Jamshidpey et al. BMC Genomics 2014, 15(Suppl 6):S3
http://www.biomedcentral.com/1471-2164/15/S6/S3

BMC Genomics

RESEARCH Open Access

Sets of medians in the non-geodesic pseudometric space of unsigned genomes with breakpoints

Arash Jamshidpey[1], Aryo Jamshidpey[2], David Sankoff[1*]

From Twelfth Annual Research in Computational Molecular Biology (RECOMB) Satellite Workshop on Comparative Genomics
Cold Spring Harbor, NY, USA. 19-22 October 2014

Abstract

Background: The breakpoint median in the set S_n of permutations on n terms is known to have some unusual behavior, especially if the input genomes are maximally different to each other. The mathematical study of the set of medians is complicated by the facts that breakpoint distance is not a metric but a pseudo-metric, and that it does not define a geodesic space.

Results: We introduce the notion of partial geodesic, or geodesic patch between two permutations, and show that if two permutations are medians, then every permutation on a geodesic patch between them is also a median. We also prove the conjecture that the input permutations themselves are medians.

Backgound

Among the common measures of gene order difference between two genomes, the edit distances, such as reversal distance or double-cut-and-join distance, contrast with the breakpoint distance in that the former are defined in a geodesic space while the latter is not. Another characteristic of breakpoint distance that it does not share with most other genomic distances is that it is a pseudometric rather than a metric.

A problem in computational comparative genomics that has been extensively studied under many definitions of genomic distance is the gene order median problem [1], the archetypical instance of the gene order small phylogeny problem. The median genome is meant, in the first instance, to embody the information in common among $k \geq 3$ given genomes, and second, to estimate the ancestral genome of these k genomes. We have shown that the second goal becomes unattainable as $n \to \infty$, where n is the length of the genomes, if there are more than $0.5n$ mutational steps changing the gene order [2]. Moreover, we have conjectured, and demonstrated in simulation studies, that where there is little or nothing in common among the k input genomes, the median tends to reflect only one (actually, any one) of them, with no incorporation of information from the other $k - 1$ [3].

In the present paper, we investigate this conjecture mathematically in the context of a wider study of medians for the breakpoint distance between unsigned linear unichromosomal genomes, although the methods and results are equally valid for genomes with signed and/or circular chromosomes, as well as those with $\chi > 1$ chromosomes, where χ is a fixed parameter. Our approach involves first a rigorous treatment of the pseudometric character of the breakpoint distance. Then, given the non-geodesic nature of the space we are able to define a weaker concept of geodesic patch, that we use later, given two or more medians, to locate further medians. We also prove the conjecture that for k genomes containing no gene order information among them, the normalized (divided by n) median score tends to $k - 1$, with high probability.

* Correspondence: sankoff@uottawa.ca
[1]Department of Mathematics and Statistics, University of Ottawa, 585 King Edward Avenue, Ottawa, Canada, K1N 6N5
Full list of author information is available at the end of the article

Results

From pseudometric to metric

We denote by S_n the set of all permutations of length n. Each permutation represents a unichromosomal linear genome where the numbers all represent different genes. For a permutation $\pi := \pi_1 \dots \pi_n$ we define the set of adjacencies of π to be all the unordered pairs $\{\pi_i, \pi_{i+1}\} = \{\pi_{i+1}, \pi_i\}$ for $i = 1, \dots, n-1$. For $I \subseteq S_n$ we denote by $\mathcal{A}_I := \mathcal{A}_I^{(n)}$ the set of all common adjacencies of the elements of I. Then $\mathcal{A}_{S_n} = \emptyset$, and we also write $\mathcal{A}_\emptyset$ for the set of all pairs $\{i, j\}$, $i \neq j$. For any $I, J \subseteq S_n$ $\mathcal{A}_{I \cup J} = \mathcal{A}_I \cap \mathcal{A}_J$. It will sometimes be convenient to write $\mathcal{A}_I$, the set of common adjacencies in $I = \{x_1, \dots, x_k\}$, as $\mathcal{A}_{x1, \dots, x_k}$. For example $A_{x,y,z}$ represents the set of adjacencies common to permutations x, y and z.

