Jamshidpey and Sankoff *BMC Bioinformatics* 2013, **14**(Suppl 15):S7
http://www.biomedcentral.com/1471-2105/14/S15/S7

BMC Bioinformatics

PROCEEDINGS Open Access

Phase change for the accuracy of the median value in estimating divergence time

Arash Jamshidpey*, David Sankoff

From Eleventh Annual Research in Computational Molecular Biology (RECOMB) Satellite Workshop on Comparative Genomics
Lyon, France. 17-19 October 2013

Abstract

We prove that for general models of random gene-order evolution of $k \geq 3$ genomes, as the number of genes n goes to ∞, the median value approximates k times the divergence time if the number of rearrangements is less than $cn/4$ for any $c < 1$. For some $c^* \geq 1$, if the number of rearrangements is greater than $c^*n/4$, this approximation does not hold.

Introduction

The iterative improvement of approximate solutions to the Steiner tree problem by optimizing one internal vertex at a time has a substantial history in the "small phylogeny" problem for parsimony-based phylogenetics, both at the sequence level [1] and the gene order level [2]. It has been generalized to iterative local subtree optimization methods such as "tree-window-hill" [3] and "disc covering" [4,5]. Here we focus on the "median problem" for gene order where we estimate the location of a single point (the median) in a metric space given the location of the three or more points connected to the median by an edge of the tree. Given $k \geq 3$ signed gene orders G_1, ..., G_k on a single chromosome or several chromosomes, and a metric d such as breakpoints [6], inversions [7], inversions and translocations [8], or double-cut-and-join [9], find the gene order M such that $\sum_{i=1}^{k} d(G_i, M)$ is minimized.

Although it plays a central role in gene order phylogeny, the median suffers from several liabilities. One is that it is hard to calculate in most metric spaces. Not only is it NP-hard [10], but exhaustive methods are costly for most instances, namely unless $G_1 \ldots, G_k$ are all relatively similar to each other, which we will refer to generically as the *similar genomes condition*. Another problem is that heuristics tend to produce inaccurate results unless a suitable similar genomes condition holds [11]. Still another, is the tendency in some metric spaces to degenerate solutions [12] unless the same conditions prevails.

In this paper we add to this litany of difficulties by showing that as k genomes evolve over time, as modeled by any one of several biologically-motivated random walks, there is a phase change after $n/4$ steps, where n is the number of genes. With $u < n/4$ steps, the sum of the normalized distances $\sum_k d/n$ from each of the genomes to the starting point - the ancestor - converges to ku/n in probability, and this is the median value. When $u > c^*n/4$ steps, for a constant $c^* \geq 1$, the sum of the normalized distances to the median converges in probability to a value less than ku/n, and that the ancestor is no longer the median.

Our proof is inspired by a result of Berestycki and Durrett [13] in showing that the reversal distance between two signed permutations converges in probability to the actual number of steps, after rescaling, if and only if $u < n/2$. The technique is to construct a graph with genes as vertices and edges added between vertices according to how they are affected by transpositions. Properties of the number of components of random Erdös-Renyi graphs can then be invoked to prove the result.

* Correspondence: ajams015@uottawa.ca
Department of Mathematics and Statistics, University of Ottawa, 585 King Edward Avenue, Ottawa, Canada, K1N 6N5

Definitions

We represent a unichromosomal genome by a signed permutation, where the sign indicates whether the gene is "read" from left to right (tail-to-head) or from right to left (head to tail) on the chromosome. Let $S_n^{\pm}$ be the signed symmetric group of order n, i.e. the space of all signed permutations of length n. A *reversal* operation applied to a signed permutation reverses the order, and changes the signs, of one or more adjacent terms in the permutation. A DCJ operation, which can apply not only to signed permutations but to more general genomes containing linear and circular chromosomes, *cuts* the genome in two places and rejoins pairs of the four "loose ends" in one of two possible new ways (one of which may be equivalent to a reversal). We define the reversal and DCJ distances, d_r and dcj, to be the minimum number of reversal and DCJ operations, respectively, needed to transform one genome to another.

