Average-case complexity analysis of perfect sorting by reversals

Mathilde Bouvel[1] Cedric Chauve[2] Marni Mishna[2]
Dominique Rossin[1]

Groupe de Travail Combinatoire du Liafa, 18 mars 2010

[1] LIAFA, Université Paris Diderot and CNRS, Paris, France.

[2] Department of Mathematics, Simon Fraser University, Burnaby, Canada.

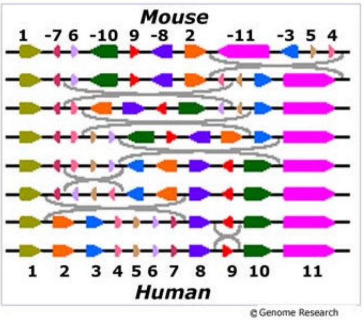
Outline of the talk

1. The context: Sorting by reversals
2. The problem we consider: Perfect sorting by reversals
3. Average-case complexity analysis
4. Restriction to the class of commuting permutations
5. Conclusion

Biological motivations

General framework: Reconstruction of evolution scenarios

$\hookrightarrow$ Only one operation on genome : reversal



© Genome Research

- Model for genome = signed permutation
- Reversal = reverse a window of the permutation while changing the signs

$$1\ \overline{7}\ 6\ \overline{10}\ 9\ \overline{8}\ 2\ \overline{11}\ \overline{3}\ 5\ 4$$

$$\Downarrow \text{ Reversal } \Downarrow$$

$$1\ \overline{7}\ 6\ \overline{10}\ 9\ \overline{8}\ 2\ \overline{4}\ \overline{5}\ 3\ 11$$

- Scenario = sequence of reversals

Sorting by reversals: the problem and solution

The problem:

- INPUT: Two signed permutations σ_1 and σ_2
- OUTPUT: A parsimonious scenario from σ_1 to σ_2 or $\overline{\sigma_2}$

Parsimonious = shortest, *i.e.* minimal number of reversals.
Without loss of generality, $\sigma_2 = Id = 1\ 2 \ldots n$

The solution:

- Hannenhalli-Pevzner theory
- Polynomial algorithms: from $\mathcal{O}(n^4)$ to $\mathcal{O}(n\sqrt{n \log n})$

Remark: the problem is *NP*-hard when permutations are unsigned.

Definition and motivation

Perfect sorting by reversals: do not break common intervals.

Common interval between σ_1 and σ_2: windows of σ_1 and σ_2 containing the same elements (with no sign)
Example: $\sigma_1 = 5\ \bar{1}\ \bar{3}\ 7\ 6\ \bar{2}\ 4$ and $\sigma_2 = 6\ \bar{4}\ 7\ 1\ \bar{3}\ 2\ \bar{5}$

When $\sigma_2 = Id$, interval of σ_1 = window forming a range (in $\mathbb{N}$)
Example: $\sigma_1 = 4\ \bar{7}\ \bar{5}\ 6\ 3\ \bar{1}\ 2$

Biological argument: groups of identical (or homologous) genes appearing together in two species are likely

- together in the common ancestor
- never separated during evolution

Algorithm and complexity

The problem:

- INPUT: Two signed permutations σ_1 and σ_2
- OUTPUT: A parsimonious perfect scenario (=shortest among perfect) from σ_1 to σ_2 or $\overline{\sigma_2}$

