The complexity of the breakpoint median problem

David Bryant

Communicating author:

David Bryant

Centre de recherches mathematiques

Université de Montréal

C.P. 6128, Succursale Centre Ville

Montreal, Quebec H3C 3J7

Canada

email: `bryant@CRM.UMontreal.ca`

phone: (514) 343-7501. fax: (514) 343-2254.

Keywords: Comparative genomics–breakpoint distance–breakpoint median problem–complexity

Abstract

The breakpoint median problems arise in the problem of determining phylogenetic history from comparative genome data. We prove that the breakpoint median problems, and a number of related and constrained versions, are all NP-hard.

1 Introduction

The growing number of complete genome maps enables the extraction of phylogenetic information from global rearrangements over the whole genome, rather than just local nucleotide or amino acid patterns. As genomes evolve, genes are inserted or deleted, segments of the genome are reversed, or removed and re-inserted at a new position. This leads to different genomes having homologous genes arranged in different orders on their gene maps, and it is these orderings that become the input data for comparative genomics.

Initially, the divergence between two genomes was estimated by calculating a minimum edit distance, such as a weighted combination of reversals and transpositions (Blachette et al. 1996, Gu et al. 1997, Sankoff 1992, Sankoff et al. 1992). There are several problems with this approach: the calculation of edit distances is generally NP-hard (Caprara 1997); the huge number of optimal solutions possible introduces an undesirable degree of ambiguity; and the minimal edit distance tends to grossly underestimate the actual amount of divergence in simulations. These difficulties led Sankoff and Blanchette to study the breakpoint distance between two genomes (Sankoff and Blanchette 1997). Intuitively, a breakpoint is a pair of genes that are adjacent in one genome but not in the other, and the breakpoint distance is the number of these breakpoints. (We give a rigourous definition in Section 2.)

Suppose that T is an unrooted tree with leaves labelled by genomes on a common gene set $\mathcal{G}$. The fixed topology Steiner breakpoint problem is to determine genome orders for the internal vertices so that the sum of the breakpoint distances over edges in the tree is minimum. If T has only one internal vertex then we obtain an analogue of the multiple sequence alignment problem (Sankoff and Blanchette 1998), the *breakpoint median problem*. It was shown in (Sankoff and Blanchette 1997) that the breakpoint median problem can be seen as a special case of the Travelling Salesman Problem (TSP). While the TSP is NP-hard, it can be readily solved for quite large instances (see, e.g. Johnson and McGeoch 1997).

The actual computational complexity of the Steiner breakpoint problem and the Breakpoint median problem was first established by Pe'er and Shamir (Pe'er and Shamir 1998). They prove NP-hardness using a reduction from the Hamiltonian cycle problem, via a Hamiltonian matching consensus problem. In this note we give a shorter, more direct proof of the NP-hardness of the breakpoint median problem, using a reduction from the directed Hamiltonian cycle problem. The new proof helps reveal aspects of the breakpoint median problem that make it NP-hard, and leads directly to NP-hardness proofs of related and constrained versions of the problem.

In Section 2 we outline the mathematical representation of gene orders, define the breakpoint distance and breakpoint median problem, and prove a number of important properties of median genomes. In Section 3 we prove that the Breakpoint median problem is NP-hard for signed genomes, even when we constrain the median genomes to include only adjacencies present in one or more of the input genomes or if all input genomes have only positive signs. As a corollary, we prove that determining whether a given median genome is unique is also NP-hard, as is the problem of determining adjacencies common to all median genomes. In Section 4 we extend these results to the case of unsigned genomes.

2 Genome order data

2.1 Genomes, adjacencies, and breakpoints

Let A be a genome with gene set $\mathcal{G} = \mathcal{G}(A)$. If we have no information on the strandedness, or direction of transcription, of each gene on the genome then we say that A is an **unsigned genome**. If A is circular, we can represent it as a Hamiltonian cycle $\langle a_1, a_2, \ldots, a_n, a_1\rangle$ of the complete undirected graph with vertex set $\mathcal{G}$. An unordered pair $\{g, h\}$ is an **adjacency** of A if $\{g, h\}$ is an edge in the corresponding Hamiltonian cycle. The set $Adj(A)$ of adjacencies of A is thus given by

$$Adj(A) := \{\{a_i, a_{i+1}\} : i = 1, \ldots, n-1\} \cup \{\{a_n, a_1\}\} \tag{1}$$

