Exemplar or Matching: Modeling DCJ Problems with Unequal Content Genome Data

Zhaoming Yin², Jijun Tang¹,³*, Stephen W. Schaeffer⁴ and David A. Bader²*

¹ School of Computer Science and Technology, Tianjin University, China
² School of Computational Science and Engineering, Georgia Institute of Technology, USA
³ Dept. of Computer Science and Engineering, University of South Carolina, USA
⁴ The Huck Institutes of Life Sciences, Pennsylvania State University, USA

Abstract. The edit distance under the DCJ model can be computed in linear time for genomes with equal content or with Indels. But it becomes NP-Hard in the presence of duplications, a problem largely unsolved especially when Indels are considered. In this paper, we compare two mainstream methods to deal with duplications and associate them with Indels: one by deletion, namely DCJ-Indel-Exemplar distance; versus the other by gene matching, namely DCJ-Indel-Matching distance. We design branch-and-bound algorithms with set of optimization methods to compute exact distances for both. Furthermore, median problems are discussed in alignment with both of these distance methods, which are to find a median genome that minimizes distances between itself and three given genomes. Lin-Kernighan (LK) heuristic is leveraged and powered up by sub-graph decomposition and search space reduction technologies to handle median computation. A wide range of experiments are conducted on synthetic data sets and real data sets to show pros and cons of these two distance metrics per se, as well as putting them in the median computation scenario.

Keywords: Genome Rearrangement, Double-cut and Join (DCJ), Lin-Kernighan Heuristic.

1 Introduction

Over the last years, many distance metrics have been introduced to calculate the dissimilarity between two genomes by genome rearrangement [2,5,30]. Among them, DCJ distance is largely studied in recent years due to its capability to model various forms of rearrangement events, with a cheap cost of linear time computation. However, when considering duplications, the distance computation becomes NP-hard [10] and APX-hard [12] for various distance models. There are two approaches to treat duplications, both are targeted at removing duplicated genes, so that existing linear algorithms can be utilized subsequently.

* Corresponding Authors
The first approach identifies the so called exemplar genes \[23\] in order to retain one copy gene in each duplicated gene family, while the other assigns one-to-one matching to every duplicated genes in each gene family \[24, 25\]. Situated in the context of duplications, gene insertion and deletion (\textit{Indels}) are also important rearrangement events that results in unequal contents \[8\]. Pioneer works were conducted to study the sorting and distance computation by reversals with \textit{Indels} \[17\]. Later on, the \textit{DCJ-Indel} distance metric was introduced to take advantages of the \textit{DCJ} model. Braga et al \[7\] proposed the first framework to compute the \textit{DCJ-Indel} distance; Compeau later simplified the problem with a much more elegant distance formula \[13\]. In this paper, we adapt the previous research results to design algorithms that procure the ability to handle both duplications and \textit{Indels} when computing \textit{DCJ} distance.

As evolutionary analysis generally involves more than two species, it is necessary to extend the above distances to deal with multiple genomes. Since three species form the smallest evolutionary tree, it is critical to study the median problem, which is to construct a genome that minimizes the sum of distances from itself to the three input genomes \[6, 15\]. The median problem is \textit{NP}-hard under most distance metrics \[4, 9, 21, 27\]. Several exact algorithms have been implemented to solve the \textit{DCJ} median problems on both circular \[27, 29\] and linear chromosomes \[26, 28\]. Some heuristics are brought forth to improve the speed of median computation, such as linear programming (\textit{LP}) \[9\], local search \[16\], evolutionary programming \[14\], or simply searching on one promising direction \[22\]. All these algorithms are intended for solving median problems with equal content genomes, which are highly unrealistic in practice. In this paper, we implement a Lin-Kernighan heuristic leveraging the aforementioned two distance metrics to compute \textit{DCJ} median when duplications and \textit{Indels} are considered.

