Exploiting the disjoint cycle decomposition in genome rearrangements

Jean-Paul Doignon   Anthony Labarre[1]
doignon@ulb.ac.be
alabarre@ulb.ac.be

Université Libre de Bruxelles

June 7th, 2007

Ordinal and Symbolic Data Analysis 2007

[1]Funded by the "Fonds pour la Formation à la Recherche dans l'Industrie et dans l'Agriculture" (F.R.I.A.).

Sequence alignment

- Comparison at the nucleotide level;
- Example:

$$\begin{array}{lcccccccccccccc} \textit{species } 1: & \cdots & T & C & C & G & C & C & A & - & - & C & T & A & \cdots \\ & & | & | & & | & & | & | & & & | & & | & \\ \textit{species } 2: & \cdots & T & C & G & G & A & C & T & G & G & C & - & A & \cdots \end{array}$$

- Matches, differences, insertions and deletions;

Genome rearrangements

- Comparison at the gene level;
- Species differ not only by “content”, but also by order:
  - genes spread over different sets of chromosomes;
  - genes ordered differently on the same chromosome;
- Example:
  - many genes in cabbage and turnip are 99% identical;

General statement of the problem

- The problem to solve can be summarized as:

Given two (or more) genomes, find a sequence of mutations that transforms one into the other and is of minimal length.

- Different assumptions yield different models:
  - gene order;
  - gene orientation;
  - duplications/deletions in the genome;
  - mutations taken into account;
  - weights given to mutations;
  - miscellaneous restrictions;

The role of permutations

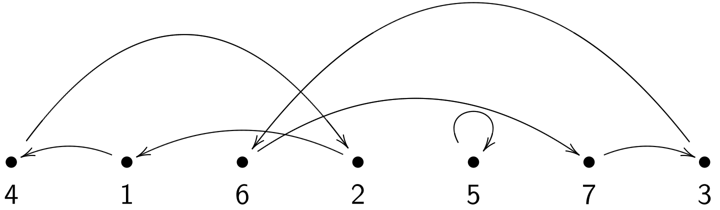
- Permutations model genomes in the case where:
  - the order of genes is known, but not their orientation;
  - each gene appears exactly once in each genome;
- Therefore:
  - $\{\text{genes}\} = \{1, 2, \ldots, n\}$;
  - genome = *permutation* of $\{1, 2, \ldots, n\}$;
- Permutations are therefore viewed as orderings, not as functions;
- One or several operations;

The disjoint cycle decomposition (DCD)

- As is well-known, permutations decompose into a product of disjoint cycles:

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

- We use a particular layout of the associated graph that we call the Γ-*graph*:



Focus of this talk

- Most results in genome rearrangements are based on a central tool called the "cycle graph" (or "breakpoint graph");
- Though the breakpoint graph is a very powerful tool, more classical notions about permutations could be useful for:
  - comparing metrics on permutations;
  - providing information and insight about a particular problem;
  - characterising tractable instances of a particular problem;
- We prove our point by using the DCD to:
  1. derive upper bounds and exhibit polynomial instances for the problem of sorting by transpositions;
  2. solve a counting problem related to the breakpoint graph;

Sorting by transpositions

- Introduced in [Bafna and Pevzner, 1995];
- *Biological* transpositions $\neq$ *algebraic* transpositions ("2-cycles")!
- A transposition displaces an interval of the permutation or, equivalently, exchanges two contiguous intervals:

$$(\ 5 \ \boxed{4 \ 3} \ \boxed{2 \ 1} \ )$$

$$\downarrow$$

$$(\ 5 \ \boxed{2 \ 1} \ \boxed{4 \ 3} \ )$$

- Our problem: transform a permutation into the identity permutation using as few transpositions as possible.

Sorting by transpositions: example

Example

$\pi = (3\ 1\ 4\ 2)$ can be sorted using two transpositions:

$$\begin{array}{rc} \pi = & (\ 3\ \boxed{1}\ \boxed{4}\ 2\ ) \\ & \downarrow \\ & (\ \boxed{3\ 4}\ \boxed{1\ 2}\ ) \\ & \downarrow \\ \iota = & (\ 1\ 2\ 3\ 4\ ) \end{array}$$

Since π cannot be sorted using only one transposition, we have $d(\pi) = 2$.

