On Exploring Genome Rearrangement Phylogenetic Patterns

Andrew Wei Xu

School of Computer and Communication Sciences
Swiss Federal Institute of Technology (EPFL)
EPFL IC LCBB, Station 14
CH-1015 Lausanne, Switzerland
wei.xu@epfl.ch

Abstract. The study of genome rearrangement is much harder than the corresponding problems on DNA and protein sequences, because of the occurrences of numerous combinatorial structures. By explicitly exploring these combinatorial structures, the recently developed adequate subgraph theory shows that a family of these structures, adequate subgraphs, are informative in finding the optimal solutions to the rearrangement median problem. Its extension gives rise to the tree scoring method GASTS, which provides quick and accurate estimation of the number of rearrangement events, for any given topology. With a similar motivation, this paper discusses and provides solid but somewhat initial results, on combinatorial structures that are informative in phylogenetic inference. These structures, called rearrangement phylogenetic patterns, provide more insights than algorithmic approaches, and may provide statistical significance for inferred phylogenies and lead to efficient and robust phylogenetic inference methods on large sets of taxa.

We explore rearrangement phylogenetic patterns with respect to both the breakpoint distance and the DCJ distance. The latter has a simple formulation and well approximates other edit distances. On four genomes, we prove that a contrasting shared adjacency, where a gene forms one adjacency in two genomes and a different adjacency in the other two genomes, is a rearrangement phylogenetic pattern. Phylogenetic inferences based on the numbers of this pattern, are very accurate and robust against short internal edges, tested on 55,000 datasets simulated by random inversions. Further analysis shows that the numbers of this pattern well explain the variations in the number of rearrangement events over different topologies.

1 Introduction

Genome rearrangement information, revealed in comparison of gene orders from related species, has been used for phylogenetic study for decades [11,8,3]. These methods, under the parsimony framework, make inferences of tree topologies and gene orders on internal nodes simultaneously. Such a detailed consideration gives extremely accurate phylogenetic inferences and provides good estimation

E. Tannier (Ed.): RECOMB-CG 2010, LNBI 6398, pp. 121–136, 2010.

of ancestral genome architectures [16,9,13,1]. However, due to the occurrences of numerous combinatorial structures, these problems are generally very difficult to solve [4,10,5,14].

We recently developed the *adequate subgraph theory* [17,15] in studying the rearrangement median problem. The theory explicitly explores the combinatorial structures and proves that a family of these structures, *adequate subgraphs*, are informative in finding optimal solutions. This theory allows us to solve the rearrangement median problem using a decomposition approach: detect the occurrence of an adequate subgraph, decompose the original problem into two smaller subproblems, and repeat this iteratively. In the case that no known adequate subgraphs can be detected, a branch-and-bound method or heuristics can be used to find optimal or heuristic solutions, respectively. An extension of this theory gives rise to a tree scoring method, *GASTS*, which minimizes the number of rearrangement events needed to explain a given data for a given tree topology [16] using a local optimization approach. GASTS scores very quickly and accurately, with typical errors within 0.1%, on trees with up to thousands of genomes and with thousands of genes in each genome.[1] To infer the phylogeny, we just need to find the topology with the smallest GASTS score.

The goal of this paper is to explore *rearrangement phylogenetic patterns* which are the combinatorial structures containing phylogenetic information. Under the parsimony framework, a combinatorial structure is said to be a rearrangement phylogenetic pattern, if we can prove that this structure always gives smaller scores (numbers of rearrangements or summarized distances over the tree) on one fixed topology but not on the others. In other words, we say a rearrangement phylogenetic pattern differentiates tree topologies. The topology with the smaller score is called the *preferred topology*.

We want to investigate whether and to what degree, we can make phylogenetic inferences by only inspecting the rearrangement phylogenetic patterns presented in the given data. Long-term goals of this topic are to analyze the probabilistic properties of these rearrangement phylogenetic patterns, to derive statistical tests for the significance of inferred results, and then to design efficient and robust phylogenetic inference methods for large numbers of taxa.

This paper focuses on problems with four signed genomes. Although only genomes with circular chromosomes are discussed, our results also can be applied to genomes with multiple linear chromosomes. We are aware that, given the computational difficulties and the numerous combinatorial structures in the rearrangement problems, this paper only presents some initial results. Hence it is not the goal of this paper to design better methods than the full parsimony methods, although our new methods do have very good accuracies.

In the rest of the section, we briefly introduce common pairwise distance measures, as they are the basis of most phylogenetic methods discussed in this paper. In Section 2, we introduce the full parsimony tree scoring method GASTS, two distance based methods and four-point metric, and phylogenetic invariants

[1] A study of the performance of GASTS can be found at http://sites.google.com/site/andrewweixu/gasts.

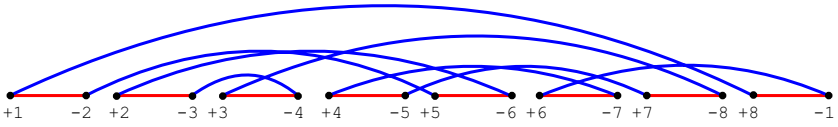


Fig. 1. The breakpoint graph of two circular genomes (1,-8,-3,4,7,5,2,6) and (1,2,3,4,5,6,7,8), in blue and red colors respectively. There are 8 genes in each genome, and there is one color-alternating cycle and no common adjacency in the breakpoint graph.

methods. In Section 3, we give results on what structures are and are not rearrangement phylogenetic patterns, and then introduce two phylogenetic scoring functions and their inference methods. In Section 4, we present comparison results of various phylogenetic inference methods.

