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

Mathilde Bouvel   Dominique Rossin

Permutation Patterns 2007

LIAFA

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 questions to be considered

Duplications and losses in the biological models of genome rearrangement

- Complete genome sequences at disposal:

$\hookrightarrow$ study molecular evolution and compute distance between genomes

- Classical models of genome rearrangement:

$\hookrightarrow$ duplications and losses of genes not taken into account

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

$\hookrightarrow$ Chaudhuri, Chen, Mihaescu and Rao isolate the duplication-loss problem

The tandem duplication - random loss model

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

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

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

The tandem duplication - random loss model

Example

$$1\ 2\ \overbrace{3\ 4\ 5\ 6}\ 7 \rightsquigarrow 1\ 2\ \overbrace{3\ 4\ 5\ 6}\ \color{red}{\overbrace{3\ 4\ 5\ 6}}\ \color{black}{7}$$

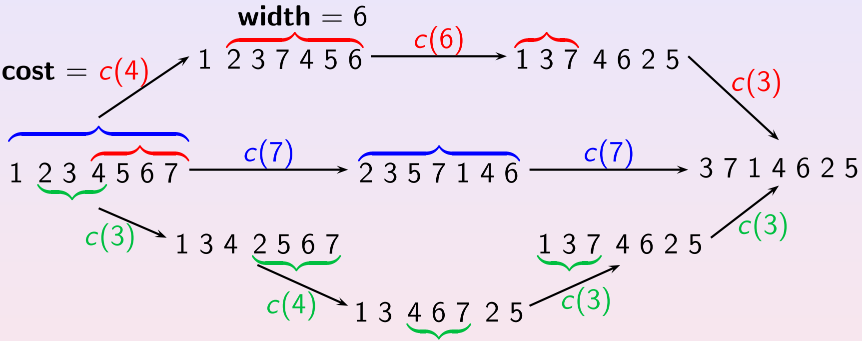
$$\rightsquigarrow 1\ 2\ \cancel{3}\ 4\ 5\ \cancel{6}\ 3\ \cancel{4}\ \cancel{5}\ 6\ 7 \rightsquigarrow 1\ 2\ 4\ 5\ 3\ 6\ 7$$

Beware ! Duplication-loss steps are not reversible !

Example

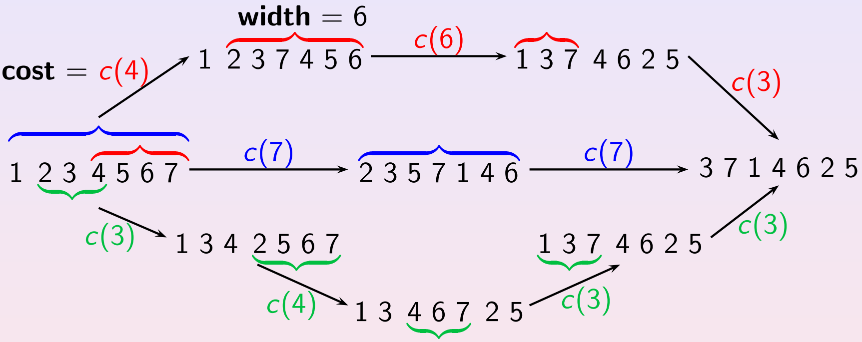
$$\overbrace{1\ 2\ 3\ 4\ 5\ 6} \rightsquigarrow 2\ 4\ 6\ 1\ 3\ 5 \not\rightsquigarrow 1\ 2\ 3\ 4\ 5\ 6$$

The tandem duplication - random loss model



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

The tandem duplication - random 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

Cost functions

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

$\hookrightarrow$ Studied by Chaudhuri, Chen, Mihaescu and Rao

- Linear or affine cost function

$\hookrightarrow$ What they suggest to study

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

$\hookrightarrow$ Where we find combinatorial properties

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

Duplication-loss from the pattern-avoidance point of view

For the whole genome duplication - random loss model:

Theorem

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

Consequence

The permutations obtainable in p steps are those having at most $2^p - 1$ descents.
$\Longrightarrow$ *a pattern-avoiding permutation class $S(B)$, with $B =$ the minimal permutations (for $\prec$) with 2^p descents.*

$\prec$ is the pattern involvement relation

The variant of the model we considered

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

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

Problems to consider:

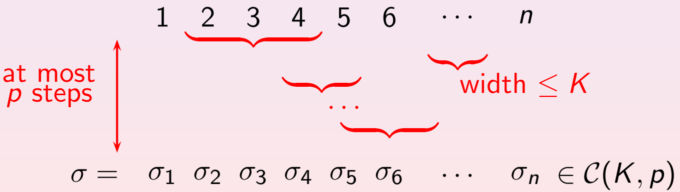
- Characterization of the permutations obtained in p steps in terms of excluded patterns ?
- Cost of obtaining a permutation ? on average ? in the worst case ?
- Finding an optimal sequence of steps from $12\dots n$ to σ, *i.e.* a sequence of minimal cost ?

