Journal of Mathematical Chemistry, Vol. 42, No. 4, November 2007 (© 2006)
DOI: 10.1007/s10910-006-9145-2

An algorithm for reversal median problem

Jianxiu Hao
Department of Mathematics, Zhejiang Normal University, Jinhua, Zhejiang 321004, P.R. China
E-mail: sx35@zjnu.cn

Received 13 April 2006; revised 10 May 2006

In this paper, we present an algorithm for reversal median problem whose performance ratio is less than 2.

KEY WORDS: Genome rearrangement, reversal, gene

1. Introduction

In order to derive evolutional and fundamental relationships between genes, sequence comparison is a useful tool. However, classical alignment algorithms deal with only local mutations (i.e., insertions, deletions, and substitutions of nucleotides) and ignore the global rearrangements (e.g., reversals, transpositions, and translocations of long fragments). In [1], Palmer and Herbon found that the rearrangements of mitochondrial genomes of Brassica (cabbage) and Brassica campestris (turnip) are with 99–99.9% identical genes. They discovered that these molecules which are almost identical in gene sequence, differ in gene order. The classical methods of sequence comparison are not very useful to analyze highly rearranged genomes [2,3]. Genome rearrangements is a common mode of molecular evolution in mitochondrial, chloroplast, viral, bacterial DNA, and human red–green color blindness [4–9].

Genome rearrangement by reversals, transpositions, and translocations has been studied widely [4–6,8,10]. Since 1990s finding algorithms to calculate the distance between genome pairs is a serious problem. Because the problem is too difficult, many works focus on studying some simplified models where all genomes contain the same set of genes and all genes appearing within a genome are pairwise different (i.e., there are no gene duplications), the most studied distance is reversal (or inversion) distance which is to find the minimum number of reversals transforming one genome into another [11]. In [8], authors found a polynomial time algorithm to compute the reversal distance when the orientation of the genes within the genomes is known. In this paper, we assume the orientation of the genes within the genomes is known.

When the orientation of the genes is known, a genome without gene duplications can be represented by a signed permutation π on $N := \{1, 2, \dots, n\}$, obtained by signing a permutation $\tau = (\tau_1, \tau_2, \dots, \tau_n)$ on N, i.e., replacing each element τ_i either by $\pi_i = +\tau_i$ or by $\pi_i = -\tau_i$. In particular, signs model the relative orientation of the genes within the genome. We denote by Σ_n the set of the $2^n n!$ signed permutations on N. A reversal of interval (i, j), $1 \leqslant i \leqslant j \leqslant n$, applied to a signed permutation π, is an operation which both inverts the subsequence $\pi_i \pi_{i+1}, \dots, \pi_{j-1} \pi_j$ and switches the signs of the elements in the subsequence, replacing $\pi_1, \dots, \pi_{i-1}\ \pi_i \pi_{i+1}, \dots, \pi_{j-1} \pi_j \pi_{j+1}, \dots, \pi_n$ by $\pi_1, \dots, \pi_{i-1} - \pi_j - \pi_{j-1}, \dots, -\pi_{i+1} - \pi_i \pi_{j+1}, \dots, \pi_n$. The minimum number of reversals needed to transform a signed permutation π^1 into a signed permutation π^2 (or viceversa) is called the reversal distance between π^1 and π^2, denoted by $d(\pi^1, \pi^2)$ [11].

In this paper, we study Reversal Median Problem (RMP for short) which is defined as follows: given q permutations $\pi^1, \pi^2, \dots, \pi^q \in \Sigma_n$, $q \geqslant 3$, representing genomes with the same set of genes, RMP calls for a permutation $\sigma \in \Sigma_n$ such that

$$\delta(\sigma) := \sum_{k=1}^{q} d(\sigma, \pi^k)$$

is minimized. Let δ^* denote the optimal solution value of RMP [11].

