# Enumerating Optimal Sequences of Reversals: Tools and Experiments

Anne Bergeron[1], Cedric Chauve[1], Tzvika Hartman[2], Karine St-Onge[1]

1. LACIM, Université du Québec à Montréal
2. Department of Applied Mathematics and Computer Science, Weizmann Institute, Tel Aviv

L. Florea, B. Walenz and S. Hannenhalli eds., Currents in Computational Molecular Biology (RECOMB 2002 posters), p 20-21, 2002.

# Enumerating Optimal Sequences of Reversals: Tools and Experiments

**Anne Bergeron** [1] **Cedric Chauve** [1] **Tzvika Hartman** [2] **Karine St-Onge** [1]

**Keywords:** reversal distance, genome rearrangements, sets of optimal sequences

## 1 Introduction

The genome of a species can be thought of as a set of ordered sequences of genes – the ordering devices being the chromosomes –, each gene having an orientation given by its location on the DNA double strand. Comparing two sets of similar genes appearing along a chromosome in two species yields two (signed) permutations. It is widely accepted that the *reversal distance* between these two permutations, defined as the minimal number of reversals that transform one permutation into the other, provides a good estimate of the evolutionary distance between the two species.

Computing the reversal distance is now a well understood computational problem that has linear complexity [1]. However, reconstructing sequences of reversals that realize this distance raises some interesting issues. In recent months, the assessment of the difficulty of the problem shifted from "Finding an optimal sequence is non trivial."([3], [4], [2]) to "There is a huge number of optimal sequences" (See, for example, [5], and Paragraph 2 below). If one accepts that some of these sequences tell the evolutionary history between two species, picking up the right history is a big challenge, and requires secondary biological constraints.

In this poster, we develop some theoretical tools that help to understand and describe sets of optimal sequences. We also present experiments that show the dramatic impact of various optimal sorting strategies on the structure of the corresponding sequence.

## 2 Almost Anything Works

A naive approach to constructing optimal sequences is to perform *oriented reversals* on one permutation, – ie. reversals that create consecutive pairs in the second one – as long as there exists one such reversal. This yields an optimal sequence if the resulting permutation is the second one. Surprisingly, most strategies work pretty well. The following table presents some of our results, obtained with 10 000 signed permutations of length 128.

| Strategy | Success Rate (in %) |
|---|---|
| Maximal Score | 100 |
| Maximal Length | 79 |
| Minimal Length | 64 |
| Random | 62 |
| Minimal Score | 12 |

The 100% success rate of the maximal score strategy [2] is a consequence of its proven success in the absence of hurdles, and hurdles are almost non-existent in random permutations of length 128. Even the 12% success rate of the minimal score strategy is amazing, considering that this strategy is almost guaranteed not to succeed.

## 3 Commutation and Cartier-Foata Normal Forms

Two reversals, respectively reversing the sets of elements $S_1$ and $S_2$, *commute* iff either $S_1 \cap S_2 = \emptyset$, or one is included in the other. When two reversals commute, the order in which they are applied to a permutation is

[1] LACIM, Université du Québec à Montréal, CP 8888 Succ. Centre-Ville, Montréal, Canada H3C 3P8. E-mail: {anne | chauve}@lacim.uqam.ca

[2] Department of Applied Mathematics and Computer Science, Weizmann Institute, Rehovot 76100, Israel. E-mail: tzvi@wisdom.weizmann.ac.il

irrelevant. This commutation law induces an equivalence relation on the set of optimal sequences that sort a given permutation: a sequence is *equivalent* to another if the second one can be obtained from the first one by exchanging adjacent commuting reversals. Each equivalence class can be described by the *Cartier-Foata normal form* of one of its members, which is unique for a class.

For example, consider the permutation $(-4\ \ 1\ -3\ \ 6\ -7\ -5\ \ 2)$. A typical sorting sequence for this permutation, described by the successive sets of reversed values, is: $\{2,5,6,7\},\{2,3\},\{1,2,3,4\},\{1,2,3\},\{6,7\},\{7\}$. Its Cartier-Foata normal form is:

$$\{2,5,6,7\},\{6,7\},\{7\}\ \mid\ \{1,2,3\},\{1,2,3,4\},\{2,3\},$$

in which the reversals of the original sequences are partitioned into lexicographically ordered subsets separated by vertical bars "|", all reversals in a subset commute, and a member of a subset has at least one non-commuting reversal in the subset at its immediate left.

We define the *stack height* of a sequence as the number of subsets in its normal form. A given permutation have, in general, more than one normal form associated to its set of optimal sorting sequences. For example, the following normal form, with stack height 4, also sorts the permutation $(-4\ \ 1\ -3\ \ 6\ -7\ -5\ \ 2)$:

$$\{2,5,7\}\ \mid\ \{1,2,3,4,6\}\ \mid\ \{1,3,4,5\},\{1,3,4,5,6\}\ \mid\ \{2,3\}\{1,2,3\}.$$

## 4 Sorting Strategies and Stack Height

The following table compares the mean stack heights obtained by sorting 10 000 random permutations, of lengths 32 to 224, according to three different strategies. The "Commutation" strategy finds a maximal subset of commuting oriented reversals.

| | Maximal Length | Maximal Score | Commutation |
|---:|---:|---:|---:|
| $n = 32$ | 19.3 | 10.2 | 7.4 |
| 64 | 42.9 | 19.6 | 11.6 |
| 96 | 66.7 | 28.3 | 14.8 |
| 128 | 91.2 | 36.7 | 17.6 |
| 160 | 115.5 | 45.3 | 20.0 |
| 192 | 139.9 | 52.8 | 22.3 |
| 224 | 164.6 | 60.1 | 24.2 |

Note that the maximal length strategy have a very negative impact on commutation, meaning that this strategy yields almost deterministic evolutionary histories. On the other hand, the strategy that promotes commutation drastically reduces stack heights, yielding a number of equivalent sequences that grows factorially with an increasing function of the length.

# References


[1] Bader, D., Moret, B., Yan, M. (2001). J. Comput. Biol. 8, 5: 483-491.

[2] Bergeron, A. (2001). CPM'01, LNCS # 2089: 106-117.

[3] Hannenhalli, S. Pevzner, P. (1999). JACM 46(1): 1-27.

[4] Kaplan, H., Shamir, R., Tarjan, R. (1999). SIAM J. Comput. 29(3): 880-892.

[5] Siepel, A. (2002). RECOMB'2002 (to appear).