

Génomique comparative

Nadia El- Mabrouk

I. Introduction

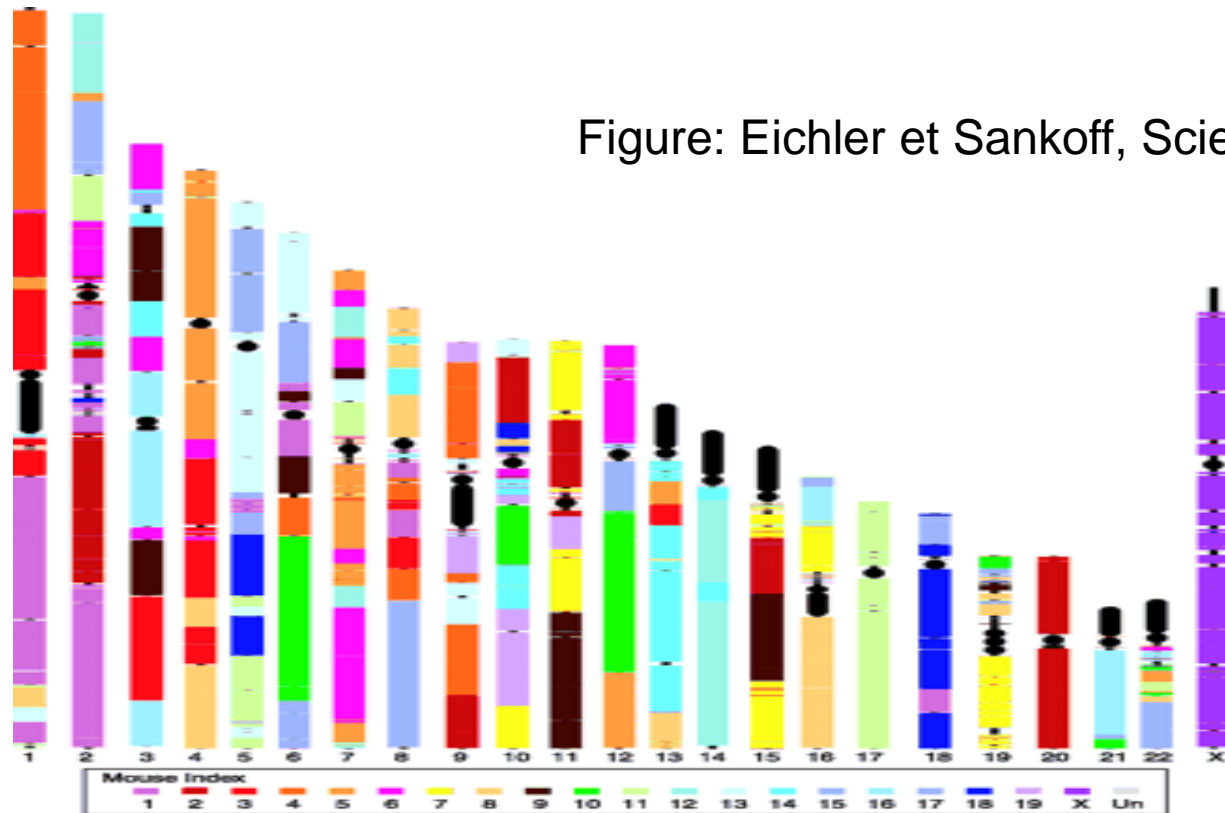
Les génomes évoluent par:

- **Mutations locales**: Au niveau de la séquence; substitutions, insertions, suppressions de nuc.
- **Mutations globales**: Au niveau du génome; insertions, suppressions, duplications, déplacements de gènes ou de fragments de chromosomes

Pour étudier les mutations globales:

- Exploiter l'information contenue dans tout le génome.
- Considérer la structure générale du génome (linéaire/circulaire, uni-chromosomique/multi-chromosomique).
- Représenter un chromosome par un **ordre de gènes** (ou autres éléments constitutifs, ou blocs conservés).
- Comparer deux génomes revient à comparer des ordres de gènes (ou des ordres de blocs).

Mutations globales



Conserved syntenic blocks from the mouse genome (MGSCv. 3.0) are overlaid on human chromosomes (April 2003, assembly). All conserved syntenic blocks >10 kb are shown.

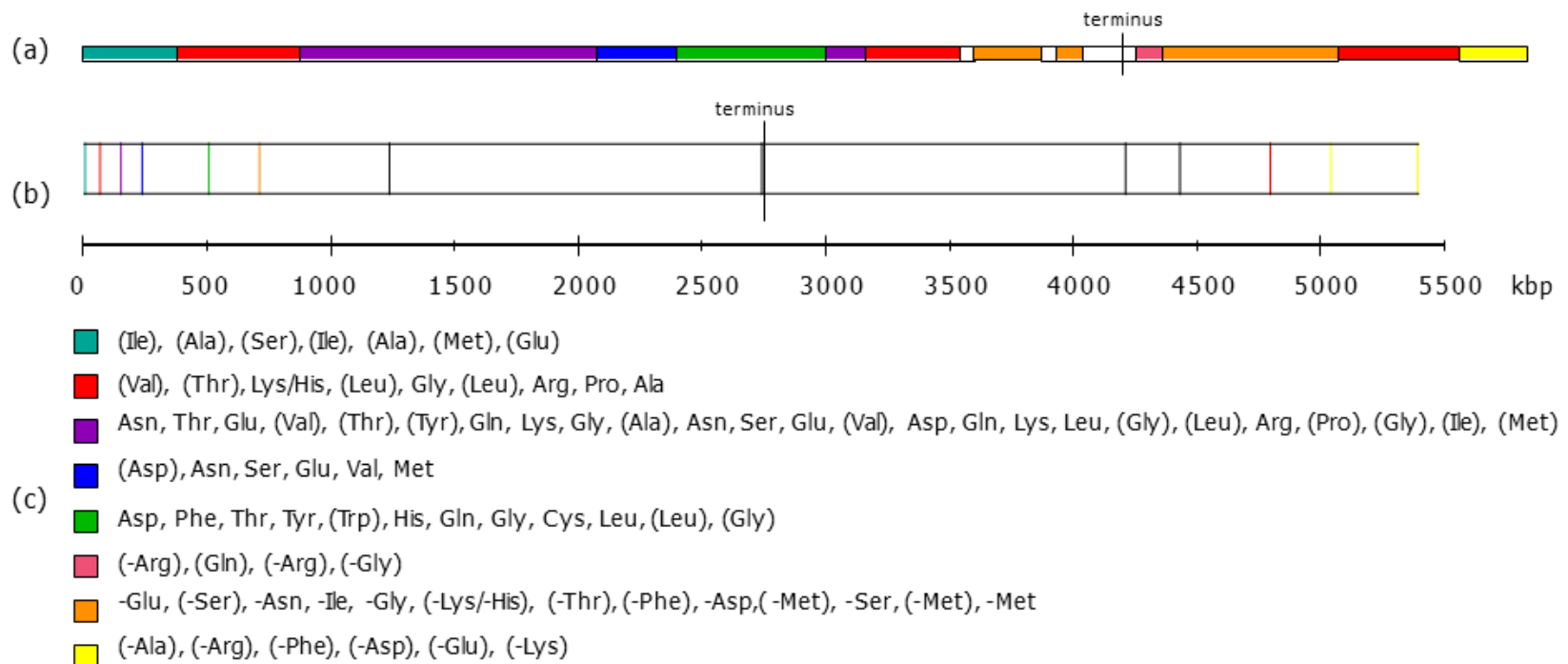
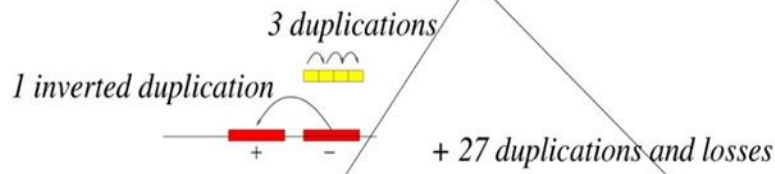
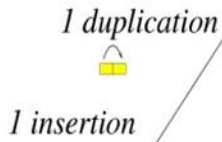


Figure 3: (a) Representation of the *Bacillus cereus* ATCC 14579 genome. (b) Locations of the tRNA gene blocks on the whole genome of *Bacillus cereus* ATCC 14579. Black bars represent the white variable blocks. (c) Description of the consensus sequence of tRNA genes contained in each syntenic block. A slash (/) between two tRNA genes indicates that one or the other can be found at that position. The tRNA genes inside parentheses can be absent from the block in some strains.

