A variant of the tandem duplication - random loss model of genome rearrangement

Mathilde Bouvel Dominique Rossin

November 21, 2007

Séminaire des thésards LIAFA-PPS

Outline of the talk

1. Biological motivations and the combinatorial model
2. Previous results: the whole genome duplication - random loss model
3. Some combinatorial properties of the classes $\mathcal{C}(K, 1)$ and $\mathcal{C}(K, p)$
4. Other algorithmic questions to be considered

Duplications and losses in the biological models of genome rearrangement

- Complete genome sequences at disposal:

↪ study molecular evolution and compute distance between genomes

- Classical models of genome rearrangement:

↪ duplications and losses of genes not taken into account

- *On the tandem duplication-random loss model of genome rearrangement* [2005]:

↪ Chaudhuri, Chen, Mihaescu and Rao isolate the duplication-loss problem

The tandem duplication - random loss model

Genes $= \{1, 2, \ldots, n\}$; Genome $=$ Permutation $\sigma = \sigma_1 \ldots \sigma_n \in S_n$

Definition

One *tandem duplication - random loss* step:

1. duplication of a contiguous fragment of the genome, inserted immediately after the original fragment
2. loss of one of the two copies of every duplicated gene

$1\,2\,\overbrace{3\,4\,5\,6}\,7 \rightsquigarrow 1\,2\,\overbrace{3\,4\,5\,6}\,\overbrace{3\,4\,5\,6}\,7 \rightsquigarrow 1\,2\,\cancel{3}\,4\,5\,\cancel{6}\,3\,\cancel{4}\,\cancel{5}\,6\,7 \rightsquigarrow 1\,2\,4\,5\,3\,6\,7$

Beware ! Duplication-loss steps are not symmetric !

$\overbrace{1\,2\,3\,4\,5\,6} \rightsquigarrow 2\,4\,6\,1\,3\,5 \not\rightsquigarrow 1\,2\,3\,4\,5\,6$

The tandem duplication - random loss model

Genes $= \{1, 2, \ldots, n\}$; Genome $=$ Permutation $\sigma = \sigma_1 \ldots \sigma_n \in S_n$

Definition

One *tandem duplication - random loss* step:

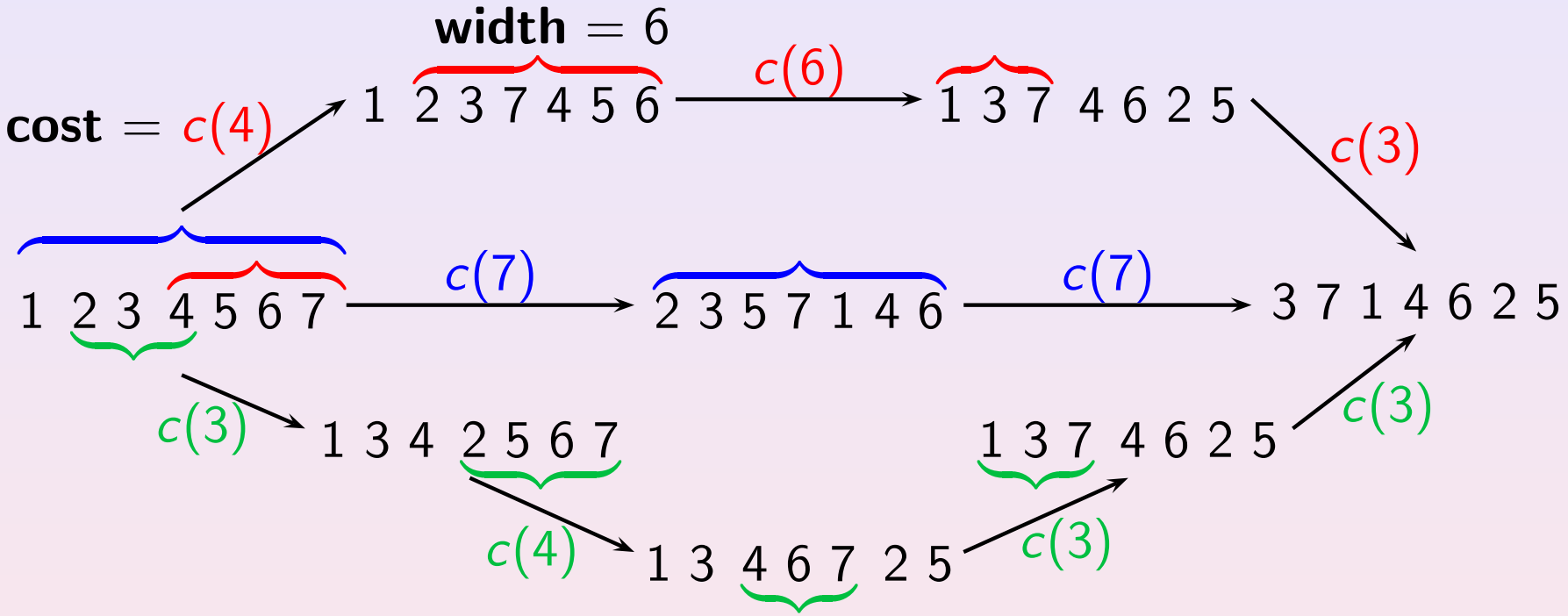
1. duplication of a contiguous fragment of the genome, inserted immediately after the original fragment
2. loss of one of the two copies of every duplicated gene