Without loss of generality, $\sigma_2 = Id = 1\ 2 \ldots n$

Beware: Parsimonious perfect $\not\Rightarrow$ parsimonious

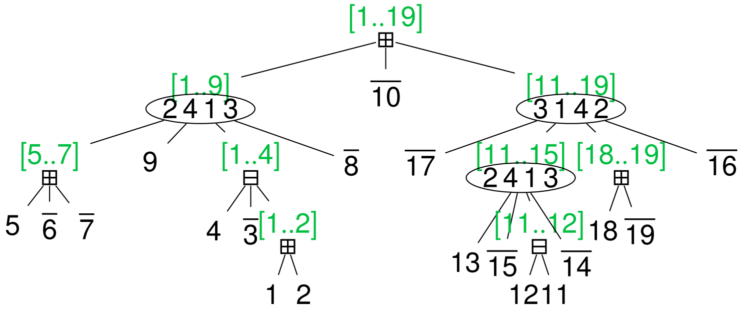
Complexity: *NP*-hard problem

Algorithm [Bérard, Bergeron, Chauve, Paul]: take advantage of decomposition trees to produce a *FPT* algorithm $\left(2^p \cdot n^{O(1)}\right)$

Decomposition trees of (signed) permutations

Also known as strong interval trees

- Strong interval = does not overlap any other interval
- Inclusion order on strong intervals: a tree-like ordering

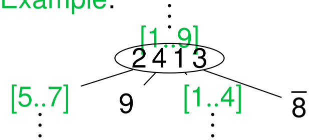


Computation: in linear time

Decomposition trees of (signed) permutations

Quotient permutation $=$
order of the children (that are intervals)

Example:

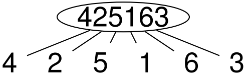


Two types of nodes:

- Linear nodes (□):
  - increasing, *i.e.* quotient permutation = 1 2 . . . k
  - ⇒ label ⊞
  - decreasing, *i.e.* quotient permutation = k $(k-1)$ . . . 2 1
  - ⇒ label ⊟
- Prime nodes (○): the quotient permutation is simple

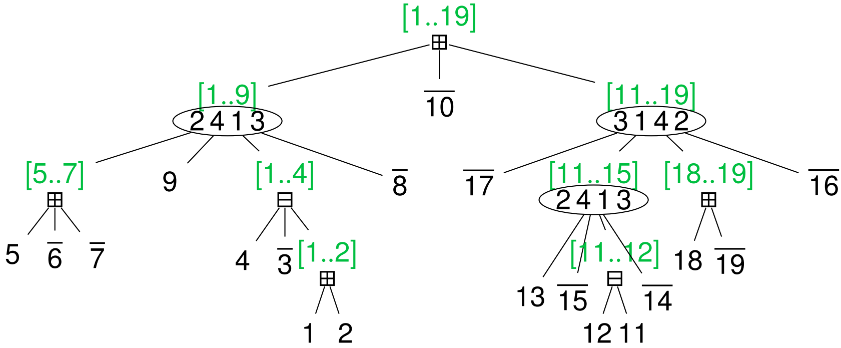
Simple permutations:
the only intervals are 1, 2,. . ., n and σ

Example: 425163, i.e.



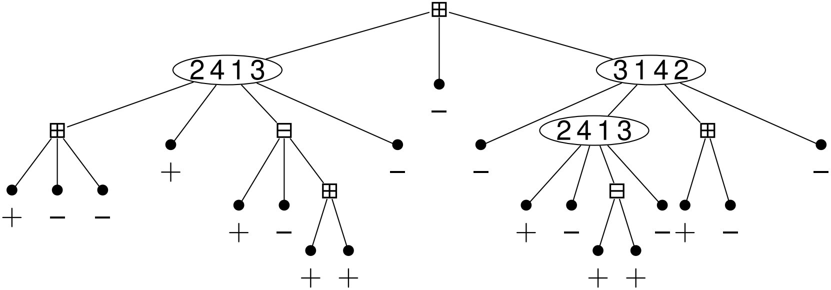
Simplified decomposition tree

Remark: redundant information ⇒ forget the leaves and intervals



Simplified decomposition tree

Remark: redundant information ⇒ forget the leaves and intervals



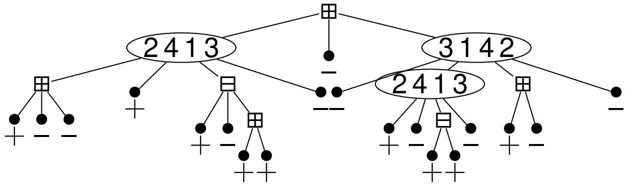
Tree uniquely defined by $\begin{cases}\text{labels of internal nodes}\\ +\text{signs of the leaves}\end{cases}$

