Decompositions of Multiple Breakpoint Graphs and Rapid Exact Solutions

Andrew Wei Xu and David Sankoff

Department of Mathematics and Statistics, University of Ottawa, Canada K1N 6N5

Abstract. The median genome problem reduces to a search for the vertex matching in the multiple breakpoint graph (MBG) that maximizes the number of alternating colour cycles formed with the matchings representing the given genomes. We describe a class of "adequate" subgraphs of MBGs that allow a decomposition of an MBG into smaller, more easily solved graphs. We enumerate all of these graphs up to a certain size and incorporate the search for them into an exhaustive algorithm for the median problem. This enables a dramatic speedup in most randomly generated instances with hundreds or even thousands of vertices, as long as the ratio of genome rearrangements to genome size is not too large.

1 Introduction

The median problem underlies one approach to phylogenetics based on genomic distance. The idea, illustrated in Figure 1, is to optimize each ancestral node of an unrooted phylogeny in terms of its three or more immediate neighbours, modern or ancestral, and to iterate across the tree until convergence of the objective function (to a local optimum) at all nodes. This approach to the "small phylogeny" problem (i.e., the graph structure of the tree is given and does not need to be inferred, in contrast to the "big phylogeny problem") has a decade of history in the study of genome rearrangement [7,6,2,1], though its use in sequence-based phylogenetics dates to the 1970s [8].

In the study of genome rearrangement, genomes are treated as signed permutations on $1, \ldots, n$, either circular or linear, sometimes fragmented into chromosomes. The metric d on the set of genomes is an edit distance that counts the minimum number of operations required to transform one genome into another. The allowed operations may include the reversal of a contiguous chromosomal fragment, which also switches the sign on each term in the scope of the reversal; translocation, which involves the exchange of suffixes or prefixes of two chromosomes; transposition, or the excision of a contiguous chromosomal fragment and its re-insertion elsewhere on the chromosome; and a limited number of other operations. While distances involving reversals and translocations only can be calculated in time linear in n [4,10], the complexity of allowing transpositions in the distance calculation, either alone or in combination with reversals and translocations, is unknown. Recently, by generalizing the operation of transposition to that of block interchange [12], it became possible to include transpositions with

K.A. Crandall and J. Lagergren (Eds.): WABI 2008, LNBI 5251, pp. 25–37, 2008.

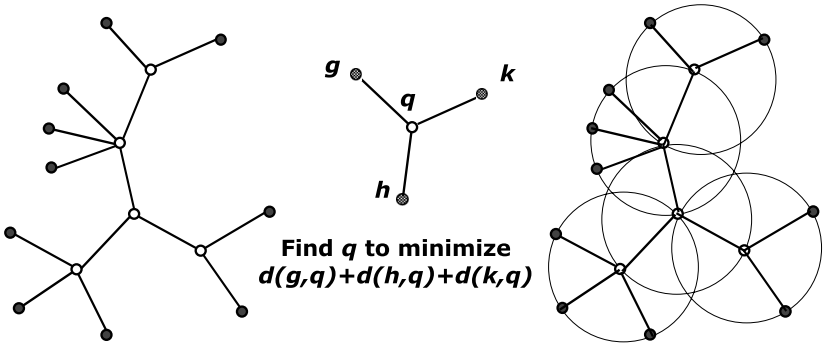

Fig. 1. Left: unrooted phylogeny with open dots representing ancestral genomes to be inferred. Middle: median problem with three given genomes g, h and k and median q to be inferred. Right: decomposition of phylogeny into overlapping median problems.

reversals and translocations in genomic distance calculations, within a framework known as "double cut and join" (DCJ). Moreover, the DCJ framework allows for substantial mathematical simplification of the distance calculation.

The median problem for genomic rearrangement distances in NP-hard [3,9]. Algorithms have been developed to find exact solutions for small instances [3,6] and there are rapid heuristics of varying degrees of efficiency and accuracy [2,1,5]. In the present paper, we explore the hypothesis that although there are no worst-case guarantees, it is worthwhile to develop methods to rapidly detect instances which are easily solved exactly.

Because of its simple structure, we choose to work with DCJ distance d as most likely to yield non-trivial mathematical results. We require genomes to consist of one or more circular chromosomes, but this is for simplicity of presentation, and our results could fairly easily be extended to genomes with multiple linear chromosomes. Then the median problem is to find a genome q with the smallest total distance $\sum_{g \in G} d(q, g)$, for a given set of genomes $\mathcal{G}$.

The mathematical analysis of genomic distances generally invokes the *breakpoint graph*, which we will describe in Section 2. For DCJ, we have $d(g, h) = n - c$, where n is the number of genes in genomes g and h, and c is the number of cycles in the breakpoint graph. We define *adequate* subgraphs of the breakpoint graph, and key graph transformations in Section 2, and we demonstrate in Section 3 how to decompose large instances of median problems into smaller instances. This effectively reduces the search space of the median problem and makes it possible to design algorithms applicable to most instances of interest to biologists. In Sections 4 and 5, we sketch some of the considerations involved in these algorithms and describe the results of simulations on various data sets. The full development of the algorithm and its application to them are detailed in reference [11].

