The Author(s) *BMC Bioinformatics* 2016, **17**(Suppl 86):473
DOI 10.1186/s12859-016-1340-y

BMC Bioinformatics

RESEARCH Open Access

# Compromise or optimize? The breakpoint anti-median

Caroline Anne Larlee, Alex Brandts and David Sankoff*

*From* 11th International Conference of the AB3C + Brazilian Symposium of Bioinformatics
São Paulo, Brazil. 3-6 November 2015

## Abstract

**Background:** The median of $k \geq 3$ genomes was originally defined to find a compromise genome indicative of a common ancestor. However, in gene order comparisons, the usual definitions based on minimizing the sum of distances to the input genomes lead to degenerate medians reflecting only one of the input genomes. "Near-medians", consisting of equal samples of gene adjacencies from all the input genomes, were designed to restore the idea of compromise to the median problem.

**Result:** We explore adjacency sampling constructions in full generality in the case $k = 3$, with given overlapping sets of adjacencies in the three genomes, where all adjacencies in two-way or three-way overlaps are included in the sample. We require the construction to be maximal, in the sense that no additional proportion of adjacencies from any of the genomes may be added without violating the local linearity of the genome. We discover that in incorporating as many adjacencies as possible, evenly from all the input genomes, we are actually maximizing, rather than minimizing, the sum of distances over all other maximal sampling schemes.

**Conclusions:** We propose to explore compromise instead of parsimony as the organizing principle for the small phylogeny problem.

**Keywords:** Median problem, Gene order, Breakpoint distance, Gene adjacency

## Background

In comparative genomics, a median genome $m$ for a set of $k \geq 3$ given genomes $g_1, \ldots, g_k$ in a metric space $(G, d)$ minimizes

$$S(m) = \sum_{i=1}^{k} d(m, g_i) \tag{1}$$

over all $m \in G$ [1]. This is meant to embody a compromise among the given genomes, usually as an inference of a common ancestor.

While the simplicity of the median concept is appealing, and it has stimulated a large literature [2], it suffers from important shortcomings: it is hard to calculate [3–5] for almost all $(G, d)$, and is not a compromise in the most important contexts. For example, for $k \geq 3$ random signed permutations of length $n$, and for $d$ the "breakpoint distance", the median tends to one or more of the given permutations as $n$ increases [6–8].

The "near median" was proposed to get around these difficulties [9]. For $k$ random genomes, the same proportion of gene adjacencies is sampled from each one, in such a way that the union of the samples is compatible – an "end" of a gene is adjacent to no more than one other gene end. The proportion of the compromise genome remaining to be constructed can be filled by any matching of the unassembled gene ends, as in Fig. 1.

If comparable proportions of the constructed genome are contributed by each of the $k$ genomes, the spirit of compromise is ensured. The sampling is rapidly carried out.

In the original paper [9], only the following, highly symmetrical cases were studied for $k = 3$: three purely random genomes, three genomes all with common adjacencies forming a proportion $\psi$ of their adjacencies, and three genomes all with a proportion $\psi$ of common

*Correspondence: sankoff@uottawa.ca
Department of Mathematics and Statistics, University of Ottawa, 585 King Edward Avenue, K1N 6N5 Ottawa, Canada

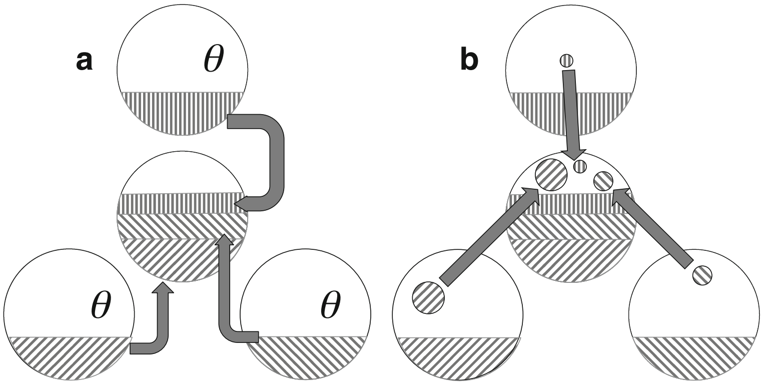

**Fig. 1 a** First sampling of $\theta n$ adjacencies from each of three genomes. **b** Supplementary sampling of residual adjacencies consisting of two free ends

adjacencies and additional proportions $\omega_{1,2}, \omega_{1,3}, \omega_{2,3}$ of adjacencies in their pairwise intersections. We only investigated the maximum $\theta$ such that the same proportion $\theta$ could be sampled from the three input genomes.