Combinatorial point of view on decomposition trees

Theorem: Every σ ($\neq 1$) is uniquely decomposed as

- $12\dots k[\alpha^{(1)},\dots,\alpha^{(k)}]$, where the $\alpha^{(i)}$ are $\oplus$-indecomposable
- $k\dots 21[\alpha^{(1)},\dots,\alpha^{(k)}]$, where the $\alpha^{(i)}$ are $\ominus$-indecomposable
- $\pi[\alpha^{(1)},\dots,\alpha^{(k)}]$, where π is simple of size $k \geq 4$

Ex.: Decomposition tree of $\sigma =$ 5 6 7 9 4 3 1 2 8 10 17 13 15 12 11 14 18 19 16



The decomposition tree and the strong interval tree of any permutation are identical.

Idea of the algorithm

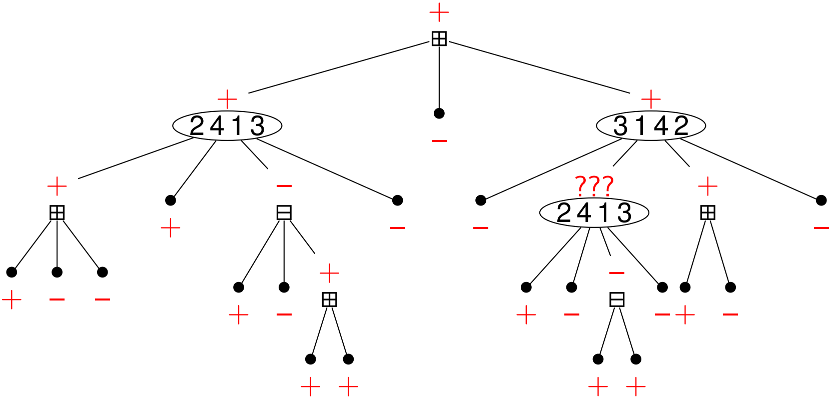
Put labels $+$ or $-$ on the nodes of the decomposition tree of σ

- Leaf: sign of the element in σ
- Linear node: $+$ for $\boxplus$ (increasing) and $-$ for $\boxminus$ (decreasing)
- Prime node whose parent is linear: sign of its parent
- Other prime node: ???
  - $\hookrightarrow$ Test labels $+$ and $-$ and choose the shortest scenario

Algorithm:

- Perform Hannenhalli-Pevzner (or improved version) on prime nodes
- Signed node belongs to scenario **iff** its sign is different from its linear parent

Example of labeled decomposition tree



Complexity results

Complexity:

- $\mathcal{O}(2^p n \sqrt{n \log n})$, with $p = \sharp$ prime nodes
- polynomial on commuting permutations ($p = 0$)

Our work:

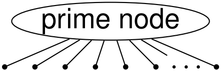
- polynomial with probability 1 asymptotically
- polynomial on average
- in a parsimonious scenario for commuting permutations
  - average number of reversals $\sim 1.2n$
  - average length of a reversal $\sim 1.02\sqrt{n}$

Probability distribution: always uniform

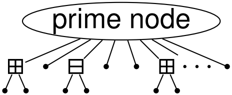
Average shape of decomposition trees

Enumeration of simple permutations: asymptotically $\frac{n!}{e^2}$

$\Rightarrow$ Asymptotically, a proportion $\frac{1}{e^2}$ of decomposition trees are reduced to one prime node.



Thm: Asymptotically, the proportion of decomposition trees made of a prime root with children that are leaves or twins is 1



twin $=$ linear node with only two children, that are leaves

Consequence: Asymptotically, with probability 1, the algorithm runs in polynomial time.