Bacillus



Dup+loss cost = 31



anthracis str. 'Ames Ancestor'



Dup+loss cost = 2



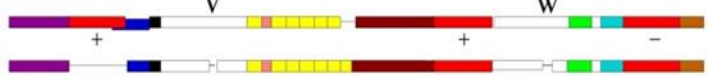
1 inverted duplication



3 insertions



cereus ATCC 14580



Dup+loss cost = 5

cereus E33L

1 duplication

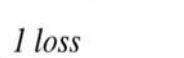


2 losses

1 inverted duplication



1 duplication

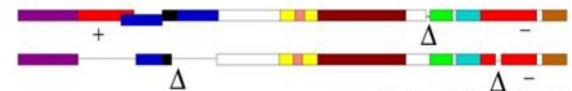


1 loss



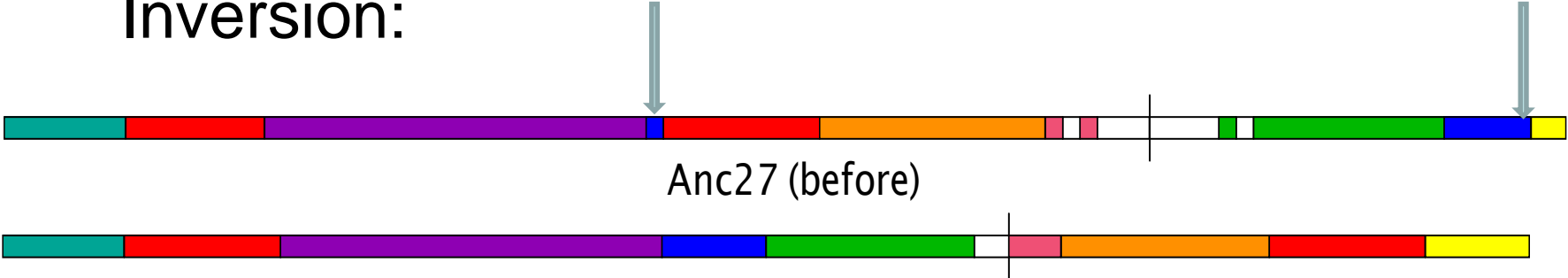
subtilis str. 168

Dup+loss cost = 5

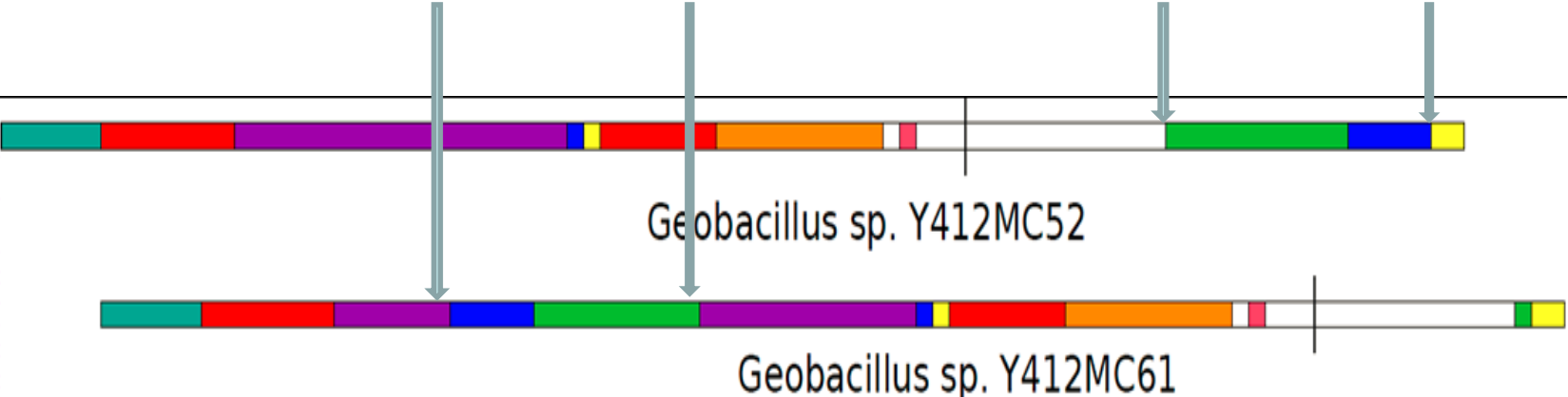


licheniformis ATCC 14580

Inversion:



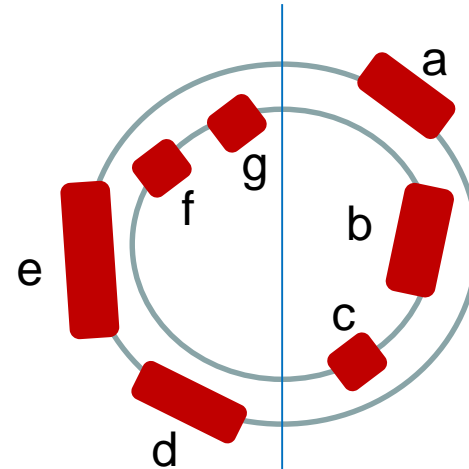
Transposition inversée:



Types de génomes

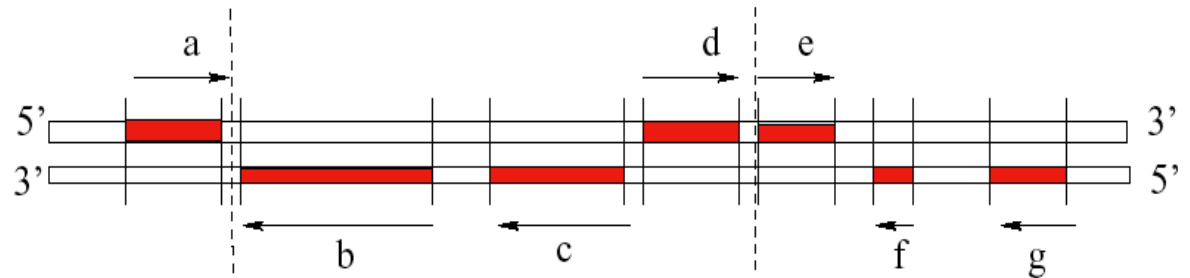
1. Génome circulaire

1. Ordre des gènes signé
2. Non signé



2. Génome linéaire

1. 1 ou plusieurs chromosomes
2. Signé
3. Non signé



+a -b -c +d +e -f -g

Types de mutations génomiques

- Réarrangements Intra-chromosomales:

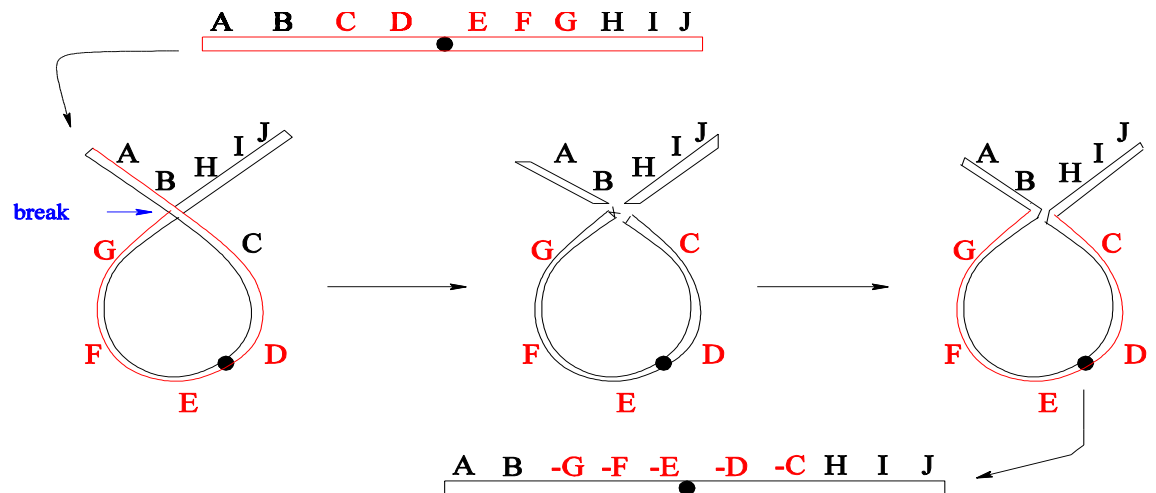
- Inversion:

a b | c d e | f g h i j

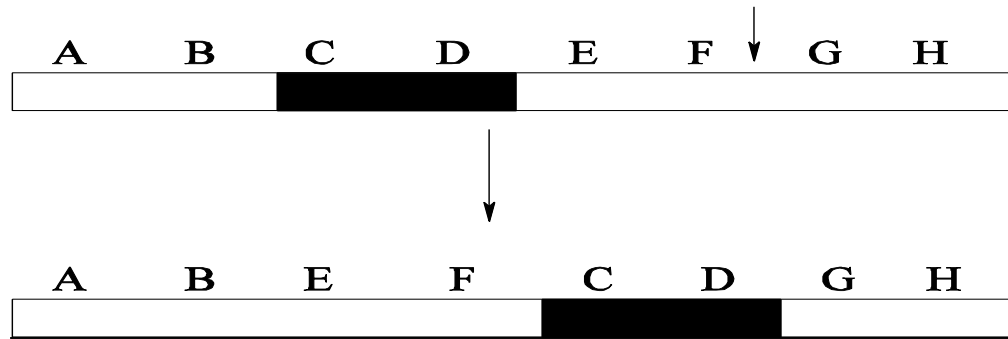


a b -e -d -c f g h i j

Origine possible:
Erreur de réplication



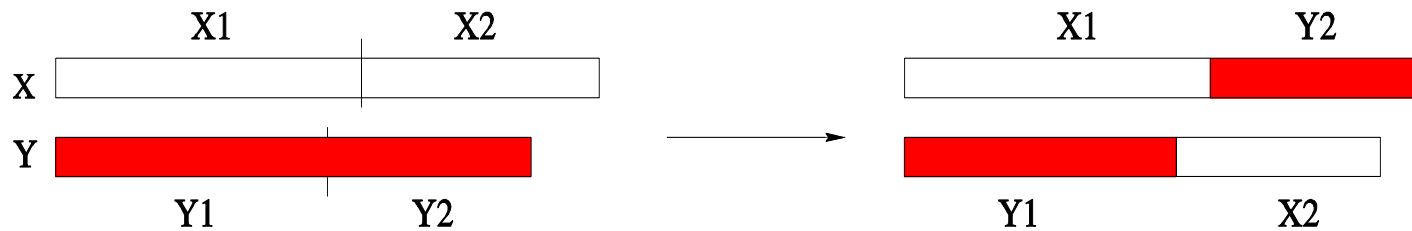
- **Transposition:** Segment supprimé et réinséré à un autre endroit dans le génome



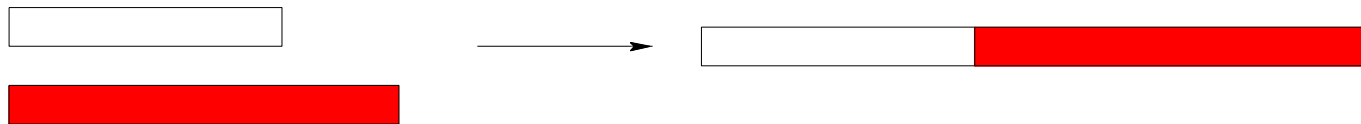
Réarrangements inter-chromosomiques:

- Translocation, fusion, fission

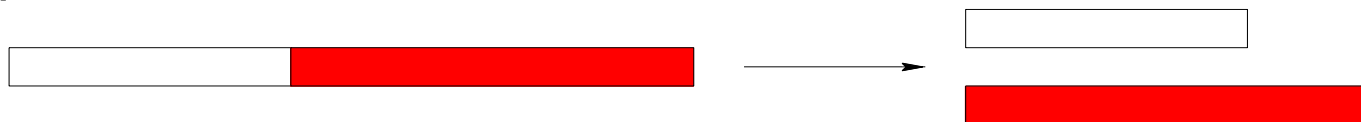
Translocation réciproque:



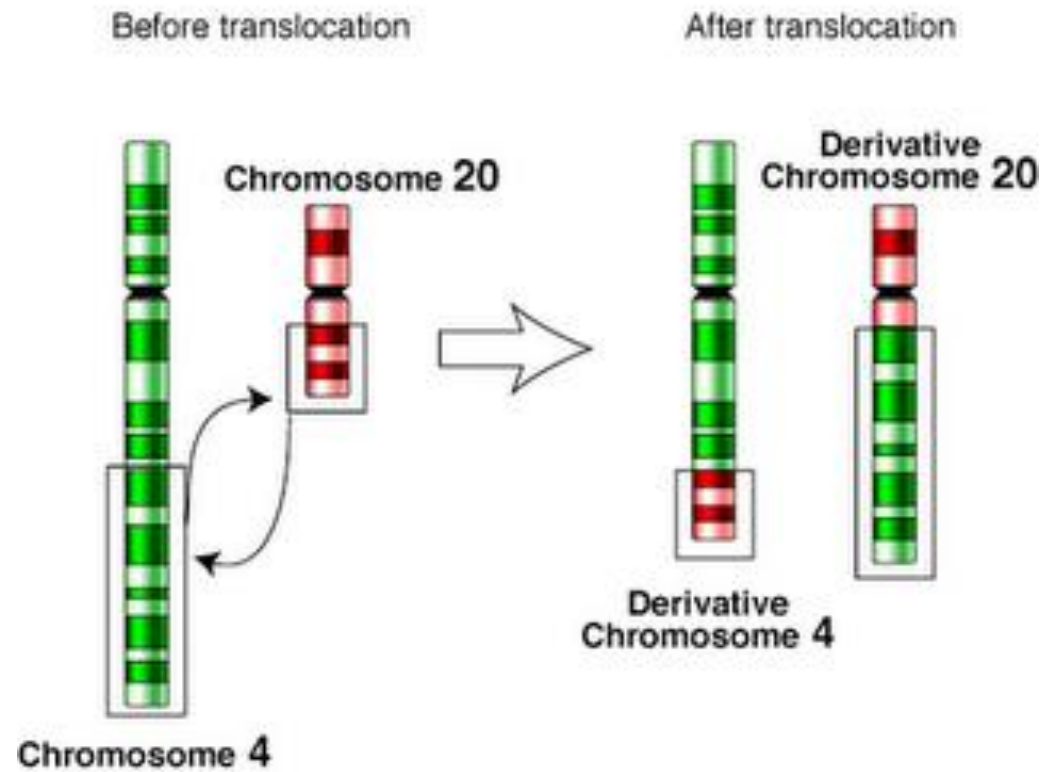
Fusion:



Fission:

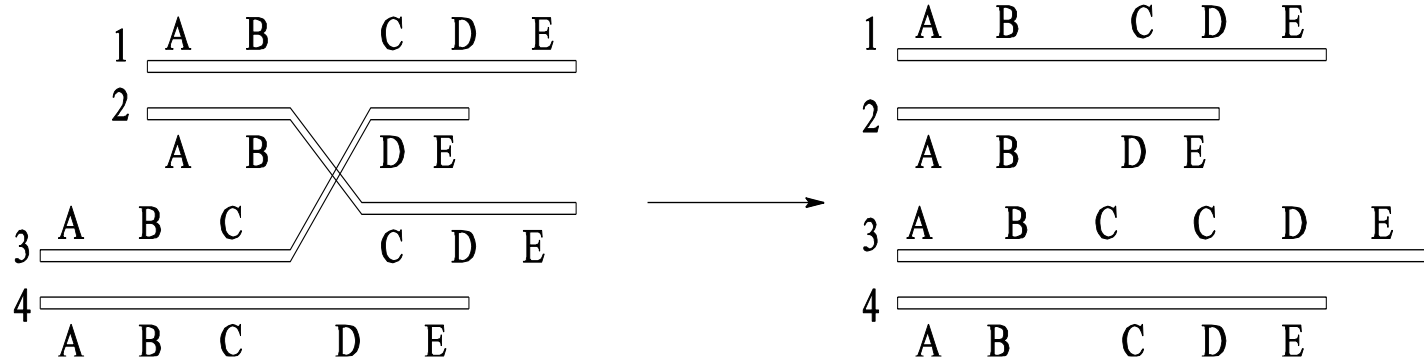


Translocation



Opérations modifiant le contenu

- **Pertes** (inactivation, dégradation, élimination). Origine possible: cross-over inégal → duplication locale et suppression

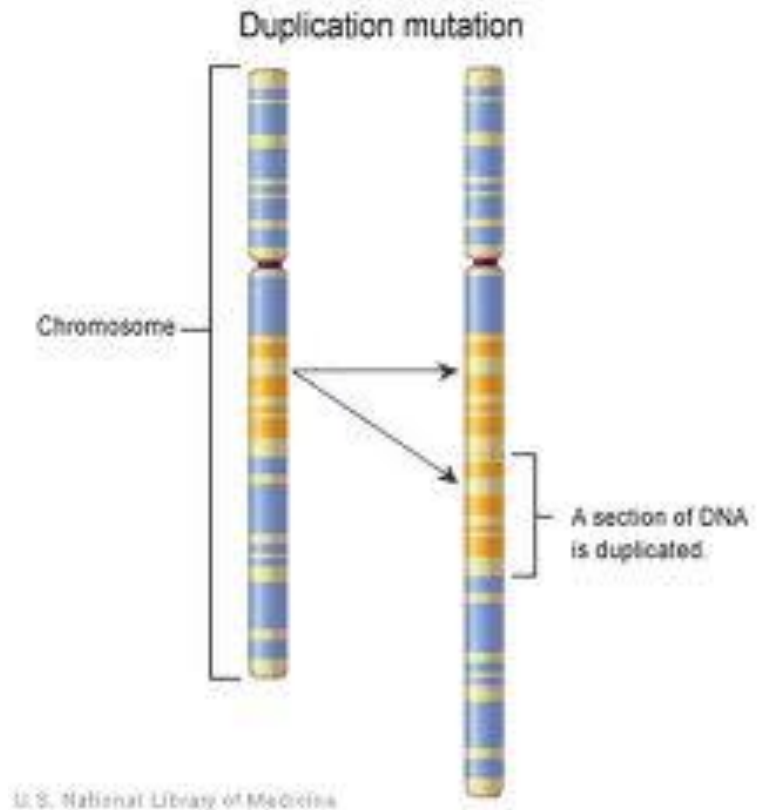


- **Duplications** (en tandem ou transposées)