The counterpart of RMP is Breakpoint Median Problem (BMP) where the breakpoint distance is used instead of the reversal distance. More precisely, BMP is the problem to find a genome which is closest to a given set of genomes such that the sum of the breakpoint distance between the finding genome and each given genome is minimized [11]. All the methods to reconstruct evolutionary trees solve BMP as a subroutine to find the best genome associated with a given tree vertex once the genomes associated with the neighbors of the vertex are fixed [12,13].

In fact, all papers dealing with BMP pointed out that RMP is a more realistic model than BMP. “For RMP, there are no algorithms available, aside from rough heuristics, for handling even three relatively short genomes ” [12,14]. “Even heuristic approaches for RMP work well only for small instances” [15].

The organization of the paper is as follows. In section 2, we present an algorithm with its performance ratio, section 3 contains some concluding remarks.

2. Main results

Definition. We call permutations π^1, π^2, π^3 share a common line if and only if there exists π^{j_1} appears in one of the shortest sequences of permutations transforming π^{j_2} into π^{j_3}, where $j_1, j_2, j_3 \in \{1, 2, 3\}$, and $\pi^{j_1}, \pi^{j_2}, \pi^{j_3}$ are different permutations.

Furthermore, we call π^{j_1} the inner permutation of π^{j_2} and π^{j_3}. For convenience, when $\pi^{j_1} = \pi^{j_2}$ or $\pi^{j_1} = \pi^{j_3}$, we call π^{j_1} the inner permutation of π^{j_2} and π^{j_3} too.

For example, let $\pi^1 = (2, -3, 1), \pi^2 = (2, -1, 3), \pi^3 = (1, -2, 3)$, the shortest sequence of reversals transforming π^1 into π^3 is shown as follows:

$$\pi^1 = 2\underline{-31},$$

$$\pi^2 = \underline{2-1}3,$$

$$\pi^3 = 1-23.$$

From above reversal sequence we know that π^2 is the inner permutation of π^1 and π^3.

Algorithm

Step 0:

$$(j_1, j_2, \ldots, j_q) = (1, 2, \ldots, q),$$

$$\delta_{Alg} = \infty,$$

$A = \{(j_1, j_2, \ldots, j_q) | (j_1, j_2, \ldots, j_q)$ is a permutation on $\{1, 2, \ldots, q\}\}$.

Step 1: If

$$\delta_{Alg} > \sum_{i=1}^{q-1} d(\pi^{j_i}, \pi^{j_{i+1}})$$

then

$$\delta_{Alg} \longleftarrow \sum_{i=1}^{q-1} d(\pi^{j_i}, \pi^{j_{i+1}})$$

else

$$\delta_{Alg} \longleftarrow \delta_{Alg}.$$

$$A \longleftarrow A - \{(j_1, j_2, \ldots, j_q)\}.$$

Draw $(j_1^{'}, j_2^{'}, \dots, j_q^{'}) \in A$,

$$(j_1, j_2, \dots, j_q) \longleftarrow (j_1^{'}, j_2^{'}, \dots, j_q^{'}).$$

Step 2: If $A = \emptyset$, stop. Else, go to step 1.

Theorem. Given an instance $\pi^1, \pi^2, \dots, \pi^q \in \Sigma_n$, let δ_{Alg} be the value provided by algorithm above and δ^* be the optimal solution value of RMP. We have

$$\frac{\delta_{Alg}}{\delta^*} < 2.$$

Proof.

Claim **1**

$$\tfrac{1}{2} \max_{(j_1, j_2, \dots, j_q)} \sum_{i=1}^{q} d(\pi^{j_i}, \pi^{j_{i+1}}) \leqslant$$

$$\delta^* \leqslant \min_{(j_1, j_2, \dots, j_q)} \sum_{i=1}^{q-1} d(\pi^{j_i}, \pi^{j_{i+1}}),$$

where $(j_1, j_2, \dots, j_q)$ takes over all permutations on $\{1, 2, \dots, q\}$.

If we proved claim 1, our theorem might follow obviously. Hence, we prove claim 1 first.