Average complexity

Average complexity on permutations of size n:

$$\frac{\displaystyle\sum_{p=0}^{n} \sharp\{\sigma \text{ with } p \text{ prime nodes}\}\ C\ 2^p n\sqrt{n\log n}}{n!}$$

Thm: When $p \geq 2$,
number of permutations of size n with p prime nodes $\leq \frac{48(n-1)!}{2^p}$

Proof: induction on p

Consequence: Average complexity on permutations of size n is $\leq 50Cn\sqrt{n\log n}$. In particular, **polynomial on average.**

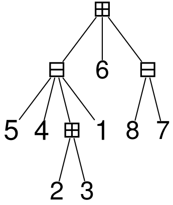
Commuting (separable) permutations

Def.: No prime node in decomposition tree

In general, in the computed perfect sorting scenario, reversals =

- linear nodes with label different from its parent
- inside prime nodes

Example: 54231687 i.e.



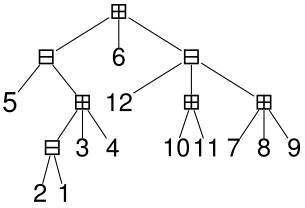
Prop.: No $\boxplus - \boxplus$ nor $\boxminus - \boxminus$ edge in decomposition trees

Consequence: For commuting permutations,

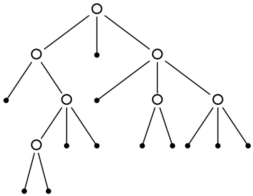
$$\text{reversals} = \begin{cases} \text{all internal nodes except the root} \\ \text{leaves with label different from its parent} \end{cases}$$

Bijection between commuting perm. and Schröder trees

Decomposition trees of unsigned commuting permutation



Schröder trees



$+$ label ⊞ on the root

size of σ	$\longleftrightarrow$	number of leaves
reversal (except leaf)	$\longleftrightarrow$	internal node (except root)
length of a reversal	$\longleftrightarrow$	number of leaves in the subtree

Average number of reversals

Average number of reversals for commuting permutations

$$= \left\{ \begin{array}{l} \text{average number of internal nodes (except root)} \\ + \text{ average number of leaves with label different from its parent} \end{array} \right.$$

$$= \quad \text{average number of internal nodes } -1 + n/2$$

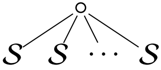
Focus on average number of internal nodes in (unsigned) Schröder trees: using bivariate generating functions.

$$S(x,y) = \sum s_{n,k} x^n y^k,$$

where $s_{n,k} =$ number of Schröder trees with n leaves and k internal nodes.

Generating function for average number of internal nodes

Definition: $S(x, y) = \sum s_{n,k} x^n y^k$,
where $s_{n,k}$ = number of Schröder trees with n leaves and k internal nodes.

$\mathcal{S} = \bullet \; + $ 

Functional equation: $S(x, y) = x + y\frac{S(x,y)^2}{1-S(x,y)}$

Solution: $S(x, y) = \frac{(x+1)-\sqrt{(x+1)^2-4x(y+1)}}{2(y+1)}$

Average number of internal nodes $= \frac{\sum_k k s_{n,k}}{\sum_k s_{n,k}} = \frac{[x^n]\frac{\partial S(x,y)}{\partial y}|_{y=1}}{[x^n]S(x,1)}$

From generating function to asymptotics

Tools: *Analytic Combinatorics* by Ph. Flajolet and R. Sedgewick

Development around singularity (here, $3 - 2\sqrt{2}$):

- $S(x,1) \sim \frac{2-\sqrt{2}}{2} - \frac{\sqrt{3\sqrt{2}-4}}{2}(1 - \frac{x}{3-2\sqrt{2}})^{1/2}$
- $\frac{\partial S(x,y)}{\partial y}|_{y=1} \sim \frac{3-2\sqrt{2}}{4\sqrt{3\sqrt{2}-4}}(1 - \frac{x}{3-2\sqrt{2}})^{-1/2}$