a b c d e
↓
a b a b a b c d e

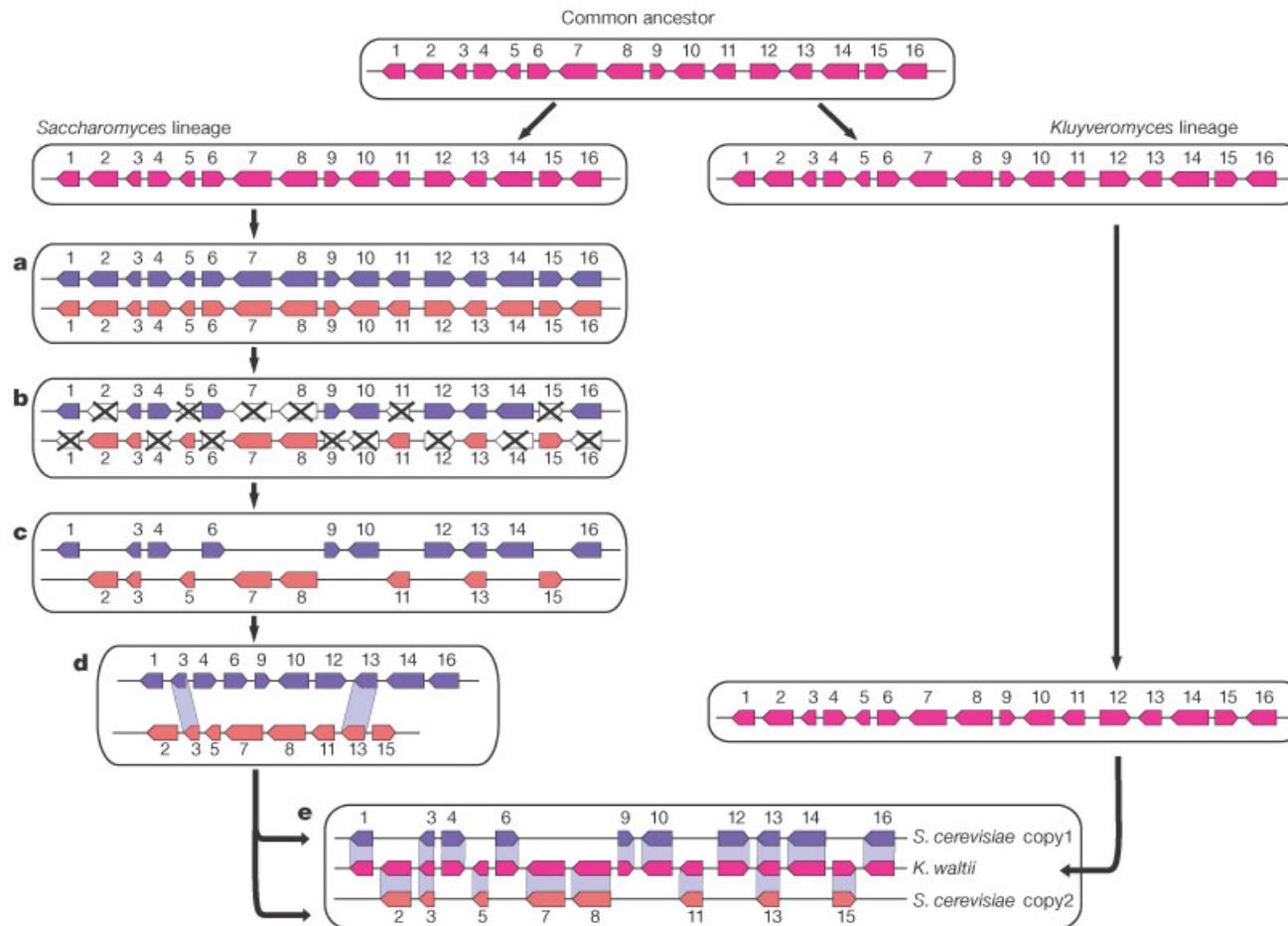
a b c d e f g h
↓
a b c d e f b c d g h

Duplication, Délétion



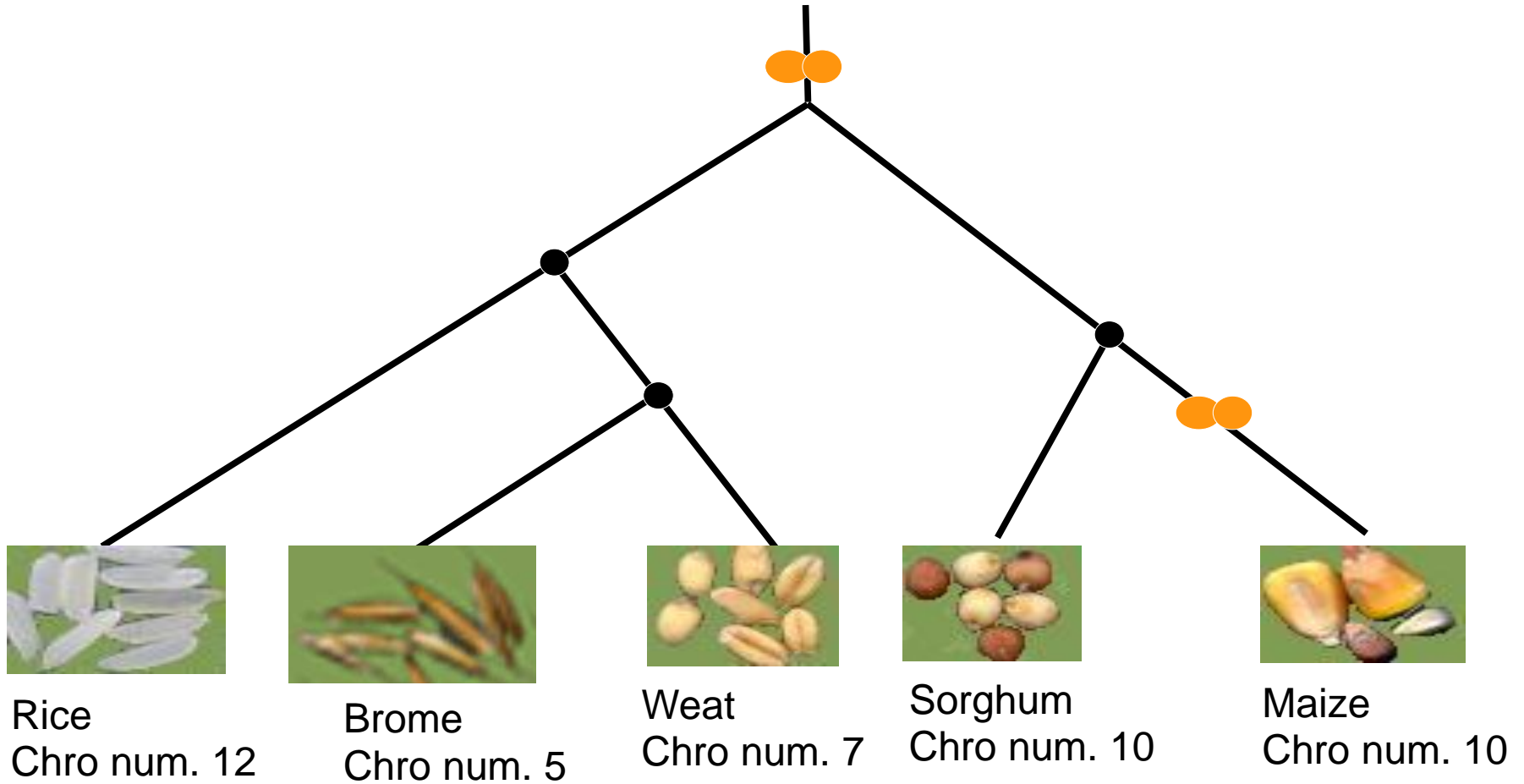
<http://www.daviddarling.info/encyclopedia/D/duplication.html>

Duplication de génome



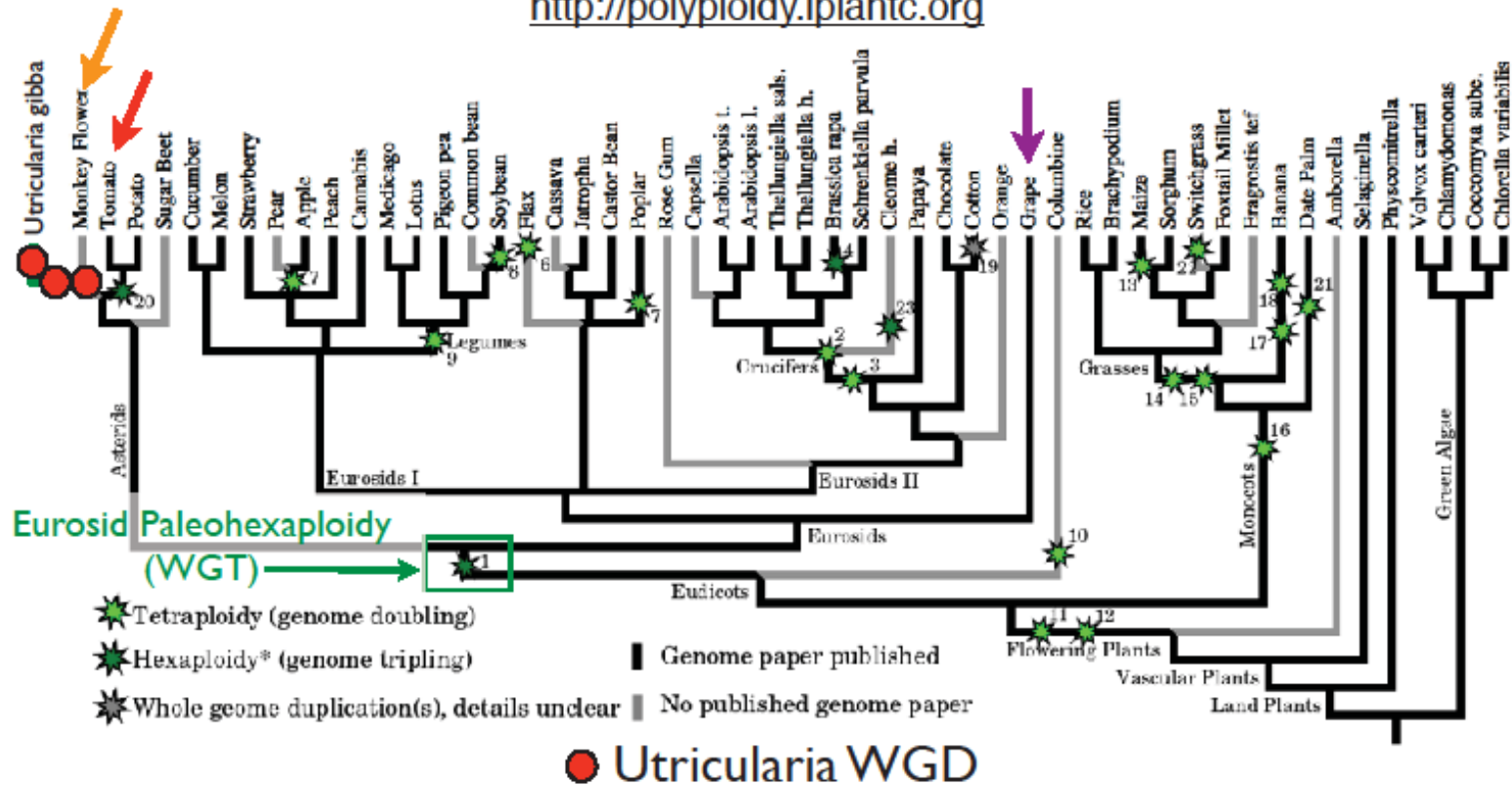
Model of WGD followed by massive gene loss predicts gene interleaving in sister regions. From Manolis Kellis, Bruce W. Birren and Eric S. Lander; *Nature* **428, 617-624, 2004**

