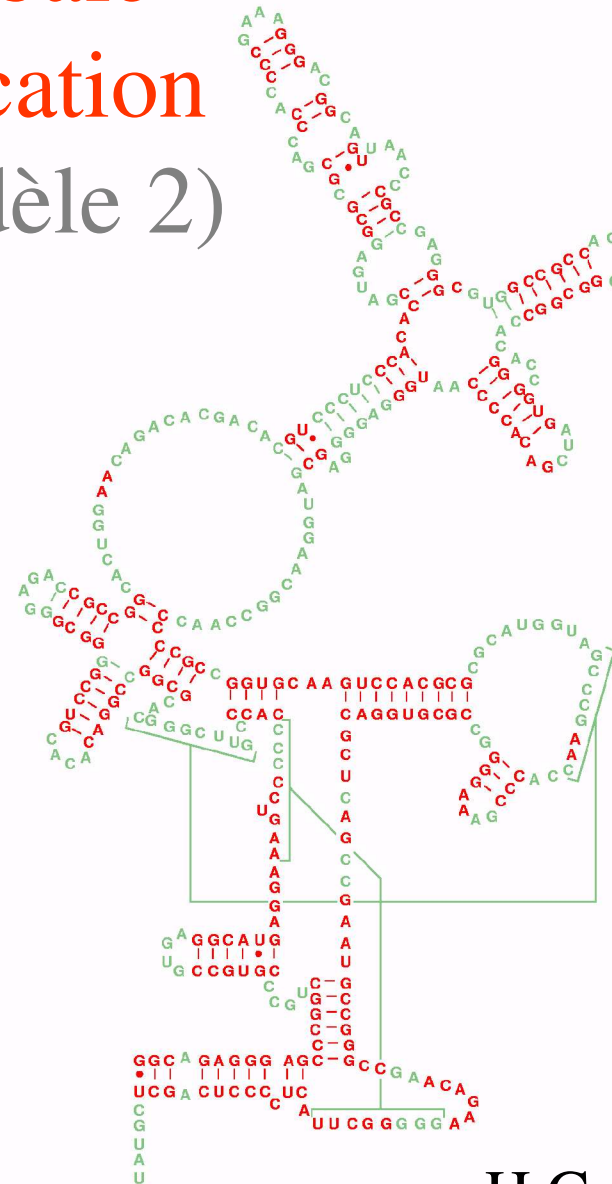
117/306 sommets

Edition globale avec densification Codage (modèle 2)

117/300 sommets



C. Trachomatis



H. Cutirubrum

Avec vs Sans Densification

L'édition globale avec densifications permet :

- de **maximiser le nombre de zones mises en évidence** (édition globale)
- d'obtenir un alignement précis
- et d'**éviter d'obtenir une trop grande dispersion des régions conservées** (due au modèle fin)

Problème principal :