In the present paper we extend our analysis to examine the entire set of compatible triples $(\theta_1, \theta_2, \theta_3)$.

In the process, we discover the surprising fact that not only does our sampling procedure fail to minimize the sum in (1), it actually maximizes it! In doing so, it illustrates that the search for optimality and compromise are at cross-purposes. In concluding we suggest how the goal of compromise may be used as a criterion for the small phylogeny problem in the place of optimality.

## Results

### Definitions

Consider three signed genomes, $g_1, g_2$ and $g_3$, each consisting of one or more chromosomes – circular orderings – containing the same $n$ genes and each containing $n$ gene adjacencies. Although we assume the chromosomes are circular for technical simplicity, the analysis is essentially the same for linear, circular, unichromosomal or multichromosomal genomes; the effect of allowing a bounded number $> 1$ of chromosomes would be $\mathcal{O}(n)$ as would be the differences between circular and linear models. We also assume $n$ is large so that for an arbitrary proportion $\theta$, the $\mathcal{O}(1)$ difference between $\theta n$ and the nearest integer to $\theta n$ may be neglected. The probabilistic justification behind these assumptions is discussed in [9].

That the genomes are “signed” means the genes have polarity, so the two ends of a gene have distinct labels. Each adjacency is thus an unordered pair of the $2n$ gene ends, chosen from among $\binom{2n}{2}$ possibilities. For a genome to be “compatible”, no gene end may be part of more than one adjacency. There is no constraint involving the two ends of the same gene, other than that both ends of all genes must eventually be included in any genome we construct. E.g., there is no constraint against the two ends of the same gene being adjacent, forming a minimal circular chromosome.

We are given that $g_1, g_2$ and $g_3$ have a proportion $\psi$ of common adjacencies and proportions $\omega_{1,2}, \omega_{1,3}, \omega_{2,3}$ of adjacencies in their pairwise intersections.

The breakpoint distance between two genomes can be defined as $d = n - a$, where $a$ is the number of adjacencies they contain in common. For example $d(g_1, g_3) = n - \psi - n\omega_{1,3}$.

For a genome $x$ the sum of the normalized distances to the three input genomes,

$$s(x) = \frac{1}{n} \sum_{i=1}^{3} d(x, g_i), \qquad (2)$$

is called its score.

A sample is defined by a triple of $(\theta_1, \theta_2, \theta_3)$ each between 0 and 1 and summing to less than $1 - \psi - \omega_{1,2} - \omega_{1,3} - \omega_{2,3}$ such that a random choice of $\theta_1 n$ adjacencies from $g_1$, $n\theta_2$ from $g_2$, and $n\theta_3$ from $g_3$ are compatible with each other and with the adjacencies in the overlaps. A sample is “randomly completed” to form a genome with $n$ genes by the addition of $1 - \psi - \omega_{1,2} - \omega_{1,3} - \omega_{2,3}$ adjacencies constructed by randomly pairing gene ends that are not in any of the adjacencies in the sample or in the overlaps. In other words, to focus on the purely statistical consequences of the sampling procedure we thus do not consider the increment in the number of adjacencies obtainable in individual instances by the ad hoc matching algorithms developed in [9]. The random completion process does not add to the number of adjacencies in the sample in common with one, two or three of $g_1, g_2$ and $g_3$.

A “maximal” sample is one where none of the $\theta_i$ may be increased without causing a number (greater than $\mathcal{O}(n)$) of incompatible adjacencies.

## The construction

From the three input genomes, we construct a set containing adjacencies sampled in various proportions among $g_1, g_2$ and $g_3$ and including the adjacencies in the given two-way and three-way overlaps, randomly completed by pairs of gene ends matched from among the remaining unsampled ends. The only constraint in adding an adjacency is that it must have two "free ends"; i.e., no adjacency previously included, whether given or sampled, may contain either of these two ends.

Note that two random permutations can be expected to have virtually no adjacencies in common; the expectation of the number of adjacencies goes to a small constant as $n$ increases [10].

As an illustration, consider the case where $\psi = \omega_{1,2} = \omega_{1,3} = \omega_{2,3} = 0$. As a first step, we may select $\theta_1 n$ adjacencies from $g_1$, where $0 \leq \theta_1 \leq 1$. Then for $g_2$, the expected proportion of "two free ends", adjacencies where neither end appears in a previously selected $g_1$ adjacency, is $(1-\theta_1)^2$. As long as $\theta_1 \neq 1$, we can pick $\theta_2 n$ adjacencies from genome $g_2$ that do not conflict with any of those selected from $g_1$, where $0 \leq \theta_2 \leq (1-\theta_1)^2$.