The *breakpoint graph* $BP(\Pi, \Pi')$ of two genomes represented by Π and Π' contains vertices for the head and tail of each gene, *black edges* edges defined by the adjoining heads or tails of two adjacent genes in the genome Π and *grey edges* defined by two adjacent genes in the genome Π'. Let $id = I$, the identity permutation, and $BP(\Pi) = BP(\Pi, id)$. It is well-known that

$$dcj(\Pi) = n + 1 - |cBP(\Pi)|. \tag{1}$$

We need to define an orientation for grey (and black) edges of $BP(\Pi)$. We traverse a cycle $c \in cBP(\Pi)$ in a counter-clockwise manner if we start at the left-most vertex of $BP(\Pi)$ (in the usual representation), travel along its unique adjacent black edge and end at the same vertex through its unique adjacent grey edge. Then we say a black edge in c is positively oriented if we move along it from left to right in a counter-clockwise traversal. Otherwise we say it is negatively oriented. Similarly, for the grey edge $(i_t, (i+1)_h)$ we say it is positively oriented if during a counter-clockwise traversal we move along it from i_t to $(i+1)_h$. Otherwise it is negatively oriented. We define the orientation function ξ on the edges of $BP(\Pi)$ to be:

$$\xi(e) = \begin{cases} +1 & \text{if } e \text{ is positively oriented} \\ -1 & \text{if } e \text{ is negatively oriented.} \end{cases} \tag{2}$$

We say the black (grey) edges e, e' are parallel, denoted by $e \,||\, e'$ if $\xi(e) = \xi(e')$. Otherwise we say they are crossing. This is just a reformulation of Hannenhalli and Pevzner's original concept of oriented cycles. An oriented cycle in this definition is a cycle including at least one positively and one negatively oriented black edge. The mechanism by which a reversal affects a genome can easily be seen using the BP graph. Let ρ be a reversal acting on two black edges e, e' in $BP(\Pi)$. If they are in two different cycles we have a merger of the two to construct a new cycle. But if e, e' are in a same cycle, that cycle either splits, if $e \nparallel e'$, or does not split if $e \,||\, e'$.

Limit Behavior of the Median Value

Suppose d^n be a metric on the space of all signed permutations length n. For a set A of these permutations, define

$$g_A^{d,n} : S_n^{\pm} \to \mathbb{N}_0 = \mathbb{N} \cup \{0\}, \tag{3}$$

$$g_A^{d,n}(x) := \sum_{y \in A} d^n(x, y). \tag{4}$$

Then let

$$m^{d,n}(A) := min\{g_A^{d,n}(x) : x \in S_n^{\pm}\}. \tag{5}$$

$m^{d,n}(A)$ is called the median value of A under the metric d^n. A signed permutation which makes $g_A^{d,n}$ minimum is called a median solution of A. Denote by d_r and dcj the reversal and DCJ distances on $S_n^{\pm}$.

Let $X_0 = id$, the identity permutation, and let X_t^n be a stochastic process on $S_n^{\pm}$, where at random Poisson times τ_κ, with rate 1, we choose two elements of $X_{\tau_\kappa}^n$, namely i, j and let $\rho(i, j)$ operate on $X_{\tau_\kappa}^n$, that is

$$X_{\tau_\kappa^+}^n = X_{\tau_\kappa}^n \, o \, \rho(i, j), \tag{6}$$

where $\rho(i, j)$ is the reversal acting on i and j. We call X_t^n a reversal random walk (r.w.) on. $S_n^{\pm}$. Suppose $X_t^{1,n}, \ldots, X_t^{k,n}$ be k independent reversal r.w. all starting at the identity element, *id*. Define

$$A_t^{(n)} := \{X_t^{1,n}, \ldots, X_t^{k,n}\} \tag{7}$$

and

$$\varepsilon_t^{d,n} := g_{A_t}^{d,n}(id) - m^{d,n}(A_t). \tag{8}$$

We investigate the time up to which the median value of $X_t^{1,n}, \ldots, X_t^{k,n}$, namely $m^{d,n}(A_t)$, remains a good estimator for the total divergence time, kt, as well as to the total distance of points in A_t to id, namely $g_{A_t}^{d,n}(id)$. To answer this question we use the fact that the speed of escape of the r.w. up to some particular time, is the same from any point of the space and is close to 1, the maximum value. Berestycki and Durrett studied speed of transposition and reversal random walks with the related edit distances while in the latter they used "approximate reversal distance" instead of reversal itself, ignoring the effect of hurdles and fortresses. This turns out to be the same as DCJ distance on single chromosomes. We have