For $x, y \in S_n$we define the breakpoint distance (bp distance) between x and y by

$$d^{(n)}(x, y) := n - 1 - |\mathcal{A}_{x,y}|. \quad (1)$$

This distance is not a metric on S_n but rather a pseudometric because of nonreflexiveness: cases where $d^{(n)}(x, y) = 0$ but $x \neq y$, namely $x = \pi_1 \dots \pi_n$ and $y = \pi_n \dots \pi_1$, for any $x \in S_n$. In these cases, the permutations x and y are said to be equivalent, denoted by $x \sim y$. The equivalence class containing π is represented by $[\pi]$ and contains exactly two permutations, $\pi_1, \dots, \pi_n$ and $\pi_n, \dots, \pi_1$. The number of classes is thus $n!/2$. For any π, we denote the other element of $[\pi]$ by $\bar{\pi}$. The bp distance, a metric on the set of all equivalence classes of S_n, denoted by $\hat{S}_n := S_n/\sim$ is defined by

$$d^{(n)}([x], [y]) := d^{(n)}(x, y). \quad (2)$$

Where there is no risk of ambiguity, we can simplify the notation by using x and y instead of $[x]$ and $[y]$, and/or drop the superscript n.

It is clear that the maximum possible bp distance between two permutation classes is $n - 1$ when they have no common adjacencies. Bp distance is symmetric on S_n and hence on $\hat{S}_n$. By construction, it is reflexive on $\hat{S}_n$. To verify the triangle inequality, consider three permutations x, y, z. We have

$$\mathcal{A}_{x,z} \supseteq \mathcal{A}_{x,y,z} = \mathcal{A}_{x,y} \cap \mathcal{A}_{y,z} \quad (3)$$

Therefore

$$d(x, z) = n - 1 - |\mathcal{A}_{x,z}| \leq n - 1 - |\mathcal{A}_{x,y}| - |\mathcal{A}_{y,z}| + |\mathcal{A}_{x,y} \cup \mathcal{A}_{y,z}|. \quad (4)$$

But $|\mathcal{A}_{x,y} \cup \mathcal{A}_{y,z}| = |\mathcal{A}_y \cap (\mathcal{A}_x \cup \mathcal{A}_z)| \leq n - 1$ and hence the triangle inequality holds.

We say a pseudometric (or a metric) $\tilde{\rho}$ is right invariant on a group G if for any $x, y, z \in G$, $\tilde{\rho}(x, y) = \tilde{\rho}(xz, yz)$. The definition of the left invariance is similar. A pseudometric (metric) which is both right and left invariant is called invariant. Bp distance is an invariant pseudometric on S_n.

Definition 1 *Given a set $\{x_1, \dots, x_k\} \subseteq S$ and a pseudometric space ρ on S, a median for the set is $\mu \in S$ such that $\sum_{i=1}^{k} \rho(\mu, x_i)$is minimal.*

Defining the geodesic patch

A discrete metric space (S, ρ) is a geodesic space if for any two points $x, y \in S$ there exists a finite subset of S containing x, y that is isometric with the discrete line segment $[0, 1, \dots, \rho(x, y)]$. Any subset of S with this property, and there may be several, is called a geodesic between x and y. For example, all connected graphs are geodesic spaces. In a geodesic space the medians of two points x and y consist of all the points located on geodesics between x and y.

What can we say when the space is not a geodesic space? To answer this, we extend the concept of geodesic by introducing the concept of a geodesic patch. A geodesic patch between x and y is a maximal subset of S containing x, y which is isometric to a subsegment (not necessarily contiguous) of the line segment $[0, 1, \dots, \rho(x, y)]$. For any two points x, y in an arbitrary metric space (S, ρ) there exists at least one geodesic patch between them because x, y is isometric to $\{0, \rho(x, y)\}$. In addition, any geodesic is a geodesic patch. Any point z on a geodesic patch between x, y satisfies:

$$\rho(x, y) = \rho(x, z) + \rho(z, y). \quad (5)$$

Therefore all the medians of two points x and y must lie on a geodesic patch between them. We denote the set of all permutations lying on geodesic patches connecting $x, y \in S_n$by $\overline{[x, y]}$, as in Figure 1.

$(\hat{S}_n, d)$ is not a geodesic space. For example there is no geodesic connecting the identity permutation id and $\pi := 1\ 2\ x_1\ x_2 \dots x_{n-4}\ n-1\ n$ when $x_1\ x_2 \dots x_{n-4}$ is a non-identical permutation on $\{3, \dots, n-2\}$. The smallest change to id is to cut one of its adjacencies, say $\{i, i+1\}$, and rejoin the two segments in one of the three possible ways: 1 to n, 1 to $i+1$ or n to i. Now if we cut the adjacencies $\{1, 2\}$ or $\{n-1, n\}$ in id the distance of the new permutation to both id and π increases. If on the other hand we cut one of the other adjacencies in id all the ways of rejoining, which increase the distance to id, either increase or leave unchanged the distance to π, since $\{1, n\}$, $\{1, i+1\}$ and $\{n, i\}$ are not adjacencies in $\mathcal{A}_\pi$. Therefore there is no geodesic connecting id to π.

