Theoretical Computer Science 698 (2017) 9–13

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Asymptotic medians of random permutations sampled from reversal random walks

Arash Jamshidpey [a], David Sankoff [b,*]

[a] Instituto Nacional de Matemática Pura e Aplicada, Estrada Dona Castorina 110, Rio de Janeiro, 22460-320, Brazil
[b] Department of Mathematics and Statistics, University of Ottawa, 585 King Edward Avenue, Ottawa, K1N 6N5, Canada

ARTICLE INFO

Article history:
Received 25 February 2017
Received in revised form 17 May 2017
Accepted 11 June 2017
Available online 8 July 2017

Keywords:
Reversal distance
Median problem
Random walk
Symmetric group
Phylogeny
Comparative genomics

ABSTRACT

Medians can serve as a good estimator of the ancestor (the initial state) for k independent reversal random walks on the space of signed permutations before time $\frac{n}{4}$, that is the identity permutation is a median of k random genomes sampled from k independent random walks at time cn where $c \leq 1/4$. In this paper we relax the time scale of the individual random walks, investigate the positions of all possible medians other than the initial state, and reduce the state space necessary for median search algorithms.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

The median plays an important role in the comparative genomics study of chromosomal rearrangements [1]. It is not only the archetypical phylogenetic instance – one unknown ancestor, $k \geq 3$ observed genomes – for the small phylogeny problem using unrooted trees, it is also the innermost calculation for the iterative "steinerization" procedure for more larger instances of this problem, with several ancestral nodes [2].

In a metric space, the median is a point whose sum of distances to k given points is minimized. In the simplest case, where a genome is a signed permutation on $\{1, \ldots, n\}$, the biologically relevant metric is the reversal distance, where a reversal involves a contiguous subset of the elements in a permutation reversed in order and sign. The distance is the minimum number of reversals necessary to transform one genome to another.

The median problem is NP-hard for reversal distance [3] and the search space is very large. Reducing the search space to a much smaller subdomain would be a significant help in practice.

In a previous paper [4] we showed that the starting point of k independent reversal random walks on the signed permutation group remains approximately a median for the k current positions up to time $n/4$.

In the present paper, we improve this result to allow the time scales to differ among the k random walks. We then show how to find the relative positions of all the other medians beside the starting point. In doing this we reduce the search space considerably for median search algorithms.

* Corresponding author.
E-mail addresses: arashj@impa.br (A. Jamshidpey), sankoff@uottawa.ca (D. Sankoff).

http://dx.doi.org/10.1016/j.tcs.2017.06.014
0304-3975/© 2017 Elsevier B.V. All rights reserved.

2. Main results

In the absence of duplication, we represent unichromosomal genomes by *permutations* or *signed permutations*. A *signed permutation* is a permutation π on $\{\pm 1, ..., \pm n\}$ such that $\pi_{-i} = -\pi_i$ (see Fertin et al. [5]). Each number represents a gene in the genome while its sign indicates its orientation or polarity (called *strandedness* in the biological literature). The set of all signed permutations of length n with the composition multiplication is a group called *signed symmetric group of order n* denoted by $\sigma_n^{\pm}$. *Genome rearrangement* is the study of large scale mutations, rearrangements, over the set of genomes or (signed) permutations. An example of a rearrangement is a reversal (called *inversion* in the biological literature). Let $\pi = \pi_{-1}\pi_1 ... \pi_{-n}\pi_n$ be a signed permutation. A *reversal* is a permutation multiplying by π from right which reverses a segment $\pi_{-i}\pi_i, ..., \pi_{-j}\pi_j$ and keeps the other positions unchanged. In other words, for any $1 \leq i \leq j \leq n$, a reversal permutation reversing segment $\pi_{-i}\pi_i, ..., \pi_{-j}\pi_j$ is

$$-1\ +1\ ...\ -(i-1)\ +(i-1)\ +j\ -j\ ...\ +i\ -i\ -(j+1)\ +(j+1)\ ...\ -n\ +n \tag{1}$$

The reversal distance between two signed permutations π and π' is the minimum number of reversal permutations needed to transform π into π'. We denote by $d^{(n)}$ the reversal distance on $\sigma_n^{\pm}$. In fact, $d^{(n)}$ is a metric on $\sigma_n^{\pm}$. Alternatively, one can define the reversal distance as the distance on the *Cayley graph* of $\sigma_n^{\pm}$ with respect to reversals. More explicitly, the set of all reversal permutations generates $\sigma_n^{\pm}$. We denote by G_n the Cayley graph of $\sigma_n^{\pm}$ with respect to the reversal permutations as the generating set. Then the reversal distance $d^{(n)}$ is the graph distance on G_n. See [5], for more generalities on genome rearrangement, distances, and Cayley graphs.