1.1 Pairwise Distance Measures

In measuring the genomic distance between two genomes, there are two types of measures: observational distances, such as the breakpoint distance; edit distances, which are the smallest numbers of operations needed to transform one genome into the other given a set of allowed operations, such as the inversion distance [7], its generalization—the HP distance [6] and a further generalization—the DCJ distance [18,2]. This paper focus on the breakpoint distance and the DCJ distance, as the latter well approximates the other edit distances.

Breakpoint graph. The breakpoint graph is an important graph tool to study a pair of genomes. For each gene g, a pair of vertices $-g$ and $+g$ are used to represent its two endpoints, also called *extremities*. For each genome, *adjacencies*, which are pairs of extremities from neighboring genes, are represented by edges connecting the corresponding vertices; each genome is assigned a different color, and all edges from each genome are given that color. The breakpoint graph naturally decomposes into a set of *color-alternating cycles*, and we use c to denote their total number. Fig. 1 shows the breakpoint graph for the two genomes $(1, -8, -3, 4, 7, 5, 2, 6)$ and $(1, 2, 3, 4, 5, 6, 7, 8)$.

The breakpoint distance. A breakpoint occurs when an adjacency exists in one genome but not in the other. The number of breakpoints between two genomes is their breakpoint distance. The breakpoint distance has an intuitive explanation. To transform one genome into the other, we can cut the first genome into small fragments, rearrange and paste them back into the second genome. And the breakpoint distance is just the minimum number of cuts we need in this process. If n denotes the number of genes in each genome and a denotes the number of adjacencies shared by both genomes, then the breakpoint distance is simply $d_{BP} = n - a$.

The DCJ distance. The DCJ operation was first introduced in [18] and further studied in [2]. The DCJ distance provides a general rearrangement framework, including all common rearrangement operations, and a simple mathematical formula: $d_{DCJ} = n - c$.

2 Full Parsimony Methods, Distance Based Methods and Phylogeny Invariants

In this section, we briefly introduce three types of existing phylogenetic inference methods. The first type of methods are the full parsimony methods, which infer phylogeny and gene orders of internal nodes simultaneously, under either the breakpoint distance or some edit distance. In this paper, we consider the DCJ distance, as there exists a very quick and accurate tree scoring method, GASTS. Phylogenetic methods based on GASTS have been shown to be very accurate [16]. The second type of methods are distance-based methods. Given a data with n extant species and a distance measure, we can easily convert these sequences or gene orders into a n by n distance matrix with $\frac{n(n-1)}{2}$ independent variables, and reconstruct the phylogeny from this matrix. This kind of method has a charm of simplicity. To another end, the third type of methods, phylogeny invariants, explicitly explore various patterns and their occurrence frequencies in the data. Given an evolution model, algebraic relations (invariants) of the probabilities of these patterns can be derived, and are used to make phylogenetic inferences.

2.1 Full Parsimony Methods

The full parsimony methods are computationally challenging. But due to the recent development of the adequate subgraph theory and its extension, the scoring method GASTS can quickly find accurate estimation of the minimum numbers of rearrangements on thousands of genomes with thousands of genes in each genome. To infer the correct phylogeny for a given quartet problem with four taxa: A, B, C and D, we just need to compute the tree scores S_{GASTS} (we call them *phylogenetic score functions* in this paper) for the three topologies $AB|CD$, $AC|BD$ and $AD|BC$ and return the topology $\hat{T}_{\text{GASTS}}$ with the smallest GASTS score.

2.2 Distance Based Methods and Four-Point Metric

Assume the correct topology of the quartet problem is depicted by Fig. 2. If the distance between any two species is equal to the length of their path on the tree, e.g. $d_{AD} = v_1 + v_5 + v_4$, then the following *four point metric* holds:

$$d_{AB} + d_{CD} < d_{AC} + d_{BD} = d_{AD} + d_{BC}. \tag{1}$$

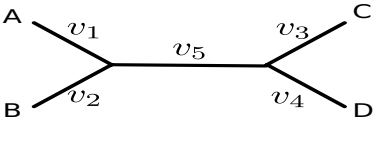


Fig. 2. The underlying tree for the four species A, B, C and D

This relation can be easily verified and it can be used to test whether the distance is additive. A weak version of this relation not requiring the equality can be used to make phylogenetic inference. The distance sum can be thought of as the sum of the distances between sibling species on the tree. Denote $d_{AB|CD} = d_{AB} + d_{CD}$, $d_{AC|BD} = d_{AC} + d_{BD}$ and $d_{AD|BC} = d_{AD} + d_{BC}$ as the distance sums for the corresponding three topologies. The inferred topology is the one with the smallest distance sum, as the others use the internal edge twice and hence have larger values.

Furthermore, the length of the internal edge can be estimated by:

$$\hat{v}_5 = \frac{1}{2}\Big[\max\{d_{AB}+d_{CD}, d_{AC}+d_{BD}, d_{AD}+d_{CB}\} - \min\{d_{AB}+d_{CD}, d_{AC}+d_{BD}, d_{AD}+d_{CB}\}\Big]. \quad (2)$$

The estimated internal length also indicates the significance of the inferred phylogeny: a small internal edge is likely to arise from noisy data and a large internal edge is likely to show the true phylogeny. In reconstructing large trees, this information is used to resolve conflicting conclusions on internal edges or subtrees.