$1\ 2\ \overbrace{3\ 4\ 5\ 6}\ 7 \rightsquigarrow 1\ 2\ \overbrace{3\ 4\ 5\ 6}\ \overbrace{3\ 4\ 5\ 6}\ 7 \rightsquigarrow 1\ 2\ \cancel{3}\ 4\ 5\ \cancel{6}\ 3\ \cancel{4}\ \cancel{5}\ 6\ 7 \rightsquigarrow 1\ 2\ 4\ 5\ 3\ 6\ 7$

Beware ! Duplication-loss steps are not symmetric !

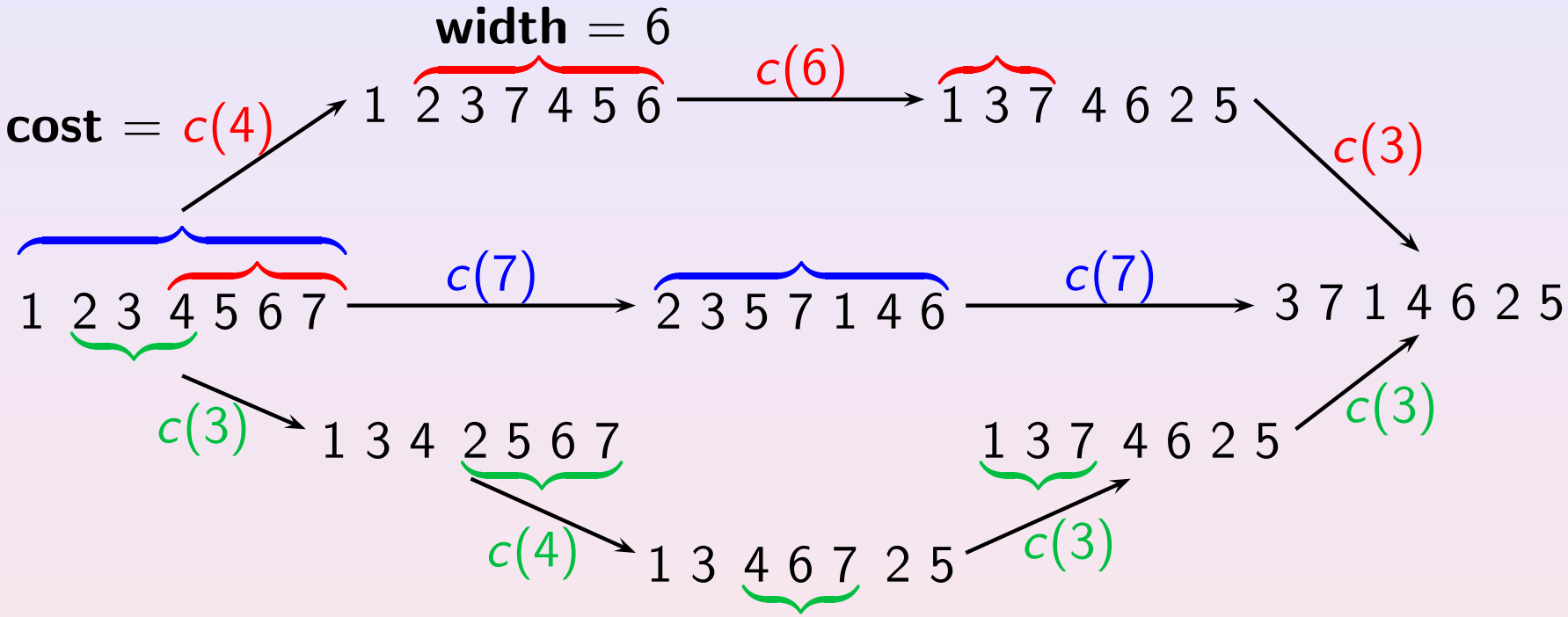
$\overbrace{1\ 2\ 3\ 4\ 5\ 6} \rightsquigarrow 2\ 4\ 6\ 1\ 3\ 5 \not\rightsquigarrow 1\ 2\ 3\ 4\ 5\ 6$

