Combinatorial aspects of genome rearrangements and haplotype networks

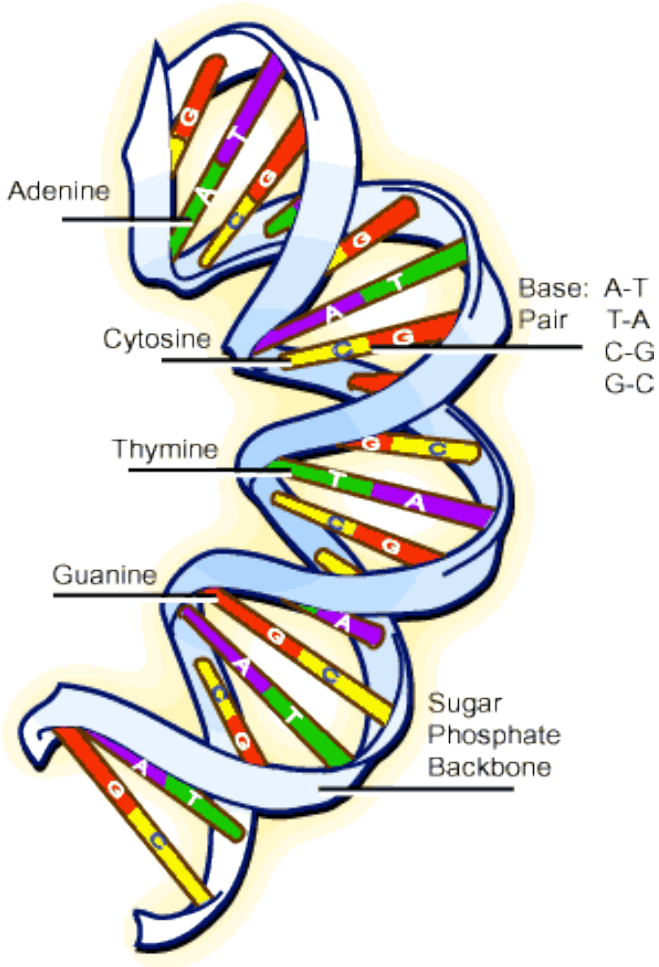
Anthony Labarre
alabarre@ulb.ac.be

Université libre de Bruxelles (U.L.B.)

12 septembre 2008

Quelques définitions

- **A**cide **d**ésoxyribo**n**ucléique : double hélice de *nucléotides* (A, C, G, T) ;
- *Gène* = séquence de nucléotides (codant pour une protéine particulière) ;
- *Chromosome* = ensemble ordonné de gènes ;
- *Génome* = ensemble de chromosomes ;
- Complémentarité (A-T, C-G) : une bande “suffit” ;



Motivations

- Réarrangements génomiques, réseaux d'haplotypes (différentes copies d'un gène) ;
- *Génomique comparative* : comparaison de génomes sous divers aspects ;
- Buts :
  - reconstruire des scénarios d'évolutions ;
  - identifier de nouveaux gènes ;
  - ⋮
- Comparaison au niveau des *nucléotides* ou des *gènes* ;

Alignement de séquences

- Comparaison au niveau des nucléotides :

Exemple 1 (alignement de séquences)

$$\begin{array}{ccccccccccccc} \cdots & T & C & C & G & C & C & A & - & - & C & T & A & \cdots \\ & | & | & & | & & | & & & & | & & | & \\ \cdots & T & C & G & G & A & C & T & G & G & C & - & A & \cdots \end{array}$$

- Correspondances, substitutions, insertions et suppressions ;
- Correspondent à des mutations ;

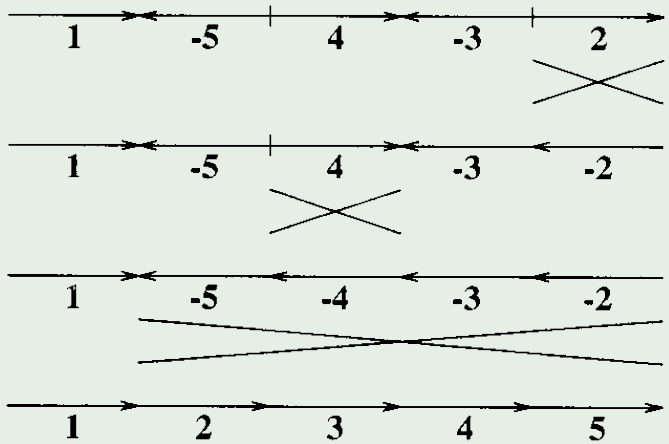
Réarrangements génomiques

- Comparaison au niveau des gènes ;
- Les espèces diffèrent par “le contenu”, mais aussi par “l’ordre” :

Exemple 2 (réarrangements génomiques [Hannenhalli et Pevzner, 1999])

B. oleracea
(cabbage)

B. campestris
(turnip)



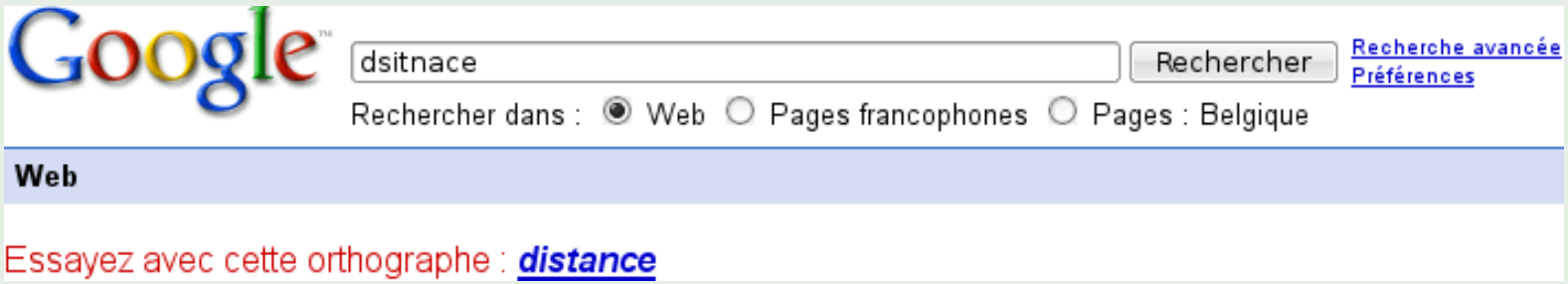
Distances d'édition

- *Opérations d'édition* : ensemble d'opérations permises ;
- *Distance d'édition* : nombre minimum d'opérations d'édition pour transformer un objet X en un objet Y ;

Distances d'édition

- *Opérations d'édition* : ensemble d'opérations permises ;
- *Distance d'édition* : nombre minimum d'opérations d'édition pour transformer un objet X en un objet Y ;

Exemple 3 (correction orthographique)

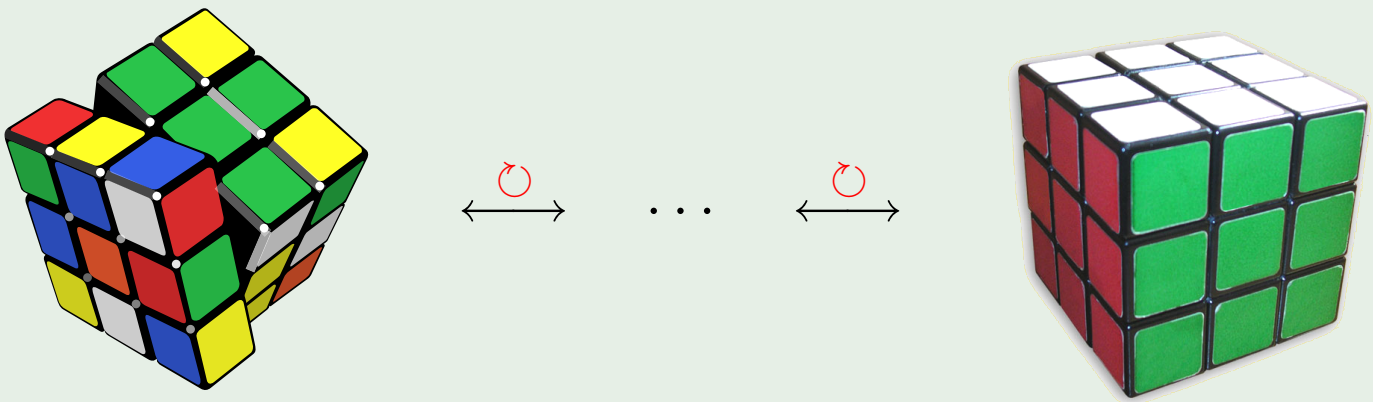


dsitnace $\longrightarrow$ distnace $\longrightarrow$ distance
(échanges de caractères adjacents)

Distances d'édition

- *Opérations d'édition* : ensemble d'opérations permises ;
- *Distance d'édition* : nombre minimum d'opérations d'édition pour transformer un objet X en un objet Y ;

Exemple 3 (cube de Rubik)

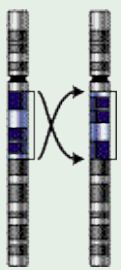
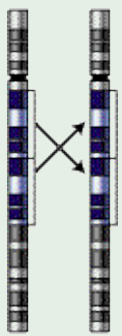
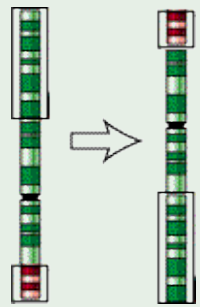


(rotations des faces)