Duplication de génome



Characterizing paleo-polyploidy

<http://polyploidy.iplantc.org>



Mimulus is closest, but under-characterized

Tomato is well characterized, but has a WGT

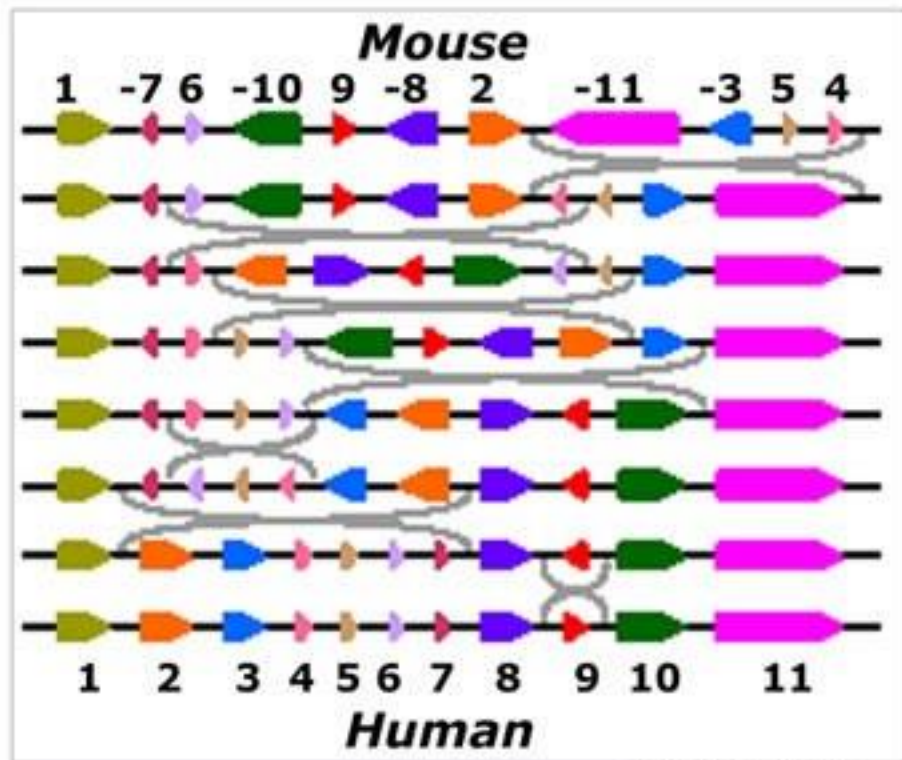
Grape is the de facto eurosid outgroup comparator

<http://www-etud.iro.umontreal.ca/~lafonman/MAGE2013/program.php>

II. Distance d'inversion

Deux génomes G et H contenant les mêmes gènes mais dans un ordre différent.

Distance d'inversion
entre G et H: Nombre minimal d'inversions pour passer de G à H.



© Genome Research

Biologists have found that the related genes in man and mouse are not chaotically distributed over the genomes, but form 'conserved blocks' instead. These conserved blocks reveal the genetic organization of the common ancestor of human and mouse, allowing Pevzner and Tesler to reconstruct a rearrangement scenario of man-mouse evolution. Genomic sequences reveal at least 11 syntenic blocks (where human and mouse genes are in the same order) of one million DNA letters or longer on the X chromosome. They provide evidence of at least 7 inversions (a type of rearrangement) which emanate from a common ancestor in the middle. Two of the 11 blocks show evidence of extensive micro-rearrangements. (Graphic by Glenn Tesler, UCSD)

8	7	6	5	4	3	2	1	11	10	9
8	7	1	2	3	4	5	6	11	10	9
4	3	2	1	7	8	5	6	11	10	9
4	3	2	8	7	1	5	6	11	10	9

Réduction: Comment transformer une permutation en l'identité?

Bibliographie:

- *Kececioglu et Sankoff, 1993*: Première heuristique, gènes non signés
- *Caprara 1997*: Problème NP-difficile pour les gènes non signés
- *Hannenhalli et Pevzner, 1995*: Algo polynomial pour les gènes signés
- *Kaplan, Shamir, Tarjan, 1999; Bader, Moret, Yan, 2001*: optimisations, algo linéaire pour calculer la distance et quadratique pour trouver un scénario d'inv.
- *Bergeron 2001; Bergeron, Mixtacki, Stoye 2005*: Représentations plus simples du problème
- ...

Points de cassure

- Distance naturelle: Distance de points de cassures (Breakpoints)

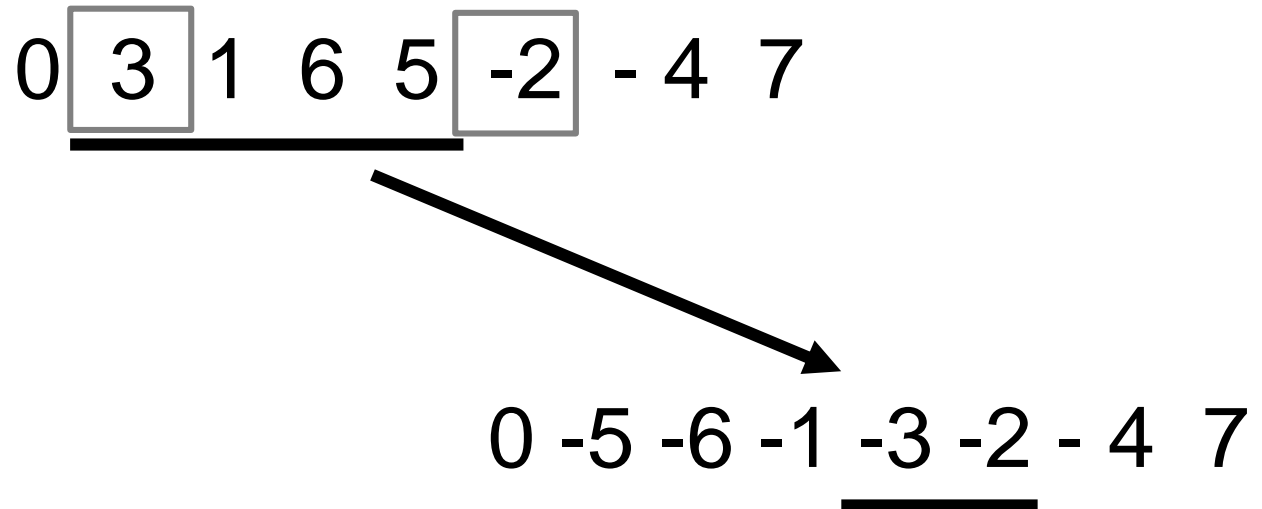
1 | 5 6 | 3 2 | 4 | 7 ← Gènes non signés

+1 | +5 +6 | +3 | +2 | +4 | +7 ← Gènes signés

- $i \bullet i+1$ ou $-(i+1) \bullet -i$: Adjacences
- Sinon: Breakpoint.

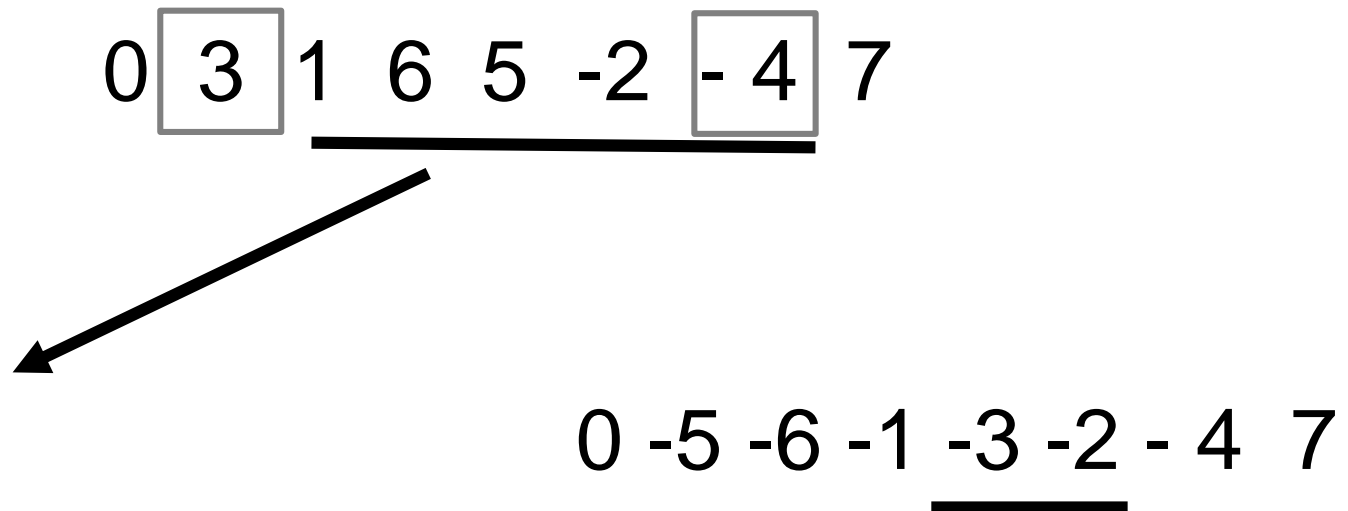
Inversion

- Une inversion d'un intervalle change **l'ordre et le signe** des gènes dans l'intervalle



Inversion

- Une inversion d'un intervalle change **l'ordre et le signe** des gènes dans l'intervalle



Inversion

- Une inversion d'un intervalle change **l'ordre et le signe** des gènes dans l'intervalle

0 3 1 6 5 -2 -4 7



0 3 4 2 -5 -2 -4 7

0 -5 -6 -1 -3 -2 -4 7



- Une **paire orientée** est une paire consécutive de gènes de signes différents.
 - Algorithme simple: Choisir, à chaque étape, une paire orientée (π_i, π_j)
 - Si $\pi_i + \pi_j = +1$, $(\pi_i \ \pi_{i+1} \ \dots \ \pi_{j-1})\pi_j$
 - Si $\pi_i + \pi_j = -1$, $\pi_i (\pi_{i+1} \ \dots \ \pi_{j-1} \ \pi_j)$
 - Une inversion créant une adjacence agit nécessairement sur une paire orientée.
- Mais pas toujours possibles, et pas toutes équivalentes.