2 Background

2.1 Genome Rearrangement Events and their Graph Representations

\textbf{Genome Rearrangement Events} The ordering of a genome can be changed through rearrangement events such as reversals and transpositions. Fig 1 shows examples of different events of a single chromosome (1 -2 3 4 -5 6 7). In the examples, we use signed numbers to represent different genes and their orientations. Genome rearrangement events involve with multiple combinatorial optimization problems and graph representation is common to abstract these problems. In this part, we will address the foundations of using the breakpoint graph to abstract genome rearrangement events.

\textbf{Breakpoint Graph} Given an alphabet \(\mathcal{A}\), two genomes \(\Gamma\) and \(\Pi\) are represented by two strings of signed (+ or −) numbers (representing genes) from \(\mathcal{A}\). Each gene \(a \in \mathcal{A}\) is represented by a pair of vertices head \(a_h\) and tail \(a_t\); If \(a\) is positive
Fig. 1. Example of different rearrangement events.

(a) Example of BPG
(b) Example of DCJ

Fig. 2. Examples of BPG; and DCJ operations.

$a_h$ is putted in front of $a_t$, otherwise $a_t$ is putted in front of $a_h$. For $a, b \in A$, if $a, b \in \Gamma$ and are adjacent to each other, their adjacent vertices will be connected by an edge. For a telomere genes, if it exists in a circular chromosome, two end vertices will be connected by an edge; if it exists in a linear chromosome, two end vertices will be connected to a special vertex called CAP vertex. If we use one type of edges to represent adjacencies of gene order $\Gamma$ and another type of edges to represent adjacencies of gene order $\Pi$, the resulting graph with two types of edges is called breakpoint graph (BPG). Fig 2(a) shows the BPG for gene order $\Gamma (1, -2, 3, -6, 5)$ (edge type: solid edges) which has one circular chromosome and $\Pi (1, 2, 3, 7, 4)$ (edge type: dashed edges) which has one linear chromosome.

**DCJ operation** Double-cut and join (DCJ) operations are able to simulate all rearrangement events. In a BPG, these operations cut two edges (within one genome) and rejoin them using two possible combinations of end vertices (shown in Fig 2(b)).

2.2 Distance computation

**DCJ distance** DCJ distance of genomes with the same content can be easily calculated by enumerating the number of cycles/paths in the BPG [30], which is of linear complexity.

**DCJ-Indel distance** When Indels are introduced in BPG, with two genomes $\Gamma$ and $\Pi$, the vertices and edges of a closed walk form a cycle. In Fig 2(a) the
walk \((1^t, (1^t; 2^h), 2^h, (2^h; 3^h), 3^h, (3^h; 2^t), 2^t, (2^t; 1^t), 1^t)\) is a cycle. A vertex \(v\) is \(\pi\)-open (\(\gamma\)-open) if \(v \notin \Gamma (v \notin \Pi)\). An unclosed walk in \(BPG\) is a path. Based on different kinds of ends points of paths, we can classify paths into different types. If the two ends of a path are \(CAP\) vertices, we simply denote this path as \(p^0\). If a path is ended by one open vertex and one \(CAP\), we denote it as \(p^\pi (p^\gamma)\). If a path is ended by two open vertices, we denote it by the types of its two open vertices: for instance, \(p^\pi,\gamma\) represents a path that ends with a \(\pi\)-open vertex and a \(\gamma\)-open vertex. In Fig 2(a), the walk \((5^t, (5^t; 1^h), 1^h, (1^h; CAP), CAP)\) is a \(p^\gamma\) path and the walk \((6^t, (6^t; 3^t), 3^t, (3^t; 7^h), 7^h)\) is a \(p^\gamma,\pi\) path. A path is even (odd), if it contains even (odd) number of edges. In [13], if \(|A| = N\) the \(DCJ\) distance between two genomes with Indels but without duplications is calculated by equation (1). We call this distance \(DCJ-Indel\) distance. From this equation, we can easily get the \(DCJ-Indel\) distance between \(\Gamma\) and \(\Pi\) in Fig 2(a) as 4.

\[
d_{\text{indel}}(\Gamma, \Pi) = N - |c| + |p^{\pi,\pi}| + |p^{\gamma,\gamma}| + |p^{\pi,\gamma}| + \delta
\]

Where \(\delta = 1\) only if \(|p^{\pi,\gamma}|\) is odd and either \(|p^{\pi,\gamma}| > |p^{\gamma,\gamma}|\) or \(|p^{\pi,\gamma}| < |p^{\gamma,\gamma}|\). Otherwise, \(\delta = 0\).