Distances d'édition

- *Opérations d'édition* : ensemble d'opérations permises ;
- *Distance d'édition* : nombre minimum d'opérations d'édition pour transformer un objet X en un objet Y ;

Exemple 3 (réarrangements génomiques)

	inversion	transposition	échange de blocs
			
modélisation :	$\langle 3\ \underline{2\ 5\ 4\ 1}\rangle$ $\downarrow$ $\langle 3\ 1\ 4\ 5\ 2\rangle$	$\langle 3\ \boxed{2\ 5}\ \boxed{4\ 1}\rangle$ $\downarrow$ $\langle 3\ 4\ 1\ 2\ 5\rangle$	$\langle \boxed{5}\ 4\ \boxed{3\ 2\ 1}\rangle$ $\downarrow$ $\langle 3\ 2\ 1\ 4\ 5\rangle$

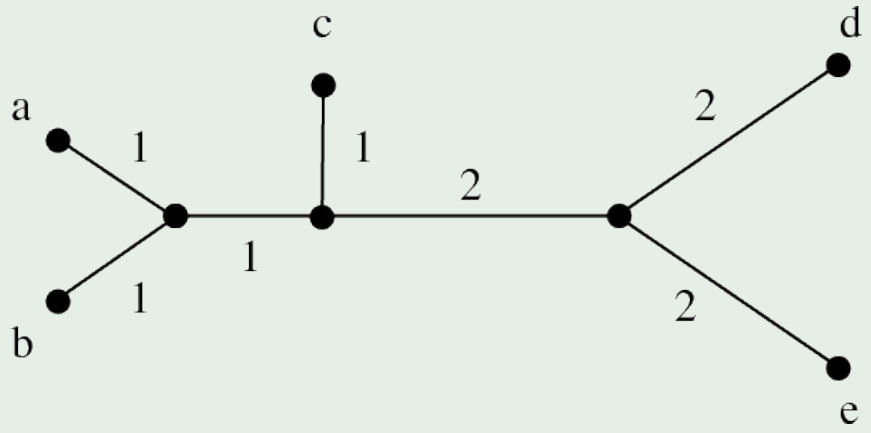
Arbres phylogéniques

- But : reconstruire des scénarios d'évolution ;

Exemple 4 (arbre à partir d'une matrice de distances [Mihaescu et al., 2006])

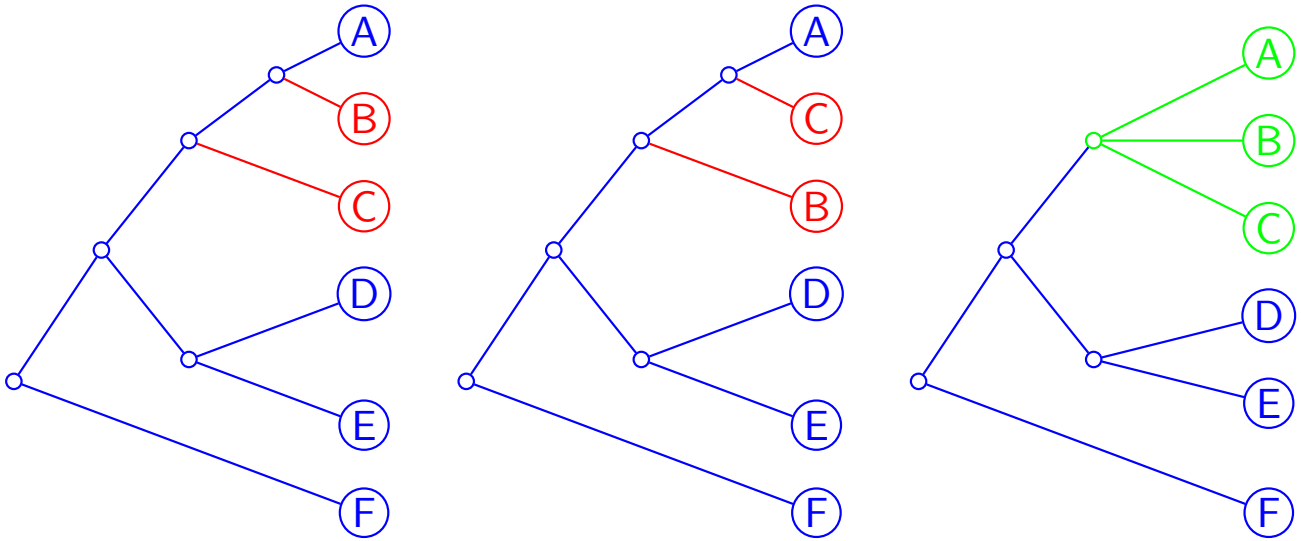
	a	b	c	d	e
a	0	2	3	6	3
b	2	0	3	6	6
c	3	3	0	5	5
d	6	6	5	0	4
e	3	6	5	4	0

$\longrightarrow$

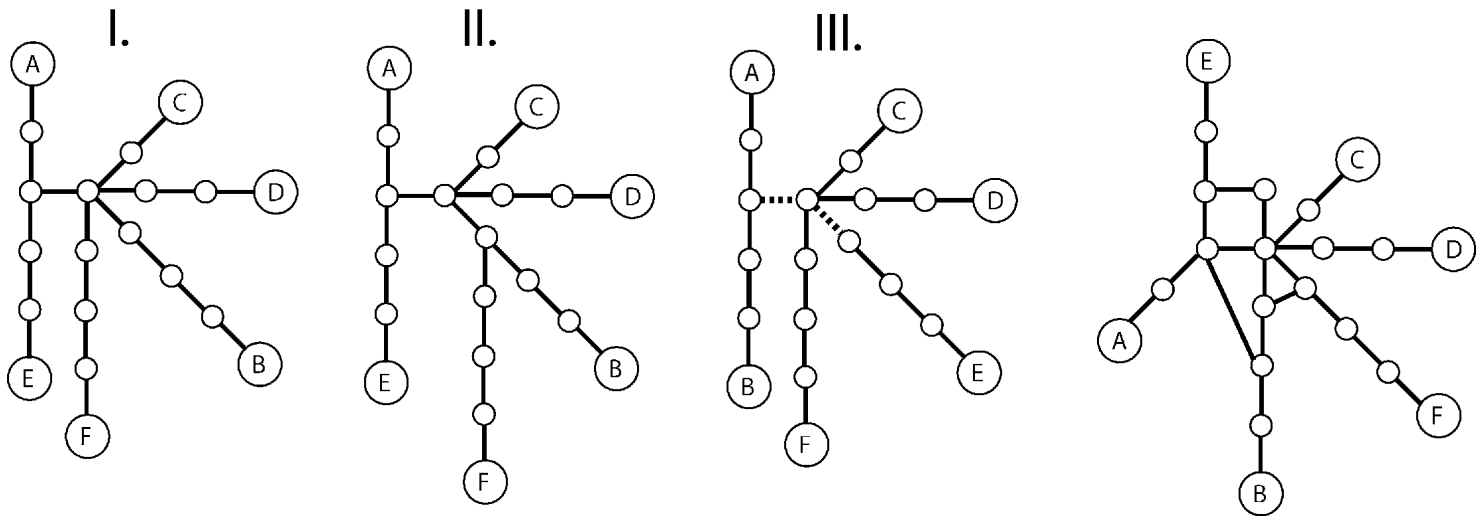


- Note : certaines conditions doivent être satisfaites [Buneman, 1971] ;
- Problème : que faire si plusieurs topologies se valent ?

1. Arbres de consensus :



2. *Réseaux* phylogéniques [Cassens et al., 2005] :



Formalisme pour les distances génomiques

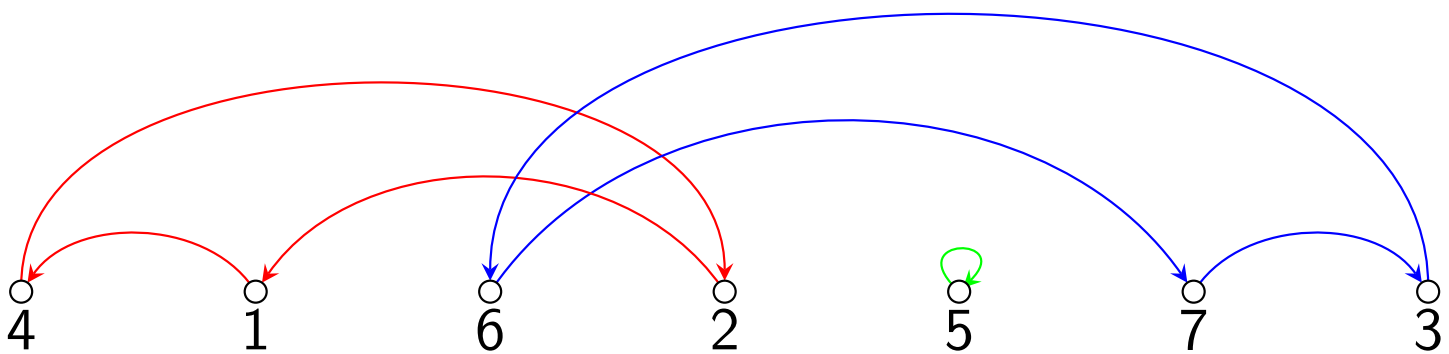
- On suppose que :
  - l'ordre des gènes est connu ;
  - chaque gène apparaît une et une seule fois dans chaque génome ;
- Alors :
  - $\{\text{gènes}\} = \{1, 2, \ldots, n\}$;
  - génome = *permutation* de $\{1, 2, \ldots, n\}$;
- Les réarrangements seront également représentés par des permutations ;