A reversal random walk on $\sigma_n^{\pm}$ is a continuous-time simple random walk on G_n starting at identity element e_n where the jumps occur at rate 1, that is starting at e_n, at each position in $\sigma_n^{\pm}$ the random walker chooses one of its $\binom{n}{2}+n$ neighbours with equal probability, and jumps to it in a Poisson time rate 1. Berestycki et al. proved that, up to time $n/2$, the speed of escape of reversal random walk on $\sigma_n^{\pm}$, endowed with approximate reversal metric (without considering *hurdles* and *fortresses*) is near to its maximum value 1, after a convenient rescaling of the metric. In other words, the rate of escape of the reversal random walk is linear [6]. Deriving an upper bound for the expected number of "hurdles", we extended this result to the reversal random walk on $\sigma_n^{\pm}$ endowed with the (exact) reversal metric [4]. We will see that the linearity of rate of escape plays an important role in the proof of the main theorems in this paper.

For a finite metric space (S, d), we define the total distance function to B by

$$d_T(., B) : S \to IR_+,$$
$$d_T(., B) := \sum_{b \in B} d(., b),$$

where the subscript T stands for "*Total*". A median of a finite subset B (with possible multiplicities) of S is a point of S (not necessarily unique) that minimizes the total distance function $d_T(., B)$. The set of all medians of B is called *median set of B*. The total distance of a median of B to B is called *median value of B*. In other words, the median value of B is the minimum value of $d_T(x, B)$ over all $x \in S$.

The median problem has played an important role in phylogeny reconstruction to approximate the true ancestor, and more generally, to reconstruct the ancestral tree. The question that arises is: When does the median approximate the true ancestor? For $(\sigma_n^{\pm}, d^{(n)})$ we denote by $d_T^{(n)}$ the total distance for $d^{(n)}$. Also, we denote by $M_n(B)$ and $m_n(B)$ the (reversal) median set and median value of $B \subset \sigma_n^{\pm}$, respectively.

Let $(\Omega, \mathcal{F}, P)$ be a Borel probability space. We denote by σ the Cartesian product $\sigma_1^{\pm} \times \sigma_2^{\pm} \times \ldots$. Let k be an integer. For a sequence of random sets of finite points on signed symmetric groups we define the concept of asymptotic almost surely median as follows.

Definition 1. For any $n \in IN$ and $i \in \{1, ..., k\}$, let x_{in} be a random element of $\sigma_n^{\pm}$, i.e. $x_{in} : \Omega \to \sigma_n^{\pm}$ is a Borel measurable function. Let $A_n = \{x_{1n}, ..., x_{kn}\} \subset \sigma_n^{\pm}$, and let $A = (A_n)_{n \in IN}$. We say $y := (y_n)_{n \in IN} \in \sigma$ is a median of A asymptotically almost surely (a.a.s.) if

1. For any $i \in \{1, ..., k\}$ there exists a function $f_i : IN \to IR_+$ such that

$$\frac{d^{(n)}(y_n, x_{in}) - f_i(n)}{\gamma_n \sqrt{n}} \to 0 \text{ in probability,} \tag{2}$$

for any sequence of real numbers $(\gamma_n)_{n \in IN}$ diverging to infinity, and $\lim_{n \to \infty} \frac{f_i(n)}{n}$ exists.

2.

$$\frac{d_T^{(n)}(y_n, A_n) - m_n(A_n)}{\gamma_n \sqrt{n}} \to 0 \text{ in probability,} \tag{3}$$

for any sequence $(\gamma_n)_{n \in IN}$ diverging to infinity.