$$d_r(\pi, I) = n + 1 - c(\pi) + h(\pi) + \tilde{f}(\pi) \tag{9}$$

while

$$dcj(\pi, I) = n + 1 - c(\pi), \tag{10}$$

where $h(\pi)$ and $\tilde{f}(\pi)$ are the number of hurdles and fortresses, respectively.

Although Berestycki and Durrett only proved their theorem for the random transposition r.w. on S_n, they suggested that same method should carry over to reversal r.w. The following proposition is proved in [13] for *approximate* reversal distance (i.e., DCJ distance).

In this result and in the ensuing discussion a_n is an arbitrary sequence such that $a_n \to \infty$ as $n \to 0$. When it is unambiguous we drop n from $A_t^{(n)}$ and X_t^n.

Propostition 1 *[Berestycki-Durrett] Let c be fixed and let Xt be a reversal r.w. on $S_n^{\pm}$ starting at id. Then*

$$dcj(id, X_{cn/2}) = (1 - f(c))n + w(n), \tag{11}$$

where

$$f(c) := \frac{1}{c}\sum_{k=1}^{\infty}\frac{k^{k-2}}{k!}(ce^{-c})^k \tag{12}$$

and $\frac{w(n)}{a_n\sqrt{n}} \to 0$ in probability.

Remark 1 *The function 1 - f is linear for c <1, f (c) = 1 - c/2, and sublinear for c >1, 1 - f (c) < c/2 This means that for c ≤ 1*

$$dcj(id, X_{cn/2}) - \frac{cn}{2} = w(n) \tag{13}$$

and r.w. travels on an approximate geodesic (or parsimonious path) asymptotically almost surely. f is the function counting the number of tree components of an Erdös-Renyi random graph with n vertices for which the probability of having each edge is $\frac{c}{n}$, denoted by G(c, n). See Theorem 12 in [14], *Chapter V.*

We extend the above theorem for the bonafide reversal distance. To do so we need to estimate the number of hurdles of $X_{\frac{cn}{2}}$. Recall that an oriented cycle in a breakpoint graph is a cycle including an orientation edge, that is a grey edge with two black adjacency edges *e*, *e'*, where a reversal involving *e* and *e'* splits the cycle [15]. As we discussed this is equivalent to saying $e \nparallel e'$. It is not difficult to show

Lemma 1 *Let C ∈ cBP(π), then C is oriented if and only if there exists exactly two equivalence classes of black edges, that is there exist at least two black edges with different signs.*

Then

Theorem 1 *Let c >0 be fixed and let X_t be a reversal r.w.starting at id. Define $h_t := h(X_t)$ to be the number of hurdles in BP (X_t). Then*

$$\frac{h_{cn/2}}{a_n\sqrt{n}} \to 0 \text{ in probability.} \tag{14}$$

Proof. Cycles of the BP that have never been involved in a fragmentation event must be oriented, since the two rejoined black edges resulting from an inversion-induced merger of cycles cannot be parallel.

Therefore we need only to count the number of edges that have been involved in a fragmentation event. To do so we apply the method of counting cycles in [13], Theorem 3. Hurdles occur only in those cycles with length more than one that have been involved in a fragmentation up to time $\frac{cn}{2}$. We call such cycles fragmented cycles. The number of fragmented cycles with length more than $\sqrt{n}$ is always less than $\sqrt{n}$. But to count all fragmented cycles in $X_{\frac{cn}{2}}$ with size less than $\sqrt{n}$ we need to find an upper bound for the rate of a fragmentation up to time $\frac{cn}{2}$. Since a fragmentation occurs when two black edges in one cycle are chosen, to fragment a cycle in BP, for any chosen black edge *e* we only can pick another black edge *e'* in the same cycle whose graph distance in the breakpoint graph is less than $2\sqrt{n}$. (The coefficient 2 arises from the fact that the cycles are alternating in BP.)