Permutations : les bases

- Permutation : arrangement linéaire de $\{1, 2, \ldots, n\}$;
- Décomposition en cycles disjoints :

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 4 & 1 & 6 & 2 & 5 & 7 & 3 \end{pmatrix} = (1, 4, 2)(3, 6, 7)(5).$$

- Le graphe de la permutation π, noté $\Gamma(\pi)$:



Le(s) problème(s)

- Soit :
  - π une permutation de $\{1, 2, \ldots, n\}$;
  - $S = \{s_1, s_2, \ldots\}$ un ensemble de *réarrangements* (permutations de $\{1, 2, \ldots, n\}$) ;
  - la *permutation identité* $\iota = \langle 1\ 2\ \cdots\ n \rangle$;
- On veut :
  1. "**trier par** S" : trouver une séquence d'éléments de S triant π de longueur minimale :

     $$\pi \circ x_1 \circ x_2 \circ \cdots \circ x_t = \iota \text{ où } x_1, \ldots, x_t \in S \text{ et } t \text{ est minimal}$$

  2. "**calculer la distance** $d_S(\pi, \iota)$" : trouver la longueur d'une telle séquence ;

Exemples de tri

- Tri par inversions :

$$\langle 3\ 2\ \underline{5\ 4\ 1}\rangle \rightarrow \langle \underline{3\ 2\ 1}\ 4\ 5\rangle \rightarrow \langle 1\ 2\ 3\ 4\ 5\rangle$$

- Tri par transpositions :

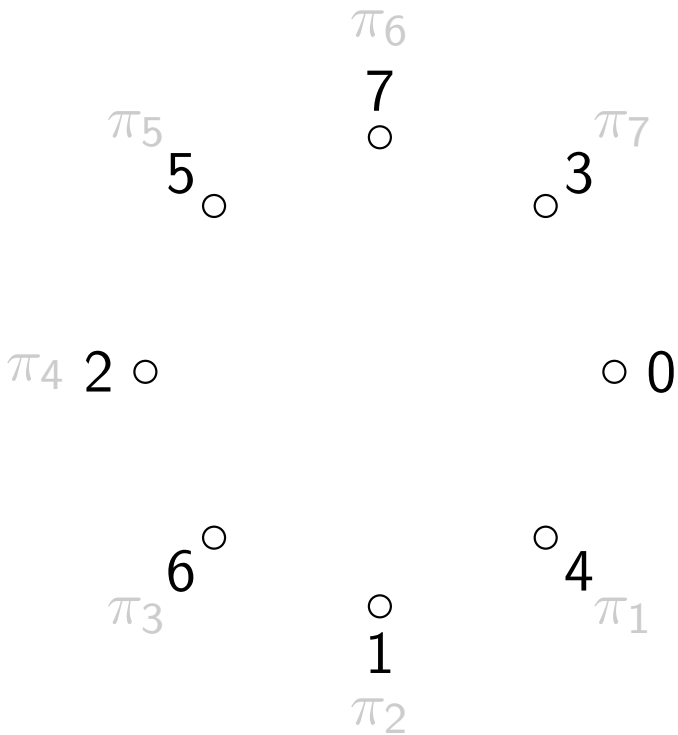
$$\langle 3\ \boxed{2\ 5}\ \boxed{4\ 1}\rangle \rightarrow \langle \boxed{3\ 4}\ \boxed{1\ 2}\ 5\rangle \rightarrow \langle 1\ 2\ 3\ 4\ 5\rangle$$

- Tri par échanges de blocs :

$$\langle \boxed{5}\ 4\ \boxed{3}\ 2\ 1\rangle \rightarrow \langle \boxed{3\ 4\ 5}\ 2\ \boxed{1}\rangle \rightarrow \langle 1\ 2\ 3\ 4\ 5\rangle$$

Le "graphe des cycles" [Bafna et Pevzner, 1998]

- Le "graphe des cycles" de π, noté $G(\pi)$:

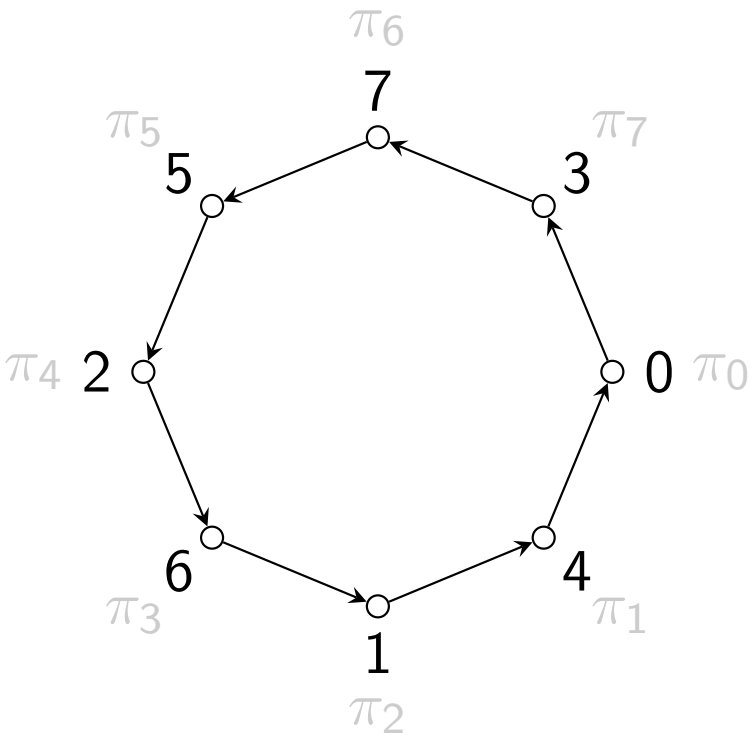


(ici, $\pi = \langle 4\ 1\ 6\ 2\ 5\ 7\ 3 \rangle$)

1. $V(G) = (\pi_0 = 0, \pi_1, \pi_2, \ldots, \pi_n)$;
2. $E(G) =$

Le "graphe des cycles" [Bafna et Pevzner, 1998]

- Le "graphe des cycles" de π, noté $G(\pi)$:

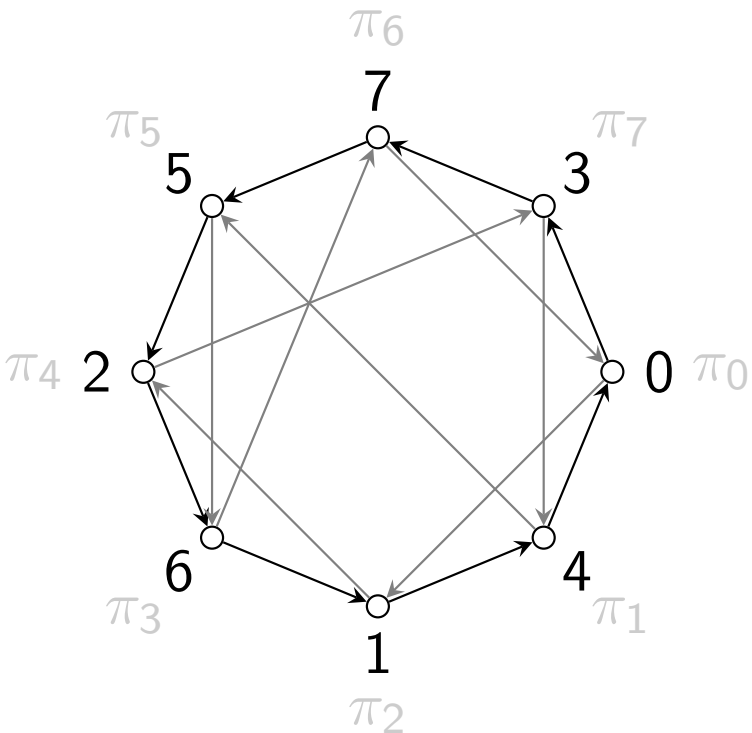


(ici, $\pi = \langle 4\ 1\ 6\ 2\ 5\ 7\ 3 \rangle$)

1. $V(G) = (\pi_0 = 0, \pi_1, \pi_2, \ldots, \pi_n)$;
2. $E(G) =$ {arcs *noirs*}

Le "graphe des cycles" [Bafna et Pevzner, 1998]

- Le "graphe des cycles" de π, noté $G(\pi)$:

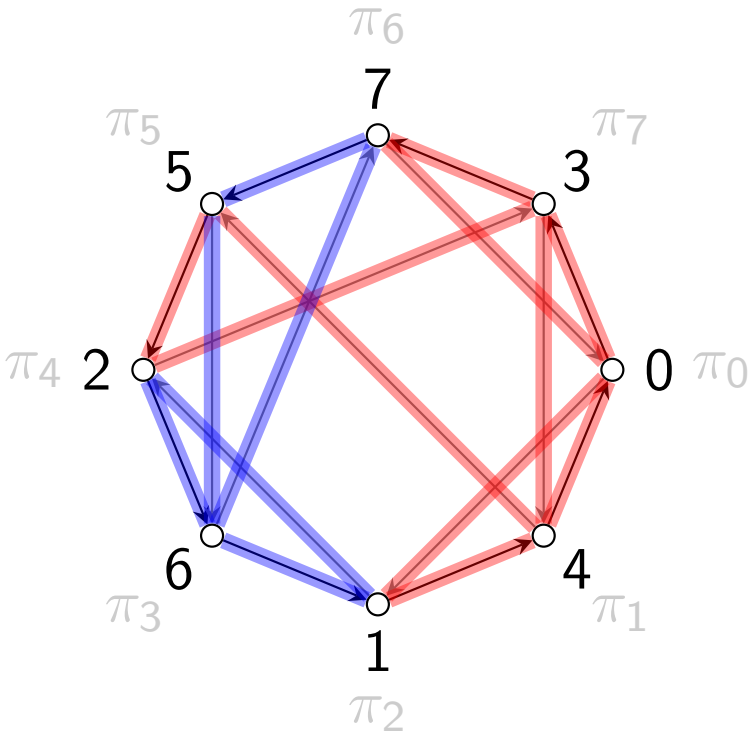


(ici, $\pi = \langle 4\ 1\ 6\ 2\ 5\ 7\ 3 \rangle$)