Distances and costs in the duplication-loss model



- "Oriented distance" = minimum cost of a path from σ_1 to σ_2
- Compute $cost(12\ldots n \hookrightarrow \sigma) = cost(\sigma)$ = the minimum cost of a duplication-loss scenario from $12\ldots n$ to σ

Distances and costs in the duplication-loss model



- "Oriented distance" = minimum cost of a path from σ_1 to σ_2
- Compute $cost(12\ldots n \hookrightarrow \sigma) = cost(\sigma)$ = the minimum cost of a duplication-loss scenario from $12\ldots n$ to σ

Some possible cost functions c

- Power cost function:
  width $k \Rightarrow$ cost α^k for some $\alpha \geq 1$

↪ Paper of Chaudhuri, Chen, Mihaescu and Rao ($\alpha = 1$ or ≥ 2)

- Linear or affine cost function

↪ What they suggest to study

- Piecewise constant cost function:

$$\text{width } k \Rightarrow \text{cost } \begin{cases} 1 \text{ if } k \leq K \\ \infty \text{ if } k > K \end{cases}$$

$1 \quad 2 \quad 3 \quad \underbrace{4 \quad 5 \quad 6}_{\text{width} \leq K} \cdots n$

↪ Where we find combinatorial properties

Previous results on the model with power cost function

Duplication-loss on a fragment of width $k \Rightarrow$ cost α^k

- $\alpha = 1$: *whole genome duplication*-random loss model

$\hookrightarrow$ the cost of any step is 1

$\hookrightarrow$ $cost(\sigma)$ is known, together with a corresponding scenario (radix sort algorithm)

- $\alpha \geq 2$: reduces to width $= 2$

$\hookrightarrow$ $cost(\sigma) = \alpha^2 \times$number of inversions in σ (Kendall-Tau or bubblesort distance)

- $1 < \alpha < 2$: open question

Whole genome duplication - random loss model

Definition

There is a *descent* at position i in σ if $\sigma_i > \sigma_{i+1}$.
$desc(\sigma) =$ number of descents of σ.

number of maximal increasing substrings of $\sigma = desc(\sigma) + 1$

Theorem

$cost(\sigma) = \lceil \log_2(desc(\sigma) + 1) \rceil$

- Lower bound: in one duplication-loss step, each maximal increasing substring splits in at most two maximal increasing substrings
- Upper bound: radix sort algorithm

Whole genome duplication - random loss model

Definition

There is a *descent* at position i in σ if $\sigma_i > \sigma_{i+1}$.
$desc(\sigma) =$ number of descents of σ.

number of maximal increasing substrings of $\sigma = desc(\sigma) + 1$

Theorem

$cost(\sigma) = \lceil \log_2(desc(\sigma) + 1) \rceil$

- Lower bound: in one duplication-loss step, each maximal increasing substring splits in at most two maximal increasing substrings
- Upper bound: radix sort algorithm

Computing an optimal duplication-loss scenario for $\sigma \in S_n$

Radix sort algorithm:

1. $\pi = 12 \dots n$
2. Partition σ into maximal increasing substrings
3. $j \in i^{th}$ maximal increasing substring gets label $binary(i-1)$
4. For $j = 1$ to $\lceil \log_2(desc(\sigma)+1) \rceil$, perform a duplication-loss step on π: first copy = elements with a 0 in the j^{th} least significant bit of its label

Example: $\sigma = 78563412 = \overbrace{78}^{00}\ \overbrace{56}^{01}\ \overbrace{34}^{10}\ \overbrace{12}^{11}$ is obtained in 2 steps through the scenario:

$$12345678 = \overbrace{12}^{11}\ \overbrace{34}^{10}\ \overbrace{56}^{01}\ \overbrace{78}^{00} \rightsquigarrow \overbrace{3478}^{*0}\ \overbrace{1256}^{*1}$$

$$\overbrace{34}^{10}\ \overbrace{78}^{00}\ \overbrace{12}^{11}\ \overbrace{56}^{01} \rightsquigarrow \overbrace{7856}^{0*}\ \overbrace{3412}^{1*} = 78563412$$