We construct a weighted complete graph K_q as follows:

$$V(K_q) = \{\pi^1, \pi^2, \dots, \pi^q\},$$

$\pi^i \pi^j \in E(K_q)$ if and only if $i \neq j$, where $1 \leqslant i, j \leqslant q$. Let

$$w(\pi^i \pi^j) = d(\pi^i, \pi^j),$$

where $w(\pi^i \pi^j)$ denotes the weight of edge $\pi^i \pi^j$. We call $\sum_{i=1}^{q-1} d(\pi^{j_i}, \pi^{j_{i+1}})$ the weight of path $\pi^{j_1} \pi^{j_2}, \dots, \pi^{j_q}$.

In the following we use mathematical induction to prove

$$\delta^* \leqslant \min_{(j_1, j_2, \dots, j_q)} \sum_{i=1}^{q-1} d(\pi^{j_i}, \pi^{j_{i+1}}),$$

where $(j_1, j_2, \dots, j_q)$ takes over all permutations on $\{1, 2, \dots, q\}$.

(1) Suppose $q = 3$, let σ be an optimal solution of δ^*.

Case **1.** Suppose $\sigma = \pi^1$. Hence, $\delta^* = \delta(\pi^1)$. Obviously, δ^* equals the weight of path $\pi^2\pi^1\pi^3$.

Similarly, we can discuss the cases when $\delta = \pi^2$ and $\delta = \pi^3$. Hence, in case 1 we have

$$\delta^* \leqslant \min_{(j_1,j_2,j_3)} \sum_{i=1}^{2} d(\pi^{j_i}, \pi^{j_{i+1}}),$$

where (j_1, j_2, j_3) takes over all permutations on $\{1, 2, 3\}$.

Case **2.** Suppose $\sigma \in \Sigma_n$, note that it is possible that $\sigma \neq \pi^1$, $\sigma \neq \pi^2$, and $\sigma \neq \pi^3$. By the definition of δ^* we have

$$\delta^* \leqslant \min\{\delta(\pi^1), \delta(\pi^2), \delta(\pi^3)\}.$$

By the conclusion of case 1, we have

$$\delta^* \leqslant \min_{(j_1,j_2,j_3)} \sum_{i=1}^{2} d(\pi^{j_i}, \pi^{j_{i+1}}),$$

where (j_1, j_2, j_3) takes over all permutations on $\{1, 2, 3\}$.

(2) Suppose

$$\delta^* \leqslant \min_{(j_1,j_2,\ldots,j_q)} \sum_{i=1}^{q-1} d(\pi^{j_i}, \pi^{j_{i+1}})$$

holds for $q = k$. In the following we assume $q = k + 1$, and let σ be an optimal solution for δ^*.

Case **3.** Suppose there exists π^i such that $\sigma = \pi^i$. Hence,

$$\begin{aligned}\delta^* &= \delta(\pi^i) \\ &= \sum_{m=1}^{q} d(\pi^i, \pi^m) \\ &= \left[-d(\pi^i, \pi^t) + \sum_{m=1}^{q} d(\pi^i, \pi^m)\right] + d(\pi^i, \pi^t),\end{aligned}$$

where π^t is any vertex belonging to $\{\pi^1, \pi^2, \ldots, \pi^q\}$ and $\pi^t \neq \pi^i$.

By hypothesis in (2) we know that

$$-d(\pi^i, \pi^t) + \sum_{m=1}^{q} d(\pi^i, \pi^m)$$

is no more than the weight of any path P_1 which starts from π^i with k vertices in $\{\pi^1, \pi^2, \dots, \pi^q\} - \{\pi^t\}$.

Let $P = \pi^i\pi^t \cup P_1$. Clearly, the weight of P is

$$\sum_{i=1}^{q-1} d(\pi^{j_i}, \pi^{j_{i+1}}).$$

Hence, $\delta^* = \delta(\pi^i)$ is no more than the weight of P. Further, we have

$$\delta^* \leqslant \min_{(j_1, j_2, \dots, j_q)} \sum_{i=1}^{q-1} d(\pi^{j_i}, \pi^{j_{i+1}}),$$

where $(j_1, j_2, \dots, j_q)$ takes over all permutations on $\{1, 2, \dots, q\}$.