1. $V(G) = (\pi_0 = 0, \pi_1, \pi_2, \ldots, \pi_n)$;
2. $E(G) =$ {arcs *noirs*} $\cup$ {arcs *gris*} ;

Le "graphe des cycles" [Bafna et Pevzner, 1998]

- Le "graphe des cycles" de π, noté $G(\pi)$:



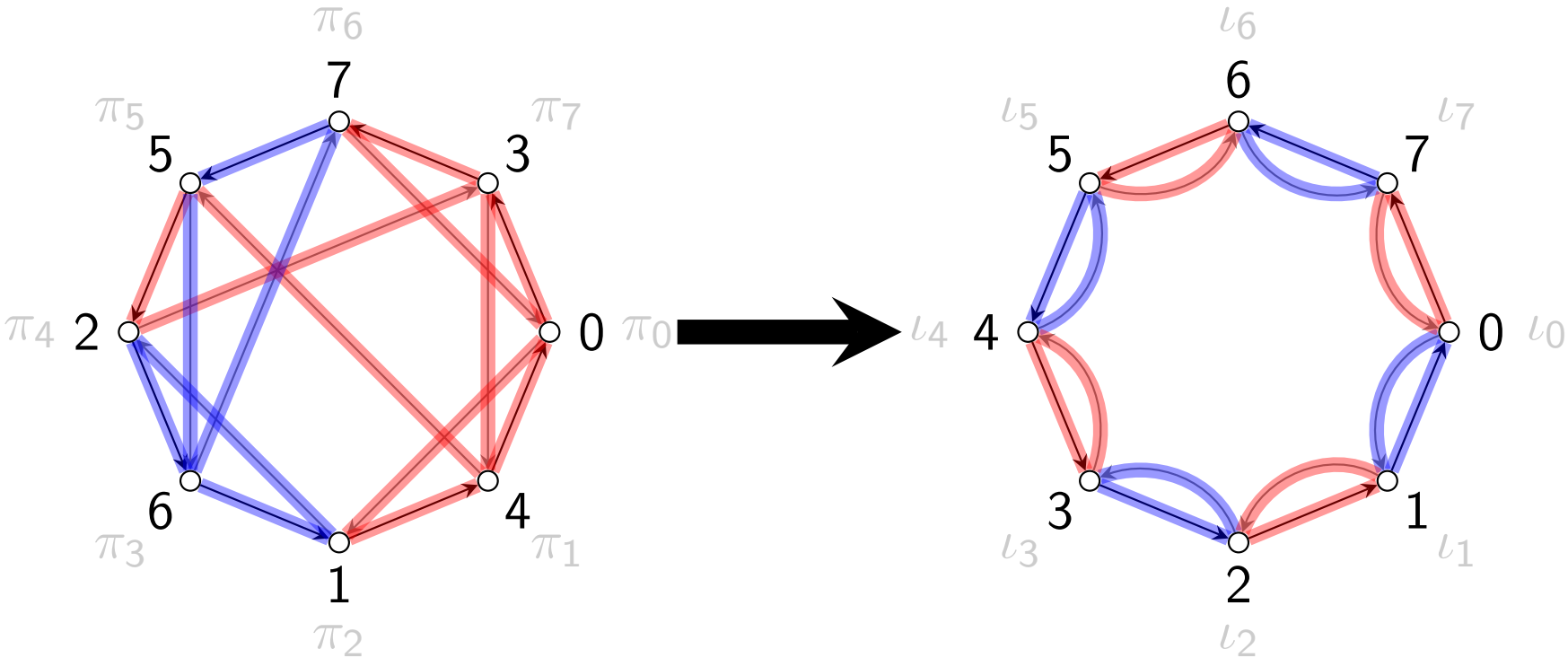
(ici, $\pi = \langle 4\ 1\ 6\ 2\ 5\ 7\ 3 \rangle$)

1. $V(G) = (\pi_0 = 0, \pi_1, \pi_2, \ldots, \pi_n)$;
2. $E(G) =$ {arcs *noirs*} $\cup$ {arcs *gris*} ;

- Décomposition unique en "cycles alternés" ;

Le "graphe des cycles" [Bafna et Pevzner, 1998]

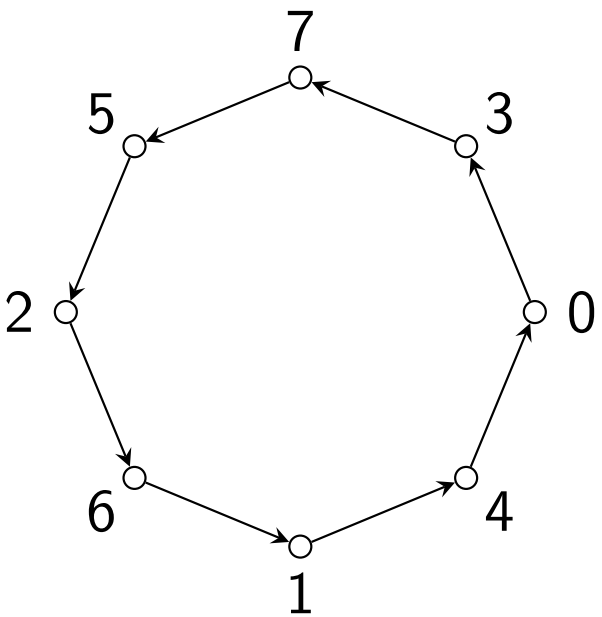
- Le "graphe des cycles" de π, noté $G(\pi)$:

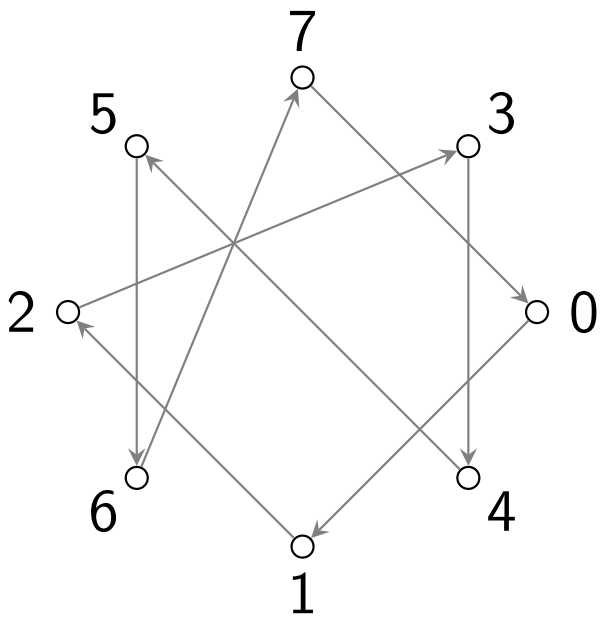


- Décomposition unique en "cycles alternés" ;

Expression de $G(\pi)$ par une nouvelle permutation

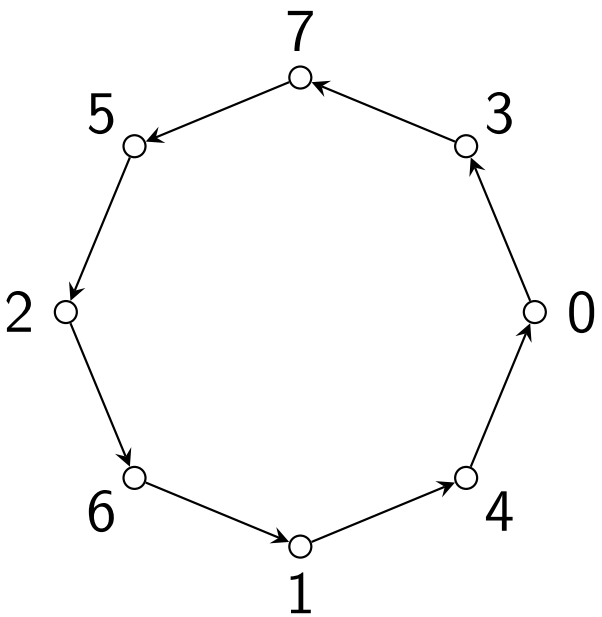
- Isolons le "cycle noir" et le "cycle gris" :





Expression de $G(\pi)$ par une nouvelle permutation

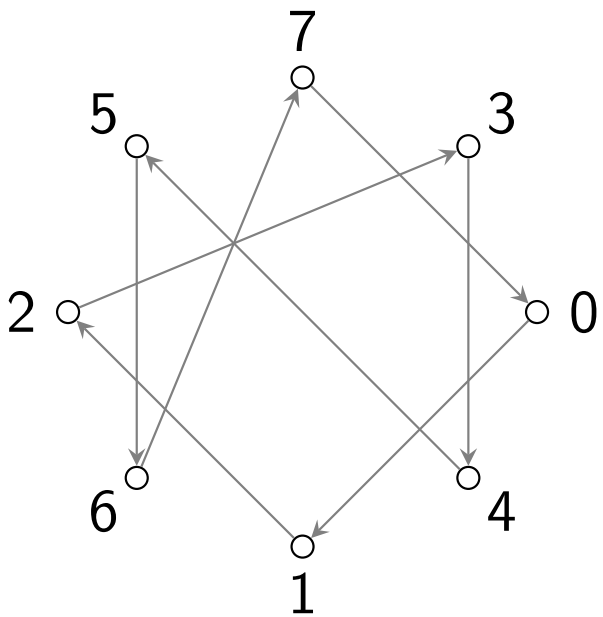
- Isolons le "cycle noir" et le "cycle gris" :