Patterns in permutations

$\sigma \in S_n$, $\tau \in S_k$ with $k \leq n$

- Permutation σ *involves* pattern τ ($\tau \prec \sigma$) if $\exists$ $1 \leq i_1 < i_2 < \ldots < i_k \leq n$ such that $\sigma_{i_1}\sigma_{i_2}\ldots\sigma_{i_k}$ is order isomorphic to τ : $\sigma_{i_p} < \sigma_{i_q}$ if and only if $\tau_p < \tau_q$
- Otherwise, σ *avoids* τ
- For example, $\mathbf{1}35\mathbf{6}\mathbf{2}\mathbf{4}$ involves 132 and avoids 321

Notation: $S(B) =$ the set of all permutations avoiding simultaneously all the patterns in the basis B

Proposition: a set S of permutations stable for $\prec$ is a class of pattern-avoiding permutations of basis $B =$ the minimal permutations not in $S = \{\sigma \notin S : \forall \tau \prec \sigma, \tau \neq \sigma, \tau \in S\}$

Patterns in permutations

$\sigma \in S_n$, $\tau \in S_k$ with $k \leq n$

- Permutation σ *involves* pattern τ ($\tau \prec \sigma$) if $\exists$ $1 \leq i_1 < i_2 < \ldots < i_k \leq n$ such that $\sigma_{i_1}\sigma_{i_2}\ldots\sigma_{i_k}$ is order isomorphic to τ : $\sigma_{i_p} < \sigma_{i_q}$ if and only if $\tau_p < \tau_q$
- Otherwise, σ *avoids* τ
- For example, $\mathbf{1}35\mathbf{6}\mathbf{2}\mathbf{4}$ involves 132 and avoids 321

Notation: $S(B)$ = the set of all permutations avoiding simultaneously all the patterns in the basis B

Proposition: a set S of permutations stable for $\prec$ is a class of pattern-avoiding permutations of basis B = the minimal permutations not in $S = \{\sigma \notin S : \forall \tau \prec \sigma, \tau \neq \sigma, \tau \in S\}$

Duplication-loss from the pattern-avoidance point of view

For the whole genome duplication - random loss model:

Theorem

$cost(\sigma) = \lceil \log_2(desc(\sigma) + 1) \rceil$

Equivalently: permutations obtainable in at most p steps = permutations with at most $2^p - 1$ descents

Fact: permutations obtainable in at most p steps: set stable for $\prec$

Consequence (Pattern-avoiding permutation class)

Permutations obtainable in at most p steps $= S(B)$
with $B =$ the minimal permutations (for $\prec$) with 2^p descents.

+ local characterization of the permutations in the basis B

Duplication-loss from the pattern-avoidance point of view

For the whole genome duplication - random loss model:

Theorem

$cost(\sigma) = \lceil \log_2(desc(\sigma) + 1) \rceil$

Equivalently: permutations obtainable in at most p steps = permutations with at most $2^p - 1$ descents

Fact: permutations obtainable in at most p steps: set stable for $\prec$

Consequence (Pattern-avoiding permutation class)

Permutations obtainable in at most p steps $= S(B)$
with $B =$ the minimal permutations (for $\prec$) with 2^p descents.

+ local characterization of the permutations in the basis B

Duplication-loss from the pattern-avoidance point of view

For the whole genome duplication - random loss model:

Theorem

$cost(\sigma) = \lceil \log_2(desc(\sigma) + 1) \rceil$

Equivalently: permutations obtainable in at most p steps = permutations with at most $2^p - 1$ descents

Fact: permutations obtainable in at most p steps: set stable for $\prec$

Consequence (Pattern-avoiding permutation class)

Permutations obtainable in at most p steps $= S(B)$
with $B =$ the minimal permutations (for $\prec$) with 2^p descents.

$+$ local characterization of the permutations in the basis B

The variant of the model we considered

Piecewise constant cost function: width $k \Rightarrow$ cost $\begin{cases} 1 \text{ if } k \leq K \\ \infty \text{ if } k > K \end{cases}$

Equivalently: Duplication of fragments of width at most K
Cost = number of steps

Problems to consider:

- Characterization of the permutations obtained in p steps in terms of excluded patterns ?
- Cost of obtaining a permutation ? on average ? in the worst case ?
- Finding an optimal sequence of steps from $12\ldots n$ to σ, *i.e.* a sequence of minimal cost ?