Status of the problem

- The following problems are open:
  1. the complexity of the sorting problem;
  2. the complexity of computing the associated distance;
  3. determining the maximal value the transposition distance can reach;
- Best approximation ratio for the sorting problem has long been 3/2;
- Improving it down to 11/8 required a computer assisted proof checking over 80 000 cases [Elias and Hartman, 2006] (that algorithm has $O(n^2)$ running time);

The breakpoint graph

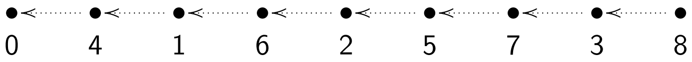
- Given a permutation π, construct the *breakpoint graph* $G(\pi)$ as follows:

0 4 1 6 2 5 7 3 8

1. $V(G) = \{\pi_0 = 0, \pi_1, \pi_2, \ldots, \pi_n, \pi_{n+1} = n+1\}$;
2. $E(G) =$

The breakpoint graph

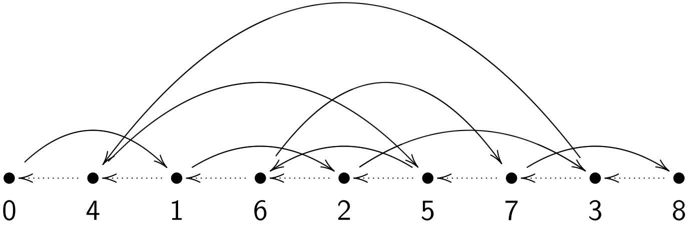
- Given a permutation π, construct the *breakpoint graph* $G(\pi)$ as follows:



1. $V(G) = \{\pi_0 = 0, \pi_1, \pi_2, \ldots, \pi_n, \pi_{n+1} = n+1\}$;
2. $E(G) = \{$*dotted* edges$\}$

The breakpoint graph

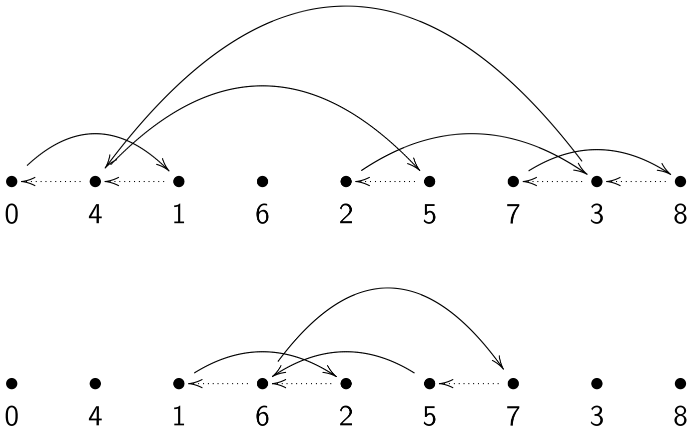
- Given a permutation π, construct the *breakpoint graph* $G(\pi)$ as follows:



1. $V(G) = \{\pi_0 = 0, \pi_1, \pi_2, \ldots, \pi_n, \pi_{n+1} = n+1\}$;
2. $E(G) = \{$*dotted* edges$\} \cup \{$*black* edges$\}$;

The alternating cycle decomposition of $G(\pi)$

- $G(\pi)$ decomposes into *alternating* cycles:



The alternating cycle decomposition of $G(\pi)$

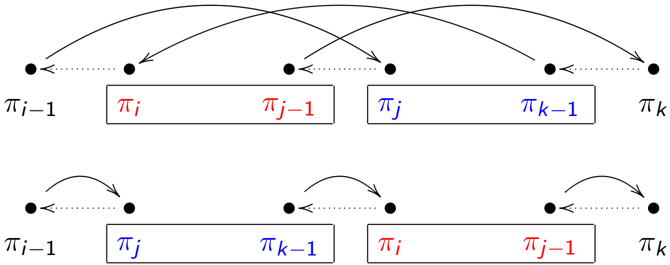
- That decomposition yields a graphical framework for sorting by transpositions:
- The identity $\iota = (1\ 2\ \cdots\ \ n)$ is the only permutation with $c(G(\iota)) = n + 1 = c_{odd}(G(\iota))$;

$G(\iota)$:

- Therefore sorting by transpositions comes down to creating odd alternating cycles "as fast as possible";

A lower bound for sorting by transpositions

- Best case: two new odd cycles in one move:



Theorem

[Bafna and Pevzner, 1995] $\forall\ \pi \in S_n : d(\pi) \geq \frac{n+1-c_{odd}(G(\pi))}{2}$.

Our results [Labarre, 2006]

- A nice correspondence between the Γ-graph and the breakpoint graph for a certain class of permutations called γ-*permutations*;
- $O(n)$ time and space computation of the transposition distance of γ-permutations, without the need of any graph structure;
- A new upper bound on the transposition distance, tight for γ-permutations.
- Even tighter bounds for many other cases.