**DCJ-Exemplar(Matching) distance** There are in general two approaches to cope with duplicated genes. One is by removing all but keeping one copy in a gene family to generate an exemplar pair [23] and the other is by relabeling duplicated genes to ensure that every duplicated gene has unique number [24, 25]. Both of these two distances can be computed with \(BPG\) using branch-and-bound methods. For both of the distance metrics, the upper bound can be easily derived by assigning an arbitrary mapping to two genomes then computing their mutual distance. In paper [23] regarding exemplar distance, it’s proved that by removing all occurrences of unfixed duplicated gene families, the resulting distance is monotony decreasing, hence the resulting distance can be served as a lower bound. In paper [11] regarding matching distance, the authors proposed a way for computing lower bounds by measuring the number of breakpoints between two genomes, which might not directly imply the lower bound between genomes with Indels. However, it is still possible to use this method to find a ‘relaxed’ lower bound.

**Distance Estimation** Note that mathematically optimized distance might not reflect the true number of biological events, thus several estimation methods such as \(EDE\) or \(IEBP\) are used to rescale these computed distances [19] to better fit true evolutionary history.

### 2.3 Median Computation

If there are three given genomes, the graph constructed by pre-defined \(BPG\) rule is called a Multiple Breakpoint Graph (\(MBG\)). Figure 3(a) shows an ex-
ample of $MBG$ with three input genomes. When there are only equal content genomes, the DCJ median problem can be briefly described by finding a maximum matching (which is called 0-matching) in $MBG$. Figure 3(b) shows an example of 0-matching which is represented by gray edges. In [29], it is proven that a type of sub-graph called adequate sub-graph (AS) could be used to decompose the graph with edge shrinking operations, which are shown in Figure 3(c).

Unfortunately, there is no branch-and-bound based median algorithm that deals with unequal content genomes. In the following section, we will show that it is actually difficult to design such algorithm.

3 Approaches

3.1 Proposed Distance Metrics

We have discussed DCJ, DCJ-Indel and DCJ-Exemplar(Matching) distances, here we formally define the DCJ-Indel-Exemplar(Matching) distances as follows:

**Definition 1.** An exemplar string is constructed by deleting all but one occurrence of each gene family. Among all possible exemplar strings, the minimum distance that one exemplar string returns is the DCJ-Indel-Exemplar distance.

**Definition 2.** A matching string is constructed by assigning a one-to-one mapping to each occurrence of genes in a gene family and relabel them to distinct markers. Among all possible matching strings, the minimum distance that one matching string returns is the DCJ-Indel-Matching distance.
Figure 4 shows examples of $BPG$ representation of exemplar mapping from genome $\Gamma_1$ (1, -2, 3, 2, -6, 5) and genome $\Pi_1$ (1, 2, 3, 7, 2, 4) to $\Gamma_2$ (1, 3, 2, -6, 5) and genome $\Pi_2$ (1, 3, 7, 2, 4), and a matching that mapping from genome $\Gamma_1$ (1, -2, 3, 2, -6, 5) and genome $\Pi_1$ (1, 2, 3, 7, 2, 4) to $\Gamma_2$ (1, -2, 3, 2', -6, 5) and genome $\Pi_2$ (1, 2', 3, 7, 2, 4).

We can use branch-and-bound methods which are applied in $DCJ$-$Exemplar$ ($Matching$) distances to solve these two distances.

### 3.2 Optimization Methods

**Optimal Assignments** Although branch-and-bound algorithms are based on enumerating the number of cycles/path in $BPG$, it is not necessary to enumerate every component in the graph, as both $[11,25]$ indicated that there are some specific patterns in $BPG$ which can be fixed before the distance computation. In this paper, we will extend their result in our optimization methods for $DCJ$-$Indel$-$Exemplar$ ($Matching$) distances.