Case **4.** Suppose $\sigma \in \Sigma_n$. By the definition of δ^* we have

$$\delta^* \leqslant \min\{\delta(\pi^1), \delta(\pi^2), \dots, \delta(\pi^q)\}.$$

By the conclusion of case 3, case 4 follows.

In the following, we want to prove

$$\delta^* \geqslant \frac{1}{2} \max_{(j_1, j_2, \dots, j_q)} \sum_{i=1}^{q} d(\pi^{j_i}, \pi^{j_{i+1}}).$$

At first, we prove claim 2 as a stepping-stone.

Claim **2.** Let $\sigma, \pi^1, \pi^2 \in \Sigma_n$, we have

$$d(\pi^1, \sigma) + d(\sigma, \pi^2) \geqslant d(\pi^1, \pi^2).$$

Further, $d(\pi^1, \sigma) + d(\sigma, \pi^2) = d(\pi^1, \pi^2)$ if and only if σ is the inner permutation of π^1 and π^2.

In fact, let A be the set of permutations appearing in one of the shortest sequences of permutations transforming π^1 into σ and B be the set of permutations appearing in one of the shortest sequences of permutations transforming σ into π^2. Then, $A \cup B$ contains the set of permutations transforming π^1 into π^2. By the definition of $d(\pi^1, \pi^2)$ we have

$$d(\pi^1, \sigma) + d(\sigma, \pi^2) \geqslant d(\pi^1, \pi^2).$$

When σ is the inner permutation of π^1 and π^2, by definition, we know that $A \cup B$ is the set of the shortest sequence of permutations transforming π^1 into π^2. Then,

$$d(\pi^1, \sigma) + d(\sigma, \pi^2) = d(\pi^1, \pi^2).$$

On the other hand, suppose

$$d(\pi^1, \sigma) + d(\sigma, \pi^2) = d(\pi^1, \pi^2)$$

holds. If σ is not the inner permutation of π^1 and π^2, from above proof we know that $A \cup B$ is not the set of the shortest sequence of permutations transforming π^1 into π^2 . Hence, we have

$$d(\pi^1, \sigma) + d(\sigma, \pi^2) > d(\pi^1, \pi^2),$$

which is a contradiction. Claim 2 follows.

Let $(j_1, j_2, \ldots, j_q)$ be a permutation on $\{1, 2, \ldots, q\}$ and $\sigma \in \Sigma_n$ be an optimal permutation of δ^*. By claim 2, we have

$$d(\pi^{j_i}, \sigma) + d(\sigma, \pi^{j_{i+1}}) \geqslant d(\pi^{j_i}, \pi^{j_{i+1}}),$$

where $i = 1, 2, \ldots, q$, $\pi^{j_{q+1}} = \pi^{j_1}$.

Therefore, we have

$$\sum_{i=1}^{q} [d(\pi^{j_i}, \sigma) + d(\sigma, \pi^{j_{i+1}})] \geqslant \sum_{i=1}^{q} d(\pi^{j_i}, \pi^{j_{i+1}}).$$

Thus, we have

$$\delta^* \geqslant \frac{1}{2} \sum_{i=1}^{q} d(\pi^{j_i}, \pi^{j_{i+1}}).$$

Further, we have

$$\delta^* \geqslant \frac{1}{2} \max_{(j_1, j_2, \ldots, j_q)} \sum_{i=1}^{q} d(\pi^{j_i}, \pi^{j_{i+1}}).$$

From above proof claim 1 follows. By claim 1 the theorem follows.

3. Concluding remarks

In order to solve RMP, we must find an optimal permutation $\sigma \in \Sigma_n$. By [11], we know that RMP is NP-hard even for $q = 3$. Hence, we can not find σ easily. However, if we regard π^i as σ, by claim 1 we can find an upper bound very closely and this upper bound can be found easily, because we can compute $d(\pi^{j_i}, \pi^{j_{i+1}})$ in polynomial time and $\pi^1, \pi^2, \ldots, \pi^q$ are known.