2 Graph and Subgraph Structures

2.1 Breakpoint Graph

We construct the breakpoint graph of two genomes as in Figure 2 by representing each gene by an ordered pair of vertices, adding coloured edges to represent the adjacencies between two genes, red edges for one genome and blue for the other.

In a genome, every gene has two adjacencies, one incident to each of its two endpoints, since it appears exactly once in that genome. Then in the breakpoint graph, every vertex is incident to one red edge and one blue one. Thus the breakpoint graph is a 2-regular graph which automatically decomposes into a set of alternating-colour cycles.

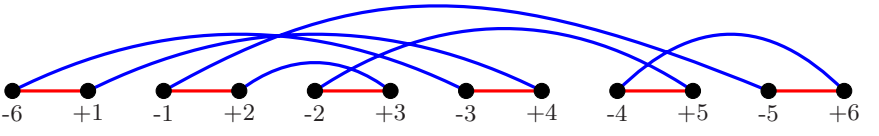

Fig. 2. Breakpoint graph for blue genome 1 -5 -2 3 -6 -4 (in gray) and red genome 1 2 3 4 5 6 (in black)

The edges of one colour form a perfect matching of the breakpoint graph, which we will simply refer to as a *matching*, unless otherwise specified. By the red matching, we mean the matching consisting of all the red edges.

The size for breakpoint graphs, multiple breakpoint graphs and median graphs is defined as half the number of vertices in it, which also equals to the number of genes in each genome and the size of each perfect matching.

2.2 Multiple Breakpoint Graph and Median Graph

The breakpoint graph extends naturally to a multiple breakpoint graph (MBG), representing a set $\mathcal{G}$ of three or more genomes. The number of genomes $N_{\mathcal{G}} \geq 3$ in $\mathcal{G}$ is also the edge chromatic number of the MBG. The colours assigned to the genomes are labeled by the integers from 1 to $N_{\mathcal{G}}$. We will use $B(\mathcal{G})$ or B throughout to refer to the MBG of the genomes $\mathcal{G}$.

For a given distance d, the median problem for $\mathcal{G} = \{g_1, \ldots, g_{N_{\mathcal{G}}}\}$ is to find a genome q which minimizes $\sum_{i=1}^{N_{\mathcal{G}}} d(g_i, q)$. For a candidate median genome, we use a different colour for its matching E, namely colour 0. Adding E to the MBG $B(\mathcal{G})$ results in the *median graph* $M_E(\mathcal{G}) = B(\mathcal{G}) \cup E$.

The set of all possible candidate matchings is denoted by $\mathcal{E}$. The set of all possible median graphs is $\mathcal{M}(\mathcal{G}) = \{M = B(\mathcal{G}) \cup E : E \in \mathcal{E}\}$.

The 0-i cycles in a median graph with matching E, numbering $c(0, i)$ in all, are the cycles where 0-edges and i edges alternate. Let $c_E(B) = \sum_{i=1}^{N_{\mathcal{G}}} c(0, i)$. Then $c_{\max}(B) = \max\{c_E(B) : E \in \mathcal{E}\}$ is the maximum number of cycles that can be constructed from B.

Minimizing the total distance in the median problem is equivalent to finding an optimal matching E, i.e., with $c_E(B) = c_{\max}(B)$. Let $\mathcal{E}^\star(B)$ be the set of all optimal matchings.

2.3 MBG Subgraphs and Connecting Edges

Let $\mathbf{V}(G)$ and $\mathbf{E}(G)$ be the sets of vertices and edges of a regular graph G. A *proper subgraph* H of G is one where $\mathbf{V}(H) = \mathbf{V}(G)$ and $\mathbf{E}(H) = \mathbf{E}(G)$ do not both hold at the same time. An *induced subgraph* H of G is the subgraph which satisfies the property that if $x, y \in \mathbf{V}(H)$ and $(x, y) \in \mathbf{E}(G)$, then $(x, y) \in \mathbf{E}(H)$.

In this paper, we will focus on the induced proper subgraphs, with an even number of vertices, of an MBG. Half of the number of these vertices is defined as the size of the subgraph H, denoted by m. $\mathcal{E}(H)$ is the set of all perfect 0-matchings $E(H)$, the cycle number determined by H and $E(H)$ is $c_{E(H)}(H)$, and $c_{\max}(H)$ is the maximum number of cycles that can be constructed from H by adding some $E(H)$. A 0-matching $E^\star(H)$ with $c_{E^\star(H)}(H) = c_{\max}(H)$ is called an optimal local matching, and $\mathcal{E}^\star(H)$ is the set of such matchings.

The *connecting edges* of a subgraph H in an MBG $B(\mathcal{G})$ are the edges of $B(\mathcal{G})$ incident to H exactly once, and are denoted by $K(H)$. The complementary induced subgraph of H in $B(\mathcal{G})$, denoted as $\overline{H}$, is the subgraph of $B(\mathcal{G})$ induced by $\mathbf{V}(B) - \mathbf{V}(H)$. Note that $B(\mathcal{G}) = H + K(H) + \overline{H}$, as illustrated in Figure 3.