Although $\hat{S}_n$ is not a geodesic space there may still exist permutations with a geodesic between them. For example

$$\{id = 123456,\ 213456,\ 312456,\ 421356,\ 531246,\ \pi = 135246) \quad (6)$$

is a geodesic between id and π. Note $d(id, \pi) = 5$, the maximum possible distance in $\hat{S}_6$.

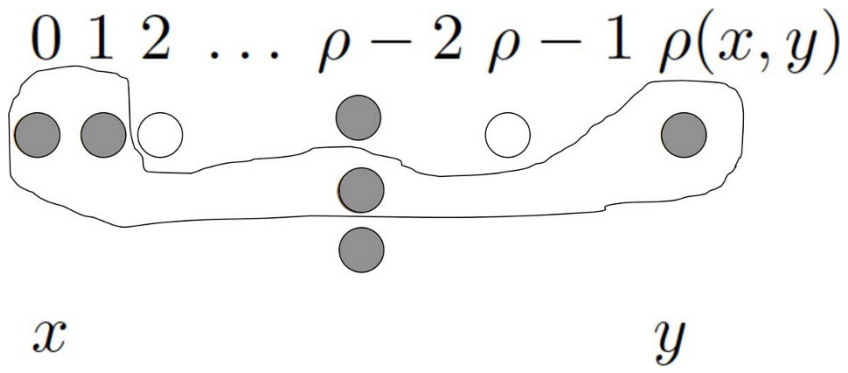

Figure 1 Geodesic patch. The open dots are situated beneath distances for which there is no permutation. The filled dots are in $\overline{[x,\ y]}$; they represent permutations that can lie on a geodesic patch, and the dots within the outlined area represent a specific geodesic patch.

The median value and medians of permutations with maximum pairwise distances

In this section we investigate the bp median problem in the case of k permutations with maximum pairwise distances. As we shall see later, this situation is very similar to the case of k uniformly random permutations. Let (S, ρ) be a pseudometric space.

The total distance of a point $x \in S$ to a finite subset $\varnothing \neq B \subseteq S$ is defined to be

$$\rho(x,\ B) := \sum_{y\ \in\ B} \rho(x,\ y). \tag{7}$$

The median value of B, $m^{S,\rho}(B)$, is the infimum of the total distance when the infimum is over all the points $x \in S$, that is

$$m^{S,\rho}(B) := \inf_{x\ \in\ S} \rho(x, B). \tag{8}$$

We can extend this definition to sets with multiplicities. Let $\varnothing \neq B \subseteq S$. We define a multiplicity function n_B from B to $\mathbb{N}$ and write $n_B(x) = n_x$. We call $A = (B, n_B)$ a set with multiplicities. We define the total distance of a point $x \in S$ to A to be

$$\rho(x,\ A) := \sum_{y\ \in\ B} n_y \rho(x,\ y). \tag{9}$$

The definition of median value in Equation (8) can be extended in an analogous way to the median value of a set with multiplicity A. When S is finite then the total distance function takes its minimum on S and "inf" turns into "min" in the above formulation. The points of the space S that minimize the total distance to A are called the median points or medians of A and the set of all these medians is called the median set of A, denoted by $M^{S,\rho}(A)$.

Let B and $A = (B, n_B)$ be a subset and a subset with multiplicities of S_n. We define $[B]$ to be the set of all permutation classes of S_n that have at least one of their permutations in B. That is

$$[B] = \{[x]\ \in\ \hat{S}_n \text{ such that } \exists y\ \in\ B \text{ with } x\ \sim\ y\}. \tag{10}$$

Two nonempty subsets $B, B' \subseteq S_n$ are said to be equivalent, denoted by $B \sim B'$, if $[B] = [B']$. Also we define $[n_B]$ to be a function from $[B]$ to $\mathbb{N}$ with

$$[n_B]([x]) = n_{[x]} := \sum_{x \sim y \in B} n_y. \tag{11}$$

Then the definition of $[A]$ is straightforward:

$$[A] := ([B], [n_B]), \tag{12}$$

and we say two nonempty subsets of $S_n$with multiplicities, namely A and A' are equivalent, denoted by $A \sim A'$, if $[A] = [A']$. In fact $[A]$ is the equivalence class containing A. We call $[A]$ a subset of $\hat{S}_n$ with multiplicities. We use the notations "[ ]" and " ~ " for all the above concepts without restriction.