Thus the rate of fragmentation at an arbitrary time t is not more than $\frac{n}{n}\cdot\frac{2(\sqrt{n})}{n} = \frac{2}{\sqrt{n}}$. Integrating up to time t, this gives us the expected number of fragmented cycles at time t is $\frac{2}{\sqrt{n}}$. For $t = \frac{cn}{2}$ this expectation is $c\sqrt{n}$. Now, dividing by $a_n\sqrt{n}$, the result follows from Chebyshev's inequality and the fact that hurdles only occurs in fragmented cycles. ■

Theorem 2 *let c >0 be fixed and let X_t be a reversal r.w. on $S_n^{\pm}$ starting at id and let $d_r := d_r^{(n)}$ denote the reversal distance on $S_n^{\pm}$. Then*

$$d_r(id, X_{cn/2}) = (1 - f(c))n + w'(n) \tag{15}$$

where f is the same function as in the statement of Proposition 1 and w' (n) is a function with $\frac{w'(n)}{a_n\sqrt{n}} \to 0$ in probability.

Proof. Since $d_r(\Pi) = dcj(\Pi) + h(\Pi) + f^*(\Pi)$ by the proposition we have $d_r(X_{cn/2}) = (1 - f(c))n + w(n) + h_{cn/2} + f^*(X_{cn/2})$. But

$$\frac{w'(n)}{a_n\sqrt{n}} := \frac{w(n) + h_{cn/2} + \tilde{f}(X_{cn/2})}{a_n\sqrt{n}} \to 0 \tag{16}$$

in probability, by the convergence of $\frac{w(n)}{a_n\sqrt{n}}$ and $\frac{h_{cn/2}}{a_n\sqrt{n}}$ in Proposition 1 and Theorem 2 and $\tilde{f}(X_{cn/2}) \le 1$.■

Theorem 3 *Let $X_t^{1,n}, \ldots, X_t^{k,n}$ be k independent reversal r.w in $S_n^{\pm}$ starting at id. Suppose either*

a) d := dcj dcj distance

or

b) $d := d_r^{(n)}$ reversal distance.

Then for c $<\frac{1}{4}$ *we have $\frac{\varepsilon_{cn}^{d,n}}{a_n\sqrt{n}} \to 0$ in probability.*

Proof. We prove the theorem only for d_r. The proof of the DCJ case is similar. For all $i, j \in \{1, ..., k\}$ and for a median solution x of $A_t^{(n)}$

$$d_r^{(n)}(X_t^{i,n}, X_t^{j,n}) \leq d_r^{(n)}(x, X_t^{i,n}) + d_r^{(n)}(x, X_t^{j,n}). \quad (17)$$

Therefore,

$$\sum_{i \neq j} d_r^{(n)}(X_t^{i,n}, X_t^{j,n}) \leq \sum_{i \neq j} (d_r^{(n)}(x, X_t^{i,n}) + d_r^{(n)}(x, X_t^{j,n})). \quad (18)$$

We conclude

$$\sum d_r^n(X_t^{i,n}, X_t^{j,n}) \leq (k-1)m^{d,n}(A_t^{(n)}) \leq (k-1)g_{A_t^{(n)}}(id). \quad (19)$$

Let $c \leq \frac{1}{4}$. Then by Theorem 2 we have for all $i, j\ i \neq j$

$$d_r^{(n)}(X_{cn}^{i,n}, X_{cn}^{j,n}) = 2cn - w(n) \quad (20)$$

and

$$d_r^{(n)}(id, X_{cn}^{i,n}) = cn - w(n) \quad (21)$$

where $\frac{w(n)}{(a_n\sqrt{n})} \to 0$ in probability. Thus

$$\binom{k}{2}(2cn - w(n)) \leq (k-1)m^{d,n}(A_{cn}^{(n)}) \leq (k-1)k(cn - w(n)). \quad (22)$$