The variant of the model we considered

Piecewise constant cost function: width $k \Rightarrow$ cost $\begin{cases} 1 \text{ if } k \leq K \\ \infty \text{ if } k > K \end{cases}$

Equivalently: Duplication of fragments of width at most K
Cost = number of steps

Problems to consider:

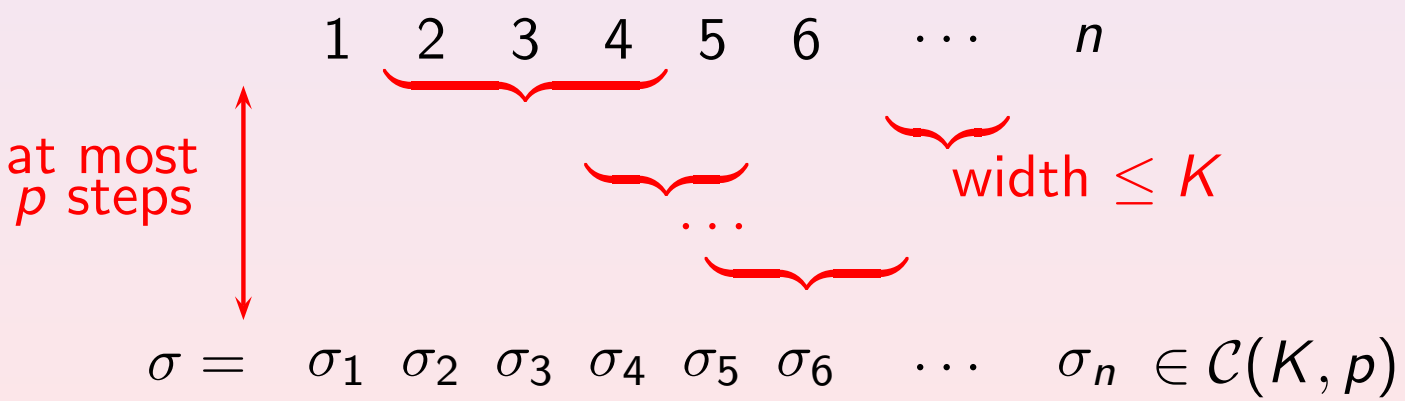
- Characterization of the permutations obtained in p steps in terms of excluded patterns ?
- Cost of obtaining a permutation ? on average ? in the worst case ?
- Finding an optimal sequence of steps from $12\ldots n$ to σ, *i.e.* a sequence of minimal cost ?

Definition of the classes $\mathcal{C}(K,p)$

Definition

$\mathcal{C}(K,p)$ = the class of all permutations obtained from $12\dots n$ (for any n) after p duplication-loss steps of width at most K.

Notice: $\mathcal{C}(K,p)$ is stable for $\prec$



$\mathcal{C}(K,1)$ is a class of pattern-avoiding permutations

Focus on $\mathcal{C}(K,1)$: one duplication-loss step from $12\dots n$

Example: $1\ \overbrace{2\ 3\ 4\ 5}\ 6\ 7\ 8 \rightsquigarrow 1\ 2\ 5\ 3\ 4\ 6\ 7\ 8$

Theorem

$\mathcal{C}(K,1) = S(B)$.

The basis B is $\{321, 3142, 2143\} \cup D$, D being the set of all permutations of S_{K+1} that do not start with 1 nor end with $K+1$, and containing exactly one descent.

$\mathcal{C}(K,1)$ is stable for $\prec \Rightarrow$ the excluded patterns are the minimal permutations not in $\mathcal{C}(K,1)$ (minimal in the sense of $\prec$):

$$B = \{\sigma \notin \mathcal{C}(K,1) : \forall \tau \prec \sigma, \tau \neq \sigma, \tau \in \mathcal{C}(K,1)\}$$

$\mathcal{C}(K,1)$ is a class of pattern-avoiding permutations

Focus on $\mathcal{C}(K,1)$: one duplication-loss step from $12\dots n$

Example: $1\ \overbrace{2\ 3\ 4\ 5}\ 6\ 7\ 8 \rightsquigarrow 1\ 2\ 5\ 3\ 4\ 6\ 7\ 8$

Theorem

$\mathcal{C}(K,1) = S(B)$.

The basis B is $\{321, 3142, 2143\} \cup D$, D being the set of all permutations of S_{K+1} that do not start with 1 nor end with $K+1$, and containing exactly one descent.