0	3	1	6	5	-2	4	7		
		0	-5	-6	-1	-3	-2	4	7
		0	-5	-6	-1	2	3	4	7
		0	-5	-6	1	2	3	4	7
		0	-5	-4	-3	-2	-1	6	7
		0	1	2	3	4	5	6	7

5 inversions

0 3 1 6 5 -2 4 7

0 -5 -6 -1 -3 -2 4 7

0 -5 -6 -1 2 3 4 7

0 -5 -6 1 2 3 4 7

0 -5 -4 -3 -2 -1 6 7

0 1 2 3 4 5 6 7

0 3 1 6 5 -2 4 7

0 -5 -6 -1 -3 -2 4 7

0 -5 -6 -1 2 3 4 7

0 -5 -6 1 2 3 4 7

0 3 1 6 5 -2 4 7

0 -5 -6 -1 -3 -2 4 7

0 -5 -6 -1 2 3 4 7

0 -5 -6 1 2 3 4 7

$$\begin{array}{cccccccc}
 0 & 3 & 1 & 6 & 5 & -2 & 4 & 7 \\
 \hline
 0 & -5 & -6 & -1 & -3 & -2 & 4 & 7 \\
 & & & & \hline
 0 & -5 & -6 & -1 & 2 & 3 & 4 & 7 \\
 \hline
 0 & 1 & 6 & 5 & 2 & 3 & 4 & 7
 \end{array}$$

Impossible de continuer

0 3 1 6 5 -2 4 7

0 -5 -6 -1 -3 -2 4 7

0 3 1 6 5 -2 4 7

0 -5 -6 -1 -3 -2 4 7

0 3 1 6 5 -2 4 7

0 -5 -6 -1 -3 -2 4 7

0 -5 -4 2 3 1 6 7

0 -1 -3 -2 -4 5 6 7

0 -1 -3 -2 4 5 6 7

0 -1 2 3 4 5 6 7

0 1 2 3 4 5 6 7

6 inversions au lieu de 5: Pas minimal

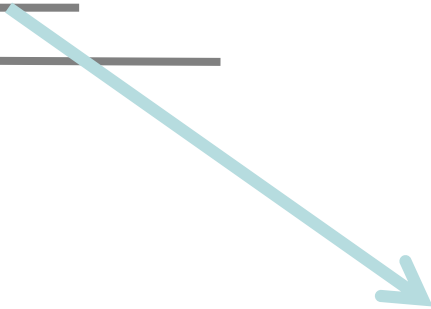
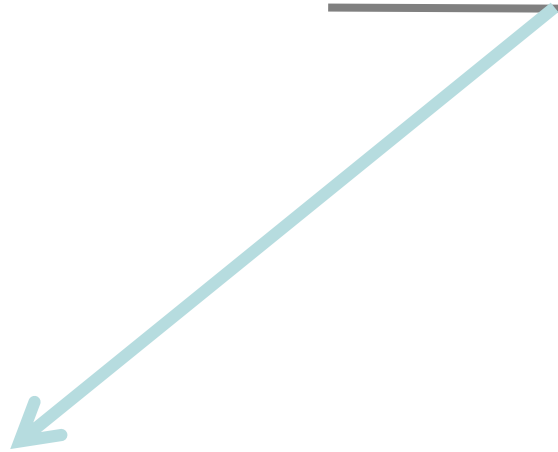
Résultat, Bergeron 2001

- Le score d'une inversion est le nombre de paires orientées dans la permutation résultante.

Algorithme: Choisir, à chaque étape, une paire orientée (π_i, π_j) de score maximal.

0 3 1 6 5 -2 4 7

0 -5 -6 -1 -3 -2 4 7



0 -5 -6 -1 2 3 4 7

0 -5 -4 2 3 1 6 7

Score = 2

Score = 4

Résultat, Bergeron 2001

- Le score d'une inversion est le nombre de paires orientées dans la permutation résultante.

Algorithme: Choisir, à chaque étape, une paire orientée (π_i, π_j) de score maximal.

Théorème: Si Algorithme applique k inversions à une permutation π donnant lieu à une permutation π' , alors

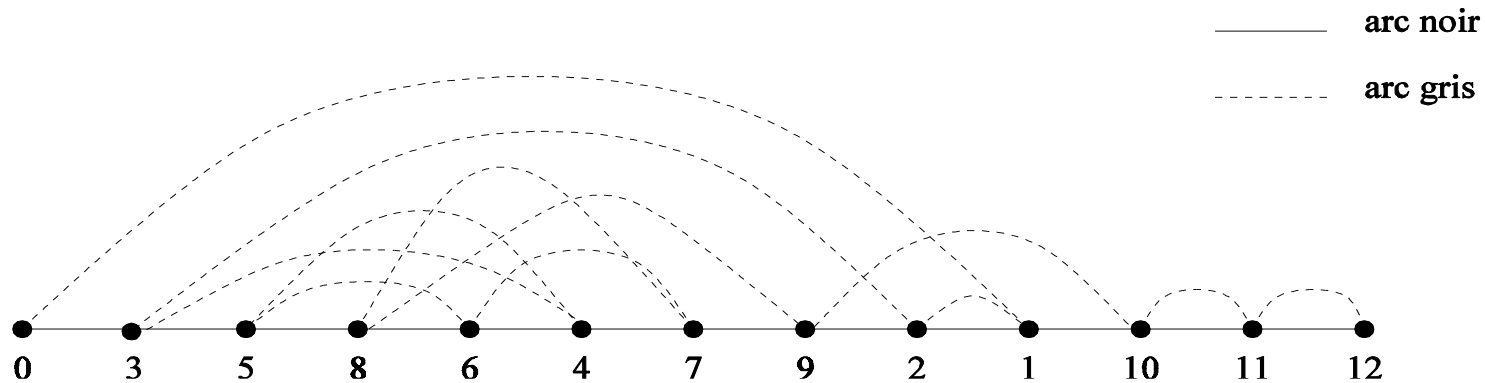
$$d(\pi) = d(\pi') + k.$$

Mais on est bloqué si on n'a pas de paire orientée!

Graphe de points de cassure, gènes non signés (Bafna 1995)

G: 0 3 5 8 6 4 7 9 2 1 10 11 12

H: 0 1 2 3 4 5 6 7 8 9 10 11 12



- Décomposition maximale en c cycles alternés d'arcs disjoints
- $d(G,H)$: distance d'inversion; b : nb d'arcs noirs (gènes)

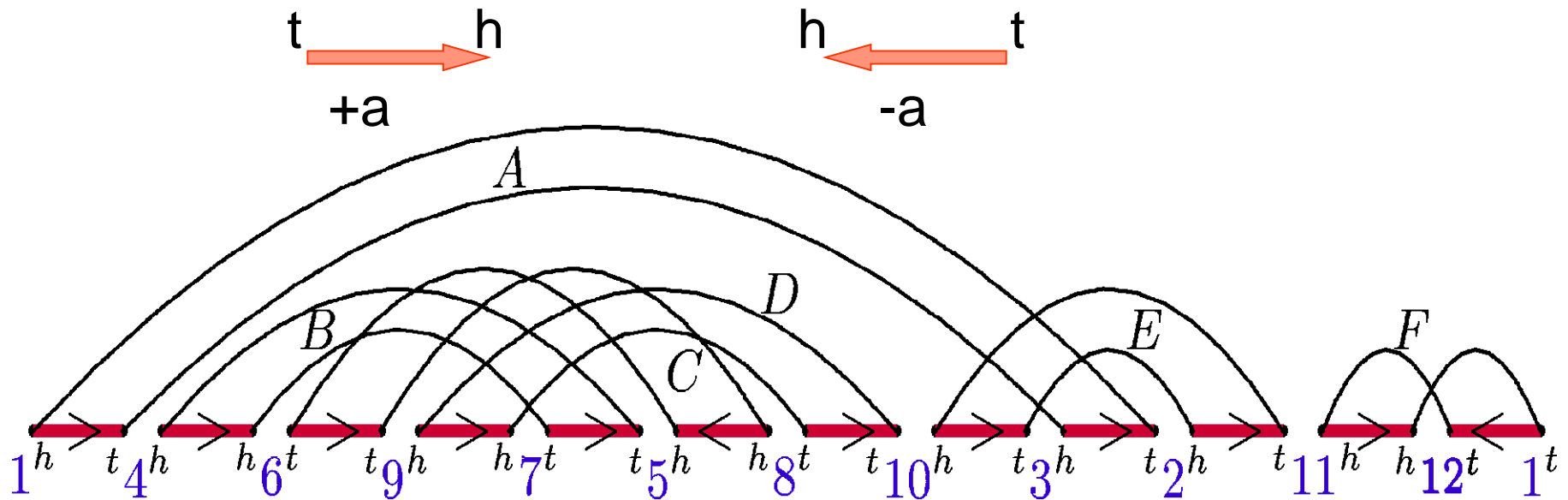
$$d(G,H) \geq b - c$$

- Problème de la décomposition d'un graphe en un maximum de cycles disjoints: NP-difficile

Gènes signés – Graphe de Hannenhalli et Pevzner (1995)

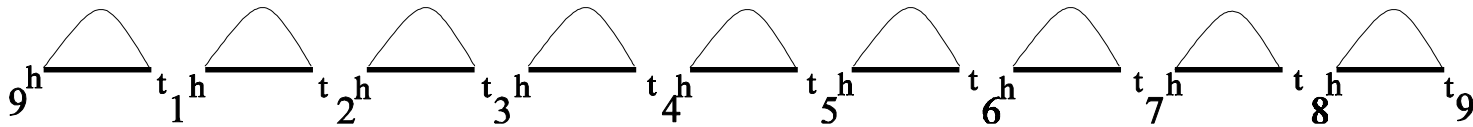
G = +1 +4 -6 +9 -7 +5 -8 +10 +3 +2 +11 -12

H = +1 +2 +3 +4 +5 +6 +7 +8 +9 +10 +11 +12



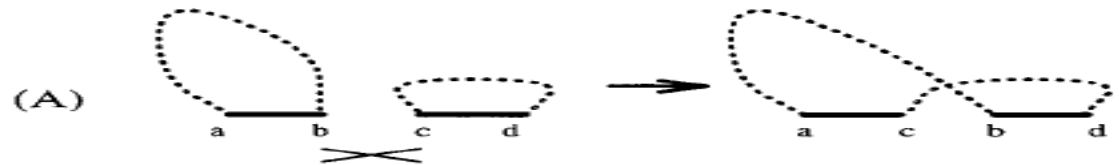
Si génome non-circulaire, rajouter des bornes fictives