Then

$$|m^{d,n}(A_{cn}^{(n)}) - kcn| \leq k'w(n) \quad (23)$$

for a constant k'. Also $|g_{A_{cn}^{(n)}}(id) - kcn| \leq kw(n)$. Therefore, there exists a constant k^* such that

$$|m^{d,n}(A_{cn}^{(n)}) - g_{A_{cn}^{(n)}}^{d,n}(id)| \leq k^*w(n). \quad (24)$$

This implies

$$\frac{\varepsilon_{cn}}{a_n\sqrt{n}} = \frac{m^{d,n}(A_{cn}^{(n)}) - g_{A_{cn}^{(n)}}^{d,n}(id)}{a_n\sqrt{n}} \to 0 \text{ in probability.} \quad (25)$$

This proves the theorem. ■

Remark 2 *The statement of the theorem suggests ignoring the error of order $o(a_n\sqrt{n})$ for $a_n \to \infty$. id remains as the median of leaves of k independent stochastic processes $X_t^{1,n}, \ldots, X_t^{k,n}$ up to time $\frac{n}{4}$ asymptotically almost surely.*

Theorem 4 *Let $c \leq \frac{1}{4}$ be fixed. Suppose d is either DCJ or reversal distance. Then by the hypothesis of Theorem 3*

$$\frac{kcn - m^{d,n}(A_{cn})}{a_n\sqrt{n}} \to 0 \text{ in probability as } n \to \infty. \quad (26)$$

Proof. This follows directly from the fact that

$$\frac{kcn - g_{A_{cn}}^{d,n}(id)}{a_n\sqrt{n}} \to 0 \quad (27)$$

in probability. ■

Now, it is natural to ask whether the statement of Theorem 4 also holds for some time after $\frac{n}{4}$. In other words, is the median value kcn a fair estimator for the total time of divergence? We conjecture not, that the property is lost after time $\frac{n}{4}$, but for now can only prove a weaker upper bound for this time.

Theorem 5 *Let $c > \frac{1}{2}$ be fixed. Suppose d is either DCJ or reversal distance. Then by the same hypothesis as in Theorem 3*

$$\frac{kcn - m^{d,n}(A_{cn})}{n} \to \alpha_c \quad (28)$$

where

$$\alpha_c := k(1 - f(2c)) \quad (29)$$

is strictly positive for $c > \frac{1}{2}$

Remark 3 *This theorem shows after time $\frac{n}{2}$ the error is of order n and so the median value is not a good estimate of k times the divergence time.*

Proof.

$$kcn - m^{d,n}(A_{cn}) \geq kcn - g_{A_{cn}}^{d,n}(id) = k(1 - f(2c))n + w(n), \quad (30)$$

where $\frac{w(n)}{a_n\sqrt{n}} \to 0$ in probability. Dividing by n, the result follows. ■

In fact, since $f(c)$, $c > 0$ is decreasing and for $c < 1$, $f(c) = 1 - \frac{c}{2}$, it is easy to see that in the case $k = 3$, for $c > 0.75$, $\varepsilon_{cn}^{d,n}$ is of order $\beta_c^d n$ for some $\beta_c^d \geq 0$.

Theorem 6 *Let $k = 3$ and d be either dcj or dr. Consider the same hypothesis in Theorem 3. Assume c^* be solution of*

$$f(\frac{x}{2}) = \frac{1}{3}. \quad (31)$$

Then for all $c > c^*$ there exists β_c^d such that

$$\varepsilon_{\frac{cn}{4}}^{d,n} = o(\beta_c^d n). \quad (32)$$

Proof.

$$m^{d,n}(A_{\frac{cn}{4}}) \leq d(X_{\frac{cn}{4}}^{1,n}, X_{\frac{cn}{4}}^{2,n}) + d(X_{\frac{cn}{4}}^{1,n}, X_{\frac{cn}{4}}^{3,n}). \quad (33)$$