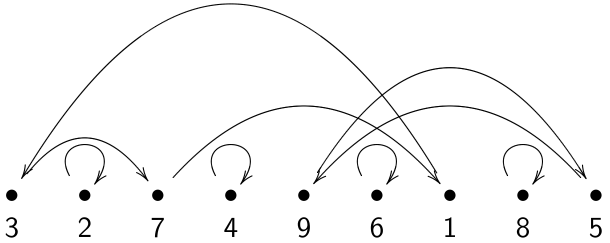
γ-permutations

Definition

For n odd, a permutation π is a γ-*permutation* if:

1. it fixes all even elements, and
2. there is no position i such that $\pi_{i+1} = \pi_i + 1$, for $1 \leq i \leq n-1$;

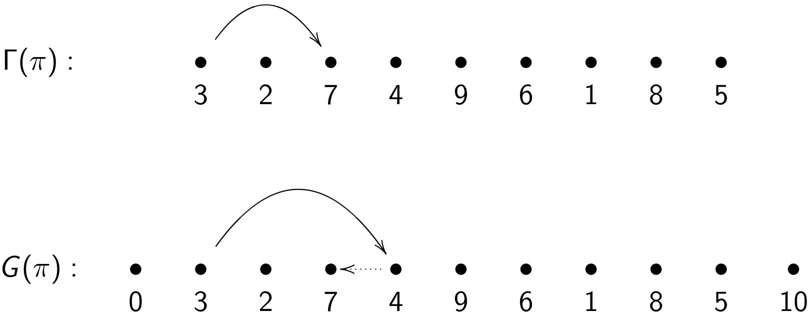
Example



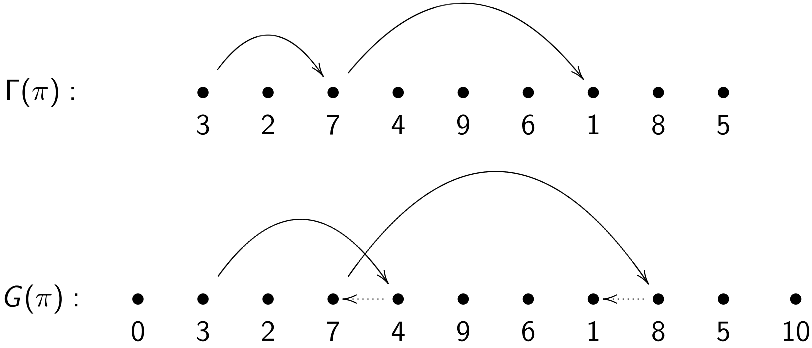
Correspondence between Γ and G for γ-permutations: example

$\Gamma(\pi)$:		3	2	7	4	9	6	1	8	5	
$G(\pi)$:	0	3	2	7	4	9	6	1	8	5	10

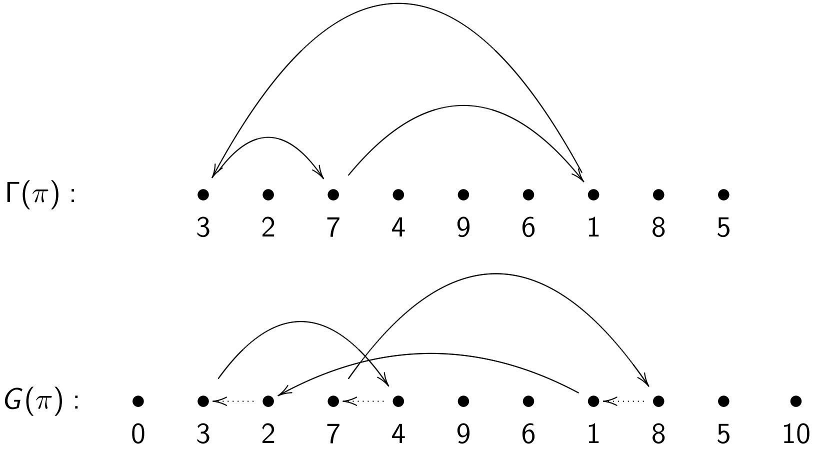
Correspondence between Γ and G for γ-permutations: example



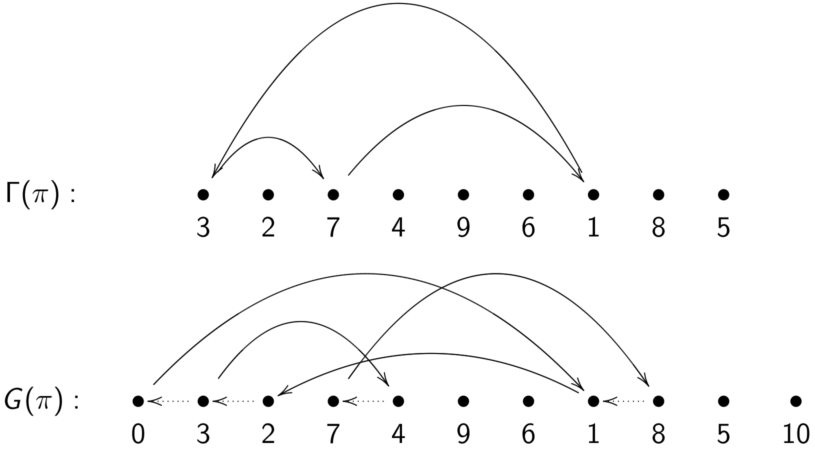
Correspondence between Γ and G for γ-permutations: example