$\mathcal{C}(K,1)$ is stable for $\prec \Rightarrow$ the excluded patterns are the minimal permutations not in $\mathcal{C}(K,1)$ (minimal in the sense of $\prec$):

$$B = \{\sigma \notin \mathcal{C}(K,1) : \forall \tau \prec \sigma, \tau \neq \sigma, \tau \in \mathcal{C}(K,1)\}$$

Is $\mathcal{C}(K,p)$ also a pattern-avoiding class ?

Theorem

The class $\mathcal{C}(K,p)$ is a class of pattern-avoiding permutations $S(B)$. Its basis B is finite and contains only patterns of size at most $(Kp+2)^2-2$.

$\mathcal{C}(K,p)$ is stable for the pattern relation $\prec$
$\Rightarrow$ show that the basis is finite + bound the size of the patterns

Idea of the proof:
Consider the minimal permutations $\sigma \notin \mathcal{C}(K,p)$, and bound the necessary moves of elements to go from $12\dots n$ to σ

How many steps from $12\dots n$ to σ ?

- Lower bound: $\Omega(\log n + \frac{n^2}{K^2})$ steps in the worst case and on average
  - $\hookrightarrow$ $\log n$ from the whole genome duplication - random loss model
  - $\hookrightarrow$ $\frac{n^2}{K^2}$ considering the number of invertions
- Algorithm (upper bound): $\Theta(\frac{n}{K}\log K + \frac{n^2}{K^2})$ steps in the worst case and on average
  - $\hookrightarrow$ $\frac{n^2}{K^2}$ for long moves
  - $\hookrightarrow$ $\frac{n}{K}\log K$ for local reorderings

Lower and upper bound **coincide** up to a constant factor except when $\frac{n}{\log n} \ll K = K(n) \ll n$ (in particular when $K =$ constant)

On a given σ our algorithm may yield a scenario **far from optimal** (*e.g.* n steps instead of $\sqrt{n}$ when $K(n) = \sqrt{n}$)

How many steps from $12\dots n$ to σ ?

- Lower bound: $\Omega(\log n + \frac{n^2}{K^2})$ steps in the worst case and on average
  - $\hookrightarrow$ $\log n$ from the whole genome duplication - random loss model
  - $\hookrightarrow$ $\frac{n^2}{K^2}$ considering the number of invertions
- Algorithm (upper bound): $\Theta(\frac{n}{K}\log K + \frac{n^2}{K^2})$ steps in the worst case and on average
  - $\hookrightarrow$ $\frac{n^2}{K^2}$ for long moves
  - $\hookrightarrow$ $\frac{n}{K}\log K$ for local reorderings

Lower and upper bound **coincide** up to a constant factor except when $\frac{n}{\log n} \ll K = K(n) \ll n$ (in particular when $K =$ constant)

On a given σ our algorithm may yield a scenario **far from optimal** (*e.g.* n steps instead of $\sqrt{n}$ when $K(n) = \sqrt{n}$)

How many steps from $12\dots n$ to σ ?

- Lower bound: $\Omega(\log n + \frac{n^2}{K^2})$ steps in the worst case and on average
  - $\hookrightarrow$ $\log n$ from the whole genome duplication - random loss model
  - $\hookrightarrow$ $\frac{n^2}{K^2}$ considering the number of invertions
- Algorithm (upper bound): $\Theta(\frac{n}{K}\log K + \frac{n^2}{K^2})$ steps in the worst case and on average
  - $\hookrightarrow$ $\frac{n^2}{K^2}$ for long moves
  - $\hookrightarrow$ $\frac{n}{K}\log K$ for local reorderings

Lower and upper bound **coincide** up to a constant factor except when $\frac{n}{\log n} \ll K = K(n) \ll n$ (in particular when $K =$ constant)

On a given σ our algorithm may yield a scenario **far from optimal** (*e.g.* n steps instead of $\sqrt{n}$ when $K(n) = \sqrt{n}$)

Open questions

Algorithmic:

- Formula for $cost(\sigma)$?
- Optimal sequence of steps from $12\dots n$ to σ ?
- Characterization of those sequences ? with a decreasing energy function ?
- Does our algorithm compute a $f(K)$-approximation of such an optimal scenario ?

Combinatorics:

- Description of the excluded patterns in $\mathcal{C}(K,p)$?
- Order of the cardinality of $\mathcal{C}(K,1)$ and $\mathcal{C}(K,p)$?