Similarly, having then selected $\theta_1 n$ pairs of gene ends from $g_1$ and $\theta_2 n$ pairs of gene ends from $g_2$, the expected proportion of pairs in $g_3$ with two free ends is $(1-\theta_1-\theta_2)^2$. As long as this quantity is greater than zero, we can chose some $\theta_3 n$ compatible pairs from $g_3$.

For a maximal sample we should take the maximum number of pairs from $g_3$, i.e., the maximum $\theta_3$, given $\theta_1$ and $\theta_2$, i.e.,

$$\theta_3 = (1-\theta_1-\theta_2)^2. \tag{3}$$

Adding the remainder of the gene ends not in any adjacency in $g_1, g_2$ or $g_3$ using any matching to form pairs, we obtain a genome $x$, and Eqs. (2) and (3) give

$$s(x) = 3-(\theta_1+\theta_2+(1-2\theta_1-2\theta_2+2\theta_1\theta_2+\theta_1^2+\theta_2^2)). \tag{4}$$

Figure 2 depicts a surface described by the values of $s(x)$ of the vertices in a Delaunay triangulation of $(\theta_1, \theta_2, \theta_3)$ in barycentric coordinates [11, 12]. It appears from this depiction that "compromise" values of $(\theta_1, \theta_2, \theta_3)$, i.e., around the interior of the triangle, give the largest values, not the smallest, value of $s(x)$.

Indeed, the derivative of the expression in (4) with respect to either $\theta_1$ or $\theta_2$,

$$s'(x) = 1-2\theta_1-2\theta_2, \tag{5}$$

is zero iff $\theta_1 + \theta_2 = 0.5$. The second derivatives are negative, so the surface is convex.

Examining some values of $\max\theta_3$ and $s(x)$ in Table 1, we confirm that the maximum value of $s(x)$ occur for a genome $x$ where $\theta_1 + \theta_2 = 0.5$ and $\theta_3 = 0.25$.

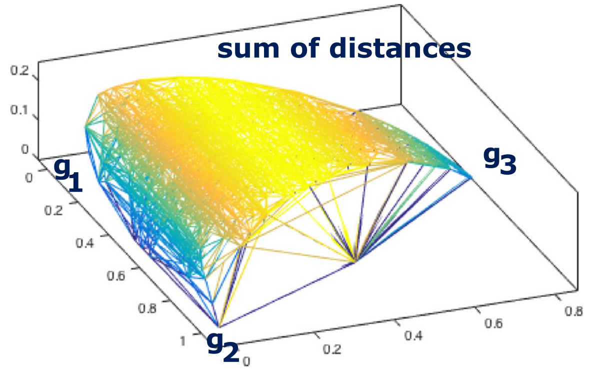

**Fig. 2** Surface of distance sum as a function of $\theta_1, \theta_2, \theta_3$, in barycentric coordinates when $\theta_3$ is maximized as a function of $\theta_1$ and $\theta_2$

By symmetry, we can obtain all of:

$$\begin{aligned} \theta_1 + \theta_2 &= 0.5 \\ \theta_2 + \theta_3 &= 0.5 \\ \theta_3 + \theta_1 &= 0.5. \end{aligned} \tag{6}$$

The unique solution of all three equations is $\theta_1 = \theta_2 = \theta_3 = 0.25$.

Turning to the more general case where $\psi$ and the $\omega_{i,j}$ are not required to be zero, as illustrated in Fig. 3, Eq. (4) becomes

$$s(x) = 3 - (\theta_1 + \theta_2 + \max\theta_3 + \sum_{i \neq j}^{3} \omega_{i,j} + \psi). \tag{7}$$

and Eq. (6) become

$$\begin{aligned} \theta_1 + \theta_2 &= 0.5 - \omega_{1,2} - \omega_{2,3} - \omega_{1,3} - \psi \\ \theta_2 + \theta_3 &= 0.5 - \omega_{1,2} - \omega_{2,3} - \omega_{1,3} - \psi \\ \theta_3 + \theta_1 &= 0.5 - \omega_{1,2} - \omega_{2,3} - \omega_{1,3} - \psi\ . \end{aligned} \tag{8}$$

The unique solution of all three equations is

$$\theta_1 = \theta_2 = \theta_3 = 0.25 - 0.5(\omega_{1,2} + \omega_{2,3} + \omega_{1,3} + \psi) \tag{9}$$

which maximizes $s(x)$ over all maximal samples.