$\hat{\pi}$

$=$

$(0, 3, 7, 5, 2, 6, 1, 4)$



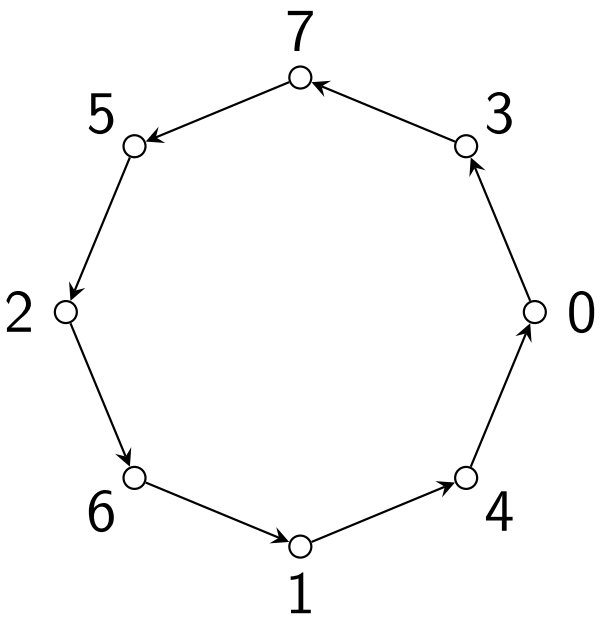
$\dot{\pi}$

$=$

$(0, 1, 2, 3, 4, 5, 6, 7)$

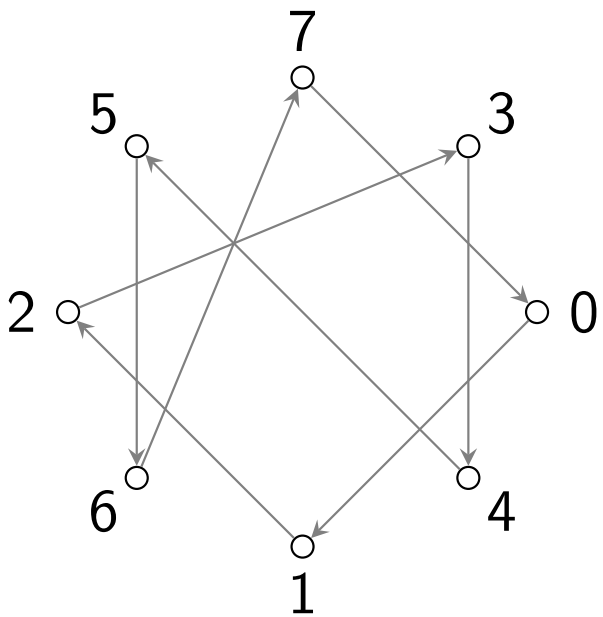
Expression de $G(\pi)$ par une nouvelle permutation

- Isolons le "cycle noir" et le "cycle gris" :



$\hat{\pi}$

$=$



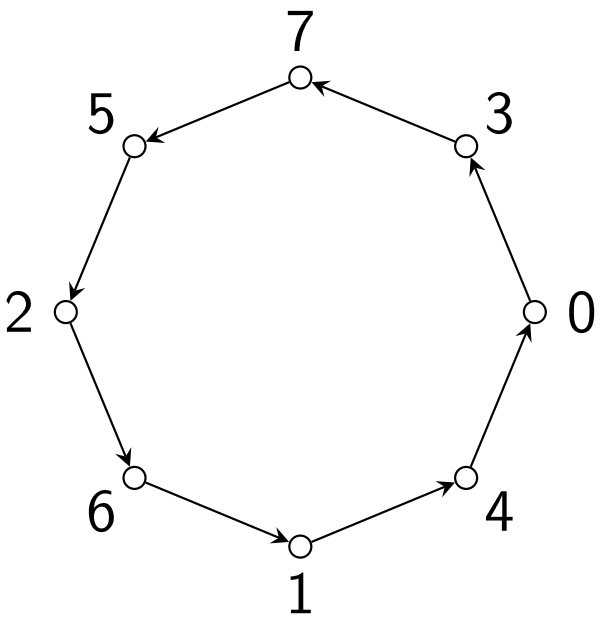
$\dot{\pi}$

$=$

$$\overline{\pi} = (0, 3, 7, 5, 2, 6, 1, 4) \circ (0, 1, 2, 3, 4, 5, 6, 7)$$

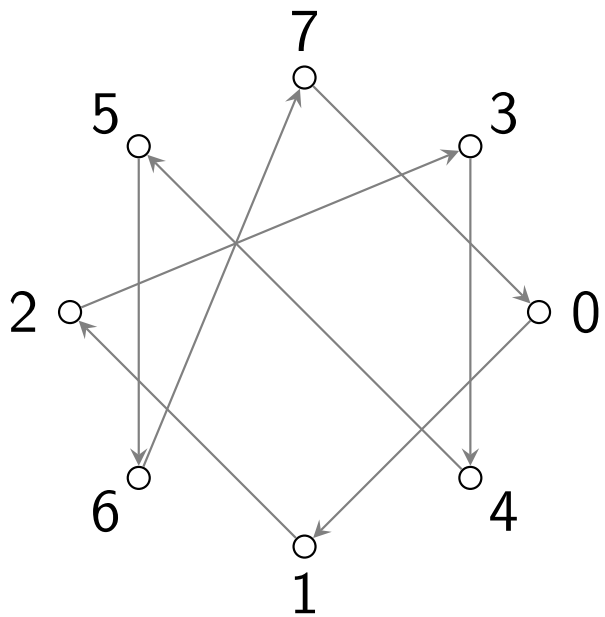
Expression de $G(\pi)$ par une nouvelle permutation

- Isolons le "cycle noir" et le "cycle gris" :



$\hat{\pi}$

$=$

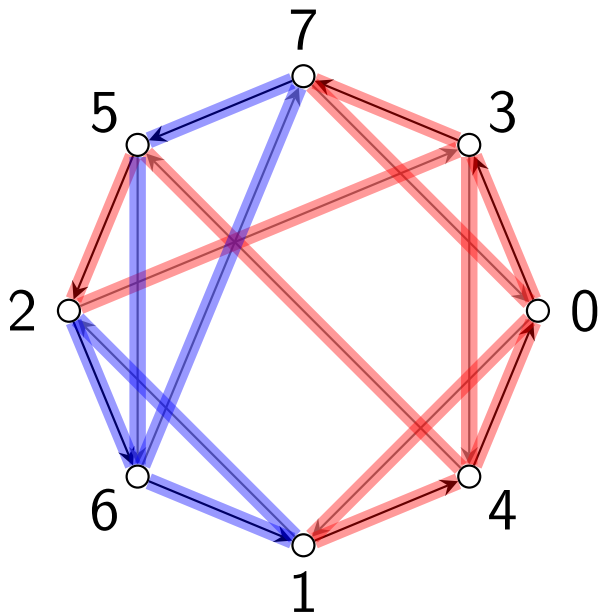


$\dot{\pi}$

$$\overline{\pi} = (0,3,7,5,2,6,1,4) \circ (0,1,2,3,4,5,6,7)$$
$$= (0,4,2,7,3)(1,6,5)$$

Expression de $G(\pi)$ par une nouvelle permutation

- On a $\overline{\pi} = \hat{\pi} \circ \dot{\pi}$, avec $\Gamma(\overline{\pi}) \simeq G(\pi)$; en effet :



$$\begin{array}{rccc} & \hat{\pi} & & \dot{\pi} \\ & = & & = \\ \overline{\pi} = & (0,3,7,5,2,6,1,4) & \circ & (0,1,2,3,4,5,6,7) \\ = & \multicolumn{3}{c}{(0,4,2,7,3)(1,6,5)} \end{array}$$

Application 1 : bornes sur diverses distances

- Remarque : "trier par S" revient à "factoriser par S" :

$$\pi \circ \underbrace{x_1 \circ x_2 \circ \cdots \circ x_t}_{x_1,x_2,\ldots,x_t \in S} = \iota \Leftrightarrow \pi = \underbrace{x_t^{-1} \circ x_{t-1}^{-1} \circ \cdots \circ x_1^{-1}}_{x_1^{-1},x_2^{-1},\ldots,x_t^{-1} \in S}$$

Théorème 1 ([Labarre, 2008])

Soit :

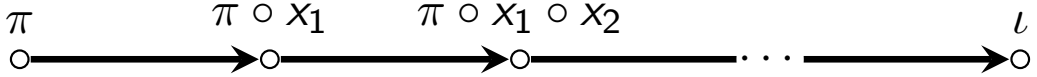
1. $S \subset S_n$, *où* $S = \{s_1, s_2, \ldots\}$,
2. $S' = \{\overline{s_1}, \overline{s_2}, \ldots\}$,
3. $\mathcal{C}$ *l'ensemble des classes de conjugaison intersectant* S'.