A linear genome A can be represented as a Hamiltonian path in the complete undirected graph with vertex set $\mathcal{G}$. The set $Adj(A)$ of adjacencies of a linear genome is given by

$$Adj(A) := \{\{a_i, a_{i+1}\} : i = 1, \ldots, n-1\}. \tag{2}$$

If A and B are two circular or linear genomes on the same gene set $\mathcal{G}$ then the unordered pairs in $Adj(A) - Adj(B)$ are called the **breakpoints** of A with respect to B. The **breakpoint distance** between A and B is defined

$$d(A, B) = |Adj(A) - Adj(B)| \tag{3}$$

which is clearly symmetric.

We modify the notion of breakpoints and breakpoint distance when we are given information about the directionality of the genes in the genome. The genomes are signed to indicate polarity, and adjacency is defined in terms of ordered pairs. A signed circular genome on gene set $\mathcal{G}$ can be represented as a cycle $\langle a_1, a_2, \ldots, a_n, a_1\rangle$ in the complete directed graph with vertex set $\mathcal{G}^{\pm} = \{g : |g| \in \mathcal{G}\}$ that passes through exactly one of $-g, g$ for each $g \in \mathcal{G}$. An ordered pair (x, y) is an **adjacency** of A if either (x, y) or $(-y, -x)$ is an edge in the cycle. The set of adjacencies of A is denoted $Adj(A)$. Note that $|Adj(A)| = 2|\mathcal{G}|$. The breakpoint distance between two circular signed genomes A, B on the same gene set is

$$d(A, B) = \tfrac{1}{2}|Adj(A) - Adj(B)|. \tag{4}$$

A signed *linear* genome A can be represented as a path that passes through exactly one of $-g, g$ for each $g \in \mathcal{G}$. The set of adjacencies of $A = \langle a_1, a_2, \ldots, a_n\rangle$ is given by

$$Adj(A) = \{(a_i, a_{i+1}) : i = 1, \ldots, n-1\} \cup \{(-a_{i+1}, -a_i) : i = 1, \ldots, n-1\} \tag{5}$$

and the breakpoint distance is given by $d(A, B) = \frac{1}{2}|Adj(A) - Adj(B)|$.

2.2 The breakpoint median problem

Given three genomes A, B, C of the same type on the same gene set $\mathcal{G}$, we wish to find a genome S on $\mathcal{G}$ that minimizes $\Psi(S) := d(A, S) + d(B, S) + d(C, S)$. Such a genome is called a **median genome** for A, B, C. We put

$$med(A, B, C) = \min_S\{\Psi(S)\} \tag{6}$$

and let

$$MED(A, B, C) = \{S : \Psi(S) = med(A, B, C)\} \tag{7}$$

denote the set of median genomes for A, B, C.

Sankoff and Blanchette (1997) provide a simple reduction from the breakpoint median problem to the traveling salesman problem (TSP). Given three unsigned circular genomes A, B, C on gene set $\mathcal{G}$, the weight

$w(x,y)$ of an unordered pair of genes x, y is defined

$$w(x,y) = |\{X \in \{A,B,C\} : \{x,y\} \in Adj(X)\}|. \tag{8}$$

Then $\Psi(S) = \sum_{\{x,y\} \in Adj(S)} (3 - w(x,y))$ and determining S that minimizes $\Psi(S)$ is equivalent to determining a tour of minimum length with respect to distance matrix δ with $\delta_{x,y} = 3 - w(x,y)$. In the same way, the problem of determining the breakpoint median of linear unsigned genomes reduces to the non-cyclic TSP problem.

Note that the set $MED(A,B,C)$ can be exponentially large: if $Adj(A)$, $Adj(B)$, and $Adj(C)$ are pairwise disjoint then it can be easily shown that $S \in MED(A,B,C)$ if and only if $Adj(S) \subseteq Adj(A) \cup Adj(B) \cup Adj(C)$, giving a possibly exponential number of median genomes. It is therefore desirable to compute adjacencies common to all median genomes, the *unambiguously reconstructed segments* (Blanchette and Sankoff 1998). We prove that determining whether an adjacency belongs to all median genomes is NP-hard (Corollary 8 and Theorem 10).