We might imagine that it would be "fairer" to distribute adjacencies among the $\theta$'s in the proportions:

$$\begin{aligned} \theta_1 : \theta_2 : \theta_3 = \frac{1}{2}(\omega_{1,2} + \omega_{1,3}) + \frac{\psi}{3} : \frac{1}{2}(\omega_{1,2} + \omega_{2,3}) \\ + \frac{\psi}{3} : \frac{1}{2}(\omega_{1,3} + \omega_{2,3}) + \frac{\psi}{3}, \end{aligned} \tag{10}$$

where each genome would contribute a number of adjacencies in proportion to the number it has already contributed in $\psi$ and the $\omega$'s. However, this is not a solution for the equations in (8) for general values of $\omega_{1,2}, \omega_{1,3}$ and $\omega_{2,3}$, and upon reflection, there is no reason to consider

**Table 1** Maximizing $\theta_3$ for various combinations of $\theta_1$ and $\theta_2$

| $\theta_1$ | $\theta_2$ | max $\theta_3$ | $s$ |
|---|---|---|---|
| 0.15 | 0.15 | 0.4900 | 2.2100 |
| | 0.20 | 0.4225 | 2.2275 |
| | 0.25 | 0.3600 | 2.2400 |
| | 0.30 | 0.3025 | 2.2475 |
| | 0.35 | 0.2500 | 2.2500 |
| 0.20 | 0.15 | 0.4225 | 2.2275 |
| | 0.20 | 0.3600 | 2.2400 |
| | 0.25 | 0.3025 | 2.2475 |
| | 0.30 | 0.2500 | 2.2500 |
| | 0.35 | 0.2025 | 2.2475 |
| 0.25 | 0.15 | 0.3600 | 2.2400 |
| | 0.20 | 0.3025 | 2.2475 |
| | 0.25 | 0.2500 | 2.2500 |
| | 0.30 | 0.2025 | 2.2475 |
| | 0.35 | 0.1600 | 2.2400 |
| 0.30 | 0.15 | 0.3025 | 2.2475 |
| | 0.20 | 0.2500 | 2.2500 |
| | 0.25 | 0.2025 | 2.2475 |
| | 0.30 | 0.1600 | 2.2400 |
| | 0.35 | 0.1225 | 2.2275 |
| 0.35 | 0.15 | 0.2500 | 2.2500 |
| | 0.20 | 0.2025 | 2.2475 |
| | 0.25 | 0.1600 | 2.2400 |
| | 0.30 | 0.1225 | 2.2275 |
| | 0.35 | 0.0900 | 2.2100 |

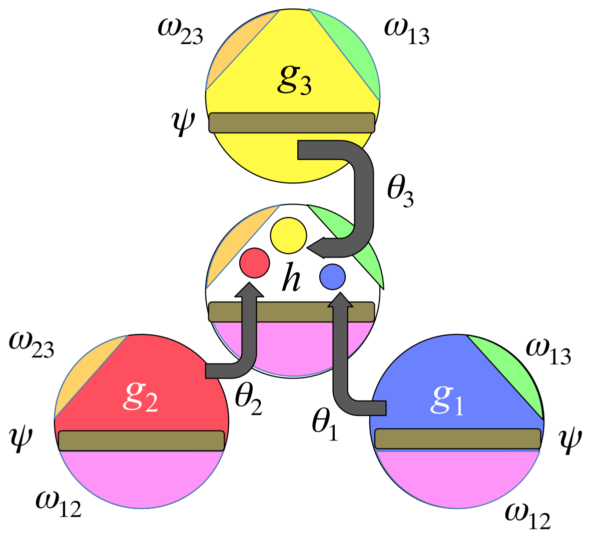

**Fig. 3** Sampling scheme showing variable proportions $\theta_1, \theta_2, \theta_3$, and given two-way intersections $\omega_1, \omega_2, \omega_3$ and three-way intersection $\psi$. All these contributions lower $s$. *White area* in genome *h* represent the randomly completed portion

this a better compromise than an equal division of adjacencies among the three genomes, beyond the unbalances already inherent in the pairwise overlaps.

## Discussion

The breakpoint median minimizes the sum of the breakpoint distance to three given genomes but in doing so foregoes any property of "compromise" among the three, despite this being the original motivation for the median. The anti-median represents a complete emphasis on "compromise" instead of on shortest distances. Somewhat surprisingly, the anti-median actually maximizes the sum of the breakpoint distance to three given genomes, in the process assuring that none of the three input genomes is disproportionately represented, other than through its given overlap with the other two genomes.

