An Easy Case of Sorting by Reversals

MS-CIS-96-25

Nicholas Tran

University of Pennsylvania
School of Engineering and Applied Science
Computer and Information Science Department
Philadelphia, PA 19104-6389

1996

An Easy Case of Sorting by Reversals

Nicholas Tran*

Abstract

We show that sorting by reversals can be performed in polynomial time when the number of breakpoints is twice the distance.

1 Introduction

A *permutation* $\pi = (\pi_1 \pi_2 \dots \pi_n)$ is a 1-1 function $\pi : [0, n+1] \mapsto [0, n+1]$, where $\pi(0) = 0, \pi(n+1) = n+1$, and $\pi(i) = \pi_i$ for $1 \le i \le n$. A *reversal of interval* $[i,j]$ is the permutation

$$\rho_{ij} = (1\ 2\ \dots\ i\ j\ j-1\ \dots\ i+2\ i+1\ j+1\ j+2\ \dots\ n).$$

Given permutations π and σ, the *reversal distance between* π *and* σ is the length of a shortest sequence of reversals $\rho_1, \rho_2, \dots, \rho_k$ such that $\pi \cdot \rho_1 \cdot \rho_2 \cdots \rho_k = \sigma$. (Note that this definition is robust since the reversals generate the permutation group S_n.) It is easy to see that this distance is at most $n-1$ [WEHM82]. *Sorting by reversals* is the problem of finding the reversal distance $d(\pi)$ between a permutation π and the identity permutation $\imath$.

Fix a permutation $\pi \in S_n$. For $0 \le i \le n$, we call (π_i, π_{i+1}) an *adjacency* of π if $\pi_i \sim \pi_{i+1}$ ($i \sim j$ means $|i-j| = 1$); otherwise, (π_i, π_{i+1}) is called a *breakpoint* of π. Let $bp(\pi)$ denote the number of breakpoints of π; note that $bp(\pi) \le n+1$, and $bp(\imath) = 0$. Two breakpoints of π (π_i, π_{i+1}) and (π_j, π_{j+1}) define an *active* interval $[i,j]$ if $\pi_i \sim \pi_j$ and $\pi_{i+1} \sim \pi_{j+1}$; similarly they define a *passive* interval $]i,j[$ if $\pi_i \sim \pi_{j+1}$ and $\pi_{i+1} \sim \pi_j$.

Let B_π be the graph whose vertices are breakpoints of π, and whose edges connect those breakpoints that form active or passive intervals. If B_π has a perfect matching M, let I_M be the graph whose vertices are the intervals defined by the edges of M, and whose edges connect intersecting intervals. Two intervals $[i,j]$ and $[k,l]$ *intersects* each other if $i < k < j < l$ or $k < i < l < j$.

Currently it is not known whether sorting by reversals can be solved in polynomial time. In fact, the complexity of a weaker question is not known: "Is $d(\pi) \le bp(\pi)/2$?"

*Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104, nick@central.cis.upenn.edu

[KS95, PW95, VP93]. In this paper, we show that the latter problem can be solved in polynomial time.

2 Main Result

We begin with an observation about permutations π that satisfy the relation $d(\pi) = bp(\pi)/2$.

Lemma 1 *Let $\pi \in S_n$ satisfy $bp(\pi) = 2d(\pi)$, and suppose $\pi \cdot \rho_1 \cdot \rho_2 \cdots \rho_{d(\pi)} = \imath$. Each reversal ρ_i can be identified with a unique interval of π.*

Proof: Since a reversal removes at most two breakpoints, it follows that each ρ_i removes exactly two breakpoints from $\pi \cdot \rho_1 \cdots \rho_{i-1}$. Thus ρ_1 reverses an active interval of π; identify ρ_1 with this interval. Furthermore, since ρ_1 does not create new intervals and can only change a remaining active interval to a passive interval and vice-versa, each interval of $\pi \cdot \rho_1$ is an interval of π. We also have $2d(\pi \cdot \rho_1) = bp(\pi \cdot \rho_1)$ and hence by the induction hypothesis, each $\rho_2, \cdots, \rho_{d(\pi)}$ is identified uniquely with an interval of $\pi \cdot \rho_1$, which is different from the one identified with ρ_1. ∎

From the lemma above, we can represent each solution $\rho_1, \ldots, \rho_{d(\pi)}$ by a sequence of intervals corresponding to the $d(\pi)$ pairs of breakpoints of π.