To begin with, we define some terms for future explanation. There are two categories of vertices in a $BPG$: one connects exactly one edge of each edge type (in this paper edge types are expressed by such as dotted, dashed edges etc.), they are called *regular* vertices; the other connects fewer or more than one edges of each edge type, they are called *irregular* vertices. A subgraph in a $BPG$ that only
contains regular vertices is defined as regular subgraph, while one that contains irregular vertices is defined as irregular subgraph. In BPG with two genomes $\Gamma$ and $\Pi$, vertices and edges of a closed walk form a cycle.

**Theorem 1.** In a BPG, an irregular subgraph which is a cycle of length 2 can be fixed before computation without losing accuracy.

**Proof.** Without loss of generality, the proof is sound for both DCJ-Indel-Exemplar and DCJ-Indel-Matching distances. We prove the theorem under two cases:

1. for the subgraph in the component which only contains cycles, this is a case that is exactly the same as mentioned in [25], proof.

2. for the subgraph in the component which contains paths, since no type of the paths has count more than one (which is the count of a cycle), following the similar proof strategy in [25], we can get the same conclusion. □

**Adopting Morph Graph Methods to Condense BPG** If a gene family has multiple copies of the gene, its corresponding two vertices (head and tail) in the BPG will have degree of more than one. In contrary, vertex representations of those singleton genes always have degree of one or zero. Once an ‘exemplar’ or ‘matching’ is fixed, only edges incident to vertices that have degree of more than one have been changed. We can view the computation of exemplar or matching distance as the process of morphing (or streaming) [32] the BPG in order to find an ad hoc shape of the BPG that achieves optimality. Following this hint, we can bridge out all vertices that are stable and just investigate these dynamically changing vertices without losing accuracy. Suppose there are $V$ vertices in the BPG, where $V_s$ are stable and $V_d$ are dynamic, the asymptotic speedup for this morph BPG strategy will be $O\left(\frac{V}{V_d}\right)$.

**Harness the Power of Divide-and-Conquer Approach to Reduce the Problem Space** In the paper by Nguyen et al [20], the authors proposed a divide and conquer method to quickly calculate the exemplar distance. Inspired by their idea, we propose the following divide-and-conquer method to compute the above two distances based on the BPG. We have the follow observation:

**Theorem 2.** The DCJ-Indel-Exemplar (Matching) distance is optimal iff the choices of exemplar edges (cycle decomposition) in each connected components of BPG are optimal.

**Proof.** Since it’s obvious that for regular connected component of BPG, there is only one choice of edges, the proof under this case is trivial. For irregular connected component of BPG, we prove by contrary: suppose there is another edge selection that can result in a better distance, based on the corresponding BPG, there must be at least one connected component that has a better edge selection, replacing it with a better edge selection will result in a better distance, which violates the assumption. □
Algorithm 1: DCJ\text{IndelExem}(\text{Matching})Distance

Input: $G_1$ and $G_2$

Output: Minimum distance $d$

1. optimization methods on $G_1$, $G_2$;
2. $G_1', G_2' \leftarrow$ randomly init exemplar(matching) of all duplicated genes of $G_1$, $G_2$;
3. $G_1^*, G_2^* \leftarrow$ remove all duplicated genes of $G_1$, $G_2$;
4. $min\_ub \leftarrow DCJ\text{Indel}(G_1', G_2')$;
5. $min\_lb \leftarrow DCJ\text{Indel}(G_1^*, G_2^*)$;
6. Init search list $L$ of size $min\_ub - min\_lb$ and insert $G_1, G_2$;
7. while $min\_ub > min\_lb$ do
8. $G_1^+, G_2^+ \leftarrow$ pop from $L[min\_lb]$;
9. for pair $\in$ all mappings of next available duplicated gene do
10. $G_1', G_2' \leftarrow$ fix the exemplar(matching) of pair $;
11. $G_1^{*'}, G_2^{*'} \leftarrow$ randomly init exemplar(matching) of rest duplicated genes $G_1^*, G_2^*$;
12. $G_1^+, G_2^+ \leftarrow$ remove rest duplicated genes $G_1^+, G_2^+$;
13. $ub \leftarrow DCJ\text{Indel}(G_1^{*'}, G_2^{*'})$;
14. $lb \leftarrow DCJ\text{Indel}(G_1^{*}, G_2^{*})$;
15. if $lb > min\_ub$ then
16. discard $G_1^+, G_2^+$
17. if $ub < min\_ub$ then
18. $min\_ub = ub$;
19. else if $ub = max\_lb$ then
20. return $d = ub$;
21. else
22. insert $G_1^+, G_2^+$ into $L[ub]$;
23. return $d = min\_lb$;