At the other extreme, if there are adjacencies common to A, B and C then these can be assumed to be in a median genome.

Lemma 1 *1. If A, B, C are unsigned circular (or linear) genomes on the same gene set, then there is $S \in MED(A,B,C)$ such that $Adj(A) \cap Adj(B) \cap Adj(C) \subseteq Adj(S)$. However, there can also be $S \in MED(A,B,C)$ such that $Adj(A) \cap Adj(B) \cap Adj(C) \not\subseteq Adj(S)$.*

2. If A, B, C are signed circular (or linear) genomes and $S \in MED(A,B,C)$ then $Adj(A) \cap Adj(B) \cap Adj(C) \subseteq Adj(S)$.

Proof
Suppose that $X = \langle x_1, x_2, \ldots, x_n, x_1 \rangle$ is a breakpoint median for unsigned circular genomes A, B, C and $\{x_i, x_j\}$ is a pair in $Adj(A) \cap Adj(B) \cap Adj(C) - Adj(X)$. Put $Y = Adj(X) \cup \{\{x_i, x_j\}\}$. If $w(x_{i-1}, x_i) \leq w(x_i, x_{i+1})$ then remove $\{x_{i-1}, x_i\}$ from Y, otherwise remove $\{x_i, x_{i+1}\}$. Likewise, if $w(x_{j-1}, x_j) \leq w(x_j, x_{j+1})$ then remove $\{x_{j-1}, x_j\}$ from Y, otherwise remove $\{x_j, x_{j+1}\}$.

Consider the subgraph formed with edges Y. There are two possibilities: either the graph is a single (Hamiltonian) path, or the graph contains a single cycle and a single path. In the first case we can add one adjacency to give a genome X' with $\Psi(X') < \Psi(X)$, a contradiction. In the second case there must be an edge $\{x, x'\}$ on the cycle such that $\{x, x'\} \notin Adj(A) \cap Adj(B) \cap Adj(C)$. If y, y' are the endpoints of the path, then removing adjacency $\{x, x'\}$ and adding adjacencies $\{y, x\}$ and $\{y', x'\}$ gives a genome X' with $\Psi(X') \leq \Psi(X)$ and one more weight three adjacency. Repeating the process gives the genome S required.

If A, B, C, and X are the genomes given by

$$A = \langle 1,2,3,4,5,6,9,8,7,11,10,12,13,14,1 \rangle \tag{9}$$
$$B = \langle 1,5,2,3,4,7,6,9,8,11,12,13,10,14,1 \rangle \tag{10}$$
$$C = \langle 1,2,3,5,4,8,7,6,9,10,11,12,13,14,1 \rangle \tag{11}$$
$$X = \langle 1,2,3,4,5,6,7,8,9,10,11,12,13,14,1 \rangle \tag{12}$$

then $X \in MED(A,B,C)$ but $\{6,9\} \in Adj(A) \cap Adj(B) \cap Adj(C) - Adj(X)$.

Now suppose that X is a breakpoint median for signed circular genomes A, B, C and $(g,h) \in Adj(A) \cap Adj(B) \cap Adj(C) - Adj(X)$. Cut X after g and before h, and add the edge (g,h). This way we obtain a cycle on a subset of the gene set and a single segment containing the deleted genes. The cycle must contain an adjacency (k,k') of weight $w(k,k') \leq 2$. Insert the deleted segment between k and k' to obtain a genome X' such that $\Psi(X') < \Psi(X)$, contradicting $X \in MED(A,B,C)$. Thus $X \in MED(A,B,C)$ implies $Adj(A) \cap Adj(B) \cap Adj(C) \subseteq Adj(X)$.

The linear case for signed and unsigned genomes can be proved in the same manner. □

The inclusion of weight three edges is fundamental to the reduction of the signed problem to the unsigned problem. We exploit a doubling-up technique introduced for similar problems (Hannenhalli and Pevzner 1995, Sankoff and Blanchette 1997).