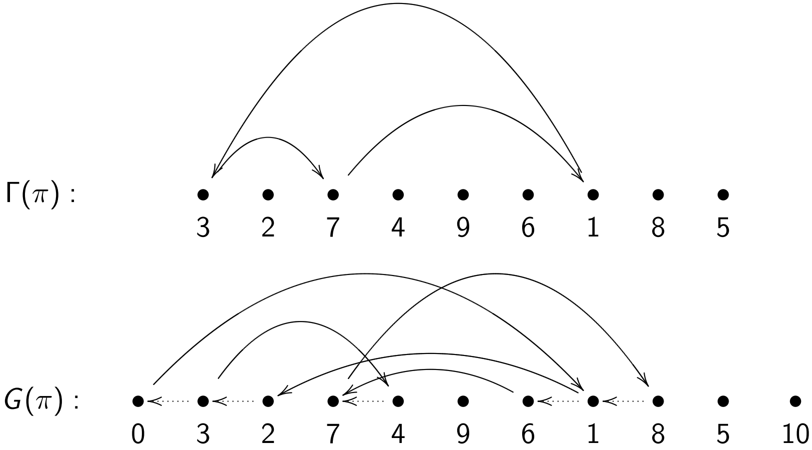
Correspondence between Γ and G for γ-permutations: example



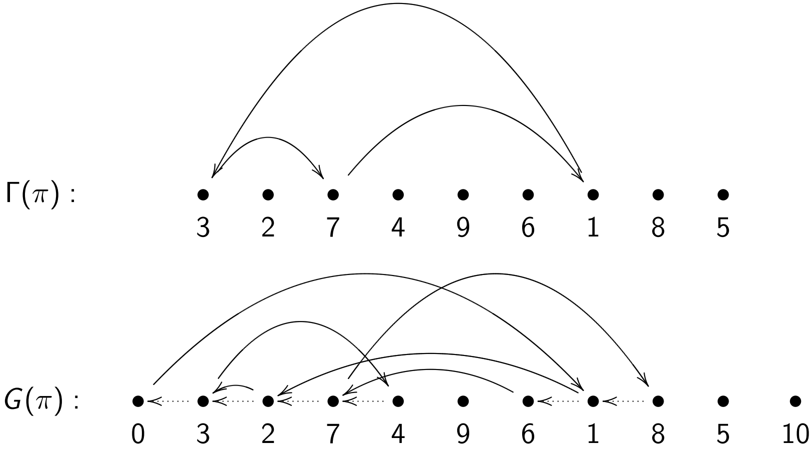
Correspondence between Γ and G for γ-permutations: example



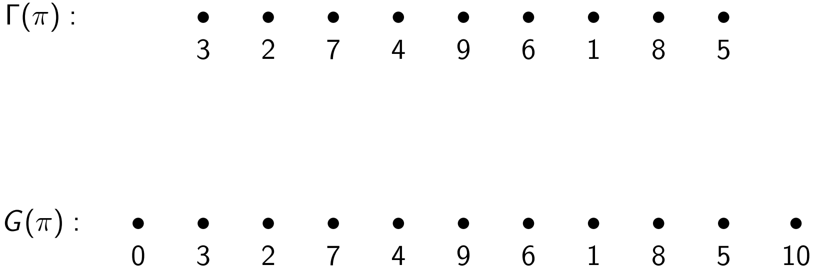
Correspondence between Γ and G for γ-permutations: example



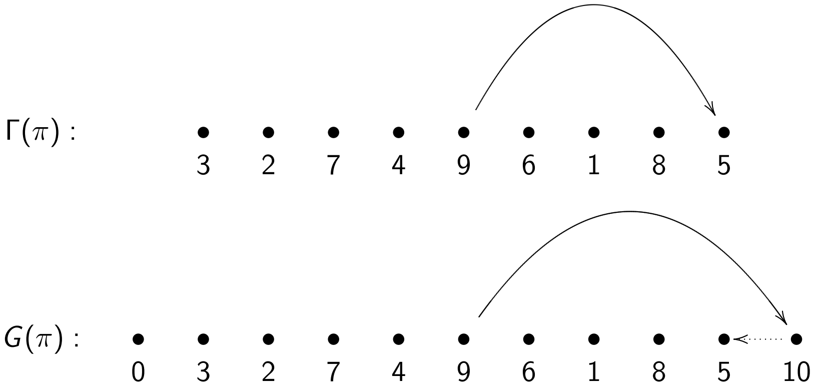
Correspondence between Γ and G for γ-permutations: example