We apply both the breakpoint distance d_{BP} and the DCJ distance d_{DCJ} for genome rearrangement data. For a topology $T = AB|CD$, the two phylogenetic score functions $S_{\mathrm{BP},T}$ and $S_{\mathrm{DCJ},T}$ denote $d_{\mathrm{BP},AB} + d_{\mathrm{BP},CD}$ and $d_{\mathrm{DCJ},AB} + d_{\mathrm{DCJ},CD}$ respectively. Phylogenetic inferences $\hat{T}_{\mathrm{BP}}$ and $\hat{T}_{\mathrm{DCJ}}$ give the topologies which minimize $S_{\mathrm{BP},T}$ and $S_{\mathrm{DCJ},T}$. In many situations, it is more convenient to compute the number of shared adjacencies a and the number of color-alternating cycles c. We use S_a and S_c to denote the corresponding phylogenetic score functions; the inferred phylogeny is the one maximizing S_a or S_c.

Under the breakpoint distance, an adjacency shared by three or all four genomes contributes once or twice to the S_a scores of all three topologies. So it does not provide phylogenetic information, and it is sufficient to only count the number of adjacencies shared by exactly two of the four genomes. Then, for each topology, say $T = AB|CD$, $S_{a,T}$ is redefined as the number of adjacencies shared only by A and B or C and D. Alekseyev and Pevzner [1] suggested using S_a to infer phylogeny.

2.3 Phylogeny Invariants

Phylogeny invariant methods directly examine patterns and their occurrence frequencies presented in the data. These methods are easier to understand on sequence data. Given 4 DNA sequences A, B, C and D, the pattern presented at site i can be, the first two nucleotides of A_i, B_i, C_i and D_i take the same value, say G, and the last two nucleotides take another value, say T. Given a topology (on which, edge lengths are unknown and irrelevant), a set of algebraic relations (*invariants*) about the expected frequencies can be derived. To infer phylogeny, we find the topology that best fits these invariants in a statistical sense.

On 4 DNA or protein sequences, we may observe $4^4 = 256$ or $20^4 = 1.6 \times 10^5$ different configurations. By the following encoding method, these configurations

can be reduced to 15 patterns[2]. For a site i, we assign x to A_i, y to the first character different from x among B_i to D_i, z to the first character different from x and y, and so on. The 15 patterns on four sequences are:

$$xxxx$$
$$xyxx, xxyx, xxxy, xyyy$$
$$xxyy, xyxy, xyyx$$
$$xxyz, xyxz, xyzx, xyyz, xyzy, xyzz$$
$$xyzw.$$

Sankoff and Blanchette [12] explored how to apply invariants methods on gene order data. The characters they used are the adjacencies in the gene orders: the successors for any gene extremity g (an endpoint of a gene, represented by a vertex in the breakpoint graph) on different gene orders. Here g plays a similar role as a site i in sequence data, and its successors in the gene orders have a similar role as nucleotides or amino acids. Sankoff and Blanchette first derived stochastic probabilities for a gene h to take position i after k random inversions. From these probabilities, they found 11 phylogenetic invariants in the case of five species.

Phylogeny invariants methods and our new methods are similar, in the way that both methods explore directly the patterns presented in the data. But phylogeny invariants methods are probabilistic methods, which concern about expected probabilities and statistics tests. Our new methods are under the parsimony framework.

3 Rearrangement Phylogenetic Patterns

In this section, we explore rearrangement phylogenetic patterns on four genomes. Under the parsimony framework, given a distance measure, a rearrangement phylogenetic pattern is a pattern that always gives smaller scores on one fixed topology but not on the others. In other words, a rearrangement phylogenetic pattern differentiates different topologies. We consider rearrangement phylogenetic patterns with regard to both the breakpoint distance and the DCJ distance, with the latter closely related to the inversion distance and its generalization, the HP distance.

Genome rearrangement problems have far more combinatorial structures than the corresponding problems on sequences, thus they are much harder to study, especially for the ones concerning three or more genomes [4,10,5,14]. This paper initiates the discussion on rearrangement phylogenetic patterns, with the emphasis on small sized patterns. As a learnt experience from adequate subgraphs for the median problem, combinatorial structures of small sizes alone may provide sufficient information on solving the problem. Furthermore, rearrangement phylogenetic patterns of small sizes will be amenable for further probabilistic analysis, which will provide statistical significance for phylogenetic inference.

[2] This number is given by the Bell number, $B(n)$, with the leading terms 1, 2, 5, 15, 52, 203 and the recurrence equation $B(n+1) = \sum_{k=0}^{n} \binom{n}{k} B(n)$.

In Subsection 3.1 we present some negative results, which exclude some combinatorial structures from being rearrangement phylogenetic patterns. Surprisingly, adjacencies shared by two or more genomes, which constitute the basis of S_{BP} or S_a, are not rearrangement phylogenetic patterns. This conclusion is also verified by the observations in Section 4. In Subsection 3.2, we introduce a class of rearrangement phylogenetic patterns and prove that they differentiate different topologies with regard to both the breakpoint distance and the DCJ distance. Then we introduce two related phylogenetic score functions and their phylogenetic inference methods.

3.1 Patterns Which Are Not Rearrangement Phylogenetic Patterns

Theorem 1. *Out of the 15 patterns discussed in Subsection 2.3, the following 6 patterns are not rearrangement phylogenetic patterns with regard to either the breakpoint distance or the DCJ distance: xxxx, xyxx, xxyx, xxxy, xyyy, and xyzw.*

Proof. Under the breakpoint distance (or the DCJ distance), the first, the next four and the last patterns contribute 5, 4 and 2 pairwise shared common adjacencies (or same numbers of color-alternating cycles), respectively. Hence these patterns do not differentiate the three topologies. □

This theorem tells us that an adjacency shared by 1, 3 or 4 genomes does not contain any phylogenetic information. The next theorem states that an adjacency shared only by 2 genomes, described by the six patterns $xxyz$, $xyxz$, $xyzx$, $xyyz$, $xyzy$, and $xyzz$, does not contain phylogenetic information either, with regard to the breakpoint distance or the DCJ distance.