Alors pour tout π *dans* S_n, *toute factorisation de* π *en* t *éléments de* S *fournit une factorisation de* $\overline{\pi}$ *en* t *éléments de* $\mathcal{C}$.

Application 1 : minorations sur diverses distances

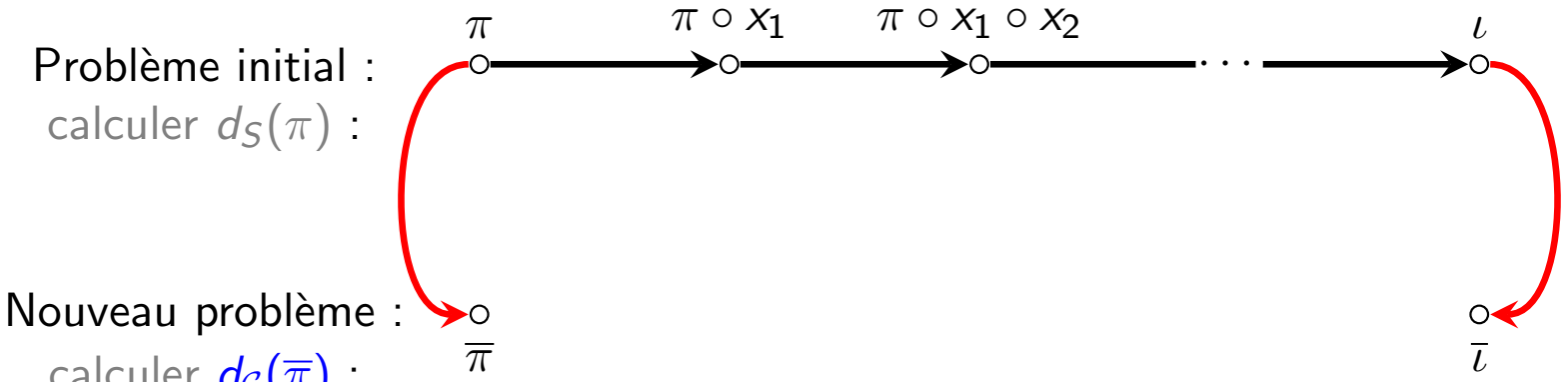
- *Minoration* : fonction m telle que $m(\pi) \leq d_S(\pi)$ pour tout π ;
- Utilisation du Théorème 1 :

Problème initial :
calculer $d_S(\pi)$:



Application 1 : minorations sur diverses distances

- *Minoration* : fonction m telle que $m(\pi) \leq d_S(\pi)$ pour tout π ;
- Utilisation du Théorème 1 :



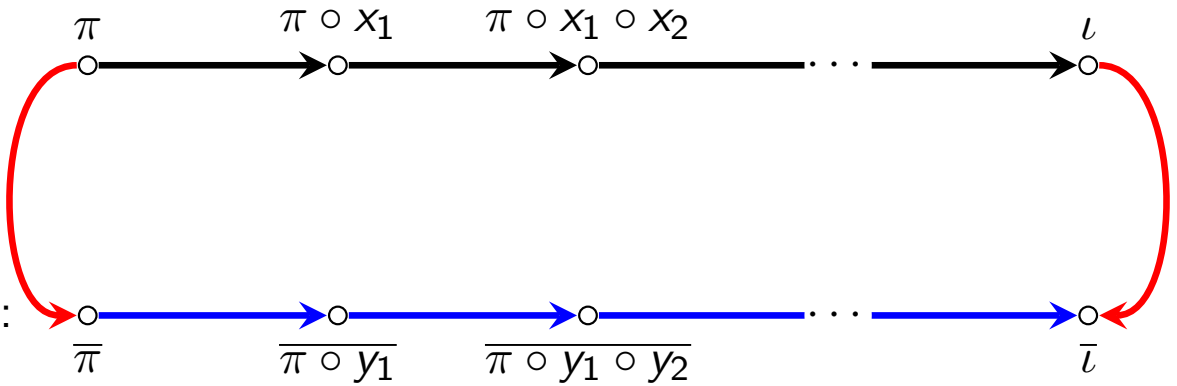
Application 1 : minorations sur diverses distances

- *Minoration* : fonction m telle que $m(\pi) \leq d_S(\pi)$ pour tout π ;
- Utilisation du Théorème 1 :

Problème initial :
calculer $d_S(\pi)$:

$\boxed{d_{\mathcal{C}}(\overline{\pi}) \leq d_S(\pi)}$

Nouveau problème :
calculer $d_{\mathcal{C}}(\overline{\pi})$:



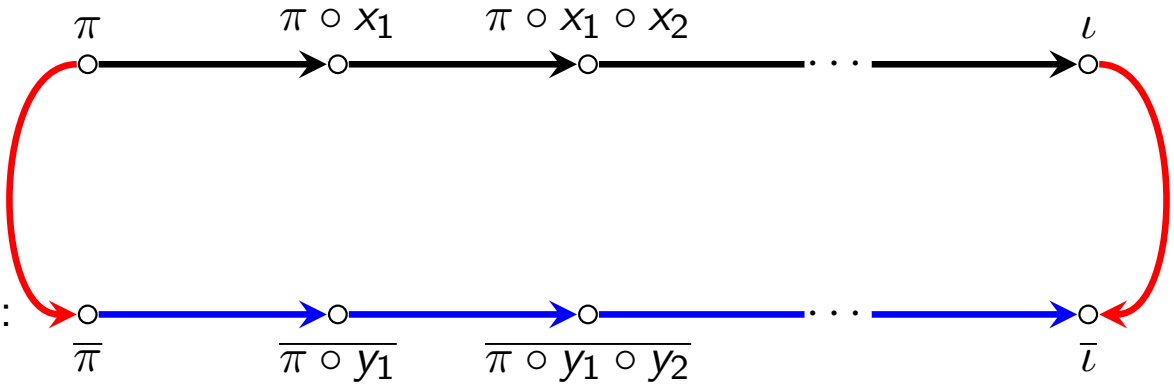
Application 1 : minorations sur diverses distances

- *Minoration* : fonction m telle que $m(\pi) \leq d_S(\pi)$ pour tout π ;
- Utilisation du Théorème 1 :

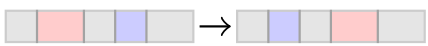
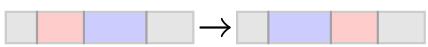
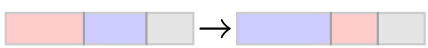
Problème initial : calculer $d_S(\pi)$:

$\boxed{d_{\mathcal{C}}(\overline{\pi}) \leq d_S(\pi)}$

Nouveau problème : calculer $d_{\mathcal{C}}(\overline{\pi})$:



- Corollaire(s) : minorations sur les distances induites par :

1. les échanges de blocs (connue [Christie, 1996]) ;
    (égale la distance exacte)
2. les transpositions (connue [Bafna et Pevzner, 1998]) ;
   
3. les transpositions préfixes (nouvelle [Labarre, 2008]) ;
    (surpasse tous les résultats précédents)

Application 2 : dénombrement

- Stirling : nombre $\mathcal{S}(n,k)$ de permutations π dans S_n avec $c(\Gamma(\pi)) = k$ (nombre de Stirling de première espèce) ;
- [Hultman, 1999] : quel est le nombre $\mathcal{S}_H(n,k)$ de permutations π dans S_n avec $c(G(\pi)) = k$?

Théorème 2 ([Doignon et Labarre, 2007])

$\mathcal{S}_H(n,k)$ est le nombre de factorisations d'un $(n+1)$-cycle fixé en le produit d'un $(n+1)$-cycle et d'une permutation ayant k cycles :

$$\frac{(n+1)!}{2^{n+1-k}} \sum_{(\mu_1,\ldots,\mu_k)\vdash(n+1)} \frac{1}{z_\mu} \sum_{i=0}^{\frac{n+1-k}{2}} \frac{1}{2i+1} \sum_{(j_1,\ldots,j_k)\models\frac{n+1-k}{2}-i} \prod_{h=1}^{k} \binom{\mu_h}{2j_h+1}.$$

- $\longrightarrow$ Caractérisation/estimation de la distribution de diverses distances ;

Tri par transpositions

- *Transposition* : déplacement d'un bloc d'éléments (ou encore : échange de deux intervalles adjacents) ;
- But : trier la permutation en utilisant le moins de transpositions possible ;

Exemple 5 (tri par transpositions)

La permutation suivante peut être triée en deux transpositions, mais pas moins :

$$\langle 3 \boxed{2\ 5} \boxed{4\ 1} \rangle \rightarrow \langle \boxed{3\ 4} \boxed{1\ 2}\ 5 \rangle \rightarrow \langle 1\ 2\ 3\ 4\ 5 \rangle ;$$

sa distance est donc 2.

Etat des connaissances et contributions personnelles

- Les problèmes suivants sont ouverts depuis 1995 :
  1. déterminer la complexité du tri et du calcul de la distance ;
  2. déterminer la valeur maximale de la distance ;
- Meilleure approximation polynomiale : 11/8 [Elias et Hartman, 2006] ;
- Résultats basés sur $G(\pi)$;
- En utilisant plutôt $\Gamma(\pi)$, on obtient :
  1. de nouvelles majorations ;
  2. des instances dont la distance est calculable en temps polynomial ;

Distance des γ-permutations et majoration sur *td*

- *Majoration* : fonction M telle que $d_S(\pi) \leq M(\pi)$ pour tout π ;
- Stratégie : trier chaque cycle de $\Gamma(\pi)$ séparément ;
- Ceci requiert au plus $k - (k \mod 2)$ transpositions par k-cycle ;
- Or : $\sum_{C \in \Gamma(\pi)} |C| - (|C| \mod 2) = n - c_{odd}(\Gamma(\pi))$;
- Ainsi :

Théorème 3 ([Labarre, 2006])

Pour toute permutation π dans S_n, on a

$$td(\pi) \leq n - c_{odd}(\Gamma(\pi)),$$

et il y a égalité pour les γ-permutations (entre autres).