Computing $d(X_{\frac{cn}{4}}^{1,n}, X_{\frac{cn}{4}}^{i,n})$ for $i = 2, 3$ is the same as $d(id, X_{\frac{cn}{2}}^{1,n})$. This is true since the Cayley graph of $S_n^{\pm}$ w.r.t. reversals is symmetric and regular and so $P(X_0 = id, X_{\frac{cn}{4}} = \Pi) = P(X_0 = \Pi, X_{\frac{cn}{4}} = id)$. But therefore by symmetry of the Cayley graph we can just consider $d(id, X_{\frac{cn}{2}}^{1,n})$. Hence,

$$m^{d,n}(A_{\frac{cn}{4}}) \leq 2(1 - f(c))n + 2w(n). \quad (34)$$

Let $x > 0$ be so that

$$0 < -2(1 - f(x)) + 3(1 - f(\frac{x}{2})). \tag{35}$$

This means

$$g^{d,n}_{A_{\frac{cn}{4}}}(id) > m^{d,n}(A_{\frac{cn}{4}}). \tag{36}$$

So it suffices to prove above inequality for $x = c > c^*$. Since $f(x) > 0$ for all $x > 0$

$$1 + 2f(x) - 3f(\frac{x}{2}) > 1 - 3f(\frac{x}{2}) \tag{37}$$

in which the right hand side is strictly increasing, Therefore for all $c \geq c^*$

$$1 + 2f(c) - 3f(\frac{c}{2}) > 1 - 3f(\frac{c^*}{2}) = 0. \tag{38}$$

This proves the statement. ■

Now, we would like to measure the volume of that part of the space $S_n^{\pm}$ for which median does well, compared with the whole space. The ratio of the two converges to 0 as n goes to ∞, showing that the median is only useful in a highly restricted region of the space.. The following theorem is entailed by a theorem in [16]. Let $c_n = c_n(\Pi)$ be the number of cycles in the BP graph of a random $\Pi \in S_n^{\pm}$. Let d_n be a distance (metric) on $S_n^{\pm}$. Define

$$B^d_{cn} = B^{d,n}_{cn} := \{\Pi \in S_n^{\pm},\ d(\Pi,\ id) \leq cn\} \tag{39}$$

to be the ball of radius cn in $S_n^{\pm}$.

Theorem 7 *Let* $0 < c < 1$ *be fixed. Then*

$$a)\gamma_n = \frac{|B^{dcj}_{cn}|}{|S_n^{\pm}|} \to 0 \ \ as \ \ n \to \infty, \tag{40}$$

$$b)\gamma'_n = \frac{|B^{d_r}_{cn}|}{|S_n^{\pm}|} \to 0 \ \ as \ \ n \to \infty. \tag{41}$$

Proof.

a)

$$For\ all\ \Pi \in B^{dcj}_{cn}\ ,\ |cBP(\Pi)| \ \geq (1 - c)n. \tag{42}$$

Suppose γ_n does not converge to 0. Therefore there exists a subsequence $\{n_i\}_{i\in\mathbb{N}}$ such that $\gamma_{n_i} \geq \varepsilon$ for a constant $\varepsilon > 0$. This implies

$$E(c_{n_i}) \geq \varepsilon(1 - c)n_i. \tag{43}$$

But by Theorem 2.2 in [16], we have

$$\frac{E(c_{n_i})}{n_i} \to 0 \ \ as \ \ n_i \to \infty. \tag{44}$$

That is in contradiction with the above inequality since

$$\frac{\varepsilon(1 - c)n_i}{n_i} \to \varepsilon(1 - c) > 0. \tag{45}$$

b) For the second part it suffices to observe that for all $\Pi \in S_n^{\pm}$ we have

$$d_r(\Pi) \geq dcj(\Pi). \tag{46}$$

Therefore

$$B^{d_r}(\Pi) \subset B^{dcj}(\Pi) \tag{47}$$

and the result follows part (a) since

$$\gamma'_n \leq \gamma_n \to 0 \ as \ n \to \infty. \tag{48}$$

■

Conclusion

We have shown that the median value for DCJ and for reversal distance for a reversal r.w.has good limiting properties if the number of steps remains below $cn/4$, for any $c < 1$, but for some value $c > 1$, more than this number of steps destroys these limiting properties. The critical value may indeed be $c = 1$, but for now we can only show that for $c > 3$ (and $c > 2$) the median value is no longer a good estimator of the distance between the *id* and the current position of the r.w. (and k times the divergence time, respectively).