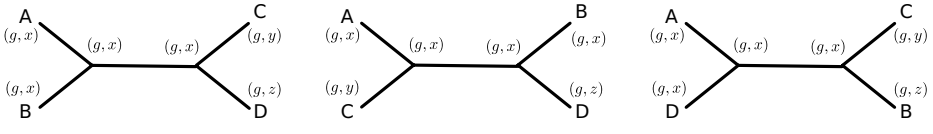


Fig. 3. A counter example showing an adjacency shared by two genomes is not a rearrangement phylogenetic pattern. The optimal configurations for the three topologies are shown next to the nodes.

Theorem 2. *A pairwise shared adjacency is not a rearrangement phylogenetic pattern.*

Proof. We prove this by showing two counter examples: a simple one and a complicated one.

Counter Example I. In Fig. 3 genome A and B share a common adjacency (g, x), genome C has the adjacency (g, y), and genome D has (g, z), where g, x, y and z are gene extremities. In the optimal solutions, the two internal nodes both contain (g, x). Hence all three topologies contribute 3 pairwise shared adjacencies or 3 color-alternating cycles.

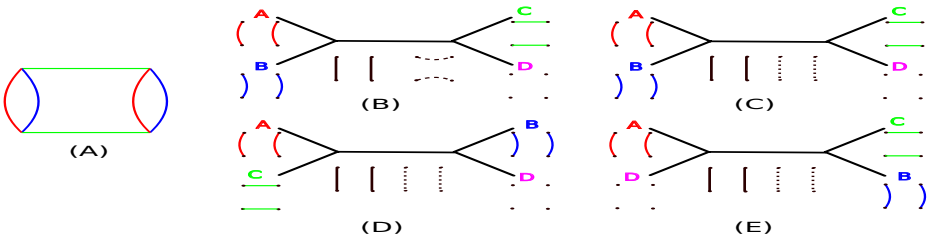


Fig. 4. Another counter example showing adjacencies shared by two genomes are not rearrangement phylogenetic patterns. Genome A and B share two adjacencies and Genome C form two different adjacencies on the same four extremities (A). (B), (C), (D) and (E) show optimal configurations of the internal nodes for the three topologies $AB|CD$, $AC|BD$ and $AD|BC$. Note that $AB|CD$ has two optimal configurations. The configurations are shown next to the nodes. Solid black edges represent adjacencies for the first internal node and dashed black edges represent adjacencies for the second internal node.

Counter Example II. In Fig. 4.(b) genome A and B share two common adjacencies and genome C has two different adjacencies on these four gene extremities. Subfigures (B)–(E) show the configurations taken by the two internal nodes, where the optimality of these configurations can be easily proved (either by the adequate subgraph theory or similar techniques applicable for the breakpoint distance). Note that for the first topology $AB|CD$, there are two locally optimal configurations, and at least one of them is part of a global optimal configuration. All three topologies contribute 6 pairwise shared adjacencies and 7 color-alternating cycles.

Therefore, adjacencies shared by only two genomes are not rearrangement phylogenetic patterns. □

3.2 Rearrangement Phylogenetic Patterns and Contrasting Shared Adjacencies

A *contrasting shared adjacency*, is when a gene extremity g forms one adjacency on two genomes and another adjacency on the other two genomes. Related to the patterns discussed in Subsection 2.3, a contrasting shared adjacency corresponds to $xxyy, xyxy, xyyx$.

In the two counter examples used to prove Theorem 2, the two internal nodes take the same configuration in the optimal configurations. When this happens on all three topologies, the two scores S_a and S_c will always be the same. For a pattern to be phylogenetic, the internal nodes have to be different in the optimal configurations. The next theorem shows that a contrasting shared adjacency has such a property.

Theorem 3. *If a contrasting shared adjacency forms an adjacency* (g, x) *on genome A and B and* (g, y) *on genome C and D, then we have the following conclusions:*

1. *the configurations of the two internal node on the topology $AB|CD$ are different;*
2. *the contrasting shared adjacency differentiates the three topologies and $AB|CD$ is the preferred one, with regard to both the breakpoint distance and the DCJ distance;*
3. *the contrasting shared adjacency is a rearrangement phylogenetic pattern.*

Proof. Fig. 5 shows the optimal configurations of the internal nodes for all three topologies. The optimality can be easily shown by the adequate subgraph theory (or similar techniques for the breakpoint distance) and so is the fact that the two nodes on $AB|CD$ take different configurations. Furthermore, on $AB|CD$ there are 4 pairwise shared adjacencies and color-alternating cycles; while on the other two topologies, there are only 3 pairwise shared adjacencies and color-alternating cycles. Hence a contrasting shared adjacency is a rearrangement phylogenetic pattern and $AB|CD$ is the preferred topology with regard to both the breakpoint distance and the DCJ distance. □