- Trouver les « bons » scores et pénalités à appliquer

Comparaison d'arbres

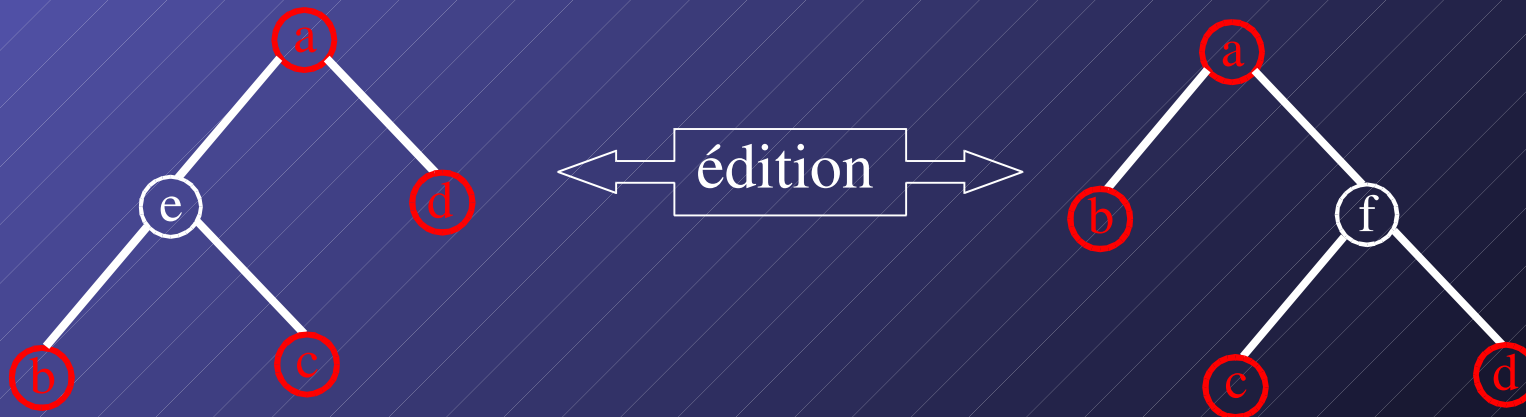
⇒ Édition d'arbres

⇒ Alignement d'arbres

Ces deux notions, **identiques** dans le cas des **séquences**, conduisent à deux problèmes **différents** dans le cas des **arbres**...