Definition

Definition

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

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



First theorem

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

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.

An important property

Notice:
$\sigma \in \mathcal{C}(K,1) \Rightarrow desc(\sigma) \leq 1$
$|\sigma| \leq K, desc(\sigma) \leq 1 \Rightarrow \sigma \in \mathcal{C}(K,1)$

Proposition

For the permutations σ of size $K+1$ having exactly one descent we have: $\sigma \notin \mathcal{C}(K,1) \Leftrightarrow \sigma$ does not start with 1 nor end with $K+1$.

$\sigma \in S_{K+1}$ with 1 descent

- $\sigma = 1\sigma_2 \dots \sigma_{K+1}$ or $\sigma = \sigma_1 \dots \sigma_K K+1 \Rightarrow \sigma \in \mathcal{C}(K,1)$
- $\sigma_1 \neq 1$ and $\sigma_{K+1} \neq K+1 \Rightarrow \sigma \notin \mathcal{C}(K,1)$

An important property

Notice:
$\sigma \in \mathcal{C}(K,1) \Rightarrow desc(\sigma) \leq 1$
$|\sigma| \leq K, desc(\sigma) \leq 1 \Rightarrow \sigma \in \mathcal{C}(K,1)$

Proposition

For the permutations σ of size $K+1$ having exactly one descent we have: $\sigma \notin \mathcal{C}(K,1) \Leftrightarrow \sigma$ does not start with 1 nor end with $K+1$.

$\sigma \in S_{K+1}$ with 1 descent

- $\sigma = 1\sigma_2 \ldots \sigma_{K+1}$ or $\sigma = \sigma_1 \ldots \sigma_K K+1 \Rightarrow \sigma \in \mathcal{C}(K,1)$
- $\sigma_1 \neq 1$ and $\sigma_{K+1} \neq K+1 \Rightarrow \sigma \notin \mathcal{C}(K,1)$

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

Key Proposition to the Theorem

Proposition

If $\sigma \notin \mathcal{C}(K,p)$, then either $|\sigma| \leq (Kp+2)^2 - 2$, or there exists a strict pattern τ of σ, $\tau \notin \mathcal{C}(K,p)$.

Proposition $\Rightarrow$ Theorem: stability for $\prec$

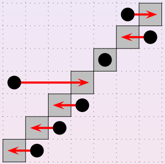
Idea of the proof of the Proposition:
Consider the minimal permutations $\sigma \notin \mathcal{C}(K,p)$, and bound the necessary moves of elements to go from $12\dots n$ to σ

vp-vectors and *vp*-domain

vp = value$\rightarrow$position

$\sigma = 4\ 1\ 2\ 3\ 5\ 7\ 6$

vp-domain of σ
$= \{1, 2, 3, 4, 6, 7\}$



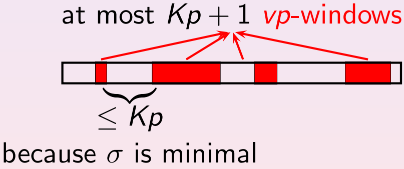
Represents the necessary moves from σ to $12\ldots n$, or when reversing the arrows from $12\ldots n$ to σ

If $\sigma \in \mathcal{C}(K,p)$, then its *vp*-domain contains at most Kp elements

What does minimal $\sigma \notin \mathcal{C}(K,p)$ look like ?

Previously: If $\sigma \in \mathcal{C}(K,p)$, then its *vp*-domain contains at most Kp elements

Consequence: If $\sigma \notin \mathcal{C}(K,p)$ is minimal, then its *vp*-domain contains at most $2Kp+2$ elements



Conclusion: $|\sigma| \leq (Kp+2)Kp + 2Kp + 2 = (Kp+2)^2 - 2$

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

Duality between "long moves" and "local reordering"

- Lower bound: $\Omega(\frac{n}{K}\log K + \frac{n^2}{K^2})$ steps in the worst case and on average
- Algorithm (upper bound): $\Theta(\frac{n}{K}\log K + \frac{n^2}{K^2})$ steps in the worst case and on average

What about $cost(\sigma)$? Our algorithm gives an K-approximation of an optimal duplication-loss scenario

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 ?

Combinatorics:

- Characterization of the minimal permutations with $d = 2^p$ descents (excluded patterns for the whole genome duplication - random loss model) ?
- Description of the excluded patterns in $\mathcal{C}(K, p)$?
- Order of the cardinality of $\mathcal{C}(K, 1)$ and $\mathcal{C}(K, p)$?

Biology:

- How can the knowledge of pattern-avoidance be of use to compute probable evolution scenarios ?