Equivalent of coefficients:

- $[x^n]S(x,1) \sim \frac{\sqrt{3\sqrt{2}-4}}{4}(3+2\sqrt{2})^n \frac{1}{\sqrt{\pi n^3}}$
- $[x^n]\frac{\partial S(x,y)}{\partial y}|_{y=1} \sim \frac{3-2\sqrt{2}}{4\sqrt{3\sqrt{2}-4}}(3+2\sqrt{2})^n \frac{1}{\sqrt{\pi n}}$

Conclusion: $\frac{[x^n]\frac{\partial S(x,y)}{\partial y}|_{y=1}}{[x^n]S(x,1)} \sim \frac{3-2\sqrt{2}}{3\sqrt{2}-4} n \sim \frac{n}{\sqrt{2}}$

Results (1) : Average number of reversals

Average number of internal nodes in Schröder trees:

$$\frac{n}{\sqrt{2}} \text{ asymptotically}$$

This result is valid both for unsigned Schröder trees and for Schröder trees with a sign (⊞ or ⊟) on the root.

Average number of reversals for commuting permutations:

$$\frac{n}{\sqrt{2}} - 1 + \frac{n}{2} \text{ i.e. } \frac{1+\sqrt{2}}{2} n \text{ asymptotically}$$

Remark: Many reversals of length 1: confirm biological experiments.

Average length of a reversal

$$\text{Average length of a reversal} = \frac{\text{average sum of the lengths of all reversals}}{\text{average number of reversals}}$$

$$\text{A reversal} = \begin{cases} \text{an internal node (except root)} \\ \text{a leaf with label different from its parent} \end{cases}$$

Hence, average sum of the lengths of all reversals for commuting permutations

$$= \text{average sum of the sizes of all subtrees in a Schröder tree} \\ -n \text{ (for the root) } -n/2 \text{ (for the leaves)}$$

Focus on this average sum.

Average sum of the lengths of all reversals

Use another bivariate generating function:

$$S(x,y) = \sum s_{n,k} x^n y^k,$$

where $s_{n,k}$ = number of Schröder trees with n leaves and size of subtrees summing to k,

Functional equation: $S(x,y) = xy + \frac{S(xy,y)^2}{1-S(xy,y)}$

Solution: $\frac{\sum_k k s_{n,k}}{\sum_k s_{n,k}} = \frac{[x^n]\frac{\partial S(x,y)}{\partial y}|_{y=1}}{[x^n]S(x,1)} \sim 2^{3/4}\sqrt{3-2\sqrt{2}}\,\sqrt{\pi n^3}.$

Results (2) : Average length of a reversal

Average sum of the sizes of all subtrees in a Schröder tree:

$$2^{3/4}\sqrt{3-2\sqrt{2}}\;\sqrt{\pi n^3}\text{ asymptotically}$$

Average sum of the lengths of the reversals: substracting $\frac{3}{2}n$ does not change the asymptotics.

Average length of a reversal for commuting permutations:

$$\frac{2^{7/4}\sqrt{3-2\sqrt{2}}}{1+\sqrt{2}}\sqrt{\pi n} \simeq 1.02\sqrt{n}$$

Summary of results

Perfect sorting by reversals for signed permutations:

- *NP*-hard problem
- algorithm running in polynomial time
  - $\hookrightarrow$ on average
  - $\hookrightarrow$ asymptotically with probability 1

Special case of commuting permutations:

- expected length of a parsimonious perfect scenario $\sim 1.2n$
- expected length of a reversal in such a scenario $\sim 1.02\sqrt{n}$

using analytic combinatorics techniques

Perspectives

- Use further analytic combinatorics techniques to get the distribution of the two parameters on commuting permutations

$\hookrightarrow$ With C. Nicaud

- Extend this analysis of parameters to permutations with few prime nodes
- Perform a similar study of the Double-Cut-and-Join model

$\hookrightarrow$ With C. Chauve