Lemma 2 *Let $\pi \in S_n$ and suppose B_π has a perfect matching M that has no edges of the type $]i, i+2[$. Then for every interval $[i,j]$ and $]k,l[$ of π, $[i,j]$ and $[i+1, j+1]$ cannot be both edges of M, and $]k,l[$ and $]k+1, l-1[$ cannot be both edges of M. Thus, if $(\pi_i = x, \pi_{i+1})$ and $(\pi_j = x+1, \pi_{j+1})$ are breakpoints of π, then M contains exactly one of $[i,j],]i, j-1[, [i-1, j-1],]i-1, j[$.*

Proof: Suppose to the contrary that M contains such forbidden pairs of intervals. Associate with each forbidden active pair the value $v_{i,j} = max(\pi_{i+1}, \pi_{j+1})$ and each forbidden passive pair the value $v_{k,l} = max(\pi_k, \pi_{l+1})$. Let $[a,b]$ or $]a,b[$ be such that $v_{a,b}$ is maximum. Without loss of generality, say $v_{a,b} = \pi_a$. Since $\pi_a \leq n$, consider $\pi_a + 1 = \pi_c$ for some c. If (π_{c-1}, π_c) and (π_c, π_{c+1}) are two breakpoints of π, then since M is a perfect matching, it must contain another forbidden pair $[c-1, d]$ and $[c, d+1]$, or $]c-1, d[$ and $]c, d-1[$ for some d, whose value is $\pi_a + 2$, contradicting our choice of $v_{a,b}$.

Else exactly one of (π_{c-1}, π_c) and (π_c, π_{c+1}) is a breakpoint of π. Without loss of generality, say (π_{c-1}, π_c). By assumption $]c, c+1[$ cannot be an edge in M, and since M is a matching, it does not contain an edge of the form $[a-1, c-1]$ or $[c-1, a-1]$ or $]a, c-1[$ or $]c-1, a[$. Hence M has no intervals with $c-1$ as an endpoint, contradicting the assumption that M is a perfect matching. ∎

We now characterize those permutations π that satisfy $2d(\pi) = bp(\pi)$.

Theorem 1 *Let $\pi \in S_n$. Then $2d(\pi) = bp(\pi)$ iff there exists a perfect matching M of B_π such that each connected component of the graph I_M includes one active interval of π.*

Proof: Let $\rho_1, \ldots, \rho_{d(\pi)}$ be a shortest sequence of reversals reducing π to $\imath$. Then by the lemma above, each reversal can be identified with a unique interval of π. Representing each reversal as an edge of B_π we obtain a subgraph M of $d(\pi)$ edges. Furthermore, no two edges share a vertex since a breakpoint cannot be removed twice. Hence the subgraph M is a perfect matching of B_π. Finally, note that a reversal can affect only reversals in its connected component of I_M. Hence, the first reversal of each connected component reverses an active interval of π.

Conversely, suppose B_π has a perfect matching M such that each connected component of the graph I_M includes one active interval of π. In particular, M has no intervals of the type $]i, i+2[$, i.e. the condition of Lemma 2 is met. We show by induction on $bp(\pi)$ (which must be even since B_π has a perfect matching) that $2d(\pi) = bp(\pi)$.

When $bp(\pi) = 2$, we have $d(\pi) = 1$. Suppose the claim is true for $n \geq 2$, and let π be a permutation such that $bp(\pi) = n + 2$ and π satisfies the condition of this theorem. Let M be a matching of B_π. Select an active interval $[i, j]$ among the edges of M such that the permutation $\sigma = \pi \cdot \rho_{ij}$ obtained by reversing the interval $[i, j]$ of π has the most active intervals. If we can show that σ also satisfies the condition of this theorem then by the induction hypothesis $2d(\sigma) = bp(\sigma)$ and hence $2d(\pi) \leq 2(d(\sigma) + 1) = bp(\sigma) + 2 = bp(\pi)$.