2.4 Crossing Edges and Decomposers

For an MBG B and a subgraph H, a potential 0-edge would be *H-crossing* if it connected a vertex in $\mathbf{V}(H)$ to a vertex in $\mathbf{V}(\overline{H})$. A candidate matching containing one or more H-crossing 0-edges is an H-crossing candidate. A MBG subgraph H is called a *decomposer* if for any MBG containing it, there is an optimal matching that is not H-crossing. It is a *strong decomposer* if for any MBG containing it, all the optimal matchings are not H-crossing.

For an MBG B, the search space for an optimal matching is $\mathcal{E}$, which is of size $(2n-1)!! = \frac{(2n)!}{2^n n!}$. If B contains a (strong) decomposer H of size m, then the search can be limited to the smaller space $\mathcal{E}(H) \times \mathcal{E}(\overline{H}) = \{E = E_H \cup E_{\overline{H}} : E_H \in \mathcal{E}(H),\ E_{\overline{H}} \in \mathcal{E}(\overline{H})\}$, which is of size $(2m-1)!! \cdot (2n-2m-1)!!$.

2.5 Adequate and Strongly Adequate Subgraphs

In an MBG for a set of genomes $\mathcal{G}$, a connected subgraph H of size m is an *adequate subgraph* if $c_{\max}(H) \geq \frac{1}{2} m N_{\mathcal{G}}$; it is *strongly adequate* if $c_{\max}(H) > \frac{1}{2} m N_{\mathcal{G}}$.

A (strongly) adequate subgraph H is *simple* if it does not contain another (strongly) adequate subgraph as an induced subgraph; deleting any vertex from H will destroy its adequacy. In addition, a simple (strong) adequate subgraph H is *minimal* if we cannot even delete any edges without destroying its adequacy, i.e., for any edge $e \in \mathbf{E}(H)$, $c_{\max}(H - e) < \frac{1}{2} m N_{\mathcal{G}}$ ($c_{\max}(H - e) \leq \frac{1}{2} m N_{\mathcal{G}}$).

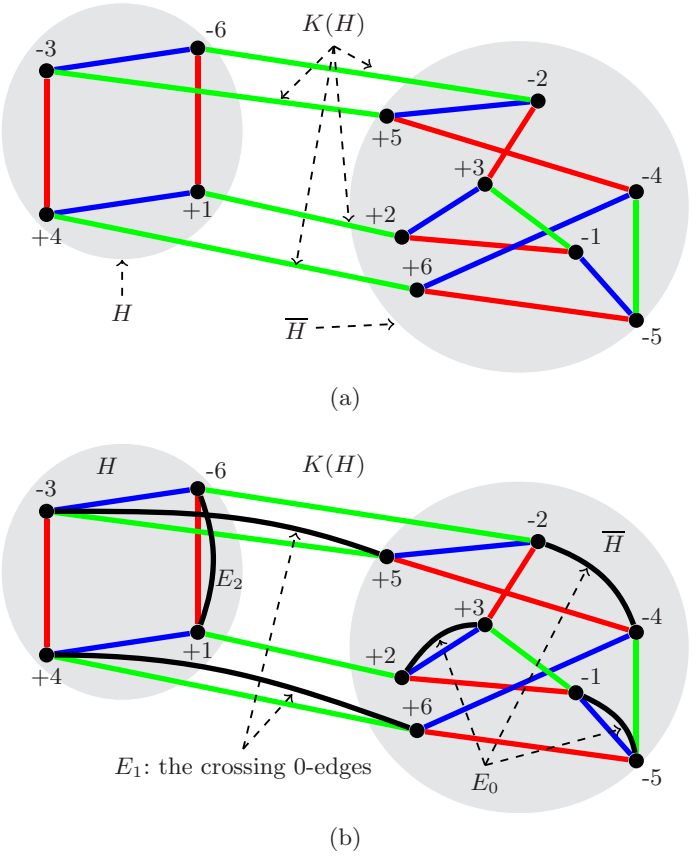

Fig. 3. MBG and median graph. Thick, gray, double and thin edges denote the edges with colours 1, 2, 3 and 0 correspondingly. (a) An MBG based on three genomes, (1 2 3 4 5 6), (1 -5 -2 3 -6 -4) and (1 3 5 -4 6 -2). A subgraph H, the connecting edge set $K(H)$ and the complementary subgraph $\overline{H}$ are illustrated. (b) A median graph. The candidate matching is divided into three 0-edge sets: E_0, E_1 and E_2.

2.6 Edge Shrinking, Expansion and Contraction

To shrink an edge e in a graph B, delete its two end vertices and any edges (including e) parallel to e, then for the edges incident to the deleted vertices, replace each pair of edges of same colour by a single edge of that colour, producing a new graph $B \circ e$, as illustrated by Fig 4(a)–(c). To shrink a set of edges A, shrink the edges in A one by one in any order, producing $B \circ A$.

To expand a 0-edge (a, b) in a graph B, remove that edge, add two new vertices i and j to the graph, connect i and j by $N_{\mathcal{G}}$ edges with colours ranging from 1 to $N_{\mathcal{G}}$, and add 0-edges (a, i) and (b, j), as illustrated by Fig 4(c) following the upward arrow.