Note that the upper bound provided by Algorithm which is the same as claim 1 is attainable. For the example provided in [11, pp. 96 and 97], by claim 1 we have

$$\delta^* \leqslant \min\{d(\pi^1, \pi^2) + d(\pi^2, \pi^3),\\ d(\pi^2, \pi^1) + d(\pi^1, \pi^3), d(\pi^2, \pi^3) + d(\pi^3, \pi^1)\} = 5.$$

Hence, $\delta_{Alg} = 5$. From [11, p. 97], we have $\delta^* = 5$ which is the same as that provided by our Algorithm.

Acknowledgments

The Project Supported by Zhejiang Provincial Natural Science Foundation of China (102055), National Nature Science Foundation of China (No. 10471131) and The Foundation of Zhejiang Universities′ Youth Teachers.

References

[1] J.D. Palmer and L.A. Herbon, Plant mitochondrial DNA evolves rapidly in structure, but slowly in sequence, J. Mol. Evol. 27 (1988) 87–97.

[2] S. Karlin, E.S. Mocarski and G.A. Schachtel, Molecular evolution of herpesviruses: Genomic and protein sequence comparisons, J. Virol. 68 (1994) 1886–1902.

[3] D.J. McGeoch, Molecular evolution of large DNA viruses of eukaryotes, Sem. Virol. 3 (1992) 399–408.

[4] V. Bafna and P. Pevzner, Sorting by reversals: Genome rearrangements in plant organelles and evolutionary history of x chromosome, Mol. Biol. Evol. 12 (1995) 239–246.

[5] V. Bafna and P. Pevzner, Sorting permutations by transpositions, in: *Proceedings 6th ACM–SIAM Ann. Symp. on Discrete Algorithms 1995*, pp. 614–623.

[6] V. Bafna and P. Pevzner, Genome rearrangements and sorting by reversals, SIAM J. Comupt. 25(2) (1996) 272–289.

[7] N. Franklin, Conservation of genome form but not sequence in the transcription antitermination determinants of bacteriophages λ, ϕ21, and p22, J. Mol. Evol. 181 (1985) 75–94.

[8] S. Hannenhalli and P. Pevzner, Transforming cabbage into turnip (polynomial algorithm for sorting signed permutation by reversals), in: *Proceedings 27th ACM Symp. on Theory of Computing (STOC'95), 1995*, pp. 178–189.

[9] D.L. Hartl and E.W. Jones, *Genetics: Analysis of Genes and Genomes, 5th edn*. (Jones and Bartlett Publishers Inc., Boston, MA), pp. 364–367.

[10] Q.-P. Gu, S. Peng and H. Sudborough, A 2−approximation algorithm for genome rearrangements by reversals and transpositions, Theor. Comput. Sci. 210 (1999) 327–339.

[11] A. Caprara, Reversal median problem, INFORMS J. Comput. 15 (2003) 93–113.

[12] M. Blanchette, G. Bourque and D. Sankoff, Breakpoint phylogenies, in: eds. *Proceedings of Genome Informatics 1997* Vol. 25. eds. S. Miyano and T. Takagi (Universal Academy Press (New York, 1997) pp. 25–34.

[13] B.M.E. Moret, S.K. Wyman, D.A. Bader, T. Warnow and M. Yan, A new implementation and detailed study of breakpoint analysis. In: *Proceedings of the Sixth Pacific Symposium on Biocomputing (PSB 2001)* (World Scientific Publishing Singapore), pp. 583–594.

[14] D. Sankoff and M. Blanchette, Multiple genome rearrangement and breakpoint plylogenies, J. Comput. Biol, 5 (2000) 555–570.

[15] D. Sankoff and N. El-Mabrouk, Duplication, rearrangement and reconciliation, in: *Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics*, eds. D. Sankoff and J.H. Nadean, (Kluwer Academic Publishers, Dordrecht, 2000) pp. 537–550.