In this case we say any function $\bar{f} : IN \to IR_+$ which satisfies in

$$\frac{\bar{f}(n) - m_n(A_n)}{\gamma_n \sqrt{n}} \to 0 \text{ in probability,} \tag{4}$$

for any sequence $(\gamma_n)_{n\in IN}$ diverging to ∞, is an asymptotic median value of A. This includes the function $\bar{f}(n) = \sum_{i=1}^{k} f_i(n)$.

In [4], Theorem 3, we proved that the sequence of identity elements $e := (e_n)_{n\in IN}$ is a median of k random signed permutations sampled from k independent reversal random walks at time cn where $c \leq 1/4$, a.a.s. By a small modification in the proof of that theorem we can extend that result as follows.

Theorem 1. *Let $0 < c_1 \leq \ldots \leq c_k$ such that $c_{k-1} + c_k \leq \frac{1}{2}$. For any natural number n, let $X^{1,n}, \ldots, X^{k,n}$ be k independent reversal random walks on $\sigma_n^{\pm}$, all starting at e_n. Let $x_{in} = X^{i,n}(c_i n)$, where $X^{i,n}(t)$ stands for the position of random walk at time t. Then $e = (e_n)_{n\in IN}$ is a median of $A := (A_n)_{n\in IN}$ a.a.s. where $A_n = \{x_{1n}, \ldots, x_{kn}\}$, and the function $\theta : IN \to IR_+$ defined by*

$$\theta(n) = (\sum_{i=1}^{k} c_i)n \tag{5}$$

is an asymptotic median value of A.

Proof. For all $i, j \in \{1, \ldots, k\}$ and for a median solution $y = (y_n)_{n\in IN}$ of A

$$d^{(n)}(x_{in}, x_{jn}) \leq d^{(n)}(y_n, x_{in}) + d^{(n)}(y_n, x_{jn}) \tag{6}$$

Therefore,

$$\sum_{i\neq j} d^{(n)}(x_{in}, x_{jn}) \leq \sum_{i\neq j} (d^{(n)}(y_n, x_{in}) + d^{(n)}(y_n, x_{jn})) \tag{7}$$

Thus

$$\sum_{i\neq j} d^{(n)}(x_{in}, x_{jn}) \leq (k-1) m_n(A_n) \leq (k-1) d_T^{(n)}(e_n, A) \tag{8}$$

By Theorem 2 in [4], for any sequence $(\gamma_n)_{n\in IN}$ diverging to infinity,

$$\frac{d^{(n)}(e_n, x_{in}) - c_i n}{\gamma_n \sqrt{n}} \to 0 \tag{9}$$

in probability, and

$$\frac{d^{(n)}(x_{in}, x_{jn}) - (c_i + c_j)n}{\gamma_n \sqrt{n}} \to 0 \tag{10}$$

in probability, for any $i \neq j \in \{1, \ldots, k\}$. Therefore

$$\frac{m_n(A_n) - \theta(n)}{\gamma_n \sqrt{n}} \to 0 \tag{11}$$

in probability, and hence e is a median of A a.a.s. since

$$\frac{d_T^{(n)}(e_n, A) - \theta(n)}{\gamma_n \sqrt{n}} \to 0 \text{ in probability.} \tag{12}$$

This proves the theorem. □

By definition of a.a.s. median, it is clear that if there exists $y = (y_n)_{n\in IN}$ such that for any $i \in \{1, \ldots, k\}$

$$\frac{d^{(n)}(y_n, x_{in}) - c_i n}{\gamma_n \sqrt{n}} \to 0 \text{ in probability,} \tag{13}$$

for any sequence $(\gamma_n)_{n\in IN}$ diverging to infinity, then y is a median of A a.a.s. Is the converse true? The next result shows that the converse is true under the hypotheses of Theorem 1.