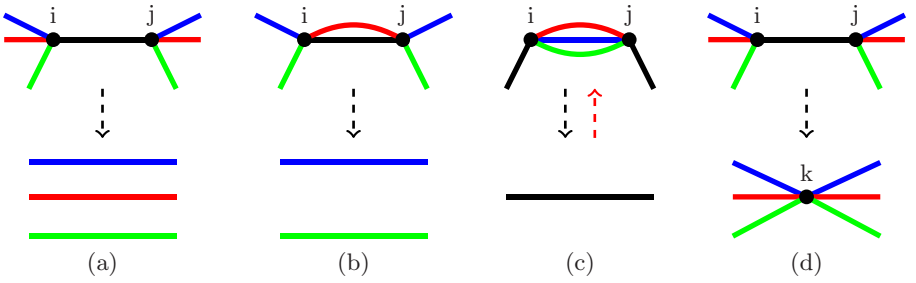

Fig. 4. Edge shrinking, expansion and contraction in a median graph based on 3 genomes: the downward arrows in (a), (b) and (c) illustrate edge shrinking in various situations; (c) the upward arrow illustrates an expansion of a thin edge; (d) illustrates a contraction of a thin edge

Proposition 1. *If median graph M' is obtained from another median graph M by expanding some 0-edge, then they contain the same number of cycles, i.e.* $c(M') = c(M)$.

To contract a 0-edge e from a graph G, delete e and merge its two end vertices, resulting in the graph G/e, as illustrated by Fig 4(d).

3 An Adequate Subgraph Is a Decomposer

In this section, we prove our main result: every (strongly) adequate subgraph is a (strong) decomposer. The general idea of the proof is that if H is a (strongly) adequate subgraph of MBG $B(\mathcal{G})$, for any H-crossing candidate matching E, we can always find another candidate matching E' that is not crossing, with $c_{E'}(B) \geq c_E(B)$ (or $c_{E'}(B) > c_E(B)$).

We partition the 0-edges in E among three sets: E_0, the set of 0-edges not incident to H; E_1, those incident to H exactly once; and E_2, those incident to H twice. In the median graph $M = B \cup E$, we shrink the 0-edge set E_0 and expand each 0-edge in E_2. The resultant median graph illustrated by Fig 5(a) is called the *twin median graph*, denoted by $\overset{\circ}{M}^{\bullet} = \overset{\circ}{B}^{\bullet} \cup \overset{\circ}{E}^{\bullet}$.

If the 0-edges of a cycle in M are all in E_0, then after shrinking all 0-edges in E_0, this cycle does not appear in $\overset{\circ}{M}^{\bullet}$. If a cycle in M contains 0-edges in E_1 or E_2, then with only part of the cycle being shrunk, this cycle does appear in $\overset{\circ}{M}^{\bullet}$. Denote $c_{E_0}(B)$ as the number of cycles formed by B and 0-edges in E_0 only. Then

Proposition 2

$$c_E(B) = c_{E_0}(B) + c_{\overset{\circ}{E}^{\bullet}}(\overset{\circ}{B}^{\bullet}) \tag{1}$$

Since E_0 is not incident to the subgraph H, shrinking E_0 does not affect H. So H remains in $\overset{\circ}{M}^{\bullet}$. Denote the subgraph in $\overset{\circ}{M}^{\bullet}$ induced by $\mathbf{V}(H)$ as F. If a pair of connecting edges with colour i in M, is connected by a 0-i alternating colour

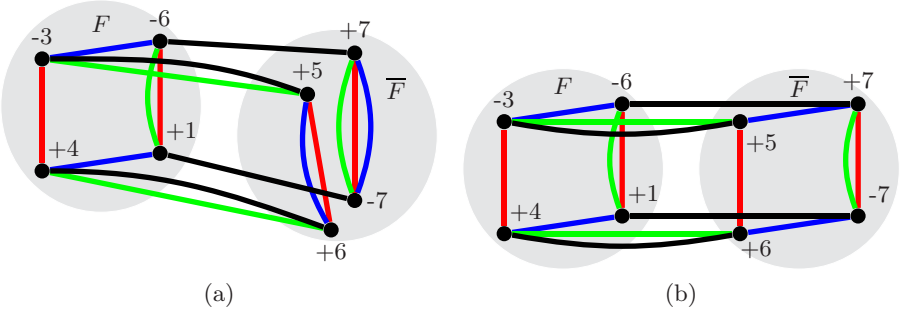

Fig. 5. Twin median graph and symmetrical median graph. (a) The twin median graph is obtained from the median graph in Figure 3b by shrinking the 0-edge set E_0 and expanding the 0-edge set E_2. (b) is the corresponding symmetric graph, with the left part mirror-symmetric to the right part.

path, with all 0-edges in E_0, then after shrinking E_0, this pair of i-edges are merged into a new i-edge e, with both ends incident to $\mathbf{V}(H)$. Edges like e are contained in F but not in H. Thus