Lemma 2 *Given a signed genome A on gene set $\mathcal{G}$, let $f(A)$ be the unsigned genome on gene set $\mathcal{G}^{\pm} = \{g : |g| \in \mathcal{G}\}$ with adjacency set*

$$Adj(f(A)) = \{\{g, -g\} : g \in \mathcal{G}\} \cup \{\{g, -h\} : \{g, h\} \in Adj(A)\}. \tag{13}$$

Then for three signed genomes A, B, C on $\mathcal{G}$ we have $med(A, B, C) = med(f(A), f(B), f(C))$. and $S \in MED(A, B, C)$ if and only if $f(S) \in MED(f(A), f(B), f(C))$.

Proof
Given any two signed genomes A and B on gene set $\mathcal{G}$ we have

$$\begin{aligned} Adj(f(A)) - Adj(f(B)) &= \{\{g, -h\} : \{g, h\} \in Adj(A)\} - \{\{g, -h\} : \{g, h\} \in Adj(B)\} && (14)\\ &= \{\{g, -h\} : \{g, h\} \in Adj(A) - Adj(B)\} && (15)\end{aligned}$$

so $d(f(A), f(B)) = d(A, B)$.

Suppose that $X \in MED(A, B, C)$. Then

$$\begin{aligned} med(f(A), f(B), f(C)) &\leq d(f(A), f(X)) + d(f(B), f(X)) + d(f(C), f(X)) && (16)\\ &= d(A, X) + d(B, X) + d(C, X) && (17)\\ &= med(A, B, C) && (18)\end{aligned}$$

By Lemma 1 there is $Y \in MED(f(A), f(B), f(C))$ such that $\{\{g, -g\} : g \in \mathcal{G}\} \subseteq Adj(Y)$. Hence there is Z such that $f(Y) = Z$ and

$$\begin{aligned} med(A, B, C) &\leq d(A, Z) + d(B, Z) + d(C, Z) && (19)\\ &= d(f(A), f(Z)) + d(f(B), f(Z)) + d(f(C), f(Z)) && (20)\\ &= med(f(A), f(B), f(C)). && (21)\end{aligned}$$

The result follows. □

Suprisingly, breakpoint median genomes can fail a complementary, apparently intuitive, inclusion property. Given three signed or unsigned genomes A, B, C we do not always have $X \in MED(A, B, C)$ such that

$$Adj(X) \subseteq Adj(A) \cup Adj(B) \cup Adj(C). \tag{22}$$

For example, if A, B, C, are the signed genomes given by

$$\begin{aligned} A &= \langle 1, 6, 7, 8, 5, 2, 3, 4, 9, 1\rangle && (23)\\ B &= \langle 1, 2, 6, 7, 5, 3, 4, 8, 9, 1\rangle && (24)\\ C &= \langle 1, 2, 3, 6, 5, 4, 7, 8, 9, 1\rangle && (25)\\ & && (26)\end{aligned}$$

then $MED(A, B, C) = \{\langle 1, 2, 3, 4, 5, 6, 7, 8, 9, 1\rangle\}$, even though $(4, 5) \notin Adj(A) \cup Adj(B) \cup Adj(C)$. It follows from Lemma 1 that the medians for unsigned genomes can also fail the inclusion property. Consider the three unsigned genomes $f(A)$, $f(B)$, and $f(C)$. If $Y \in MED(f(A), f(B), f(C))$ then $Adj(Y) \not\subseteq Adj(f(A)) \cup Adj(f(B)) \cup (f(C))$. Consequently, NP-hardness of the median breakpoint problem does not imply the

NP-hardness of the constrained problem of minimizing $\Psi(X)$ such that $Adj(X) \subseteq Adj(f(A), f(B), f(C))$. Nevertheless, the problem is still NP-hard (Theorem 7 and Theorem 10).

On a positive note we can obtain a useful lower bound for the breakpoint median problem. It works for both signed and unsigned genomes. Note that in the signed case we do not consider the adjacencies (g, h) and $(-h, -g)$ to be distinct.

Lemma 3 *(Sankoff and Blanchette 1997) Given three genomes A, B, C, let λ_i denote the number of distinct adjacencies of weight i and put*

$$L(A, B, C) = 2n - 2\lambda_3 - \lambda_2. \tag{27}$$

Then $med(A, B, C) \geq L(A, B, C)$ and this bound is realised if and only if there is a genome X with $Adj(X) \subseteq Adj(A) \cup Adj(B) \cup Adj(C)$ that contains all weight two and weight three adjacencies.