With these definitions we can readily verify that in the context of bp distance, for $A \sim A'$ and $x \sim x'$, we have

$$d(x,\ A) = d(x',\ A') = d([x], [A]). \tag{13}$$

Recall that we use d as both a metric on $\hat{S}_n$ and a pseudometric on S_n. Therefore we can conclude that

$$m^{S_n,d}(A) = m^{S_n,d}(A') = m^{\hat{S}_n,d}([A]) \tag{14}$$

and similarly

$$[M^{S_n,d}(A)] = [M^{S_n,d}(A')] = M^{\hat{S}_n,d}([A]). \tag{15}$$

Henceforward, we will simplify by replacing the notation $m^{S_n,d}(A)$ and $M^{S_n,d}(A)$ by $m_n(A)$ and $M_n(A)$, respectively. Also for a subset $[A]$ of $\hat{S}_n$ with multiplicities, we will use the notation $m_n([A])$ and $M_n([A])$ instead of $m^{\hat{S}_n,d}([A])$ and $M^{\hat{S}_n,d}([A])$ respectively. Where there is no ambiguity we will suppress the subscript n.

Proposition 1 *Suppose* $X := \{x_1, \ldots, x_k\} \subset \hat{S}_n$ *such that* $d(x_i, x_j) = n - 1$ *for any* $i \neq j$, $i \leq i, j \leq n$. *Then the bp*

median value of × *is* $(k-1)(n-1)$. *Moreover,* m^* *is a median of X,* $m^* \in M(X)$, *if and only if* $A_{m*} \subset \cup_{i=1}^{k} A_{x_i}$.

Proof Let $\pi \in \hat{S}_n$ be an arbitrary permutation class. Since $A_{\pi,x_i} \subset A_{x_i}$ and $A_{\pi,x_j} \subset A_{x_j}$ for any $1 \le i, j \le k$, we have $A_{\pi,x_i} \cap A_{\pi,x_j} = \emptyset$. Also

$$\cup_{i=1}^{k} A_{\pi,x_i} \subset A_{\pi} \tag{16}$$

Therefore

$$\sum_{i=1}^{k} |A_{\pi,x_i}| \le |A_{\pi}| = n-1 \tag{17}$$

Hence

$$\sum_{i=1}^{k} d(\pi, x_i) \ge (k-1)(n-1) \tag{18}$$

The equality holds letting $\pi = x_i$ for any $1 \le i \le k$. This proves the first part of the proposition. For the second part we know that $m^* \in M(X)$ is equivalent with the fact that the total distance of m^* to X is $(k-1)(n-1)$, and this is equivalent to $\sum_{i=1}^{k} |A_{m^*,x_i}| = n-1$ and $\cup_{i=1}^{k} A_{m^*,x_i} = A_{m^*}$ be written as $A_{m^*} \cap (\cup_{i=1}^{k} A_{x_i})$. This finishes the proof of the equivalence relation in the proposition.

Lemma 1 *Let x, y, z be three permutation classes in* $\hat{S}_n$ *that are pairwise at a maximum distance* $n-1$ *from each other. Then for any* $w \in \overline{[x,y]}$ *we have* $d(w, z) = n-1$.

Proof Having $w \in \overline{[x,y]}$ we have $A_w \subset A_x \cup A_y$. Also we know that $A_z \cap (A_x \cup A_y) = \emptyset$. This concudes the result.

The above lemma simply indicates that for any two points x_i, x_j in the set X in the proposition above $\overline{[x_i, x_j]} \subset M(X)$ since the total distance of each point in $\overline{[x_i, x_j]}$ to X is $(k-1)(n-1)$.

Corollary 1 *Suppose* $X := \{x_1, \ldots, x_k\} \subset \hat{S}_n$ *such that* $d(x_i, x_j) = n-1$ *for any* $i \ne j$. *Then* $\cup_{i,j} \overline{[x_i, x_j]} \subset M(X)$.

What more can we say about the median positions? The notion of "accessibility" will help us to keep track of some other medians of the set X that are not in $\cup_{i,j} \overline{[x_i, x_j]}$. Before defining this concept, we first need more information about the properties of $\overline{[x,y]}$ for $x, y \in \hat{S}_n$.

Lemma 2 *Let* $x, y \in \hat{S}_n$. *Then* $z \in \overline{[x,y]}$ *if and only if* $A_{x,y} \subset A_z \subset A_x \cup A_y$.