Proposition 3. *Suppose $\overset{\circ}{B}\overset{\bullet}{}$ is a twin MBG constructed from B based on a subgraph H of size m, and F is the subgraph in $\overset{\circ}{B}\overset{\bullet}{}$ induced by $\mathbf{V}(H)$. Then F is of size m and $F \supseteq H$. If H is a (strongly) adequate subgraph, then so is F.*

Suppose the number of connecting edges in $K(F)$ of the twin MBG $\overset{\circ}{B}\overset{\bullet}{}$ is $2k$. The 0-edges in $\overset{\circ}{M}\overset{\bullet}{}$ denoted by $\overset{\circ}{E}\overset{\bullet}{}$ are either from E_1 or the new added ones when expanding E_2. All of them are incident to F exactly once, so each 0-edge in $\overset{\circ}{E}\overset{\bullet}{}$ is F-crossing. Then F and $\overline{F}$ must be of the same size.

The 0-edges in $\overset{\circ}{E}\overset{\bullet}{}$ can be viewed as a mapping from the vertex set $\mathbf{V}(F)$ to $\mathbf{V}(\overline{F})$. If under this mapping, F is isomorphic to $\overline{F}$, as illustrated by Fig 5(b) then we call the twin median graph a *symmetrical median graph*, and we denote it by $\overset{\circ}{M}\overset{\circ}{}$.

In any twin median graph, the size of an alternating colour cycle is at least 1, which is only possible when a 0-edge is parallel to a connecting edge. All other cycles have minimum size 2. We have

Proposition 4. *If in a twin median graph $\overset{\circ}{M}\overset{\bullet}{}$, any cycle containing a connecting edge is of size 1 and any other cycle is of size 2, then $\overset{\circ}{M}\overset{\bullet}{}$ contains the largest possible number of cycles among all twin median graphs formed from $\overset{\circ}{B}\overset{\bullet}{}$. The maximum cycle number is $mN_{\mathcal{G}} + k$. This can be achieved only when $\overset{\circ}{M}\overset{\bullet}{}$ is a symmetrical median graph $\overset{\circ}{M}\overset{\circ}{}$.*

Proof. Since there are $2k$ connecting edges, the number of cycles of size 1 must be $2k$. Then the number of remaining non-0 edges is $2mN_{\mathcal{G}} - 2k$. Hence there are $mN_{\mathcal{G}} - k$ cycles of size 2. The maximum total number of cycles is $mN_{\mathcal{G}} + k$. Because of the symmetry of $\overset{\circ}{M}\overset{\circ}{}$, the other cycles can only be of size 2. Hence $\overset{\circ}{M}\overset{\circ}{}$ is the only twin median graph containing the maximum number of cycles. □

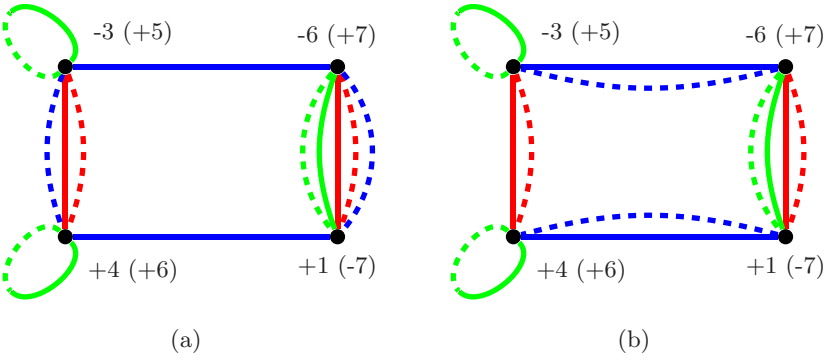

Fig. 6. The contracted twin graph (a) and contracted symmetric graph (b). The contracted graphs are generated from a twin median graph by contracting 0-edges. Dashed edges are from the complementary subgraphs and the half-solid-half-dashed ones are the connecting edges.

Next we investigate the difference between a twin median graph $\mathring{M}^{\bullet}$ and a symmetric median graph $\mathring{M}^{\circ}$, in terms of the number of DCJ operations needed to transform one into another.

Lemma 1. *If $\mathring{M}^{\bullet}$ is a twin median graph and $\mathring{M}^{\circ}$ is the symmetric median graph, then we can transform one into the other by exactly $mN_{\mathcal{G}} + k - c(\mathring{M}^{\bullet})$ DCJ operations on non 0-edges.*

Proof. We construct the *contracted graph*, illustrated in Figure 6, by contracting 0-edges of a median graph $\mathring{M}^{\bullet}$, where edges in $\overline{F}$ are represented by dashed lines and the connecting edges are represented by half-dashed, half-solid lines with the solid end incident to F and the dashed end incident to $\overline{F}$. For conciseness, when we say *solid edges* (*dashed edges*), we mean the solid (dashed) edges contained by F ($\overline{F}$) or the solid (dashed) ends of connecting edges. The contracted graph for $\mathring{M}^{\bullet}$ is denoted by $\overset{\bullet}{M}$ and the contracted graph for $\mathring{M}^{\circ}$ is denoted by $\overset{\circ}{M}$.