Theorem 2. *Let $k \geq 3$ and c_i, x_{in}, and A_n be as defined in the statement of Theorem 1 for $i \in \{1, \ldots, k\}$ and natural number n. Then $y := (y_n)_{n\in IN}$ is a median of A a.a.s. if and only if (13) holds for any sequence of real numbers $(\gamma_n)_{n\in IN}$ diverging to ∞.*

Proof. By the explanation given before the statement of the theorem, it suffices to prove the necessary part. Let $y = (y_n)_{n\in IN}$ be a median of A a.a.s. Then, by definition, there exist functions $f_i : IN \to IR_+$ for $i = 1, \ldots, n$ such that

$$\frac{d^{(n)}(y_n, x_{in}) - f_i(n)}{\gamma_n\sqrt{n}} \to 0 \text{ in probability}, \tag{14}$$

for any sequence $(\gamma_n)_{n\in IN}$ diverging to ∞, and

$$\frac{f_i(n)}{n} \to c_i' \tag{15}$$

for real numbers $c_i' \geq 0$. Letting $\gamma_n := \sqrt{n}$ and applying Theorem 2 in [4], for $0 \leq i, j \leq k$, we have

$$\frac{f_i(n) + f_j(n) - d^{(n)}(y_n, x_{in}) - d^{(n)}(y_n, x_{jn})}{n} \to 0 \text{ in probability} \tag{16}$$

and

$$\frac{d^{(n)}(x_{in}, x_{jn}) - (c_i + c_j)n}{n} \to 0 \text{ in probability} \tag{17}$$

Therefore

$$\delta_n \to (c_i' + c_j') - (c_i + c_j) \text{ in probability} \tag{18}$$

where

$$\delta_n = \frac{d^{(n)}(x_{in}, y_n) + d^{(n)}(y_n, x_{jn}) - d^{(n)}(x_{in}, x_{jn})}{n} \tag{19}$$

which is positive for any $n \in IN$. Hence

$$c_i' + c_j' \geq c_i + c_j \tag{20}$$

Now, suppose there exists $i \in \{1, \ldots, k\}$ such that $c_i' < c_i$. Then for any $j \neq i$ $c_j < c_j'$, and hence $\sum_{j=1}^k c_j < \sum_{j=1}^k c_j'$ since $k \geq 3$. This contradicts with the fact that y is a median of A a.a.s. as we know $e = (e_n)_{n\in IN}$ is a median of A a.a.s., and therefore the function θ defined by $\theta(n) = \sum_{i=1}^k c_i n$ is an asymptotic median value of A by Theorem 1. Hence $c_i \leq c_i'$. Also, if there exists $1 \leq i \leq k$ such that $c_i < c_i'$, then y can not be a median of A a.a.s. since there exists a sequence of real numbers $(\gamma_n)_{n\in IN}$ diverging to ∞ such that

$$\frac{\sum_{i=1}^k c_i' n - \theta(n)}{\gamma_n\sqrt{n}} \to \infty \tag{21}$$

which is a contradiction again. Thus $c_i = c_i'$ for any $1 \leq i \leq k$. □

To simplify the results, we define an equivalence relationship on σ. We say $x = (x_n)_{n\in IN}$ and $y = (y_n)_{n\in IN}$ in σ are equivalent and we denote it by $x \cong y$ if and only if

$$\frac{d^{(n)}(x_n, y_n)}{\gamma_n\sqrt{n}} \to 0, \tag{22}$$

for any sequence of real numbers $(\gamma_n)_{n\in IN}$ diverging to ∞. We denote the quotient space of all equivalence classes of σ by $\bar{\sigma}$, that is $\bar{\sigma} := \sigma/\cong$. It is clear that if $x \cong y$ for two random elements in σ and if x is a median of A a.a.s. then so is y. Motivated by this, we say that $\alpha \in \bar{\sigma}$ is a median of A a.a.s. if any $x \in \alpha$ be a median of A a.a.s. Let

$$M(A) := \{\alpha \in \bar{\sigma} : \alpha \text{ is a median of } A \text{ a.a.s.}\} \tag{23}$$

Note that for any two sequences of permutations $x = (x_n)_{n\in IN}$ and $y = (y_n)_{n\in IN}$ in an equivalence class $\alpha \in M(A)$, $d^{(n)}(x_n, y_n)$ is of order $o(\gamma_n\sqrt{n})$, for any arbitrary sequence $(\gamma_n)_{n\in IN}$ diverging to ∞. On the contrary, by definition, x and y in two different classes must be located far from each other.