Proof We know $z \in \overline{[x,y]}$ if and only if $d(x, z) + d(z, y) = d(x, y)$. On the other hand we can write A_z as follows

$$A_z = A_{z,x,y} \cup (A_{z,x} \setminus A_y) \cup (A_{z,y} \setminus A_x) \cup (A_z \setminus (A_x \cup A_y)) \tag{19}$$

where the pairwise intersection of the sets in the right hand side is empty. We can also write

$$d(x, z) = (n-1) - |A_{z,x,y}| - |A_{z,x} \setminus A_y| \tag{20}$$

and

$$d(z, y) = (n-1) - |A_{z,x,y}| - |A_{z,y} \setminus A_x|. \tag{21}$$

Furthermore

$$d(x, y) \le (n-1) - |A_{z,x,y}| \tag{22}$$

and

$$(n-1) - |A_{z,x,y}| - |A_{z,x} \setminus A_y| - |A_{z,y} \setminus A_x| = |A_z \setminus (A_x \cup A_y)|. \tag{23}$$

Now for "sufficiency", we have

$$(n-1) - |A_{z,x,y}| - |A_{z,x} \setminus A_y| - (n-1) - |A_{z,x,y}| - |A_{z,y} \setminus A_x| \tag{24}$$

$$= (n-1) - |A_{x,y}| \le (n-1) - |A_{x,y,z}| \tag{25}$$

Therefore by Equation (23) we have

$$(n-1) - |A_{z,x,y}| - |A_{z,x} \setminus A_y| - |A_{z,y} \setminus A_x| = |A_z \setminus (A_x \cup A_y)| \le 0 \tag{26}$$

This results in $|A_{x,y}| = |A_{x,y,z}|$ and hence in $A_{x,y} \subset A_z$. Otherwise the inequality in (26) will be strict, which is impossible. On the other hand the inequality in (26) shows $A_z \setminus (A_x \cup A_y) = \emptyset$ which concludes at $A_z \subset A_x \cup A_y$.

For "necessity", we have

$$(n-1) - |A_{z,x,y}| - |A_{z,x} \setminus A_y| - |A_{z,y} \setminus A_x| + (n-1) - |A_{x,y}| = (n-1) - |A_{x,y}| \tag{27}$$

This is true because of $A_z \subset A_x \cup A_y$ and Equation (23). But since $A_{x,y} \subset A_z \subset A_x \cup A_y$ we have $|A_{x,y}| = |A_{x,y,z}|$ and we can replace $|A_{x,y}|$ by $|A_{x,y,z}|$ in the left hand side of the last equality. This finishes the "necessity" proof.

Definition 2 *Let* $\times := \{x_1, \ldots, x_k\}$ *be a subset of* $\hat{S}_n$. *We say a permutation class* $z \in \hat{S}_n$ *is 1-accessible from X if there exists an* $m \in \mathcal{N}$, *a finite sequence* $y_1, \ldots, y_m$ *where* $y_i \in X$ *and* $z_1, \ldots, z_m$, *where* $z_i \in \hat{S}_n$ *such that* $z_1 = y_1$, $z_m = z$ *and* $z_{i+1} \in \overline{[z_i, y_{i+1}]}$ *for* $i = 1 \ldots m-1$. *See* Figure 2.

We denote the set of all 1-accessible points of X by $Z(X)$. *We define* $Z_0(X) := X$. *Also for* $r \in \mathcal{N} \cup \{0\}$, *by induction, we define* $Z_{r+1}(X)$ *to be* $Z(Z_r(X))$ *and we call it the set of all r+1-accessible permutation classes. That is* $Z_1(X) = Z(X)$, $Z_2(X) = Z(Z(X))$ *and so on. It is clear that* $Z_{r+1}(X)$ *includes* $Z_r(X)$ *and also* $\cup_{x,y \in Z_r(X)} \overline{[x,y]}$. *A permutation class z is said to be accessible from* × *if there exists* $r \in \mathcal{N}$ *such that* $z \in Z_r(X)$. *We denote the set of all accessible points by* $\bar{Z}(X) = \cup_{r \in IN \cup \{0\}} Z_r(X)$.

Note that $Z(\bar{Z}(X)) = \bar{Z}(X)$. This holds because for any 1-accessible permutation class z from $\bar{Z}(X)$, there must exist $m \in \mathcal{N}$, $r_0 \in \mathcal{N}, \cup\{0\}, y_1, \ldots, y_m \in \bar{Z}_{r_0}(X)$, (the y_i's must be in $\bar{Z}(X)$, thus there must be such an r_0) and $z_1, \ldots, z_m$ where $z_i \in \hat{S}_n$ such that $z_1 = y_1$, $z_m = z$ and $z_{i+1} \in \overline{[z_i, y_{i+1}]}$. Therefore $z \in Z_{r_0+1}(X) \subset \bar{Z}(X)$. We can then conclude that $\bar{Z}(\bar{Z}(X)) = \bar{Z}(X)$.