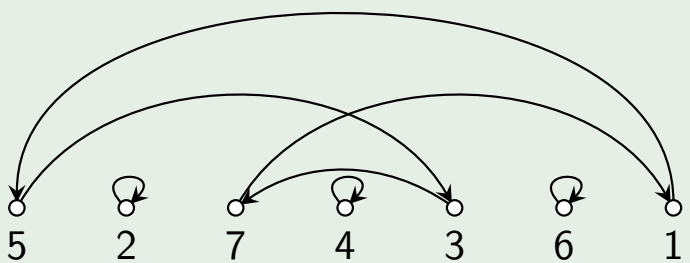
γ-permutations

Définition 4 (γ-permutation)

Permutation π de n (impair) telle que :

1. $\pi_i = i$ pour tout i pair, et
2. pour aucune position i, on n'a $\pi_{i+1} = \pi_i + 1$;

Exemple 6



$\Gamma(\pi)$

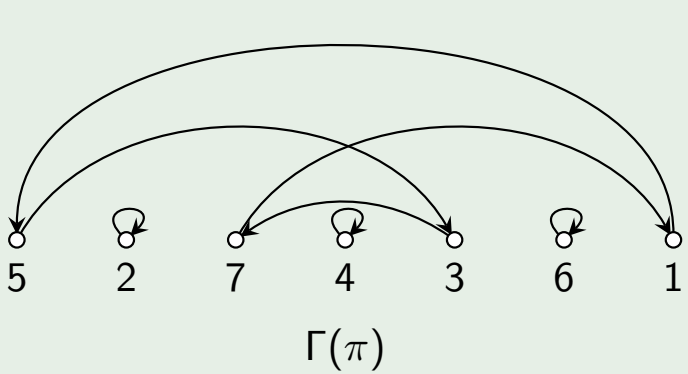
γ-permutations

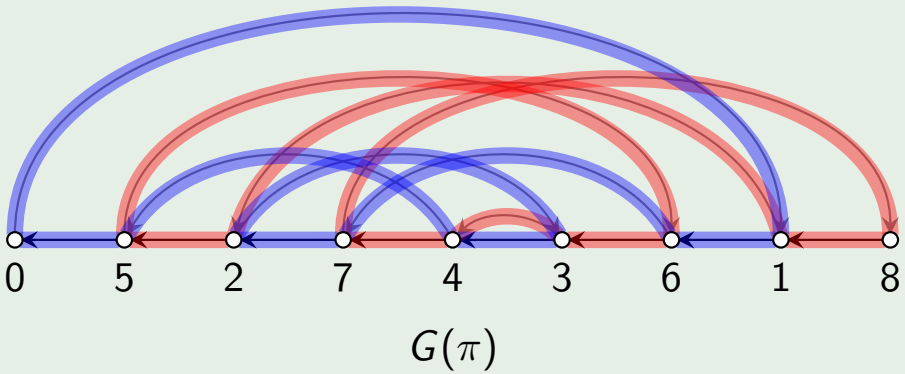
Définition 4 (γ-permutation)

Permutation π de n (impair) telle que :

1. $\pi_i = i$ pour tout i pair, et
2. pour aucune position i, on n'a $\pi_{i+1} = \pi_i + 1$;

Exemple 6





Correspondance entre $G(\pi)$ et $\Gamma(\pi)$ pour les γ-permutations

Proposition 5 ([Labarre, 2006])

Pour toute γ-permutation π de n éléments :

$$\begin{cases} c_{even}(G(\pi)) &= 2\, c_{even}(\Gamma(\pi)); \\ c_{odd}(G(\pi)) &= 2\left(c_{odd}(\Gamma(\pi)) - \frac{n-1}{2}\right). \end{cases}$$

- Par un simple calcul et la minoration connue, on obtient :

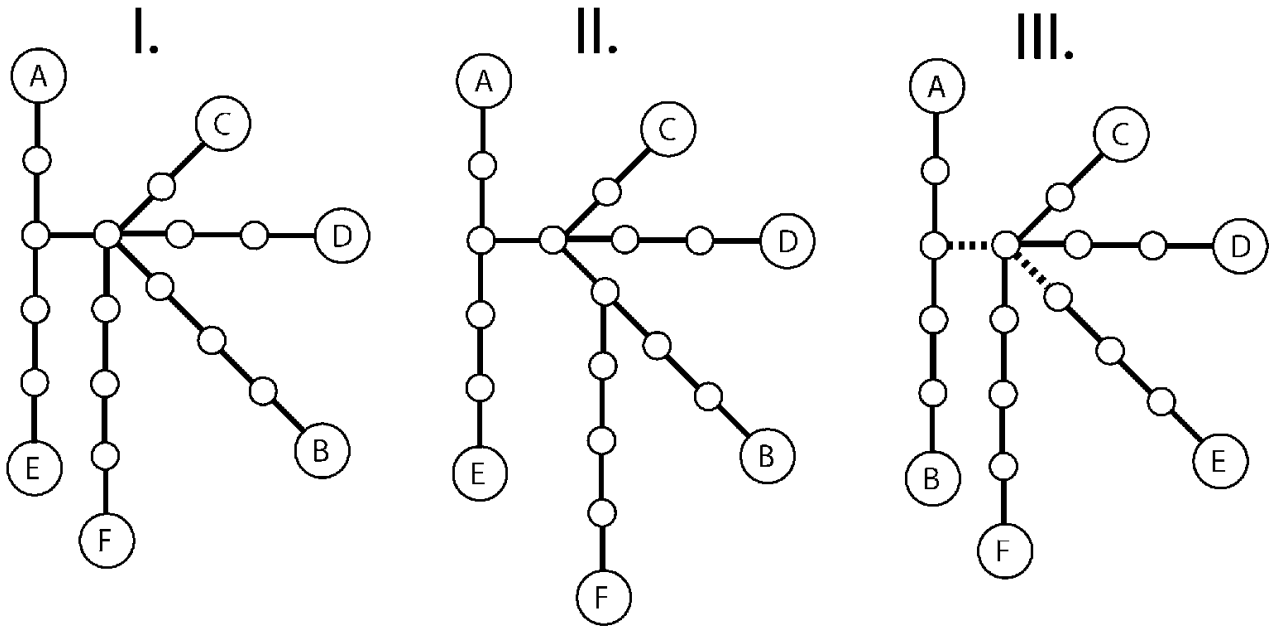
Lemme 6 ([Labarre, 2006])

Pour toute γ-permutation π dans S_n :

$$td(\pi) \geq n - c_{odd}(\Gamma(\pi)).$$

Le problème

- Topologies divergentes pour un même ensemble de gènes :



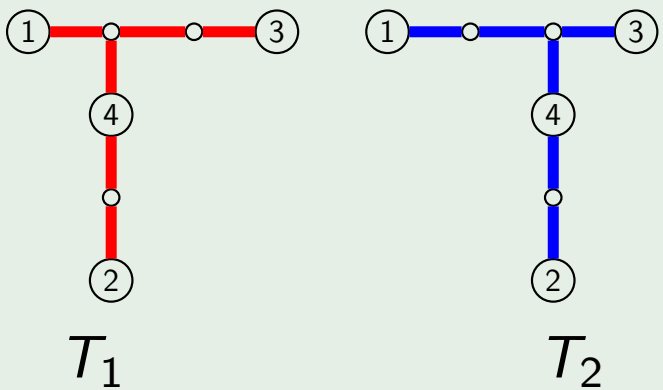
- Aucune n'est "préférable", on veut trouver un bon moyen pour toutes les représenter ;
- Idée [Cassens et al., 2005] : les fusionner en un graphe ;

Ma formalisation

Définitions 7 ((n, k)-graphe, supergraphe commun)

1. (n, k)-graphe : graphe sur n sommets dont k sont étiquetés (par $\{1, 2, \ldots, k\}$), en particulier les sommets de degré 1.

Exemple 7 ((n, k)-graphes et supergraphes communs)



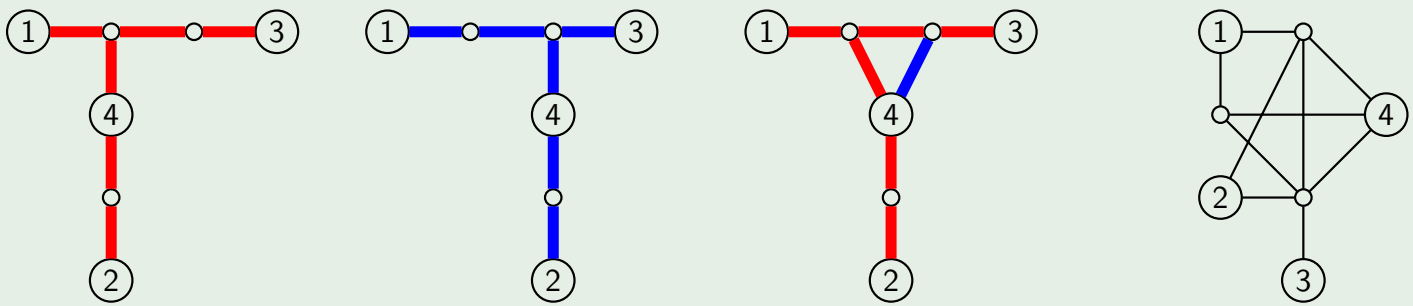
T_1 T_2

Ma formalisation

Définitions 7 ((n, k)-graphe, supergraphe commun)

1. (n, k)-graphe : graphe sur n sommets dont k sont étiquetés (par $\{1, 2, \ldots, k\}$), en particulier les sommets de degré 1.
2. supergraphe commun à $G_1, G_2, \ldots, G_t$: (n, k)-graphe G contenant G_i pour tout i ; G est *minimum* s'il n'existe aucun autre tel graphe G' avec $|E(G')| < |E(G)|$.