Comparing the median graph $\mathring{M}^{\bullet}$ and the contracted graph $\overset{\bullet}{M}$, it easy to see that each vertex in $\overset{\bullet}{M}$ has degree $2N_{\mathcal{G}}$, incident to $N_{\mathcal{G}}$ solid edges and $N_{\mathcal{G}}$ dashed edges. The 0-i alternating colour cycle in $\mathring{M}^{\bullet}$ becomes the alternating pattern (solid/dashed) cycle with colour i. The number of alternating pattern cycles is equal to the number of alternating colour cycles. Thus there are $c(\mathring{M}^{\bullet})$ pattern alternating cycles in $\overset{\bullet}{M}$ and $mN_{\mathcal{G}} + k$ cycles in $\overset{\circ}{M}$.

To transform $\overset{\bullet}{M}$ to $\overset{\circ}{M}$, we can show that there always exists a DCJ operation on two dashed edges with the same colour that increases the cycle number by one. When a connecting edge does not form a loop, apply a DCJ operation to loop it. Then arbitrarily select a solid edge from a cycle with size more than 2, apply a DCJ operation to make a dashed edge parallel to it. Thus with a number $mN_{\mathcal{G}} + k - c(\mathring{M}^{\bullet})$ DCJ operations, we can transform $\mathring{M}^{\bullet}$ to $\mathring{M}^{\circ}$ or *vice versa*. □

Proposition 5. *An arbitrary DCJ operation on non-0 edges in a median graph changes the cycle number by 1, 0, or -1.*

Proof. If the two edges belong to one cycle, it will either split into two cycles or remain as a single cycle. If the two edges belong to two cycles, then they will be joined into one cycle. □

Theorem 1. *If H is a (strongly) adequate subgraph of MBG B and E is a H-crossing candidate matching, then there is a candidate matching E' which is not H-crossing, with $c_{E'}(B) \geq c_E(B)$ (or $c_{E'}(B) > c_E(B)$).*

Proof.
1. From the median graph $M = B \cup E$, construct the twin median graph $\overset{\circ}{M}^{\bullet}$ and twin MBG $\overset{\circ}{B}^{\bullet}$ by shrinking 0-edges not incident to H (E_0) and expanding 0-edges incident to H twice (E_2). Denote the subgraph of $\overset{\circ}{M}^{\bullet}$ induced by $\mathbf{V}(H)$ as F. Then $c_E(B) = c_{E_0}(B) + c_{\overset{\circ}{E}^{\bullet}}(\overset{\circ}{B}^{\bullet})$.
2. Construct the symmetrical median graph $\overset{\circ}{M}^{\circ}$ with $F = \overline{F}$ and F also a (strongly) adequate subgraph.
3. Since F is a (strongly) adequate subgraph, there exists a 0-matching D of F satisfying $c_D(F) \geq \frac{1}{2}mN_{\mathcal{G}}$ (or $c_D(F) > \frac{1}{2}mN_{\mathcal{G}}$).
4. Replace the 0-matching in $\overset{\circ}{M}^{\circ}$ by two copies of D, one on F and one on $\overline{F}$. Denote the 0-matching as $2D$ and denote the resultant median graph as $\overset{\circ}{B}^{\circ} \cup 2D$, with $c_{2D}(\overset{\circ}{B}^{\circ}) \geq mN_{\mathcal{G}}$ (or $> mN_{\mathcal{G}}$).
5. Transform $\overset{\circ}{B}^{\circ}$ to $\overset{\circ}{B}^{\bullet}$ by $mN_{\mathcal{G}} + k - c(\overset{\circ}{M}^{\bullet})$ DCJ operations on $\overline{F}$ in $\overset{\circ}{B}^{\circ}$. So $c_{2D}(\overset{\circ}{B}^{\bullet}) \geq c_{\overset{\circ}{E}^{\bullet}}(\overset{\circ}{B}^{\bullet})$ (or $c_{2D}(\overset{\circ}{B}^{\bullet}) > c_{\overset{\circ}{E}^{\bullet}}(\overset{\circ}{B}^{\bullet})$).
6. Shrink the newly added sets of $N_{\mathcal{G}}$ parallel edges in $\overset{\circ}{B}^{\bullet}$ and reverse the shrinking operations on E_0 in step 1, to recover the MBG B. Then the 0-matching $2D$ becomes the candidate matching E' and the new median graph becomes $M' = B \cup E'$. Then $c_{E'}(B) = c_{2D\,\overset{\circ}{B}^{\bullet}} + c_{E_0}(B)$. Thus $c_{E'}(B) \geq c_E(B)$ (or $c_{E'}(B) > c_E(B)$). □

Theorem 2. *Any adequate subgraph is a decomposer. A strongly adequate subgraph is a strong decomposer.*

Proof. For an adequate subgraph there must be a optimal matching that is not crossing. Otherwise by Theorem 1, from the optimal crossing matching, we can construct a candidate matching that is not crossing and has at least as many cycles. Thus the adequate subgraph is a decomposer.