Note that the anti-median genomes are constructed to have precise normalized distances from $g_1, g_2$, and $g_3$, in the sense of their limiting behaviour as $n \to \infty$. This behaviour is predicated on the inclusion of all the adjacencies in the two-way and three-way overlap, and the completion of the sampled genome by random matching of unpaired ends. These anti-medians contrast with arbitrary random genomes whose normalized sums of scores to $g_1, g_2$, and $g_3$ approach 3. At the other extreme, they also contrast with the "near medians" [9] completed by maximum matching algorithms, whose scores are less than those of the randomly completed samples constructed here.

## Conclusions

Median constructions form the basis of the steinerization strategy for solving the small phylogeny problem, finding the ancestral genomes to populate the ancestral nodes of a given phylogeny when the genomes at the leaf nodes are known. Each ancestral node in turn is subjected to a median search, based on its three neighbors, and this is iterated until convergence. This constitutes a search for a most parsimonious solution. But if we wish ancestral nodes to reflect all three neighboring nodes (in a binary tree), there is no obstacle in using anti-medians instead of medians, and actually searching for a *least* parsimonious solution, so that compromise becomes the organizing principle in the reconstruction. Exploring this becomes the most important project for future work on this subject.

**Acknowledgements**
Research supported in part by grants from the Natural Sciences and Engineering Research Council of Canada. DS holds the Canada Research Chair in Mathematical Genomics. A preliminary version of this work was presented at the BICOB 2016 conference.

**Declarations**

This article has been published as part of *BMC Bioinformatics* Volume 17 Supplement 18, 2016. Proceedings of X-meeting 2015: 11th International Conference of the AB3C + Brazilian Symposium on Bioinformatics: bioinformatics. The full contents of the supplement are available online https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-17-supplement-18.

**Funding**

Funding for this publication comes from a Natural Sciences and Engineering Research Council of Canada Discovery Grant.

**Availability of data and materials**

Not applicable.

**Authors' contributions**

CAL and DS carried out the research and wrote the paper. AB developed the barycentric coordinates representation and implemented it in visualization package. All authors read and approved the final manuscript.

**Competing interests**

The authors declare that they have no competing interests.

**Consent for publication**

Not applicable.

**Ethics approval and consent to participate**

Not applicable.

Published: 15 December 2016

**References**

1. Sankoff D, Blanchette M. The median problem for breakpoints in comparative genomics In: Jiang T, Lee DT, editors. Computing and Combinatorics, Proceeedings of COCOON 97. Lecture Notes in Computer Science 1276: Springer Verlag; 1997. p. 251–263.
2. Tannier E, Zheng C, Sankoff D. Multichromosomal median and halving pr oblems under different genomic distances. BMC Bioinforma. 2009;10:120.
3. Bryant D. The complexity of the breakpoint median problem. Tech Rep. CRM-2579 Centre de recherches mathématiques; 1998.
4. Pe'er II, Shamir R. The median problems for breakpoints are NP-complete. Electron Colloq Comput Complex. 1998;71(5).
5. Xu AW. A fast exact algorithm for the median of three problem: A graph decomposition approach. J Comput Biol. 2009;16:1369–1381.
6. Haghighi M, Sankoff D. Medians seek the corners, and other conjectures. BMC Bioinforma. 2012;13(S19):S5.
7. Jamshidpey A, Sankoff D. Phase change for the accuracy of the median value in estimating divergence time. BMC Bioinforma. 2013;14(S15):S7.
8. Jamshidpey A, Sankoff D. Sets of medians in the non-geodesic pseudometric space of unsigned genomes with breakpoints. BMC Genomics. 2014;15(S6):S3.
9. Larlee CA, Zheng C, Sankoff D. Near-medians that avoid the corners; a combinatorial probability approach. BMC Genomics. 2014;15(S6):S1.
10. Xu W, Alain B, Sankoff D. Poisson adjacency distributions in genome comparison: multichromosomal, circular, signed and unsigned cases. Bioinformatics. 2008;24:i146–i152.
11. Weisstein EW. Barycentric Coordinates. From MathWorld–A Wolfram Web Resource http://mathworld.wolfram.com/BarycentricCoordinates.html.
12. Coxeter HSM. Barycentric Coordinates, Ch 13.7. In: Introduction to Geometry. 2nd ed. New York: Wiley; 1969. p. 216–221.