First it is clear that the matching M minus the edge $[i, j]$ is a perfect matching of B_σ, since the reversal $[i, j]$ does not destroy other reversals which do not share one of its breakpoints. Call this matching N. It remains to show each connected component of the graph I_N has an active interval. Each such connected component is either a connected component of I_M (and thus has an active interval unaffected by the reversal of $[i, j]$) or a fragment of the connected component C_{ij} of I_M that includes $[i, j]$. A connected component of the second type must have an interval $[k, l]$ or $]k, l[$ intersecting $[i, j]$. If this interval is passive in I_M, it becomes active in I_N and we are done. Similarly, if in I_M this interval intersects with an active interval which does not intersect $[i, j]$, or if in I_M it does not intersect with a passive interval which intersects $[i, j]$, then in I_N the interval intersects with an active interval, and we are done.

Thus, suppose in I_M i) the interval $[k, l]$ is active and intersects $[i, j]$, ii) each active interval intersecting $[k, l]$ also intersects $[i, j]$, and iii) each passive interval intersecting $[i, j]$ also intersects $[k, l]$. From these conditions and the choice of $[i, j]$, it follows that any interval (active or passive) intersecting $[i, j]$ also intersects $[k, l]$ and vice-versa. Without loss of generality, assume $i < k < j < l$. Let $v = \pi_r$ be the largest integer among $\pi_{i+1}, \pi_{i+2}, \ldots, \pi_k$, and $\pi_{j+1}, \pi_{j+2}, \ldots, \pi_l$. Clearly $v \leq n$, and so $v + 1 = \pi_s$ for some s. By Lemma 2, M includes exactly one of $[r, s],]r, s-1[, [r-1, s-1],]r-1, s[$. This interval cannot intersect both $[i, j]$ and $[k, l]$, contradicting the assumption at the beginning of this paragraph.

Thus we conclude every connected component of I_N has an active interval, and the theorem follows. ∎

Theorem 2 *Deciding whether $bp(\pi) = 2d(\pi)$ for any permutation $\pi \in S_n$ is in P.*

Proof: Given π, we construct the graph B_π and assign to each active interval the weight $+1$ and each passive interval the weight -1. Then find a perfect matching M that has maximum weight. If M satisfies the condition of Theorem 1 then $2d(\pi) = bp(\pi)$. Suppose I_M has a connected component C consisting of only passive intervals. Let $]i, x[$ and $]y, j[$ be the leftmost and rightmost intervals of C, respectively. It is clear that every breakpoint between i and j is an endpoint of some interval in C; otherwise, let (π_z, π_{z+1}) be such a breakpoint such that $max(\pi_z, \pi_{z+1})$ is maximum. Then C must contain an interval $]z, z'[$ and $z' > j$ or $]z', z[$ and $z' < i$, contradicting our choice of i and j.

So $\pi_{i+1}, \pi_{i+2}, \ldots, \pi_j$ form a set of consecutive integers R. If B_π has another perfect matching M' that satisfies the condition of Theorem 2, then it must have a connected component C' whose intervals are pairs of breakpoints between i and j. Furthermore, C' has at least one active interval.

Hence we can construct from M and M' a new matching N' by replacing the connected component C of M with C' of M'. But the weight of N is greater than that of M, a contradiction. Hence $2d(\pi) \neq bp(\pi)$. ∎

References

[KS95] J. Kececioglu and D. Sankoff, *Exact and approximation algorithms for sorting by reversals, with application to genome rearrangement*, Algorithmica, 13 (1995), pp. 180–210.

[PW95] P. Pevzner and M. Waterman, *Open combinatorial problems in computational molecular biology*, in Proceedings of the 3rd Israel Symposium on the Theory of Computing and Systems, 1995, pp. 148–173.

[VP93] V. Vafna and P. Pevzner, *Genome rearrangements and sorting by reversals*, in Proc. 34th FOCS, IEEE, 1993, pp. 148–157.

[WEHM82] G. A. Watterson, W. J. Ewens, T. E. Hall, and A. Morgan, *The chromosome inversion problem*, Journal of Theoretical Biology, 99 (1982), pp. 1–7.