Proof
Let A, B, C be unsigned genomes on gene set $\mathcal{G}$ and define $w(x, y)$ as above. Given any unsigned genome S on $\mathcal{G}$ we have $\Psi(S) = \sum_{\{x,y\} \in Adj(S)} (3 - w(x, y))$. For each $i = 0, 1, 2, 3$ put $l_i = |\{x, y\} \in Adj(S) : w(x, y) = i\}|$. Then

$$\begin{aligned} \Psi(S) &= 3n - (3l_3 + 2l_2 + l_1) && (28)\\ &= 2n - 2l_3 - l_2 + l_0 && (29)\\ &\geq 2n - 2\lambda_3 - \lambda_2 && (30) \end{aligned}$$

with equality if and only if $l_3 = \lambda_3$, $l_2 = \lambda_2$ and $l_0 = 0$, that is, if and only if S contains all weight two and weight three adjacencies and no adjacencies of weight zero.

The result for signed genomes follows from Lemma 2, noting that if A, B, C are signed genomes then $L(A, B, C) = L(f(A), f(B), f(C))$. □

3 The breakpoint median problem is NP-hard for signed genomes

We say that a signed genome A is *positive* if all genes in A have a positive sign. We consider now a constrained version of the breakpoint median problem for signed genomes:

BREAKPOINT MEDIAN PROBLEM FOR POSITIVE, SIGNED GENOMES
INSTANCE: Positive signed genomes A, B, C.
PROBLEM: Find a positive signed genome X minimizing $d(A,X)+d(B,X)+d(C,X)$.

Our NP-hardness proofs of the breakpoint median problems are all based on the following result.

Theorem 4 *The BREAKPOINT MEDIAN PROBLEM FOR POSITIVE SIGNED GENOMES is NP-hard.*

Proof
We provide a reduction from DIRECTED HAMILTONIAN CIRCUIT (Garey and Johnson 1979, Karp 1972). Let $G=(V,E)$ be a directed graph with maximum vertex degree 3. We show how to construct three positive signed genomes A,B,C such that G has a directed Hamiltonian circuit if and only if $med(A,B,C)=L(A,B,C)$.

Let G' be the graph with vertex set $V'=\{v_1,v_2,v_3 : v\in V\}$ and edge set

$$E'=\{(v_1,v_2),(v_2,v_3) : v\in V\}\cup\{(u_3,v_1) : (u,v)\in E\}. \tag{31}$$

This is the same as replacing each vertex in G with a two edge chain. There is a simple one to one correspondence between Hamiltonian circuits in G and Hamiltonian circuits in G'.

Given any directed graph H with edges labelled by one or more of A,B,C, and $X\in\{A,B,C\}$, let H_X denote the subgraph with the same vertex set and edge set containing edges with X in their label set. It is a straightfoward matter to label each edge of G' with one or more of A,B,C such that for each $X\in\{A,B,C\}$ the subgraph G'_X is a Hamiltonian circuit or a subset of a Hamiltonian circuit.

We perform a series of progressive modifications of G' to obtain a graph G'' with coloured edges such that the subgraphs of G'' induced by each colour are Hamiltonian cycles, and there is a one to one correspondence between Hamilton circuits of G' and Hamiltonian circuits of G'' that contain all edges with multiple labels.

Choose X such that G'_X is not a Hamiltonian circuit. Choose two vertices a,b such that adding (a,b) to G'_X either creates a Hamiltonian circuit, or gives a subgraph of a Hamiltonian circuit with fewer components. Let x be any vertex apart from a or b. Add two new vertices y and z. For each w such that (x,w) is an edge of G'_X add the edge (z,w) with the same label set as (x,w). Now remove all outgoing edges of x and add the edges (x,y) and (y,z), both labelled with $\{A,B,C\}-\{X\}$., and the edges (x,z), (a,y), (y,b) all with label set $\{X\}$ (see Figure 1).