Let f be a positive real-valued function on IN, and let $o = (o_n)_{n\in IN}$ be a random element of σ. Define the asymptotic sphere of radius f centred at o, denoted by $S(o, f)$, to be the set of all equivalence classes $\alpha \in \bar{\sigma}$ such that for any $(x_n)_{n\in IN} \in \alpha$

$$\frac{d^{(n)}(o_n, x_n) - f(n)}{\gamma_n \sqrt{n}} \to 0 \text{ in probability,} \tag{24}$$

for any sequence $(\gamma_n)_{n\in IN}$ diverging to ∞.

Let $x_i^* := (x_{in})_{n\in IN}$, and define the function $c_i^* : IN \to IR_+$ by $c_i^*(n) = c_i n$. The following is an immediate consequence of Theorem 2, which reduces the median search space.

Corollary 1. *Let $k \geq 3$ and $A = (A_n)_{n\in IN}$ be as defined in the statement of Theorem 1. Then*

$$M(A) = \cap_{i=1}^{k} S(x_i^*, c_i^*) \tag{25}$$

Proof. Trivial from Theorem 2. □

3. Conclusion

We have explored the set of all possible positions of the medians of k random genomes sampled from k independent reversal random walks starting at the identity, namely, $X^{1,n}(c_1 n), ..., X^{k,n}(c_k n)$ where $X^{1,n}, ..., X^{k,n}$ are independent reversal random walks on $\sigma_n^{\pm}$ and $0 < c_1 \leq ... \leq c_k$ such that $c_{k-1} + c_k \leq \frac{1}{2}$. Doing this makes a major difference in the volume of the median problem search space. This is normally the whole of $\sigma_n^{\pm}$, but it is now reduced to a much smaller search space (Corollary 1). More specifically, the proportion of the volume of the new search space to the volume of $\sigma_n^{\pm}$, i.e. $2^n n!$, converges to 0. By definition, it is clear that, in the case of existence of several asymptotic medians for $A = (A_n)_{n\in IN}$, the median classes are located far from each other. In other words, for $x = (x_n)_{n\in IN}$ in $\alpha_1 \in M(A)$ and $y = (y_n)_{n\in IN}$ in $\alpha_2 \in M(A)$ $(\alpha_1 \neq \alpha_2)$, there exists a sequence $(\gamma_n)_{n\in IN}$ diverging to ∞ such that

$$\frac{d^{(n)}(x_n, y_n)}{\gamma_n \sqrt{n}} \nrightarrow 0,$$

which, roughly speaking, means that, for a subsequence of natural numbers $(n_i)_{i\in IN}$, $d^{(n_i)}(x_{n_i}, y_{n_i})$ is of order greater than $\gamma_{n_i}\sqrt{n_i}$. Therefore, the other median classes are located far from the identity. Settling for a single median, then, can mislead us in the search for the position of the true ancestor. Further investigation is needed to study the number of the medians for an arbitrary value of k.

4. Acknowledgements

Funding: This work was supported in part by CNPq of Brazil [grant 150349/2016-5] and the Natural Sciences and Engineering Research Council of Canada [grant RGPIN-2016-05585]. DS holds the Canada Research Chair in Mathematical Genomics.

5. Authors' contributions

Both authors contributed the research and writing of this paper. Both have approved the final version.

References

[1] E. Tannier, C. Zheng, D. Sankoff, Multichromosomal median and halving problems under different genomic distances, BMC Bioinformatics 10 (120) (2009).
[2] C. Zheng, D. Sankoff, On the pathgroups approach to rapid small phylogeny, BMC Bioinformatics 12 (S4) (2011).
[3] A. Caprara, The reversal median problem, INFORMS J. Comput. 15 (2003) 93–113.
[4] A. Jamshidpey, D. Sankoff, Phase change for the accuracy of the median value in estimating divergence time, BMC Bioinformatics 14 (2013) S15:S7.
[5] G. Fertin, A. Labarre, I. Rusu, E. Tannier, S. Vialette, Combinatorics of Genome Rearrangements, MIT Press, 2009.
[6] N. Berestycki, R. Durrett, A phase transition in the random transposition random walk, Probab. Theory Related Fields 136 (22) (2006) 203–233.