For a strongly adequate subgraph, the non-crossing candidate matchings are always better than the corresponding crossing candidate matchings. Then the optimal matchings cannot be crossing matchings. The strongly adequate subgraph is thus a strong decomposer. □

4 Median Calculation Incorporating MBG Decomposition

As adequate subgraphs are the key to decompose the median problems, we need to inventory them before making use of them. It turns out that it is most useful to limit this project to simple adequate graphs. Non-simple adequate graphs are both harder to enumerate and harder to use, and are likely to have simple ones

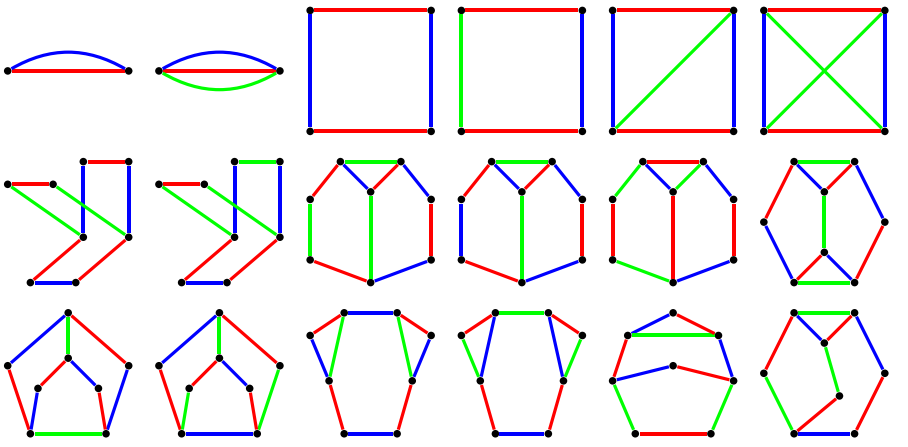

Fig. 7. Simple adequate subgraphs of size 1, 2 and 4 for MBGs on three genomes. See reference [11] for how they were identified.

embedded in them, which serve the same general purpose [11]. By exhaustive search, we have found all simple adequate graphs of size < 6; these are depicted in Figure 7. Though we have some of size 6, it would be a massive undertaking to compile the complete set with current methods.

Our basic algorithm for solving the median problem is a branch and bound, where edges of colour 0 are added at each step; we omit the details of procedures we use to increase the effectiveness of the bounds. To make use of the adequate subgraph theory we have developed, at each step we search for such an inventoried subgraph before adding edges, and if one is found, we carry out a decomposition and then solve the resulting smaller problem(s) [11].

Table 1. The number of runs, out of ten, where the median was found in less than 10 minutes on a MacBook, 2.16GHz, on one CPU

ρ/n \ n	10	20	30	40	50	60	80	100	200	300	500	1000	2000	5000
0.1	10	10	10	10	10	10	10	10	10	10	10	10	10	10
0.2	10	10	10	10	10	10	10	10	10	10	10	10	10	10
0.3	10	10	10	10	10	10	10	10	10	10	10	10	1	
0.4	10	10	10	10	10	10	10	10	0	0				
0.5	10	10	10	10	10	10	4	0						
0.6	10	10	10	10	9	6								
0.7	10	10	10	10	6									
0.8	10	10	10	10	6									
0.9	10	10	10	10	4									
1.0	10	10	10	8	2									

Table 2. Speedup due to discovery of larger adequate subgraphs (AS2, AS4). Three genomes are generated from the identity genome with $n = 100$ by 40 random reversals. Time is measured in seconds. Runs were halted after 10 hours. AS1, AS2, AS4, AS0 are the numbers of edges in the solution median constructed consequent to the detection of adequate subgraphs of sizes 1, 2, 4 and at steps where no adequate subgraphs were found, respectively.

	speedup	run time		number of edges			
run	factor	with AS1,2,4	with AS1	AS1	AS2	AS4	AS0
1	41,407	4.5×10^{-2}	1.9×10^{3}	53	39	8	0
2	85,702	3.0×10^{-2}	2.9×10^{3}	53	34	12	1
3	2,542	5.4×10^{0}	1.4×10^{4}	56	26	16	2
4	16,588	3.9×10^{-2}	6.5×10^{2}	58	42	0	0
5	$> 10^{6}$	5.9×10^{2}	stopped	52	41	4	3
6	199,076	6.0×10^{-3}	1.2×10^{3}	56	44	0	0
7	6,991	2.9×10^{-1}	2.1×10^{3}	54	33	12	1
8	$> 10^{6}$	4.2×10^{1}	stopped	57	38	0	5
9	1,734	8.7×10^{0}	1.5×10^{4}	65	22	8	5
10	855	2.1×10^{0}	1.8×10^{3}	52	38	8	2

5 Experimental Results

To see how useful our method is on a range of genomes, we undertook experiments on sets of three random genomes. Our JAVA program included a search for adequate subgraphs followed by decomposition at each step of a branch and bound algorithm to find the maximum number of cycles. We varied the parameters n and $\pi = \rho/n$, where ρ was the number of random reversals applied to the ancestor $I = 1, \ldots, n$ independently to derive three different genomes.

5.1 The Effects of n and $\pi = \rho/n$ on the Proportion of Rapidly Solvable Instances

Table 1 shows that relatively large instances can be solved if ρ/n remains at 0.3 or less. It also shows that for small n, the median is easy to find even if ρ/n is large enough to effectively scramble the genomes.