Proposition 2 *Suppose* $X := \{x_1, \ldots, x_k\} \subset \hat{S}_n$ *such that* $d(x_i, x_j) = n-1$ *for any* $i \ne j$. *Then for any permutation*

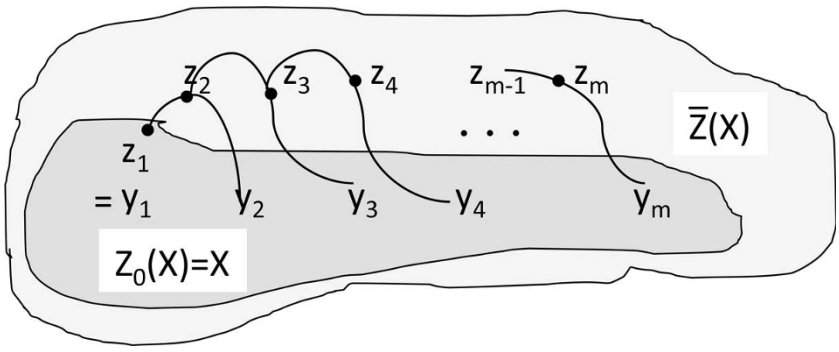

Figure 2 Accessibility. Illustration of how $\bar{Z}$ is constructed.

class $z \in \bar{Z}(X)$ *the total distance* $d(z, X)$ *between* z *and* $\times$ *is* $(k-1)(n-1)$ *and hence* $\bar{Z}(X) \subset M(X)$ *Furthermore if* $m_1, m_2 \in M(X)$ *then* $\overline{[m_1, m_2]} \subset M(X)$.

Proof Suppose $m_1, m_2 \in M(X)$ and $m^* \in \overline{[m_1, m_2]}$. By Lemma 2 and Proposition 1 we have $A_{m^*} \subset A_{m_1} \cup A_{m_2} \subset \cup_{i=1}^{k} A_{x_i}$. Applying Proposition 1 again, we have $m^* \in M(X)$. Now it suffices to show that for any $r \in IN \cup \{0\}$, $Z_r(X) \subset M(X)$. We prove this by induction. For $r = 0$ this follows from Corollary 1. Suppose $Z_r(X) \subset M(X)$. By definition we have $Z_{r+1}(X) = Z(Z_r(X))$. That is for $z \in Z_{r+1}(X)$ there exists an $m \in \mathcal{N}$, $y_1, ..., y_m \in Z_r(X)$ and $z_1, ..., z_m$, where $z_i \in \hat{S}_n$, such that $z_1 = y_1$, $z_m = z$ and $z_{i+1} \in \overline{[z_i, y_{i+1}]}$. $z_1 \in \overline{[y_1, y_2]}$ and by the fact we proved above $z_1 \in M(X)$ since $y_1, ..., y_m \in Z_r(X) \subset M(X)$. Continuing this we conclude that $z_1, z_2, ..., z_m = z \in M(X)$. Hence $Z_{r+1}(X) \subset M(X)$. This finishes the proof.

Conjecture 1 *Every median point of* X *is accessible from* X, *that is* $M(X) = \bar{Z}(X)$.

The median value and medians of k random permutations.

In this section we study the median value and median points of k independent random permutation classes uniformly chosen from $\hat{S}_n$. This is equivalent to studying the same problem for k random permutations sampled from S_n. All the results of this section carry over to permutations without any problem.

We make use of the fact that the bp distance of two independent random permutations tends to be close to its maximum value, $n - 1$. Xu et al. [4] showed that if we fix a reference linear permutation id and pick a random permutation x uniformly, the expected number and variance of $|\mathcal{A}_{id,x}^{(n)}|$ both are very close to 2 for large enough n. Because of the symmetry of the group S_n and the fact that bp distance is an invariant pseudometric the same results hold for two random permutations x and y. We first summarize the results we need from [4].

Let $\tilde{\nu}_n$ be the uniform measure on S_n. Let $\Pi : S_n \to \hat{S}_n$ be the natural surjective map sending each permutation onto its corresponding permutation class.

Define

$$\nu_n := \Pi * \tilde{\nu}_n \tag{28}$$

to be the push-forward measure of $\tilde{\nu}_n$ induced by the map Π. It is clear that ν_n is the uniform measure on $\hat{S}_n$. The following proposition is a reformulation of Theorems 6 and 7 in [4].