Note that a simulation strategy to estimate c is not available because of the hardness of calculating the median. As n increases even to moderate values all exact methods require prohibitive computing time.

These results imply that the steinerization strategy for the small phylogeny problem may lead to poor estimates of the interior nodes of a phylogeny unless the taxon sampling is sufficient to assure that a "similar genomes condition" holds for every k-tuple of genomes used in the course of of the iterative optimization search. This can be monitored prior to each step in the iterative optimization of the phylogeny through successive application of the median method.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

AJ and DS planned the study, carried out the research and wrote the article.

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. Thanks to Armin Jamshidpey and Leili Rafiee Sevyeri for help in preparation of the manuscript.

Declarations
Publication of this article was supported by the Canada Research Chair in Mathematical Genomics.
This article has been published as part of *BMC Bioinformatics* Volume 14 Supplement 15, 2013: Proceedings from the Eleventh 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/bmcbioinformatics/supplements/14/S15.

Published: 15 October 2013

References

1. Sankoff D, Cedergren RJ, Lapalme G: **Frequency of insertion/deletion, transversion and transition in the evolution of 5S ribosomal RNA.** *Journal of Molecular Evolution* 1976, **7**:133-149.
2. Blanchette M, Bourque G, Sankoff D: **Breakpoint phylogenies.** In *Genome Informatics.* Tokyo: Universal Academy Press;S. Miyano & T. Takagi 1997:25-34.
3. Sankoff D, Abel Y, Hein J: **A Tree - A Window - A Hill; Generalization of nearest-neighbour inter-change in phylogenetic optimisation.** *Journal of Classification* 1994, **11**:209-232.
4. Huson D, Nettles S, Warnow T: **Disk-covering, a fast-converging method for phylogenetic tree reconstruction.** *Journal of Computational Biology* 1999, **6**:369-386.
5. Tang J, Moret B: **Scaling up accurate phylogenetic reconstruction from gene-order data.** *Bioinformatics* 2003, **19**:i305-i312.
6. Sankoff D, Blanchette M: **The median problem for breakpoints in comparative genomics.** In *Proceedings of Computing and Combinatorics (COCOON)* T. Jiang and D.T. Lee 1997, **1276**:251-263, Lecture Notes in Computer Science.
7. Sankoff D, Sundaram G, Kececioglu J: **Steiner points in the space of genome rearrangements.** *International Journal of the Foundations of Computer Science* 1996, **7**:1-9.
8. Bourque G, Pevzner PA: **Genome-scale evolution: Reconstructing gene orders in the ancestral species.** *Genome Research* 2002, **12**:26-36.
9. Zhang M, Arndt W, Tang J: **An exact solver for the DCJ median problem.** *Pacific Symposium on Biocomputing* 2009, 138-149.
10. Tannier E, Zheng C, Sankoff D: **Multichromosomal median and halving problems under different genomic distances.** *BMC Bioinformatics* 2009, **10**:120.
11. Zheng C, Sankoff D: **On the Pathgroups approach to rapid small phylogeny.** *BMC Bioinformatics* 2011, **12**:S4.
12. Haghighi M, Sankoff D: **Medians seek the corners, and other conjectures.** *BMC Bioinformatics* 2012, **13(S19)**:S5.
13. Berestycki N, Durrett R: **A phase transition in the random transposition random walk.** *Probability Theory and Related Fields* 2006, **136**:203-233.
14. Bollobás B: *Random Graphs.* 2 edition. Cambridge University Press; 2001.
15. Hannenhalli S, Pevzner PA: **Transforming cabbage into turnip: Polynomial algorithm for sorting signed permutations by reversals.** *Journal of the ACM* 1999, **46**:1-27.
16. Székely LA, Yang Y: **On the expectation and variance of the reversal distance.** *Acta Univ. Sapientiae, Mathematica* 2009, **1**:5-20.

doi:10.1186/1471-2105-14-S15-S7
Cite this article as: Jamshidpey and Sankoff: **Phase change for the accuracy of the median value in estimating divergence time.** *BMC Bioinformatics* 2013 **14**(Suppl 15):S7.