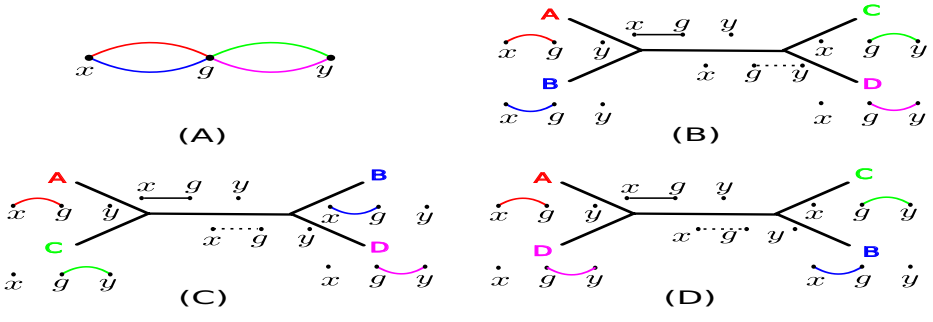


Fig. 5. Contrasting shared adjacency, where Genome A and B share one adjacency (g, x) and Genome C and D share another adjacency (g, y). (B), (C) and (D) show optimal configurations of the internal nodes for the three topologies $AB|CD$, $AC|BD$ and $AD|BC$. The configurations are shown next to the nodes. Solid black edges represent adjacencies for the first internal node and dashed black edges represent adjacencies for the second internal node.

3.3 Multi-paths, Multi-cycles and Two New Phylogenetic Score Functions

A multi-path (a multi-cycle) is a path (a cycle) only consisting of multi-edges.The size of a multi-path (a cycle) is the number of multi-edges it contains, which is denoted by l. We say a multi-path (a multi-cycle) is consistent with a topology $T = AB|CD$, if its adjacencies are pairwise shared by either genomes A, B or genomes C, D. In the breakpoint distance based method, a multi-path (a multi-cycle) of size l contributes l units toward $S_{\mathrm{BP},T}$.

We have shown, in proving Theorem 3, two contrasting adjacencies (g, x) and (g, y) together force the two internal nodes to take different configurations. In

counting how many units of phylogenetic information a multi-path (a multi-cycle) contains, we argue that it should be equal to the maximum number of non-overlapping[3] contrasting shared adjacencies contained in that path (cycle), which is $\lfloor \frac{l}{2} \rfloor$ units of phylogenetic information[4].

We define the phylogenetic score function $S_{\mathrm{CA},T}$ as the maximum number of non-overlapping contrasting shared adjacencies consistent with the topology T. This can be explicitly expressed as:

$$S_{\mathrm{CA},T} = \sum_{p \text{ consistent with } T} \left\lfloor \frac{|p|}{2} \right\rfloor, \tag{3}$$

where p is a multi-path or multi-cycle and $|p|$ is its size. And the corresponding phylogenetic inference $\hat{T}_{\mathrm{CA}}$ is the topology T with the maximum $S_{\mathrm{CA},T}$.

Under the DCJ distance, we have the following result for multi-cycles. This result means, a multi-cycle of size $l = 2k$ (always an even number) only contributes $k - 1$ units of phylogenetic information, instead of k.

Theorem 4. *A multi-cycle of size $l = 2k$ contributes $4k + 1$ color-alternating cycles on its consistent topology and contributes $3k + 2$ color-alternating cycles on other two topologies.*

As the difference in the number of cycles is $k - 1$, it will be better to assign only $k - 1$ units of phylogenetic information to the multi-cycle.

Remark 1. For a multi-cycle with $2k$ multi-edges, if we treat these multi-edges as simple edges, the DCJ distance defined on this cycle is also $k - 1$. Although the two $k - 1$s coincide, they come from two different problems. If the distance measure is additive, the distance on the wrong topology will count the internal edge twice, and the extra counting explains the difference in the total distances.

With this modification, we have a new phylogenetic score function $S_{\mathrm{MCA},T}$:

$$S_{\mathrm{MCA},T} = \sum_{\text{multi-path } p \text{ consistent with } T} \left\lfloor \frac{|p|}{2} \right\rfloor + \sum_{\text{multi-cycle } p \text{ consistent with } T} \frac{|p|}{2} - 1. \tag{4}$$

And the corresponding phylogenetic inference $\hat{T}_{\mathrm{MCA}}$ is just the topology with the largest S_{MCA}.

4 Testing the Accuracies of Phylogenetic Inference Methods on Simulated Data

We generated various groups of simulation data to compare the accuracies of various phylogenetic inference methods. Simulated genomes were generated according to the tree shown in Fig. 2 but with only two parameters: $e_1 = v_5$ and

[3] Non-edge-overlapping, to be more accurate.

[4] The notation $\lfloor x \rfloor$ denotes the largest integer which is no larger than x.

$e_2 = v_1 = v_2 = v_3 = v_4$. The edge lengths e_1 and e_2 denote the number of random inversions applied. In the first group of data, e_1 and e_2 varied among 5, 10, 20, 30 and 40. In the second group of data, to study the effect of short internal edges, we let e_1 take very short lengths: 1, 2, 3, 4 and 5, and let e_2 take values among 5, 10, 20, 30, 40, 50 and 60. We generated 1,000 datasets for each parameter combination; thus we generated a total of 55,000 simulation datasets. We used genomes of single circular chromosome containing 200 genes, so that we finished the whole test very quickly: for each dataset, all phylogenetic inferences finished within a couple of seconds.

