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 $C(K, 1)$ and $C(K, p)$
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

Beware ! Duplication-loss steps are not symmetric !
Motivations and the model

Biological motivations and the combinatorial model

The tandem duplication - random loss model

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

Definition

One \textit{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

Beware! Duplication-loss steps are not symmetric!

Mathilde Bouvel

A variant of the tandem duplication - random loss model of genome rearrangement
Distances and costs in the duplication-loss model

- "Oriented distance" = minimum cost of a path from $\sigma_1$ to $\sigma_2$
- Compute $\text{cost}(12 \ldots n \leftrightarrow \sigma) = \text{cost}(\sigma)$ = the minimum cost of a duplication-loss scenario from $12 \ldots n$ to $\sigma$

Mathilde Bouvel
A variant of the tandem duplication - random loss model of genome rearrangement
Distances and costs in the duplication-loss model

"Oriented distance" = minimum cost of a path from $\sigma_1$ to $\sigma_2$

Compute $\text{cost}(12\ldots n \leftrightarrow \sigma) = \text{cost}(\sigma) =$ the minimum cost of a duplication-loss scenario from $12\ldots n$ to $\sigma$
Some possible cost functions $c$

- **Power cost function:**
  $\text{width } k \Rightarrow \text{cost } \alpha^k$ for some $\alpha \geq 1$

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

- **Linear or affine cost function**

  What they suggest to study

- **Piecewise constant cost function:**

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

  Where we find combinatorial properties
Previous results on the model with power cost function

Duplication-loss on a fragment of width \( k \Rightarrow \text{cost} \, \alpha^k \)

- \( \alpha = 1 \): whole genome duplication-random loss model
  - the cost of any step is 1
  - \( \text{cost}(\sigma) \) is known, together with a corresponding scenario (radix sort algorithm)

- \( \alpha \geq 2 \): reduces to width = 2
  - \( \text{cost}(\sigma) = \alpha^2 \times \text{number of inversions in } \sigma \) (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 $\sigma$ if $\sigma_i > \sigma_{i+1}$.

$\text{desc}(\sigma) =$ number of descents of $\sigma$.

- number of maximal increasing substrings of $\sigma = \text{desc}(\sigma) + 1$

**Theorem**

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

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

Mathilde Bouvel

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

**Definition**

There is a *descent* at position $i$ in $\sigma$ if $\sigma_i > \sigma_{i+1}$.

$\text{desc}(\sigma) =$ number of descents of $\sigma$.

The number of maximal increasing substrings of $\sigma = \text{desc}(\sigma) + 1$

**Theorem**

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

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

Mathilde Bouvel

A variant of the tandem duplication - random loss model of genome rearrangements
Computing an optimal duplication-loss scenario for $\sigma \in S_n$

Radix sort algorithm:

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

Example: $\sigma = 78563412 = 78 \ 56 \ 34 \ 12$ is obtained in 2 steps through the scenario:

$12345678 \Rightarrow 10 \ 00 \ 11 \ 01 \ 12 = 78563412$
Patterns in permutations

\[ \sigma \in S_n, \tau \in S_k \text{ with } k \leq n \]

- Permutation \( \sigma \) involves pattern \( \tau \) (\( \tau \prec \sigma \)) 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 \( \tau : \sigma_{i_p} \prec \sigma_{i_q} \) if and only if \( \tau_p \prec \tau_q \)

- Otherwise, \( \sigma \) avoids \( \tau \)

- For example, \( 135624 \) 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 $\sigma$ involves pattern $\tau$ ($\tau \prec \sigma$) 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 $\tau$: $\sigma_{i_p} < \sigma_{i_q}$ if and only if $\tau_p < \tau_q$

- Otherwise, $\sigma$ avoids $\tau$

- For example, $135624$ 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 $S = \{\sigma \notin S : \forall \tau \prec \sigma, \tau \neq \sigma, \tau \in S\}$
Motivations and the model

Previous results

Combinatorial properties

Algorithmic questions

Previous results: the whole genome duplication - random loss model

Duplication-loss from the pattern-avoidance point of view

For the whole genome duplication - random loss model:

**Theorem**

\[
\text{cost}(\sigma) = \lceil \log_2(\text{desc}(\sigma) + 1) \rceil
\]

**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 \( B \)

Mathilde Bouvel

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

Previous results

Combinatorial properties

Algorithmic questions

Previous results: the whole genome duplication - random loss model

Duplication-loss from the pattern-avoidance point of view

For the whole genome duplication - random loss model:

**Theorem**

\[ \text{cost}(\sigma) = \lceil \log_2(\text{desc}(\sigma) + 1) \rceil \]

**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 \( B \)

Mathilde Bouvel

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

Previous results

Combinatorial properties

Algorithmic questions

Previous results: the whole genome duplication - random loss model

Duplication-loss from the pattern-avoidance point of view

For the whole genome duplication - random loss model:

**Theorem**

\[ \text{cost}(\sigma) = \lceil \log_2(\text{desc}(\sigma) + 1) \rceil \]

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

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

Problems to consider:

- Characterization of the permutations obtained in $\rho$ steps in terms of excluded patterns?
- Cost of obtaining a permutation $\sigma$ on average? in the worst case?
- Finding an optimal sequence of steps from $12\ldots n$ to $\sigma$, i.e. a sequence of minimal cost?
The variant of the model we considered

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

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 \( \sigma \) on average \( \text{?} \) in the worst case?
- Finding an optimal sequence of steps from \( 12 \ldots n \) to \( \sigma \), 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 $12\ldots n$ (for any $n$) after $p$ duplication-loss steps of width at most $K$.

Notice: $\mathcal{C}(K, p)$ is stable for $\prec$
**Motivations and the model**

Some combinatorial properties of the classes $C(K, 1)$ and $C(K, p)$

$C(K, 1)$ is a class of pattern-avoiding permutations

Focus on $C(K, 1)$: one duplication-loss step from $12\ldots n$

Example: $1\ 2\ 3\ 4\ 5\ 6\ 7\ 8 \xrightarrow{\text{duplication-loss}} 1\ 2\ 5\ 3\ 4\ 6\ 7\ 8$

**Theorem**

$C(K, 1) = S(B)$.

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.

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

$$B = \{\sigma \notin C(K, 1) : \forall \tau \prec \sigma, \tau \neq \sigma, \tau \in C(K, 1)\}$$
Some combinatorial properties of the classes $C(K,1)$ and $C(K,p)$

$C(K,1)$ is a class of pattern-avoiding permutations

Focus on $C(K,1)$: one duplication-loss step from $12\ldots n$

Example: $1\underbrace{2\ 3\ 4\ 5\ 6\ 7\ 8}_\sim\rightarrow 1\ 2\ 5\ 3\ 4\ 6\ 7\ 8$

Theorem

$C(K,1) = S(B)$.

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.

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

\[
B = \{\sigma \notin C(K,1) : \forall \tau \prec \sigma, \tau \neq \sigma, \tau \in C(K,1)\}
\]
Some combinatorial properties of the classes $C(K, 1)$ and $C(K, p)$

Is $C(K, p)$ also a pattern-avoiding class?

**Theorem**

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

$C(K, p)$ 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 $\sigma \notin C(K, p)$, and bound the necessary moves of elements to go from $12\ldots n$ to $\sigma$
How many steps from $12 \ldots n$ to $\sigma$ ?

- **Lower bound**: $\Omega(\log n + \frac{n^2}{K^2})$ steps in the worst case and on average
  - $\log n$ from the whole genome duplication - random loss model
  - $\frac{n^2}{K^2}$ considering the number of inversions

- **Algorithm (upper bound)**: $\Theta(\frac{n}{K} \log K + \frac{n^2}{K^2})$ steps in the worst case and on average
  - $\frac{n^2}{K^2}$ for long moves
  - $\frac{n}{K} \log K$ 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 $K = \text{constant}$)

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

Mathilde Bouvel

A variant of the tandem duplication - random loss model of genome rearrangement
How many steps from $12 \ldots n$ to $\sigma$?

- **Lower bound**: $\Omega(\log n + \frac{n^2}{K^2})$ steps in the worst case and on average
  - $\log n$ from the whole genome duplication - random loss model
  - $\frac{n^2}{K^2}$ considering the number of inversions

- **Algorithm (upper bound)**: $\Theta(\frac{n}{K} \log K + \frac{n^2}{K^2})$ steps in the worst case and on average
  - $\frac{n^2}{K^2}$ for long moves
  - $\frac{n}{K} \log K$ 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 $\sigma$ our algorithm may yield a scenario far from optimal (e.g. $n$ steps instead of $\sqrt{n}$ when $K(n) = \sqrt{n}$)

**Mathilde Bouvel**

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

Previous results

Combinatorial properties

Algorithmic questions

Other algorithmic questions to be considered

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

- **Lower bound**: $\Omega(\log n + \frac{n^2}{K^2})$ steps in the worst case and on average
  - $\log n$ from the whole genome duplication - random loss model
  - $\frac{n^2}{K^2}$ considering the number of inversions

- **Algorithm (upper bound)**: $\Theta(\frac{n}{K} \log K + \frac{n^2}{K^2})$ steps in the worst case and on average
  - $\frac{n^2}{K^2}$ for long moves
  - $\frac{n}{K} \log K$ 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 $K = \text{constant}$)

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

Previous results

Combinatorial properties

Algorithmic questions

Other algorithmic questions to be considered

Open questions

Algorithmic:

- Formula for $cost(\sigma)$?
- Optimal sequence of steps from 12...n to $\sigma$?
- 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 $C(K, p)$?
- Order of the cardinality of $C(K, 1)$ and $C(K, p)$?