5.2 The Effect of Adequate Subgraph Discovery on Speed-Up

Table 2 shows how the occurrence of larger adequate subgraphs (AS2 and AS4) can dramatically speed up the solution to the median problem, generally from more than a half an hour to a fraction of a second.

5.3 Time to Solution

Our results in Section 5.1 suggest a rather abrupt cut-off in performance as n or ρ/n become large. We explore this in more detail by focusing on the particular

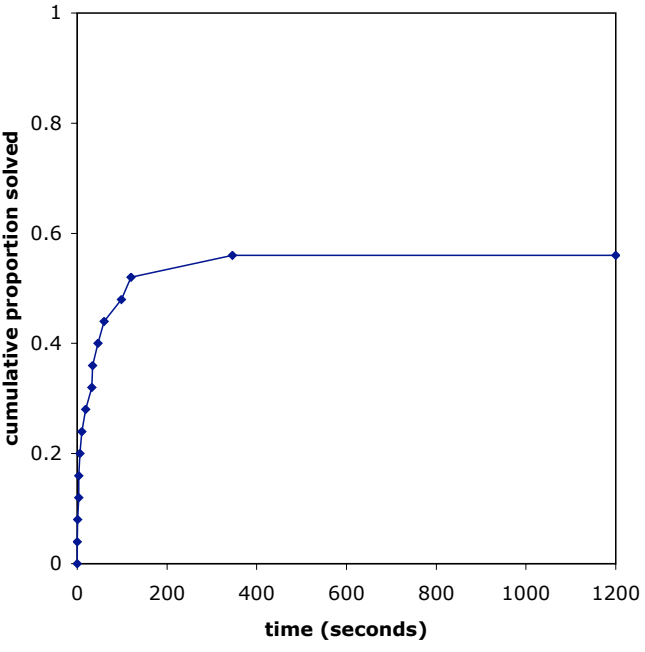

Fig. 8. Cumulative proportion of instances solved, by run time. $n = 1000$, $\rho/n = .31$. More than half are solved in less than 2 minutes; almost half take more than 20 minutes.

parameter values $n = 1000$ and $\rho/n = .31$. Figure 8 shows how the instances are divided into a rapidly solvable fraction and a relatively intractable fraction, with very few cases in between.

6 Conclusion

In this paper we have demonstrated the potential of adequate subgraphs for greatly speeding up the solution of realistic instances of the median problem. Many improvements seem possible, but questions remain. If we could inventory non-simple adequate graphs, or all simple adequate graphs of size 6 or more, could we achieve significant improvement in running time? It may well be that the computational costs of identifying larger adequate graphs within MBGs would nullify any gains due to the additional decompositions they provided.

Acknowledgments

We thank the reviewers for their helpful comments and suggestions. Research supported in part by a grant to DS from the Natural Sciences and Engineering Research Council of Canada (NSERC). DS holds the Canada Research Chair in Mathematical Genomics.

References

1. Adam, Z., Sankoff, D.: The ABCs of MGR with DCJ. Evol. Bioinform. 4, 69–74 (2008)
2. Bourque, G., Pevzner, P.: Genome-scale evolution: Reconstructing gene orders in the ancestral species. Genome Res. 12, 26–36 (2002)
3. Caprara, A.: The reversal median problem. INFORMS J. Comput. 15, 93–113 (2003)
4. Hannenhalli, S., Pevzner, P.: Transforming cabbage into turnip: Polynomial algorithm for sorting signed permutations by reversals. JACM 46, 1–27 (1999)
5. Lenne, R., Solnon, C., Stützle, T., Tannier, E., Birattari, M.: Reactive stochastic local search algorithms for the genomic median problem. In: van Hemert, J., Cotta, C. (eds.) EvoCOP 2008. LNCS, vol. 4972, pp. 266–276. Springer, Heidelberg (2008)
6. Moret, B.M.E., Siepel, A.C., Tang, J., Liu, T.: Inversion medians outperform breakpoint medians in phylogeny reconstruction from gene-order data. In: Guigó, R., Gusfield, D. (eds.) WABI 2002. LNCS, vol. 2452. Springer, Heidelberg (2002)
7. Sankoff, D., Blanchette, M.: Multiple genome rearrangement and breakpoint phylogeny. J. Comput. Biol. 5, 555–570 (1998)
8. Sankoff, D., Morel, C., Cedergren, R.: Evolution of 5S RNA and the nonrandomness of base replacement. Nature New Biol. 245, 232–234 (1973)
9. Tannier, E., Zheng, C., Sankoff, D.: Multichromosomal median and halving problems. In: WABI 2008. LNBI, vol. 5251. Springer, Heidelberg (2008)
10. Tesler, G.: Efficient algorithms for multichromosomal genome rearrangements. JCSS 65, 587–609 (2002)
11. Xu, A.W.: A fast and exact algorithm for the median of three problem—a graph decomposition approach (submitted, 2008)
12. Yancopoulos, S., Attie, O., Friedberg, R.: Efficient sorting of genomic permutations by translocation, inversion and block interchange. Bioinform. 21, 3340–3346 (2005)