4.1 Comparing Accuracies of Various Inference Methods

We compared the accuracies of five phylogenetic inference methods ($\hat{T}_{\text{MCA}}$, $\hat{T}_{\text{CA}}$, $\hat{T}_{\text{GASTS}}$, $\hat{T}_{\text{BP}}$ and $\hat{T}_{\text{DCJ}}$) on the first group of data. We used a strict criteria to calculate accuracies: a method makes a correct inference only when the true topology is given as the unique result. For example, if the inference of a method contains two topologies, even with the true one included, we still treat this inference as wrong. Under this strict criteria, accuracies would appear low, but the comparison results are not affected.

Table 1 shows the comparison results for $e_1 = 5$, 10 and 20. Results for $e_1 =$ 30 and 40 are not shown, as the accuracies were all 100%. Overall, these methods were very accurate, except when the internal edge was small and the outer edges were large. With no surprise, the full parsimony method $\hat{T}_{\text{GASTS}}$ had the best accuracy, leading the second best method by up to 7 percentage points. The two new phylogenetic inference methods $\hat{T}_{\text{MCA}}$ and $\hat{T}_{\text{CA}}$ had the second best accuracies, better than the next best method by up to another 7 percentage points. Among the two, $\hat{T}_{\text{MCA}}$ was slightly better than $\hat{T}_{\text{CA}}$. This shows that special modification regarding the DCJ distance measure has a marginal advantage. The next best method was $\hat{T}_{\text{BP}}$, leading $\hat{T}_{\text{DCJ}}$ by nearly 5 percentage points. However, there was a trend that their difference decreases as the internal edge length increases.

4.2 Tests on Challenging Cases with Short Internal Edges and Long Outer Edges

Fig. 6 shows the comparison results on cases with e_1 =1,2,3 and 5, e_2 =5, 10, 20, 30, 40, 50 and 60. Results for $e_1 = 4$ are not shown, as they are similar to the ones for $e_1 = 5$. As the cases with short internal edges and long outer edges are very difficult, it is not surprising to see accuracies as low as 40%. On the contrary, it is a little surprising to see the two new methods $\hat{T}_{\text{MCA}}$ and $\hat{T}_{\text{CA}}$ had nearly 50% accuracies on the extremely difficult cases where the outer edges were 60 times the length of the internal edge. As the chance to make a correct inference by chance is no larger than 33% (as ties are not regarded as correct), these results were fairly good.

Another obvious observation is, the accuracies increase quickly with the length of the internal edge length. When $e_1 = 1$, the accuracies started from 90%

Table 1. Comparison of 5 phylogenetic inference methods on the first group of datasets. The internal edge length e_1 and the outer edge length e_2 took values among 5, 10, 20, 30 and 40 random inversion. Accuracies are measured strictly where any tie is treated as incorrect.

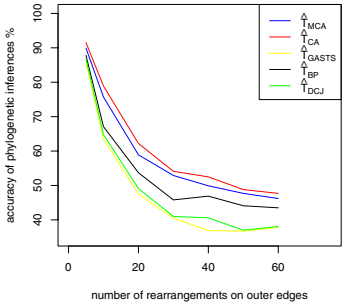
e_1	e_2	$\hat{T}_{\mathrm{MCA}}$	$\hat{T}_{\mathrm{CA}}$	$\hat{T}_{\mathrm{GASTS}}$	$\hat{T}_{\mathrm{BP}}$	$\hat{T}_{\mathrm{DCJ}}$
5	5	100	100	100	99.9	99.7
5	10	100	99.9	100	99.6	98.8
5	20	98.1	97.0	98.5	94.9	90.6
5	30	88.0	88.2	93.0	81.5	78.6
5	40	80.2	79.9	87.7	72.4	68.8
10	5	100	100	100	100	100
10	10	100	100	100	100	100
10	20	100	99.9	100	99.9	99.2
10	30	99.0	98.5	100	97.7	95.8
10	40	94.8	94.1	99.1	93.2	88.7
20	20	100	100	100	100	100
20	30	100	100	100	100	100
20	40	99.6	99.6	100	99.3	99.6

($e_2 = 5$) and quickly dropped to 40%–50% ($e_2 = 60$); when $e_1 = 5$, they increase to 100% ($e_2 = 5$) or 55%–75% ($e_2 = 60$).

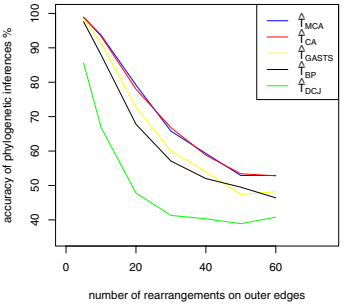
The two new methods $\hat{T}_{\mathrm{MCA}}$ and $\hat{T}_{\mathrm{CA}}$ had the best overall performance. They remained the best, until e_1 increased to 4 and 5, where they were only second to the full parsimony method $\hat{T}_{\mathrm{GASTS}}$. $\hat{T}_{\mathrm{GASTS}}$ performed poorly when $e_1 = 1$, fairly when $e_1 = 2$, and well when $e_1 \geq 3$. $\hat{T}_{\mathrm{BP}}$ had decent performance, better than $\hat{T}_{\mathrm{DCJ}}$. We can explain the fact that the three adjacency-related methods had the best performance for $e_1 = 1$ by the following argument. The single inversion on the internal edge leaves two breakpoints. The trace of these breakpoints may be erased by random inversions on the four outer edges. In order to correctly reconstruct this inversion, DCJ based methods need both breakpoints to remain. Adjacency-related methods can work even if only one breakpoint remains.