We see that there is no Hamiltonian circuit that contains all edges with multiple labels that also contains the edges (x,z), (a,y), (y,b) all with label set $\{X\}$. Hence there is a one-to-one correspondence between Hamiltonian circuits containing all multiple label edges in the modified graph and Hamiltonian circuits containing all multiple label edges in the original graph.

We repeat this process until all subgraphs G'_X are Hamiltonian circuits, obtaining the required graph G''.

For each $X\in\{A,B,C\}$ let X be the signed genome given by G''_X with gene set equal to the vertex set of G''. Given a signed genome Z we have $\Psi(X)=L(A,B,C)$ if and only if $Adj(Z)$ contains all weight two or three adjacencies and no weight zero adjacencies, if and only if the circuit in G'' corresponding to Z is a Hamiltonian circuit that contains all multiply labelled edges. □

Even when we are given a Hamiltonian circuit in a graph it is NP-hard to determine if that Hamiltonian circuit is unique (Johnson and Papadimitious 1987). In the proof we established a one to one correspondence between genomes Z such that $\Psi(Z)=L(A,B,C)$ and Hamiltonian circuits of G. We therefore have

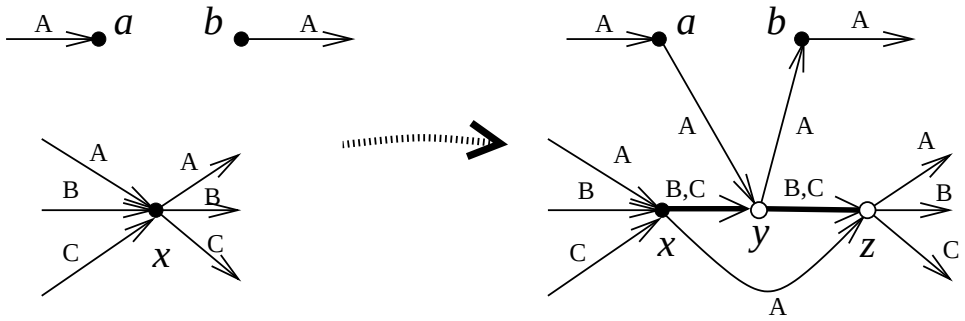

Figure 1: An example of the linking construction used in the proof of Theorem 4. Vertex a has no outgoing edge with $X = A$ in its label set, and b has no incoming vertex with A in its label set. We choose a, b such that adding (a, b) to G'_A does not give a non-Hamiltonian circuit. We choose an another vertex x and insert two new vertices y and z. The incoming edges of x in the right hand graph are the same as in the left hand graph. The outgoing edges of z are the same as the incoming edges of x in the left hand graph. The remaining edges reduce the number of components in G'_A but leave the same number of components in G'_B and G'_C.

Corollary 5 *Given signed, positive genomes A, B, C and $Z \in MED(A, B, C)$ it is an NP-hard problem to determine if there is positive $Z' \neq Z$ such that $Z' \in MED(A, B, C)$.*

The reduction from the general signed case to the positive signed case is provided by Lemma 6.

Lemma 6 *Let A, B, and C be positive signed genomes on the same gene set. If $med(A, B, C) = L(A, B, C)$ and $X \in MED(A, B, C)$ then X is a positive signed genome (or the reverse of a positive signed genome).*

Proof
Suppose $Adj(X)$ contains an adjacency (g, h) such that g and h have different signs. Then (g, h) is not an adjacency of any of A, B, C, contradicting the fact that $\Psi(X) = L(A, B, C)$. □

We have now established

Theorem 7 *The median breakpoint problem for signed genomes is NP-hard, even with the constraint that the median genome contains no adjacencies not present in one or more of the input genomes.*

Lemma 6 still maintains a one to one relationship between Hamiltonian circuits and median genomes. Hence we can extend the uniqueness result.

Corollary 8 *Given signed genomes A, B, C and $X \in MED(A, B, C)$ it is an NP-hard problem to determing if there is $X' \neq X$ such that $X' \in MED(A, B, C)$. Consequently, it is also NP-hard to determine those edges common to all median genomes.*

4 The breakpoint median problem is NP-hard for unsigned genomes

The NP-hardness of the unsigned breakpoint median problem follows directly from Lemma 2 and Theorem 7. The NP-hardness of the associated uniqueness problem requires a little more care.