Proposition 3 *[Xu-Alain-Sankoff] Let* $\times$ *and* y *be two independent random permutation classes (irpc) chosen uniformly from* $\hat{S}_n$. *Then*

$$\mathrm{E}[d(x, y)] = n - 3 - \frac{2}{n} - o(\frac{2}{n}) \tag{29}$$

$$\mathrm{Var}[d(x, y)] = 2 - \frac{2}{n} - o(\frac{2}{n}) \tag{30}$$

Define the error function for the distance of x, y by

$$\varepsilon_n(x, y) := (n-1) - d(x, y) = |\mathcal{A}_{x,y}|. \tag{31}$$

Corollary 2 *Suppose* $\times$ *and* y *are two irpc's sampled from the uniform measure* ν_n *and* a_n *is an arbitrary sequence of real numbers diverging to* $+\infty$. *Then* $\frac{\varepsilon_n(x, y)}{a_n}$ *converges to zero asymptotically* ν_n^{*2}*-almost surely (a.a.s.), that is*

$$\frac{\varepsilon_n(x, y)}{a_n} \to 0 \text{ in probability.} \tag{32}$$

Proof The proof is straightforward from [4] and Chebyshev's inequality.

Now we are ready to study the median value of k *irpc*'s. Let $[A]$ be a subset of $\hat{S}_n$ with multiplicities and with k elements. Define

$$e_n([A]) := (k-1)(n-1) - m_n([A]). \tag{33}$$

Theorem 1 *Let $X^{(n)} := \{x_1^{(n)}, x_2^{(n)}, \ldots, x_k^{(n)}\}$ be a set of k irpc in $\hat{S}_n$ sampled from the measure ν_n^{*k}. Then their breakpoint median value $m_n^* := m_n(X^{(n)})$ tends to be close to its maximum after a convenient rescaling with high probability, that is for any arbitrary sequence $a_n \to \infty$ as ν_n^{*k} in ν_n^{*k}-probability where $e_n^* := e_n(X^{(n)})$.*

Proof Let π be an arbitrary point of S_n. Let $\mathcal{A}_{\pi \setminus X} = \mathcal{A}_\pi \setminus \mathcal{A}_X$. We have

$$\sum_{i=1}^{k} |\mathcal{A}_{\pi,x_i}| \leq |\mathcal{A}_{\pi \setminus X}| + \sum_{i=1}^{k} |\mathcal{A}_{\pi,x_i}| \leq (n-1) + \binom{k}{2} \alpha_n \quad (34)$$

where α_n is $\max_{i,j} \varepsilon_n(x_i, x_j)$. On the other hand $m_n(X^{(n)}) \leq (k-1)(n-1)$. The reason is the same as has already been discussed in the proof of Proposition 1. Therefore subtracting $(k-1)(n-1)$ we have

$$0 \leq e_n^* \leq \binom{k}{2} \alpha_n. \quad (35)$$

Dividing by a_n and letting n go to ∞ the result follows from the last corollary.

Theorem 2 *Let $X^{(n)} := \{x_1^{(n)}, x_2^{(n)}, \ldots, x_k^{(n)}\}$ be a set of k irpc's in $\hat{S}_n$ sampled from the measure ν_n^{*k}. Then for any permutation class $z^{(n)} \in \bar{Z}(X^{(n)})$ the total distance of $z^{(n)}$ to $\times$ is close to $(k-1)(n-1)$ with high probability after a convenient rescaling. More explicitly, for any arbitrary sequence of real numbers a_n converging to ∞*

$$\frac{(k-1)(n-1) - d^{(n)}(z^{(n)}, X^{(n)})}{a_n} \to 0 \text{ in } \nu_n^{*k} - \text{probability.} \quad (36)$$

Therefore

$$\frac{d^{(n)}(z^{(n)}, X^{(n)}) - m_n(X^{(n)})}{a_n} \to 0 \text{ in } \nu_n^{*k} - \text{probability.} \quad (37)$$

Furthermore if $m_1^{(n)}, m_2^{(n)} \in M_n(X^{(n)})$ then for any $\frac{d^{(n)}(\tilde{m}^{(n)}, X^{(n)}) - m_n(X^{(n)})}{a_n} \to 0$ *in* ν_n^{*k} – probability.

$$\frac{d^{(n)}(\tilde{m}^{(n)}, X^{(n)}) - m_n(X^{(n)})}{a_n} \to 0 \text{ in } \nu_n^{*k} - \text{probability.} \quad (38)$$