4.3 How Do the Phylogenetic Score Functions Correlate to the Number of Rearrangements

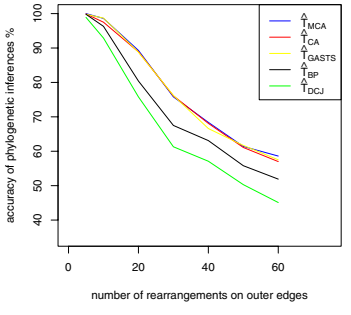
On the first group of datasets, we further investigated how the phylogenetic score functions correlate to the number of rearrangements. The number of rearrangements on the true topology can be easily calculated during simulation, however this is not feasible on the other two topologies. As GASTS can very accurately estimates the number of rearrangements on a given topology, we use the GASTS scores to approximate the real numbers of rearrangements. We then calculated how the other phylogenetic score functions deviated from the GASTS scores. We denote the three topologies as T_i with $i = 1$, 2 and 3, where T_1 is the true topology. We calculated the deviations for S_{MCA}, S_{CA}, S_{BP}, $\frac{S_{\mathrm{BP}}}{2}$, S_{DCJ} and $\frac{S_{\mathrm{DCJ}}}{2}$. The two fractional scores are considered, because of the factor of 2 in Equation 2.



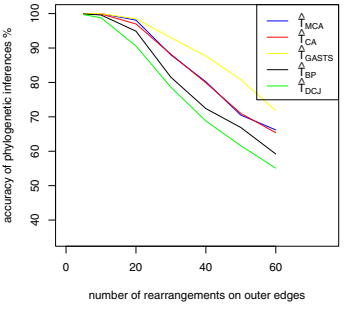
(a) 1 rearrangement on the internal edge



(b) 2 rearrangements on the internal edge



(c) 3 rearrangements on the internal edge



(d) 5 rearrangements on the internal edge

Fig. 6. Comparison of 5 phylogenetic inference methods on datasets with very small internal edges. The length e_1 of these edges varied among 1, 2, 3, 4 and 5 random inversions. Accuracies are measured strictly where any tie is treated as incorrect.

Let P represent any of the above 6 phylogenetic score functions. The deviation diff(P) calculates the discrepancy between the difference in S_{P} and the difference in S_{GASTS} over different topologies on the same datasets. It is given by the following formula:

$$\begin{aligned}\mathrm{diff}(P) = \frac{1}{2000}\sum_{i=1}^{1000}\{&|\,(S_{P,T_1} - S_{P,T_2}) - (S_{\mathrm{GASS},T_1} - S_{\mathrm{GASS},T_2})\,| \\ &+|\,(S_{P,T_1} - S_{P,T_3}) - (S_{\mathrm{GASS},T_1} - S_{\mathrm{GASS},T_3})\,|\}. \end{aligned} \quad (5)$$

Table 2 shows the results. The two score functions S_{BP} and S_{DCJ} deviated significantly from the GASTS scores; but their fractional versions had much better performance. $\frac{S_{\mathrm{DCJ}}}{2}$ had very small deviations, only second to S_{MCA}. S_{MCA} had an excellent performance, with very small deviations from the GASTS scores,

Table 2. Deviations between six phylogenetic score functions and the GASTS score. The deviation is defined by Equation 5. As the GASTS score well approximates the number of rearrangements for any given tree, these deviations show how well these six phylogenetic score functions reflect the variation of the number of rearrangements over different topologies. The deviations for S_{MCA} are in bold font, if they are smaller than 2. e_1 and e_2 are the lengths of internal and outer edges, respectively.

e_1	e_2	diff(S_{MCA})	diff(S_{CA})	diff(S_{BP})	diff($\frac{S_{\mathrm{BP}}}{2}$)	diff(S_{DCJ})	diff($\frac{S_{\mathrm{DCJ}}}{2}$)
5	5	**0.1435**	3.1985	12.465	3.8337	4.8845	0.4315
5	10	**0.3970**	2.2015	10.664	3.089	4.8190	0.8970
5	20	**1.0155**	1.5345	8.4615	2.6680	5.454	1.7460
5	30	**1.5490**	1.6580	7.2310	2.6550	6.1460	2.4405
5	40	2.1055	2.1435	6.6230	2.8687	7.0365	2.9890
10	5	**0.2410**	6.0200	24.127	7.2652	9.6765	0.5550
10	10	**0.6345**	3.8970	20.876	5.7495	9.4910	1.0505
10	20	**1.6520**	1.8600	14.825	3.4027	8.9575	1.8160
20	5	**0.4670**	10.687	44.850	12.969	19.125	0.7430
20	10	**1.2020**	6.2245	38.051	9.7300	18.644	1.2835
20	20	3.1775	2.4450	26.761	4.8037	17.483	2.0010
30	5	**0.6205**	13.855	62.395	17.191	28.235	0.8870
30	10	**1.7490**	7.9415	53.082	12.777	27.477	1.3840
40	5	**0.9020**	16.229	77.793	20.449	37.253	1.0150
40	10	2.4585	8.7070	65.984	14.852	36.190	1.5275

especially when the total number of events were small. This shows that S_{MCA} can well explain the difference in the number of rearrangements over different topologies. And the reason that its deviations increased is: the gap between S_{MCA} and S_{GASTS} was mainly caused by large rearrangement phylogenetic patterns, which occurred more frequently when the number of events got larger; these large rearrangement phylogenetic patterns are not considered in our paper. The closely related score function S_{CA} had much worse performance and this justifies our special consideration on multi-cycles.