Nombre de cycles maximal lorsque les deux génomes sont identiques

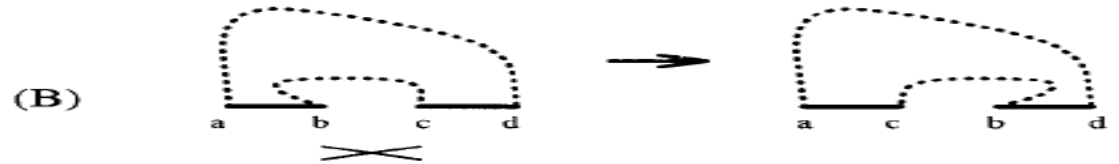


Inversions possibles:

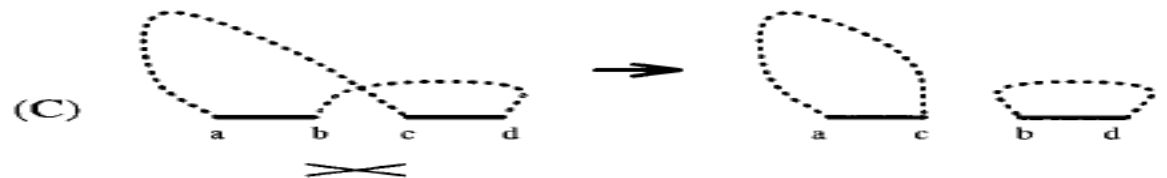
(A) Inversion sur deux arêtes de **deux cycles différents**



(B) Inversion sur une **paire non-orientée** (ou convergentes) d'arêtes

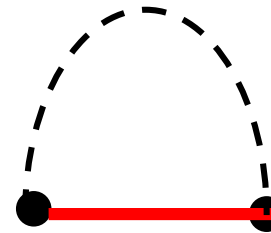
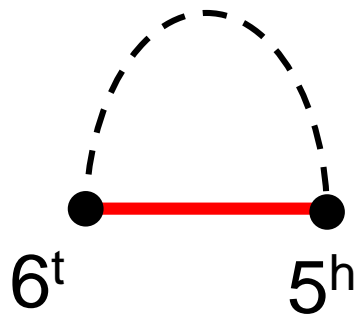
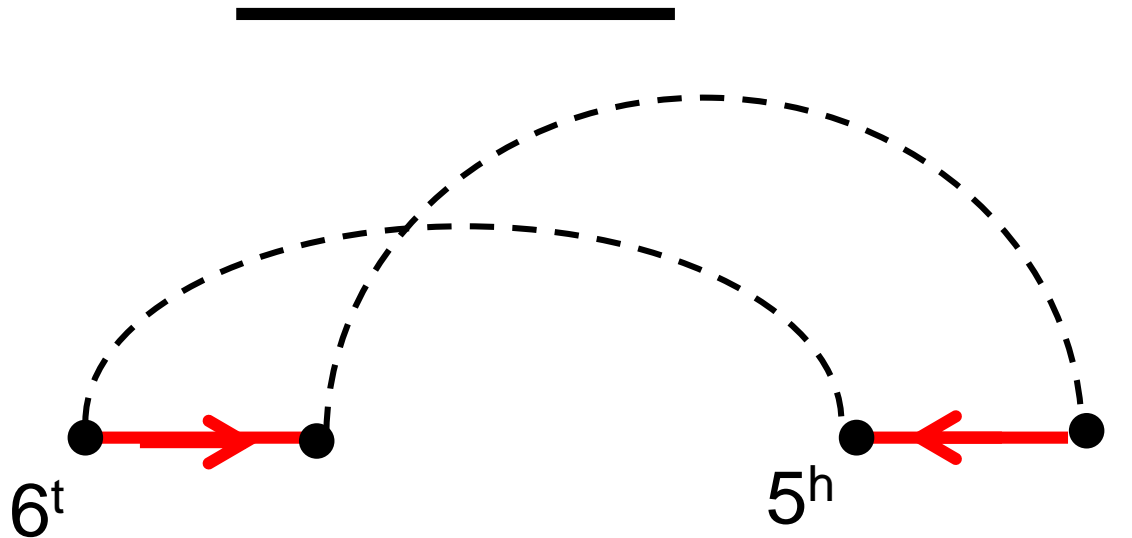


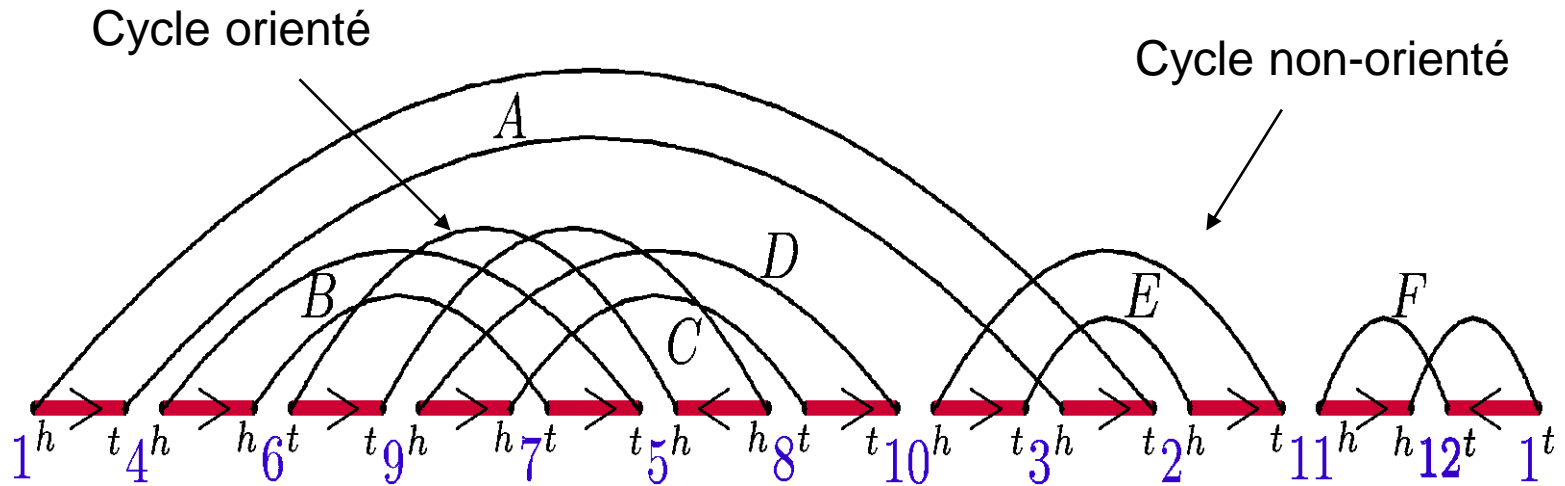
(c) Inversion sur une **paire d'arêtes orientées** (ou divergentes)



Lien avec les paires orientées

+1 +4 -6 +9 -7 +5 -8 +10 +3 +2 +11 -12





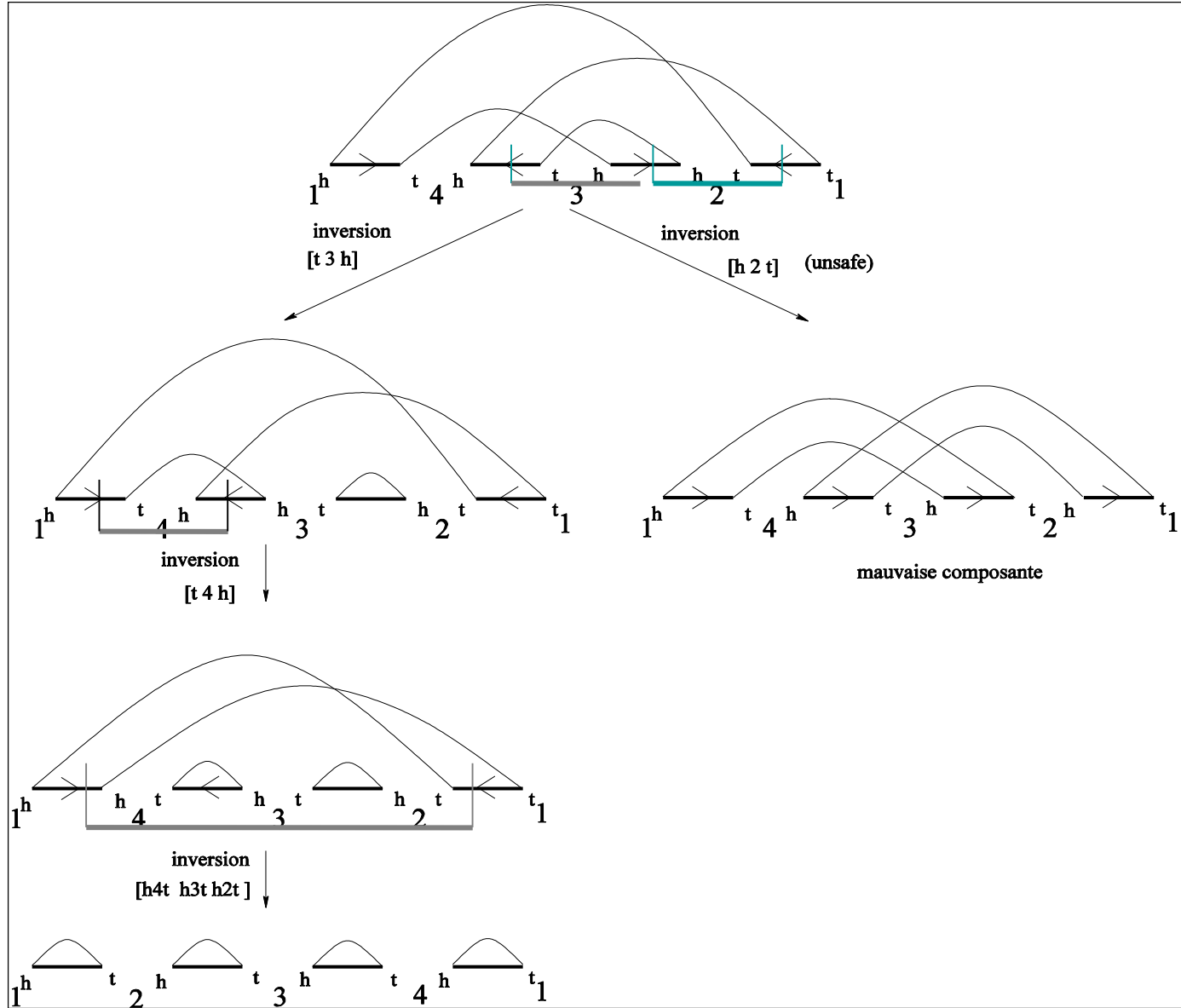
- $\{B, C, D\}$, $\{F\}$: Composantes orientées (bonne composante)
- $\{A, E\}$: Composante non-orientée