Édition d'arbre
 $\neq$
Alignement d'arbres

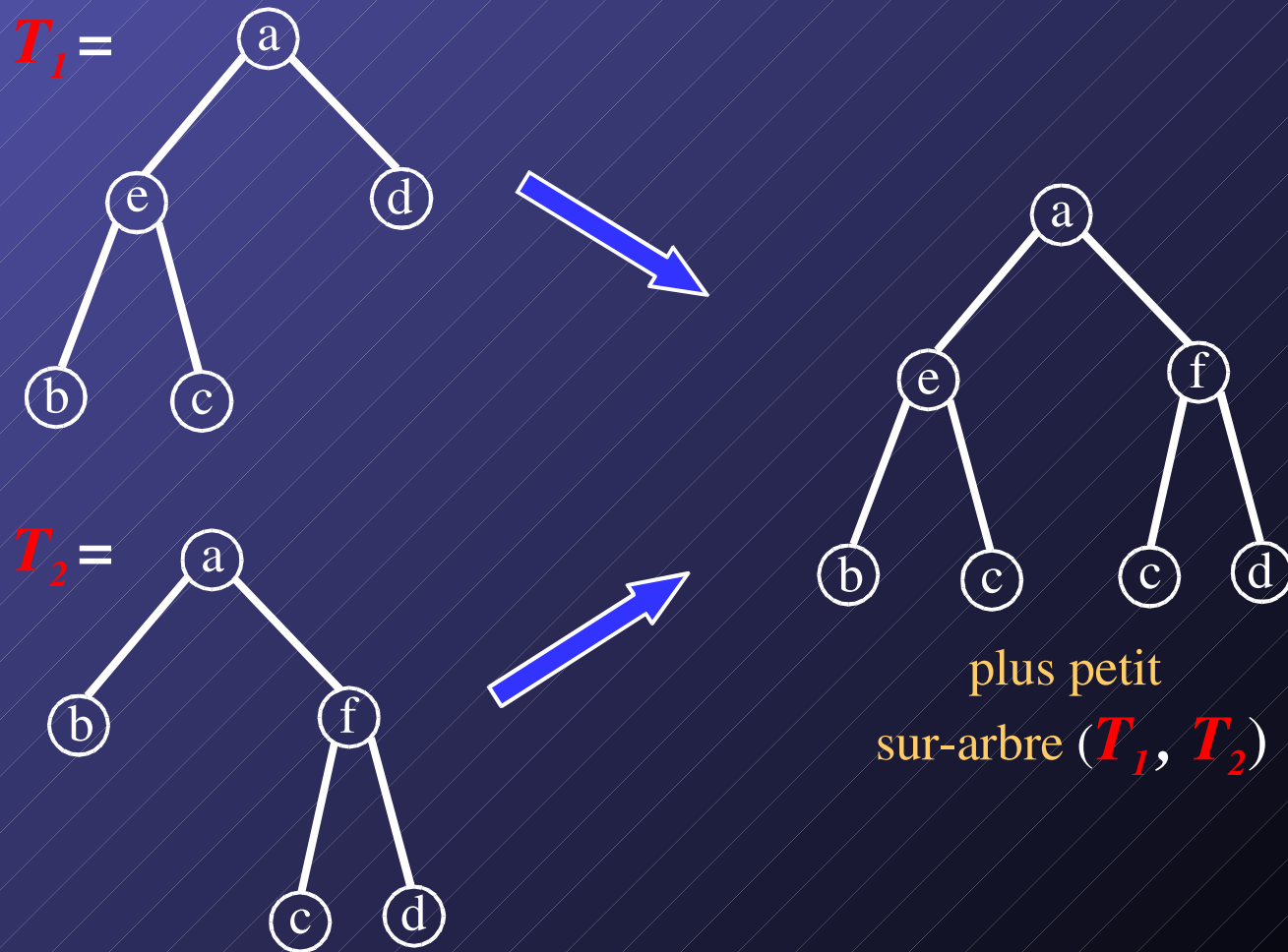
Édition d'arbres



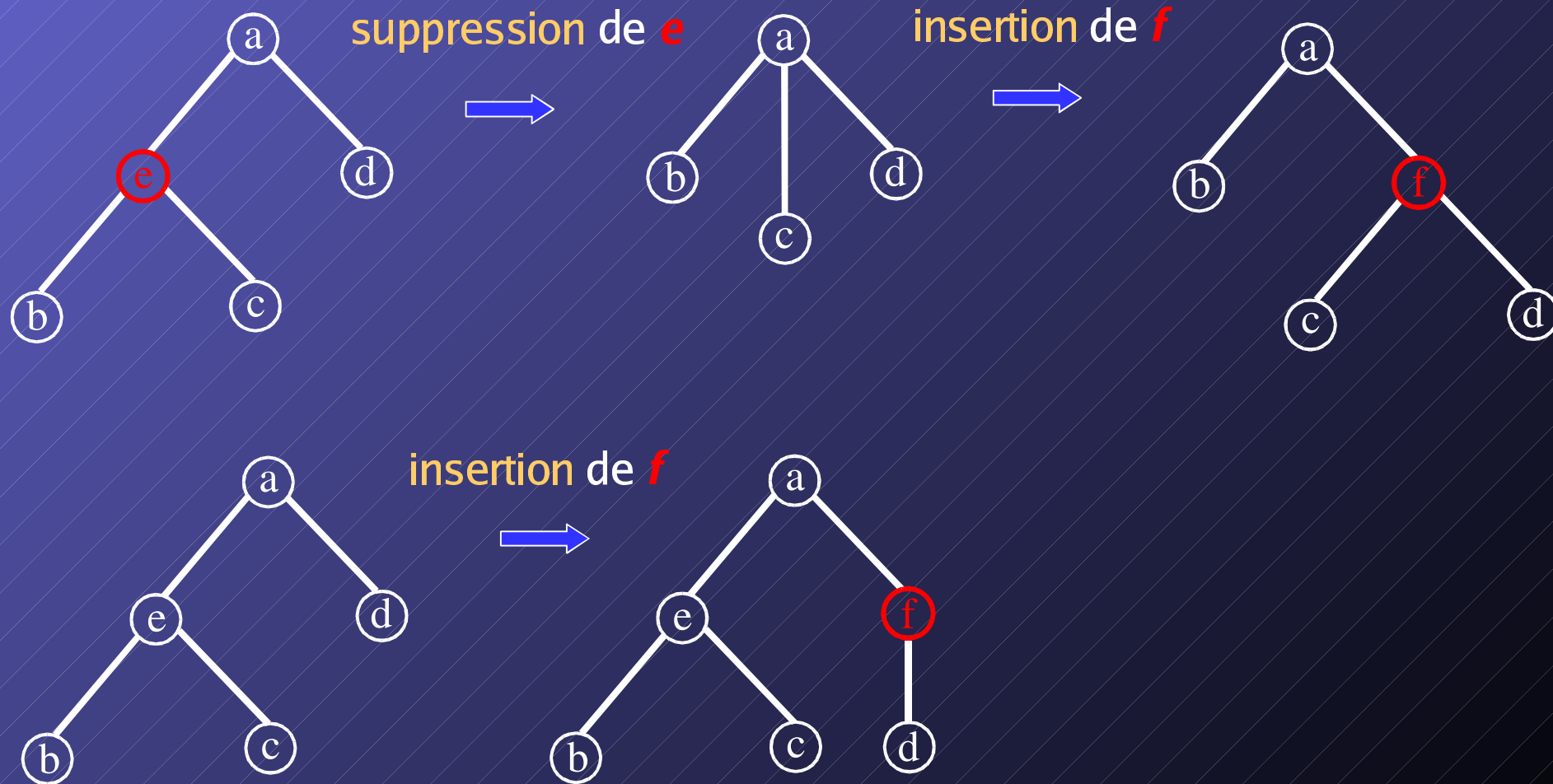
nœuds conservés entre les deux arbres