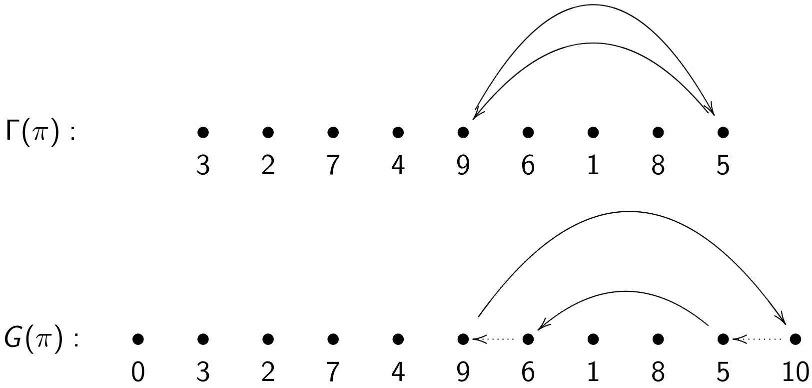
Correspondence between Γ and G for γ-permutations: example



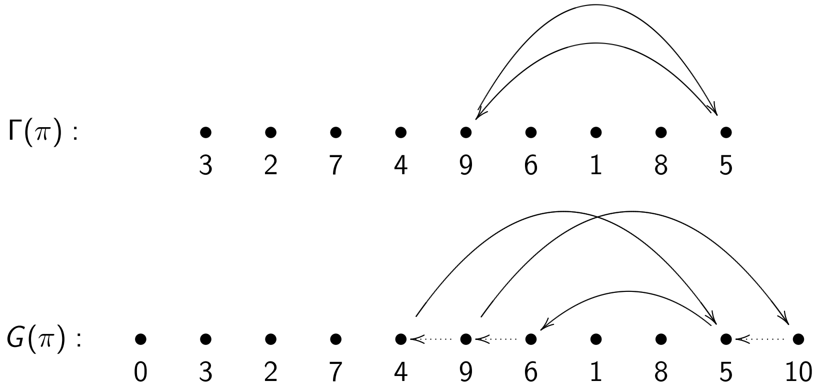
Correspondence between Γ and G for γ-permutations: example



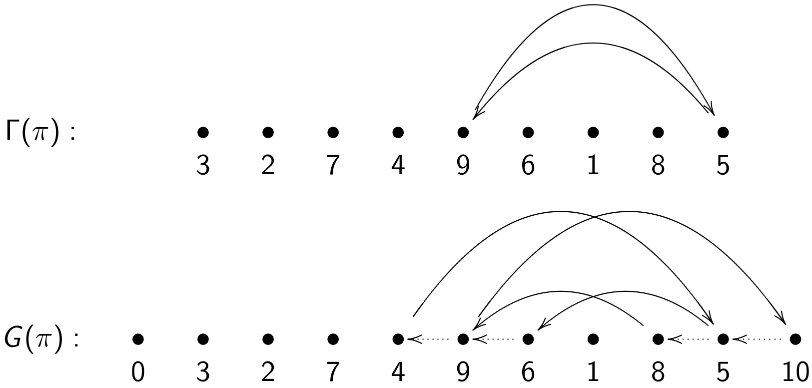
Correspondence between Γ and G for γ-permutations: example



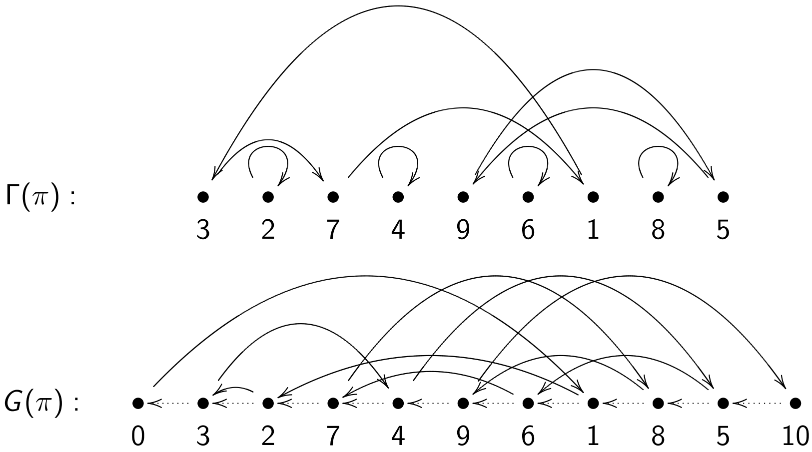
Correspondence between Γ and G for γ-permutations: example



Correspondence between Γ and G for γ-permutations: example



Correspondence between Γ and G for γ-permutations: example



Correspondence between Γ and G for γ-permutations

Proposition

[Labarre, 2006] For every $\gamma-$permutation π in S_n:

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

Lower bound based on the DCD

- Recall that $d(\pi) \geq \frac{n+1-c_{odd}(G(\pi))}{2}$ (Theorem 2);
- This and Proposition 1 yield the following result:

Lemma

[Labarre, 2006] For every $\gamma-$permutation π in S_n:

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

- This lower bound is actually reached;

Transposition distance of γ-permutations

- Strategy: sort each cycle in $\Gamma(\pi)$ independently;

Transposition distance of γ-permutations

- Strategy: sort each cycle in $\Gamma(\pi)$ independently;
- The minimal number of transpositions sorting a k-cycle in $\Gamma(\pi)$ is equal to $k - (k \mod 2)$;

Transposition distance of γ-permutations

- Strategy: sort each cycle in $\Gamma(\pi)$ independently;
- The minimal number of transpositions sorting a k-cycle in $\Gamma(\pi)$ is equal to $k - (k \mod 2)$;
- The strategy yields an upper bound on $d(\pi)$, which is $\sum_{C \in \Gamma(\pi)} |C| - (|C| \mod 2) = n - c_{odd}(\Gamma(\pi))$;

Transposition distance of γ-permutations

- Strategy: sort each cycle in $\Gamma(\pi)$ independently;
- The minimal number of transpositions sorting a k-cycle in $\Gamma(\pi)$ is equal to $k - (k \mod 2)$;
- The strategy yields an upper bound on $d(\pi)$, which is $\sum_{C \in \Gamma(\pi)} |C| - (|C| \mod 2) = n - c_{odd}(\Gamma(\pi))$;
- ... which equals the lower bound of Lemma 5, and therefore:

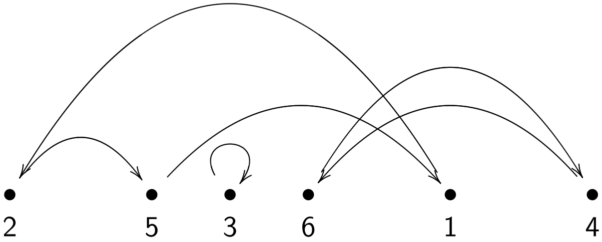
Theorem

[Labarre, 2006] For every γ-permutation in S_n, we have

$$d(\pi) = n - c_{odd}(\Gamma(\pi)).$$

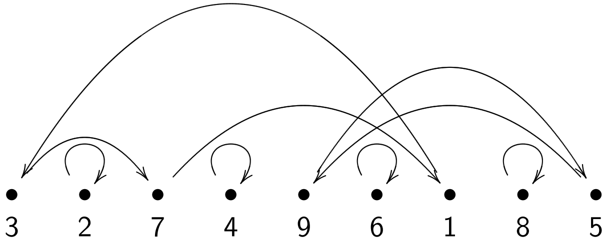
A new upper bound on the transposition distance

- Every permutation (except ι) can be obtained from a (permutation equivalent to a) γ-permutation;



A new upper bound on the transposition distance

- Every permutation (except ι) can be obtained from a (permutation equivalent to a) γ-permutation;



A new upper bound on the transposition distance

- Every permutation (except ι) can be obtained from a (permutation equivalent to a) γ-permutation;
- We can still sort each cycle in Γ independently (but this may not be an optimal strategy anymore);

A new upper bound on the transposition distance

- Every permutation (except ι) can be obtained from a (permutation equivalent to a) γ-permutation;
- We can still sort each cycle in Γ independently (but this may not be an optimal strategy anymore);
- Therefore $d(\pi) \leq d(\sigma)$, where σ is the γ-permutation from which π is obtained by removing k fixed points;

A new upper bound on the transposition distance

- Every permutation (except ι) can be obtained from a (permutation equivalent to a) γ-permutation;
- We can still sort each cycle in Γ independently (but this may not be an optimal strategy anymore);
- Therefore $d(\pi) \leq d(\sigma)$, where σ is the γ-permutation from which π is obtained by removing k fixed points;
- Finally: $d(\sigma) = n + k - c_{odd}(\Gamma(\sigma)) = n - c_{odd}(\Gamma(\pi))$;

A new upper bound on the transposition distance

- Every permutation (except ι) can be obtained from a (permutation equivalent to a) γ-permutation;
- We can still sort each cycle in Γ independently (but this may not be an optimal strategy anymore);
- Therefore $d(\pi) \leq d(\sigma)$, where σ is the γ-permutation from which π is obtained by removing k fixed points;
- Finally: $d(\sigma) = n + k - c_{odd}(\Gamma(\sigma)) = n - c_{odd}(\Gamma(\pi))$;