Proof The structure of the proof is similar to the proof of Proposition 1. Suppose $o \in \hat{S}_n$ with $\mathcal{A}_o \subset_{i=1}^{k} \cup \mathcal{A}_{x_i}$. Let α_n be as defined in the proof of Theorem 1. Then by the same discussion we have

$$n - 1 \leq \sum_{i=1}^{k} |\mathcal{A}_{o,x_i}| \leq n - 1 + \binom{k}{2} \alpha_n. \quad (39)$$

Therefore

$$(k-1)(n-1) \geq d(o, X) \geq (k-1)(n-1) - \binom{k}{2} \alpha_n \quad (40)$$

and

$$\frac{(k-1)(n-1) - d(o,X)}{a_n} \to 0 \quad \text{in probability.} \quad (41)$$

From Theorem 1 we have

$$\frac{(k-1)(n-1) - m_n(X)}{a_n} \to 0 \text{ in probability.} \quad (42)$$

Hence

$$\frac{d(o,X) - m_n(X)}{a_n} \to 0 \text{ in probability.} \quad (43)$$

It suffices to show that $z := Z^{(n)} \in \bar{Z}(X)$ has the same property, that is $\mathcal{A}_z \in \cup_{i=1}^{k} \mathcal{A}_{x_i}$. But this is clear by induction. For the second part of the theorem let $m_{1,n}^*, m_{2,n}^* \in M(X)$. Suppose $m^* \in [m_{1,n}^*, m_{2,n}^*]$. By Theorem 1 $\frac{|A_{m_{in}^* \setminus X}|}{a_n} \to 0$ in probability for $i = 1, 2$. On the other hand we have $\mathcal{A}_{m^* \setminus X} \subset \mathcal{A}_{m_{1,n}^* \setminus X} \cup \mathcal{A}_{m_{2,n}^* \setminus X}$.

Therefore

$$\frac{|\mathcal{A}_{m^* \setminus X}|}{a_n} \to 0 \text{ in probability.} \quad (44)$$

Therefore

$$(k-1)(n-1) \leq d(m^*, X) \leq (k-1)(n-1) + \binom{k}{2} \alpha_n \quad (45)$$

since

$$\frac{|\mathcal{A}_{m^*,x_i} \cap \mathcal{A}_{m^*,x_j}|}{a_n} \to 0 \text{ in probability.} \quad (46)$$

The statement follows from the last inequality.

Conclusions

We have shown that the median value for a set of random permutations tends to be close to its extreme value with high probability. Also it has been shown that every permutation accessible from a set of random permutations can be considered as a median of that set asymptotically almost surely, and conjectured that the converse is true, that every median is accessible from the original set in this way.

Further work is needed to characterize the existence and size of non-trivial geodesic patches, in order to assess how extensive the set of medians is.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

All authors participated in the research, wrote the paper, read and approved the manuscript.

Acknowledgements
Research supported in part by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC). DS holds the Canada Research Chair in Mathematical Genomics.

Declarations
The publication charges for this article were funded by the Canada Research Chair in Mathematical Genomics, and by the University of Ottawa.
This article has been published as part of *BMC Genomics* Volume 15 Supplement 6, 2014: Proceedings of the Twelfth Annual Research in Computational Molecular Biology (RECOMB) Satellite Workshop on Comparative Genomics. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcgenomics/supplements/15/S6.

Authors' details
[1]Department of Mathematics and Statistics, University of Ottawa, 585 King Edward Avenue, Ottawa, Canada, K1N 6N5. [2]Department of Computer Science and Information Technology, Institute for Advanced Studies in Basic Sciences, Gava Zang, Zanjan 45195-1159, Iran.

Published: 17 October 2014

References

1. Tannier E, Zheng C, Sankoff D: **Multichromosomal median and halving problems under different genomic distances.** *BMC Bioinformatics* 2009, **10**:120.
2. Jamshidpey A, Sankoff D: **Phase change for the accuracy of the median value in estimating divergence time.** *BMC Bioinformatics* 2013, **14**:S15:S7.
3. Haghighi M, Sankoff D: **Medians seek the corners, and other conjectures.** *BMC Bioinformatics* 2012, **13**:S19:S5.
4. Xu AW, Alain B, Sankoff D: **Poisson adjacency distributions in genome comparison: multichromosomal, circular, signed and unsigned cases.** *Bioinformatics* 2008, **24**:i146-i152.

doi:10.1186/1471-2164-15-S6-S3
Cite this article as: Jamshidpey *et al.*: Sets of medians in the non-geodesic pseudometric space of unsigned genomes with breakpoints. *BMC Genomics* 2014 **15**(Suppl 6):S3.