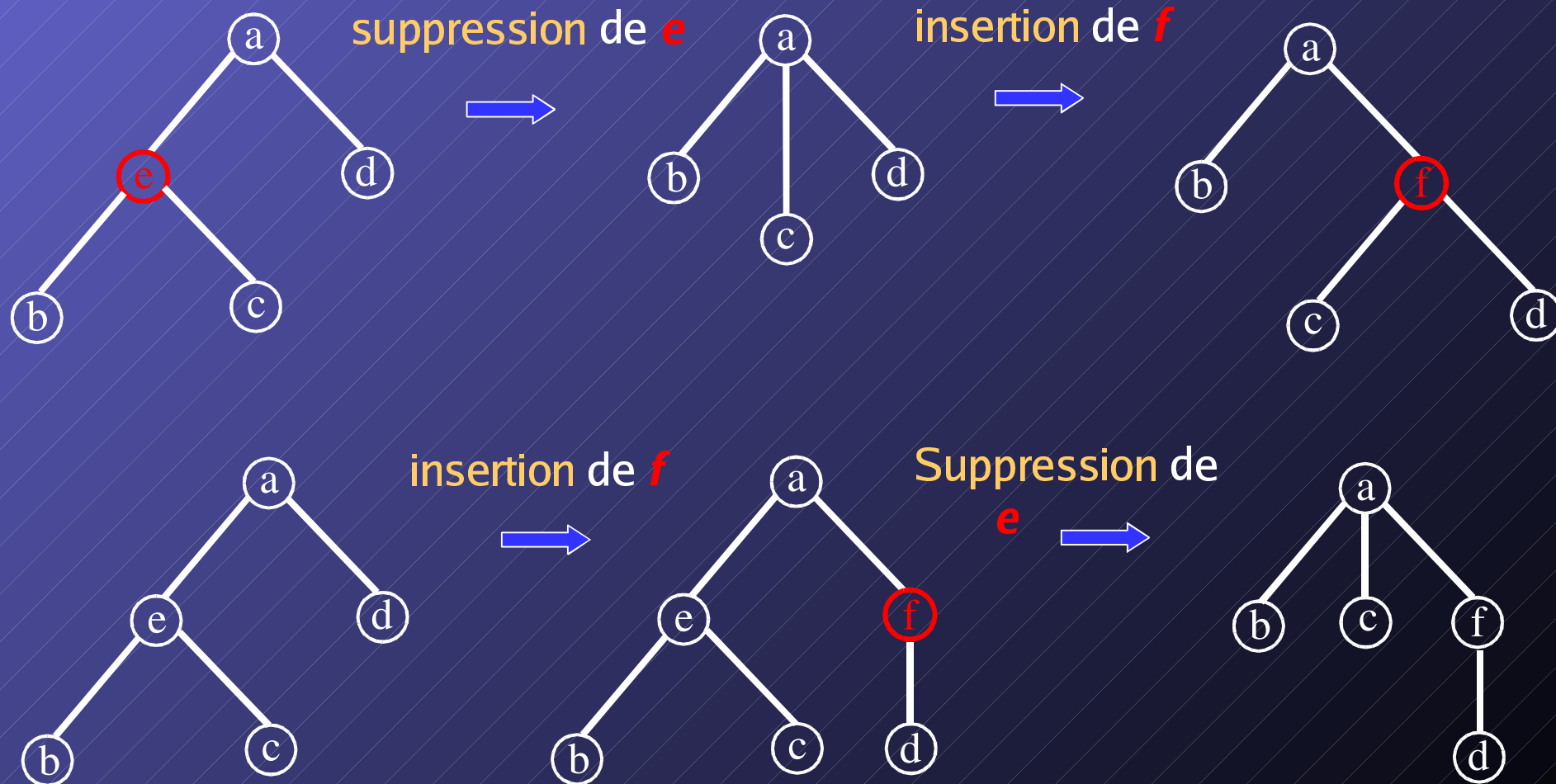
Alignement d'arbres

≡ trouver le **plus petit sur-arbre commun**

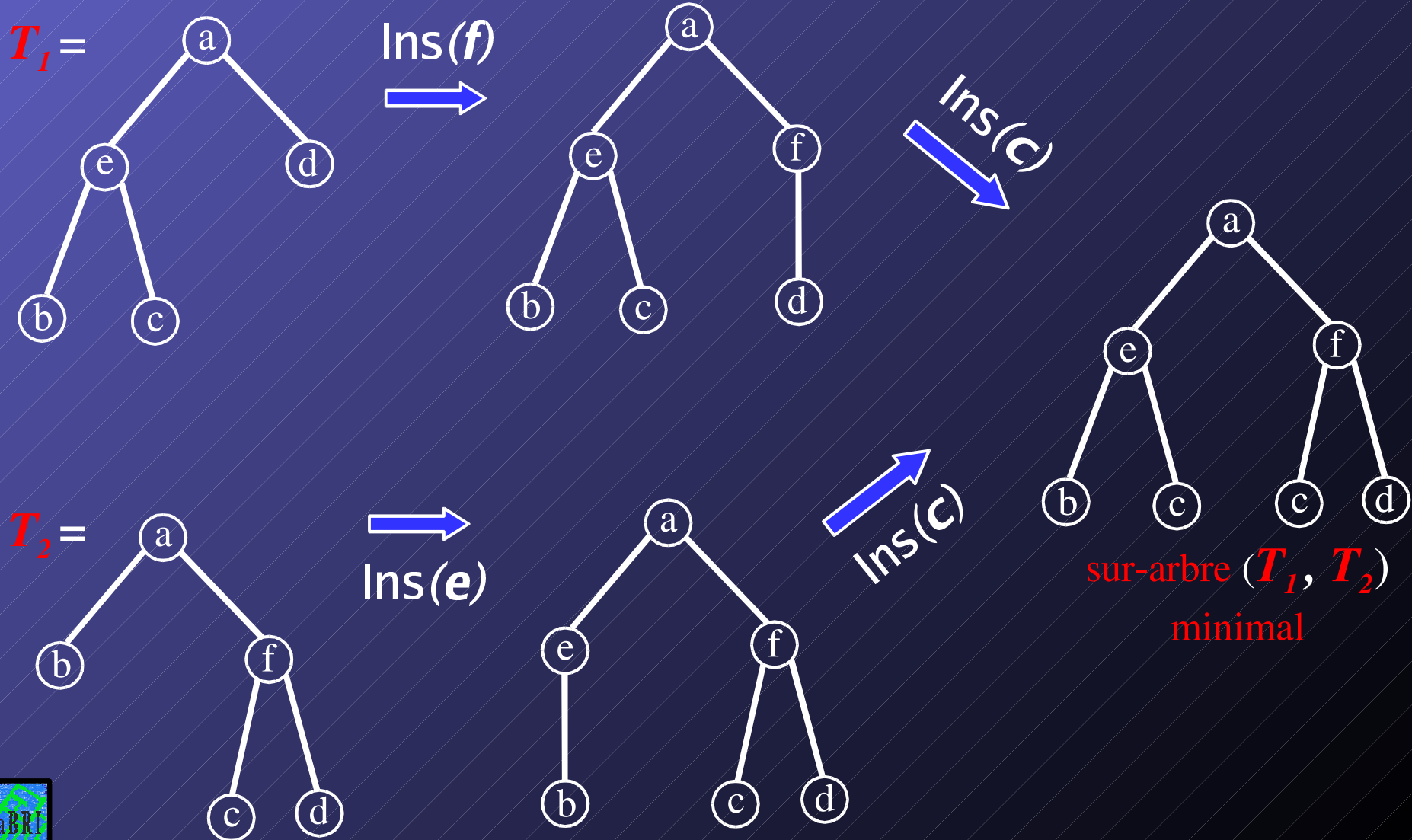


Dans le cas des arbres,
une suppression et une insertion
consécutives
ne commutent pas forcément





Alignement d'arbres $\Leftrightarrow$ "superposition"



Alignement d'arbres = les insertions suivies des suppressions