5 Conclusion and Future Work

In this paper, we explore rearrangement phylogenetic patterns for the genome rearrangement quartet problem. A rearrangement phylogenetic pattern is a combinatorial structure, which contains phylogenetic information, and by examining their occurrences, phylogenetic inferences can be made. As the first solid study of this subject, we prove what are or are not genome rearrangement phylogenetic patterns, with regard to the breakpoint distance and the DCJ distance and under the parsimony framework. We define two phylogenetic score functions as the numbers of the observed rearrangement phylogenetic patterns, and based on them we design two phylogenetic inference methods. Tested on simulated data, these new methods demonstrated good accuracies, only second to the full parsimony method, and remarkable robustness when the internal edge of the tree is

extremely short. All these observations imply that, rearrangement phylogenetic patterns indeed carry a significant amount of phylogenetic information and this is a promising alternative approach to study phylogenetic problems.

There are many open problems to explore. On the quartet problem, discovering of more rearrangement phylogenetic patterns of larger sizes or the patterns specialized for linear chromosomes will certainly increase the power for phylogenetic inferences; analyzing the probabilistic properties of these patterns will allow us to develop statistic tests for the significance of the inferences. The topic of discovering rearrangement phylogenetic patterns can go beyond four genomes. The question of how to use these rearrangement phylogenetic patterns and their probabilistic properties to design efficient and accurate methods for phylogenetic problems with large numbers of taxa, will be very interesting but challenging.

Acknowledgments

The author would like to thank the anonymous referees for their helpful suggestions on writing this paper.

References

1. Alekseyev, M.A., Pevzner, P.: Breakpoint Graphs and Ancestral Genome Reconstructions. Genome Res. 19, 943–957 (2009)
2. Bergeron, A., Mixtacki, J., Stoye, J.: A unifying view of genome rearrangements. In: Bücher, P., Moret, B.M.E. (eds.) WABI 2006. LNCS (LNBI), vol. 4175, pp. 163–173. Springer, Heidelberg (2006)
3. Bourque, G., Pevzner, P.: Genome-scale evolution: Reconstructing gene orders in the ancestral species. Genome Res. 12, 26–36 (2002)
4. Bryant, D.: The complexity of the breakpoint median problem. TR CRM-2579. Centre de recherches mathématiques, Université de Montréal (1998)
5. Caprara, A.: The reversal median problem. INFORMS J. Comput. 15, 93–113 (2003)
6. Hannenhalli, S., Pevzner, P.: Transforming men into mice (polynomial algorithm for genomic distance problem. In: Proc. 43rd IEEE Symp. on Foudations of Computer Science FOCS 1995, pp. 581–592. IEEE Computer Soc., Los Alamitos (1995)
7. Hannenhalli, S., Pevzner, P.: Transforming cabbage into turnip: Polynomial algorithm for sorting signed permutations by reversals. JACM 46, 1–27 (1999)
8. Moret, B., Siepel, A., 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, p. 521. Springer, Heidelberg (2002)
9. Murphy, W.J., Larkin, D.M., van der Wind, A.E., Bourque, G., Tesler, G., Auvil, L., Beever, J.E., Chowdhary, B.P., Galibert, F., Gatzke, L., Hitte, C., Meyers, S.N., Milan, D., Ostrander, E.A., Pape, G., Parker, H.G., Raudsepp, T., Rogatcheva, M.B., Schook, L.B., Skow, L.C., Welge, M., Womack, J.E., O'Brien, S.J., Pevzner, P.A., Lewin, H.A.: Dynamics of Mammalian Chromosome Evolution Inferred from Multispecies Comparative Maps. Science 309(5734), 613–617 (2005)
10. Pe'er, I., Shamir, R.: The median problems for breakpoints are np-complete. In: Burkhardt, H., Neumann, B. (eds.) ECCV 1998. LNCS, vol. 1407. Springer, Heidelberg (1998)

11. Sankoff, D., Blanchette, M.: Multiple genome rearrangement and breakpoint phylogeny. J. Comput. Biol. 5, 555–570 (1998)
12. Sankoff, D., Blanchette, M.: Phylogenetic invariants for genome rearrangements. Journal of computational biology: a journal of computational molecular cell biology 6(3-4), 431–445 (1999)
13. Sankoff, D., Zheng, C., Wall, P.K., DePamphilis, C., Leebens-Mack, J., Albert, V.A.: Towards improved reconstruction of ancestral gene order in angiosperm phylogeny. Journal of Computational Biology: A Journal of Computational Molecular Cell Biology 16(10), 1353–1367 (2009)
14. Tannier, E., Zheng, C., Sankoff, D.: Multichromosomal median and halving problems. In: Crandall, K.A., Lagergren, J. (eds.) WABI 2008. LNCS (LNBI), vol. 5251, pp. 1–13. Springer, Heidelberg (2008)
15. Xu, A.W.: DCJ median problems on linear multichromosomal genomes: Graph representation and fast exact solutions. In: Ciccarelli, F.D., Miklós, I. (eds.) RECOMB-CG 2009. LNCS, vol. 5817, pp. 70–83. Springer, Heidelberg (2009)
16. Xu, A.W., Moret, B.M.: Genome rearrangement analysis on thousands of taxa with thousands of synteny blocks (submitted, 2010)
17. Xu, A.W., Sankoff, D.: Decompositions of multiple breakpoint graphs and rapid exact solutions to the median problem. In: Crandall, K.A., Lagergren, J. (eds.) WABI 2008. LNCS (LNBI), vol. 5251, pp. 25–37. Springer, Heidelberg (2008)
18. Yancopoulos, S., Attie, O., Friedberg, R.: Efficient sorting of genomic permutations by translocation, inversion and block interchange. Bioinformatics 21, 3340–3346 (2005)