Exemple 7 ((n, k)-graphes et supergraphes communs)



T_1 T_2 G_1 (7 arêtes) G_2 (10 arêtes)

Problème et résultats

- Formalisation de la "fusion d'arbres" :

SUPERGRAPHE COMMUN MINIMUM :
Instance : (n, k)-arbres T_1, T_2, ..., T_t.
Problème : trouver un supergraphe commun à T_1, T_2, ..., T_t qui soit minimum.

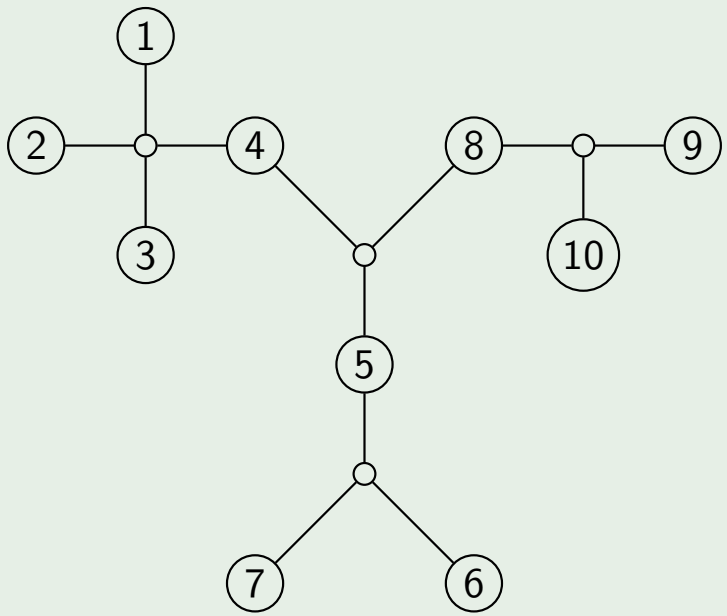
- Résultats : algorithmes exacts pour deux (n, k)-graphes :
  - polynomial ($O((n-k)^3)$) si l'un des graphes est *restreint* ;
  - exponentiel ($O((n-k)^{p+3})$) sinon.

Graphes restreints

Définition 8 ((n, k)-graphe restreint)

(n, k)-graphe dans lequel aucune paire de sommets non-étiquetés ne forme une arête.

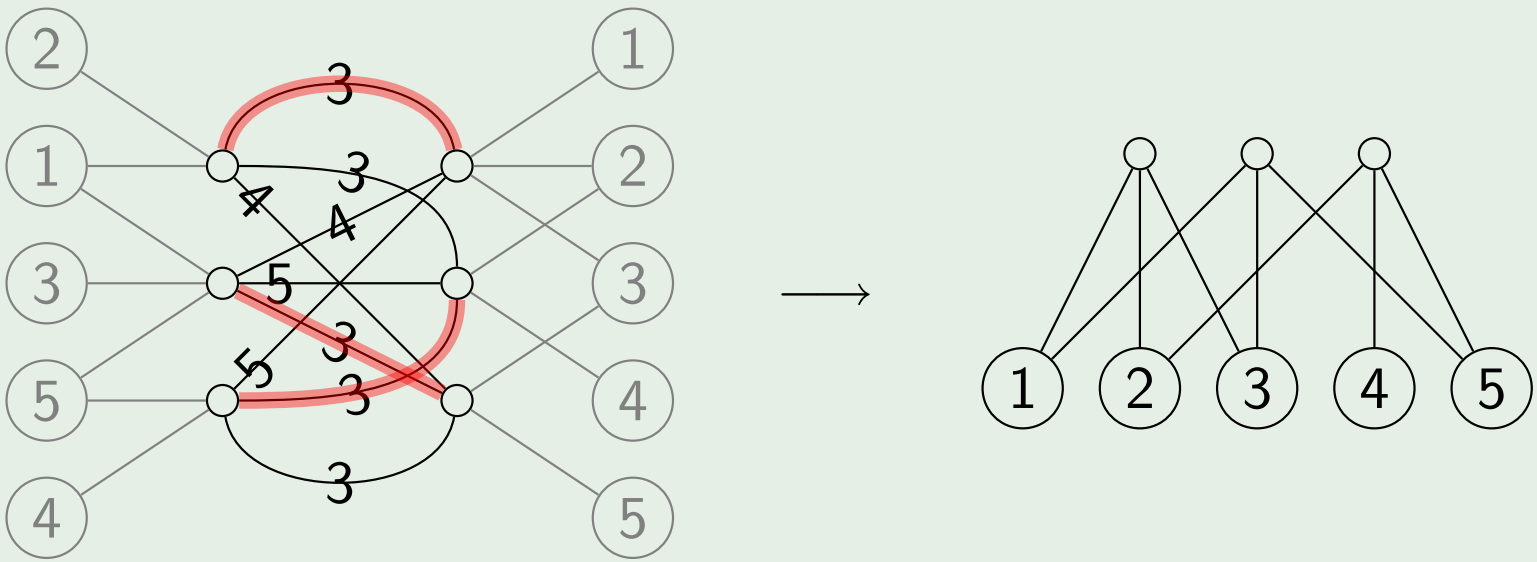
Exemple 8 (arbre restreint)



Graphes restreints

- Fusion optimale de deux (n, k)-graphes restreints en temps $O((n-k)^3)$ (couplage de poids minimum dans un graphe biparti) ;

Exemple 9 (fusion optimale de deux $(8, 5)$-arbres restreints)



- Marche encore si seul un des deux graphes est restreint ;

Cas général

- La stratégie de couplage ne marche plus ;
- Description de l'algorithme :
  1. "restreindre" un des deux graphes (disons G_1) en étiquetant artificiellement p sommets ;
  2. essayer tous les étiquetages possibles de p sommets de G_2 ;
  3. garder la meilleure solution ;

Cas général

- La stratégie de couplage ne marche plus ;
- Description de l'algorithme :
  1. "restreindre" un des deux graphes (disons G_1) en étiquetant artificiellement p sommets ;
     Recherche d'un "minimum vertex cover" sur une forêt ($O(n)$).
  2. essayer tous les étiquetages possibles de p sommets de G_2 ;
     Toute paire de graphes est "fusionnable" en temps $O((n-k)^3)$ puisque G_1 est restreint.
  3. garder la meilleure solution ;
- Temps d'exécution : $O((n-k)^{p+3})$.

Conclusions

- Questions ouvertes :
  - Complexité/approximation :
    - transpositions (+ préfixes) : tri, distance, diamètre
    - supergraphe commun minimum
- Extension/amélioration des résultats :
  - utilisation de $\Gamma(\pi)$ dans d'autres problèmes de réarrangements ? d'autres notions ?
  - le modèle $\overline{\pi}$: permutations signées ? majorations ?
  - simplification des résultats de dénombrement ?
  - fusion optimale de plus de deux graphes/arbres ?


Bafna, V. et Pevzner, P. A. (1998).
Sorting by transpositions.
SIAM Journal on Discrete Mathematics, 11(2) :224–240 (electronic).


Buneman, P. (1971).
The recovery of trees from measures of dissimilarity.
Dans Hodson, F. R., Kendall, D. G., et Tautu, P., éditeurs, *Mathematics in the Archaeological and Historical Sciences*, pages 387–395. Edinburgh University Press.


Cassens, I., Mardulyn, P., et Milinkovitch, M. C. (2005).
Evaluating intraspecific "network" construction methods using simulated sequence data : Do existing algorithms outperform the global maximum parsimony approach ?
Systematic Biology, 54(3) :363–372.


Christie, D. A. (1996).
Sorting permutations by block-interchanges.
Inform. Process. Lett., 60(4) :165–169.


Doignon, J.-P. et Labarre, A. (2007).
On Hultman numbers.
Journal of Integer Sequences, 10(6).
Article 07.6.2, 13 pages.


Elias, I. et Hartman, T. (2006).
A 1.375-approximation algorithm for sorting by transpositions.
IEEE/ACM Trans. Comput. Biol. Bioinform., 3(4) :369–379.


Hannenhalli, S. et Pevzner, P. A. (1999).
Transforming cabbage into turnip : Polynomial algorithm for sorting signed permutations by reversals.
J. ACM, 46(1) :1–27.


Hultman, A. (1999).
Toric permutations.
Mémoire, Department of Mathematics, KTH, Stockholm, Sweden.


Labarre, A. (2006).
New bounds and tractable instances for the transposition distance.
IEEE/ACM Trans. Comput. Biol. Bioinform., 3(4) :380–394.


Labarre, A. (2008).
Edit distances and factorisations of even permutations.
Dans Halperin, D. et Mehlhorn, K., éditeurs, *ESA*, vol. 5193 de *LNCS*, pages 635–646. Springer.


Mihaescu, R., Levy, D., et Pachter, L. (2006).
Why neighbor-joining works.
CoRR, abs/cs/0602041.