Theorem

[Labarre, 2006] For every permutation in S_n, we have

$$d(\pi) \leq n - c_{odd}(\Gamma(\pi)).$$

Extensions

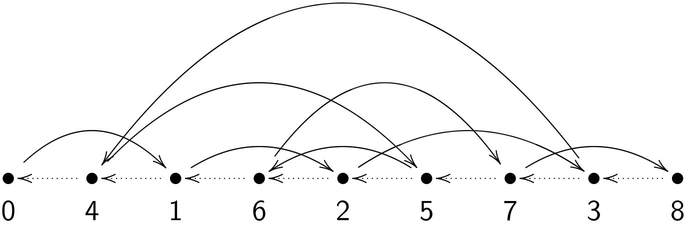
- Other results can be obtained by analysing the effect of removing fixed points on both Γ and G; we can:
  1. either compute the exact distance in polynomial time, for instance:
     1.1 if no two cycles in Γ cross and all cycles are "monotonic",
     1.2 if no two cycles in Γ cross and all cycles are odd;
  2. or lower our upper bound;
- For more examples and details, see [Labarre, 2006];

The problem

- Recall the *Stirling number of the first kind*, which counts the number of permutations in S_n with k cycles;
- [Hultman, 1999] asked for a characterisation of an analogue number, which counts the number of permutations in S_n whose breakpoint graph has k cycles;
- Using the DCD, we solved Hultman's problem and a more general question [Doignon and Labarre, 2007];

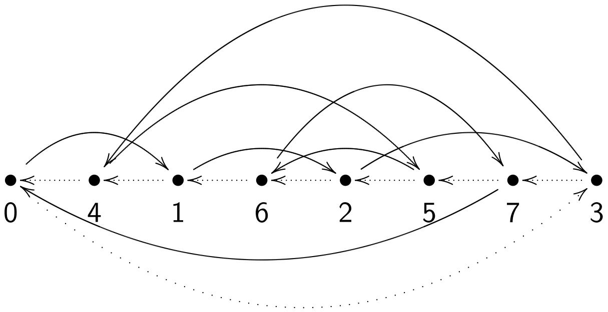
The bijection

- Let π be a permutation in S_n, and $G(\pi)$ its breakpoint graph;



The bijection

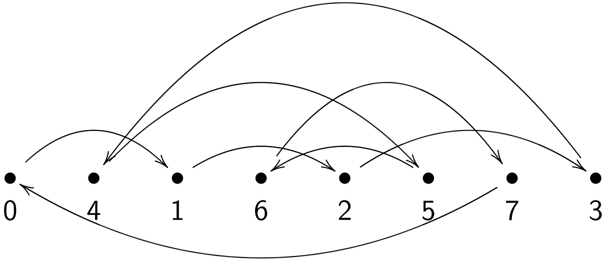
- Let π be a permutation in S_n, and $G(\pi)$ its breakpoint graph;



- We circularise $G(\pi)$ by identifying 0 and $n+1$, thus obtaining $G'(\pi)$;

The bijection

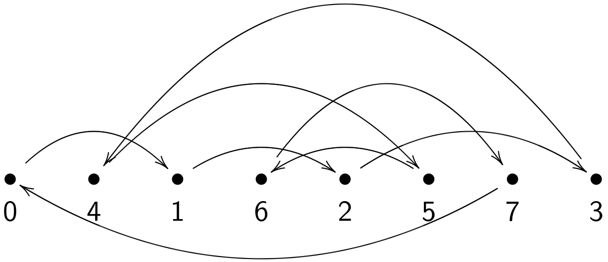
- $G'(\pi)$ yields two permutations:



1. $\alpha =$ the cycle formed by the black edges;

The bijection

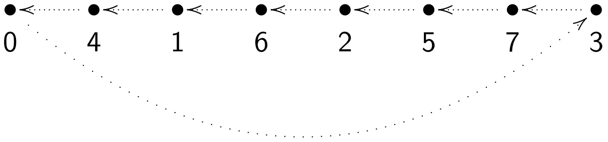
- $G'(\pi)$ yields two permutations:



1. $\alpha = (0, 1, 2, 3, 4, 5, 6, 7)$;

The bijection

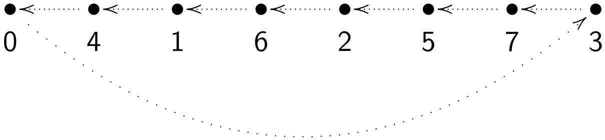
- $G'(\pi)$ yields two permutations:



1. $\alpha = (0, 1, 2, 3, 4, 5, 6, 7)$;
2. $\dot{\pi} =$ the cycle formed by the dotted edges;

The bijection

- $G'(\pi)$ yields two permutations:

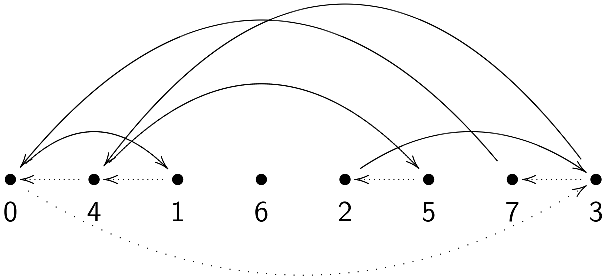


1. $\alpha = (0, 1, 2, 3, 4, 5, 6, 7)$;
2. $\dot{\pi} = (0, 3, 7, 5, 2, 6, 1, 4)$;

The bijection

- The decomposition of $G'(\pi)$ is expressed by those permutations:

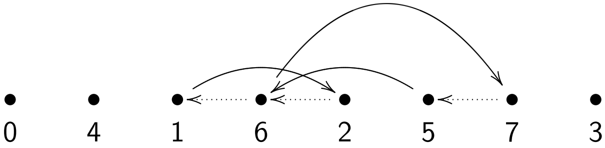
$$\begin{aligned}\dot{\pi} \circ \alpha &= (0,3,7,5,2,6,1,4) \circ (0,1,2,3,4,5,6,7) \\ &= (0,4,2,7,3)(1,6,5) = \mathring{\pi}\end{aligned}$$



The bijection

- The decomposition of $G'(\pi)$ is expressed by those permutations:

$$\begin{aligned}\dot{\pi} \circ \alpha &= (0,3,7,5,2,6,1,4) \circ (0,1,2,3,4,5,6,7) \\ &= (0,4,2,7,3)(1,6,5) = \mathring{\pi}\end{aligned}$$



The bijection

- The decomposition of $G'(\pi)$ is expressed by those permutations:

$$\begin{aligned} \dot{\pi} \circ \alpha &= (0,3,7,5,2,6,1,4) \circ (0,1,2,3,4,5,6,7) \\ &= (0,4,2,7,3)(1,6,5) = \mathring{\pi} \end{aligned}$$

- Note that

$$\mathring{\pi} = \dot{\pi} \circ \alpha \Leftrightarrow \underbrace{\alpha}_{\textit{fixed } (n+1)-\textit{cycle}} = \underbrace{\dot{\pi}^{-1}}_{(n+1)-\textit{cycle}} \circ \underbrace{\mathring{\pi}}_{k \textit{ cycles}}$$

The bijection

- The decomposition of $G'(\pi)$ is expressed by those permutations:

$$\begin{aligned}\dot{\pi} \circ \alpha &= (0,3,7,5,2,6,1,4) \circ (0,1,2,3,4,5,6,7) \\ &= (0,4,2,7,3)(1,6,5) = \mathring{\pi}\end{aligned}$$

- Note that

$$\mathring{\pi} = \dot{\pi} \circ \alpha \Leftrightarrow \underbrace{\alpha}_{\textit{fixed } (n+1)-\textit{cycle}} = \underbrace{\dot{\pi}^{-1}}_{(n+1)-\textit{cycle}} \circ \underbrace{\mathring{\pi}}_{k \textit{ cycles}}$$

Theorem

[Doignon and Labarre, 2007] The Hultman number $\mathcal{S}_H(n,k)$ is the number of factorisations of a fixed $(n+1)$-cycle into the product of an $(n+1)$-cycle and a permutation with k cycles.

Formulas for the Hultman number

- A complicated expression gives an exact formula for $\mathcal{S}_H(n,k)$ [Goupil and Schaeffer, 1998];
- Simpler formulae can be obtained for particular cases:
  - $S_H(n,1) = 2\frac{n!}{n+2}$;
  - the number of permutations whose breakpoint graph has only 2-cycles is

$$\frac{(n+1)!}{\left(\frac{n+1}{2}+1\right)!\,2^{\frac{n+1}{2}}}$$

  - the number of permutations whose breakpoint graph has only 3-cycles is

$$\frac{(n+1)!}{\left(\frac{n+1}{3}\right)!\,12^{\frac{n+1}{3}}}\sum_{i=0}^{\frac{n+1}{3}}\binom{\frac{n+1}{3}}{i}\frac{3^i}{2i+1}$$

Bafna, V. and Pevzner, P. A. (1995).
Sorting permutations by transpositions.
In *Proceedings of SODA*, pages 614–623. ACM/SIAM.

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

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

Goupil, A. and Schaeffer, G. (1998).
Factoring n-cycles and counting maps of given genus.
European Journal of Combinatorics, 19(7):819–834.

Hultman, A. (1999).
Toric permutations.
Master's thesis, 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.

The “complicated formula”

The Hultman number $\mathcal{S}_H(n,k)$ is equal to

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

where $z_\mu = \prod_i \alpha_i!\, i^{\alpha_i}$ and α_i denotes the number of occurences of part i in μ.