Cas général: $d(G, H) \geq b - c$

Si que des bonnes composantes: $d(G, H) = b - c$

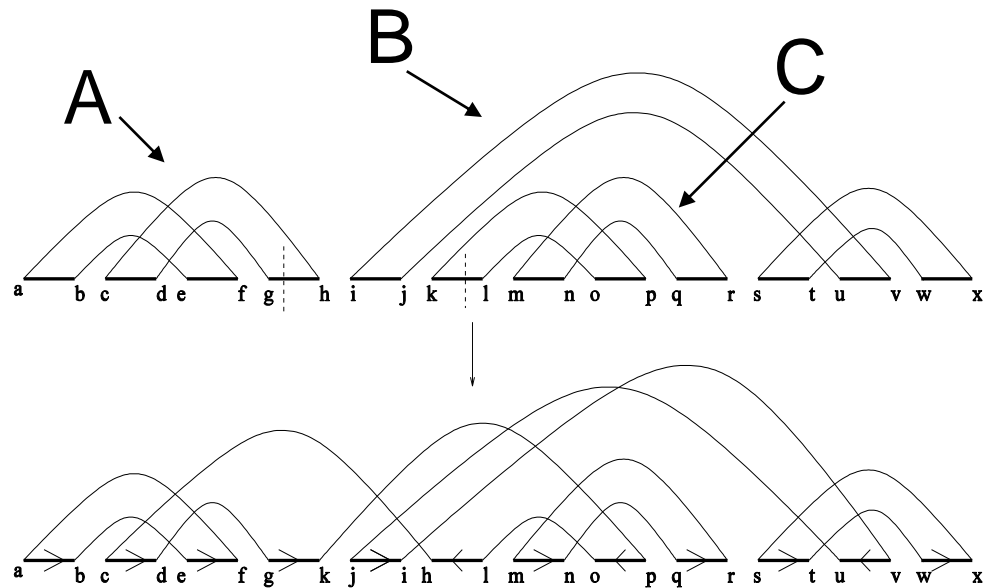
- Bonnes composantes: peuvent être résolues par *b-c* "bonnes inversions"

- Bonne inversion (safe): Inversion sur deux arêtes orientées, qui ne crée pas de mauvaise composante.



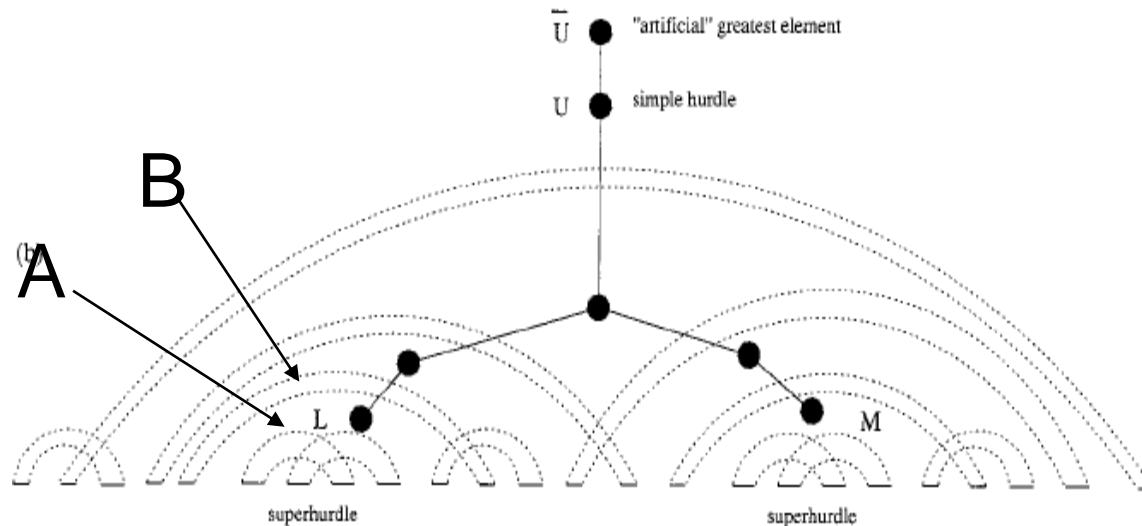
Mauvaises composantes

- Composante B sépare A et C.
- **Non-obstacle:** Mauvaise composante qui sépare deux mauvaises composantes
- **Obstacle (hurdle):** Mauvaise composante qui ne sépare pas deux mauvaises composantes



Forteresse

- Un obstacle A **protège** un non-obstacle B si la suppression de A transforme B en obstacle.
- **Super-obstacle**: Obstacle A qui protège un non-obstacle B



Forteresse: Graphe qui contient un nb impair d'obstacles, tous des super-obstacles.

Résultat de Hannenhalli et Pevzner

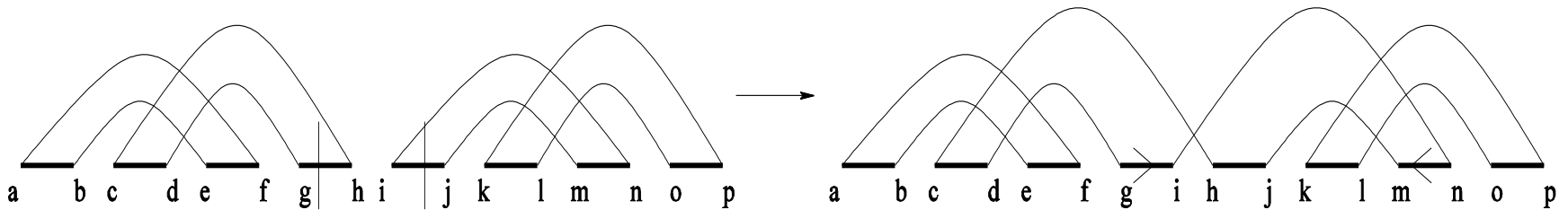
- $d(G,H)$: distance d'inversions
- $b(G,H)$: nb de gènes
- $c(G,H)$: nb de cycles du graphe
- $h(G,H)$: nb d'obstacles
- $f(G,H)$: 1 si le graphe est une forteresse, 0 sinon.

$$d(G,H) = b(G,H) - c(G,H) + h(G,H) + f(G,H)$$

Résolution des obstacles:

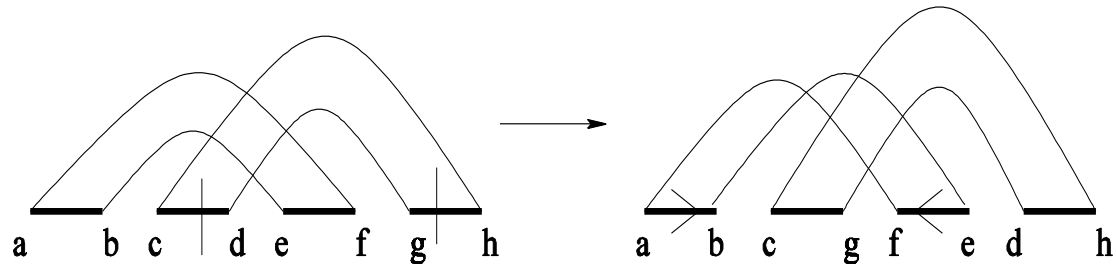
Deux opérations:

- Fusion:



Un cycle de moins, mais un obstacle de moins

- Coupure:



Même nb de cycles, mais un obstacle de moins.

Algorithme HP:

1. Si G contient $h(G,H)$ obstacles
2. Si $h(G,H)$ est pair
3. Considérer des paires d'obstacles non consécutifs, et les fusionner deux à deux;
4. Si $h(G,H)$ est impair et il existe un obstacle simple O
5. Couper O ;
6. Fusionner deux à deux les obstacles restants;
7. Sinon (forteresse)
8. Fusionner deux à deux les obstacles non-consécutifs
9. (si possible), et couper le dernier obstacle restant;
10. Pour chaque bonne composante C faire
11. Résoudre C en choisissant une inversion sûre à chaque étape.

- Une inversion est bonne si $\Delta(b-c+h+f)=-1$
- L'algorithme n'effectue que des bonnes inversions:

- Inversion sure:

$$\Delta(c)=1; \Delta(h)=0; \Delta(f)=0; \text{ donc } \Delta(b-c+h+f)=-1$$

- Fusion de deux obstacles:

$$\Delta(c)=-1; \Delta(h)=-2; \Delta(f)=0; \text{ donc } \Delta(b-c+h+f)=-1$$

- Coupure d'un obstacle:

$$\Delta(c)=0; \Delta(h)=-1; \Delta(f)=0; \text{ donc } \Delta(b-c+h+f)=-1$$

- Coupure du dernier obstacle de la forteresse:

$$\Delta(c)=0; \Delta(h)=0; \Delta(f)=-1; \text{ donc } \Delta(b-c+h+f)=-1$$

Complexité

- Construire la structure, trouver les cycles et les composantes, déterminer leurs orientations: temps $O(n^2)$
=> trouver la distance d'inversion en $O(n^2)$
- La partie la plus coûteuse: résolution des bonnes composantes.

Méthode brutale: Essayer toutes les inversions (n^2) et vérifier le graphe obtenu. Effectuer ce travail $d(G,H)$ fois => $O(n^5)$