Lemma 9 *Let A, B, C be three signed genomes on the same gene set. If there is unique X such that $\Psi(X) = L(A, B, C)$ then there is unique Y such that $\Psi(Y) = L(f(A), f(B), f(C))$.*

Proof
One such Y is given by $Y = f(X)$. If $\Psi(Y') = L(f(A), f(B), f(C))$ then $Adj(Y')$ contains all weight three adjacencies, so there is X' such that $f(X') = Y'$. Then $\Psi(X') = \Psi(Y') = L(f(A), f(B), f(C)) = L(A, B, C)$ so $X' = X$ by the uniqueness of X. □

We have now established

Theorem 10 *The breakpoint median problem is NP-hard for unsigned genomes. Given a solution to the breakpoint median problem, it is an NP-hard problem to determine if the solution is unique. Hence it is also NP-hard to determine whether a particular adjacency belongs to all median genomes.*

Acknowledgements

This work was carried out while D.Bryant held a Bioinformatics Postdoctoral Fellowship from the Canadian Institute for Advanced Research, Evolutionary Biology Program. Research supported in part by the Natural Sciences and Engineering Research Council of Canada and the Canadian Genome Analysis and Technology grants to D. Sankoff.

References

Blanchette, M. and Kunisawa, T. and Sankoff, D. 1996. Parametric genome rearrangement. *Gene* GC 11–17.

Blanchette, M. and Kunisawa, T. and Sankoff, D. 1998. Gene order breakpoint evidence in animal mitochondrial phylogeny. Tech. report, C.R.M. Université de Montréal.

Caprara, A. Sorting by Reversals is Difficult. 1997. *Proceedings of the First International Conference on Computational Molecular Biology*. ACM Press. New York.

Garey, Michael R. and Johnson, David S. 1979. *Computers and intractability, A guide to the theory of NP-completeness*. W. H. Freeman and Co. San Francisco.

Gu, Q.-P. and Iwata, K. and Peng, S. and Chen, Q.-M. 1997. A heuristic algorithm for genome rearrangements, 268–269. *In* Miyano, S. and Takagi, T., eds., *Genome Informatics 1997*, Tokyo.

Hannenhalli, S. and Pevzner, P. 1995. Transforming Cabbage into Turnip (polynomial algorithm for sorting signed permutations by reversals). *Proceedings of the Twenty-Seventh Annual ACM Symposium on the Theory of Computing*, 178–189.

Johnson, D. and Papadimitriou, C. 1987. Computational complexity, 37–85. *In* Lawler, E. and Lenstra, J. and Rinnooy Kan, A. and Shmoys, D., eds. *The traveling salesman problem, a guided tour in combinatorial optimization*, Wiley, Chichester.

Johnson, David S. and McGeoch, Lyle A. 1997. Local search in combinatorial optimization, 215–310. *In* Aarts, E.H.L. and Lenstra, J.K., eds., *Local search in combinatorial optimization*. Wiley, Chichester.

Karp, R. 1972. Reducibility Among Combinatorial Problems, 85–103. *In* R. E. Miller and J. W. Thatcher, eds., *Complexity of Computer Computations*, Plenum Press, New York.

Pe'er I. and Shamir, R. 1998. The median problems for breakpoints are NP-complete. Manuscript.

Sankoff, D. 1992. Edit Distance for Genome Comparison Based on Non-Local Operations, 121–135. *In* Apostolico, A., Crochemore, M., Galil, A. and Manber, U., eds., *Proceedings of Combinatorial Pattern Matching. Lecture notes in computer science 644*, Springer.

Sankoff, D. and Blanchette, M. 1997. The median problem for breakpoints in comparative genomics. *Computing and combinatorics (Shanghai, 1997)*, 251–263.

Sankoff, D. and Blanchette, M. 1998. Multiple genome rearrangement. *Proceedings of the Second International Conference on Computational Molecular Biology*, 243–247.

Sankoff, D., Leduc, G., Antoine, N., Paquin, B., Lang, B., and Cedergren, R. 1992. Gene order comparisons for phylogenetic inference: Evolution of the mitochondrial genome. *Proc. Natl. Acad. Sci. USA, 89*, 6575–6579.