Combining three optimization methods in tandem with the branch-and-bound framework, we can summarize our algorithm to compute $DCJ$-\text{Indel-Exemplar} (\text{Matching}) distance as outlined in Algorithm 1.

3.3 Adapting Lin-Kernighan Heuristic to Find the Median Genome

**Problem Statement** Not surprisingly, finding the median genome that minimizes the $DCJ$-\text{Indel-Exemplar} (\text{Matching}) distance is challenging. To begin with, given three input genomes, there are multiple choices of possible gene content selections for the median; however, since identifying gene content is simpler and there exists very accurate and fast methods to fulfill the task [15], we are more interested on a relaxed version of the median problem that assumes known gene content on the median genome. Which is formally defined as:
**Definition:** Given the gene content of a median genome, and gene orders of three input genomes. Find an adjacency of the genes of the median genome that minimize the DCJ-Indel-Exemplar(Matching) distance between the median genome and the three input genomes.

The DCJ-Indel-Exemplar(Matching) median problem is not even in the class of NP because there is no polynomial time algorithm to verify the results. It is hard to design an exact branch-and-bound algorithm for the DCJ-Indel-Exemplar(Matching) median problem mainly because the DCJ-Indel distance violates the property of triangular inequality which is required for a distance metrics [31]. Furthermore, when there are duplicated genes in a genome, it is possible that there are multiple edges of the same type connecting to the same vertex of a 0-matching, which leads to ambiguity in the edge shrinking step and makes the followed branch-and-bound search process very complicated and extremely hard to implement. To overcome these problems, we provide an adaption of Lin-Kernighan (LK) heuristic to help solving this challenging problem.

**Design of the Lin-Kernighan Heuristic** The LK heuristic can generally be divided into two steps: initialize the 0-matching for the median genome, and LK search to get the result.

The initialization problem can be described as: given the gene contents of three input genomes, find the gene content of the median genome that minimizes the sum of the number of Indels and duplications operations required to transfer the median gene content to the gene contents of the other three genomes. In this paper, we design a very simple rule to initialize the median gene content: given the counts of each gene family occurred in the three genomes, if two or three counts are the same, we simply select this count as the number of occurrence of the gene family in the median genome; if all three counts are different, we select the median count as the number of occurrence of the gene family in the median genome.

After fixing the gene content for the median genome, we randomly set up the 0-matching in the MBG. The followed LK heuristic selects two 0-matching edges on the MBG of a given search node and performs a DCJ operation, obtaining the MBG of a neighboring search node. We expand the search frontier by keeping all neighboring search nodes to up until the search level \( L_1 \). Then we only examine and add the most promising neighbors to the search list until level \( L_2 \). The search is continued when there is a neighbor solution yielding a better median score. This solution is then accepted and a new search is initialized from the scratch. The search will be terminated if there are no improvement on the result as the search level limits have been reached and all possible neighbors have been enumerated. If \( L_1 = L_2 = K \), the algorithm is called K-OPT algorithm.

** Adopting Adequate Sub-graphs to Simplify Problem Space ** By using the adequate subgraphs [26][29], we can prove that they are still applicable for decomposing the graph in the DCJ-Indel-Exemplar(Matching) median problem.
**Algorithm 2: DCJI\textsc{ndelExem(MATC)Median}**

**Input:** MBG $G$, Search Level $L_1$ and $L_2$

**Output:** 0-matching of $G$

1. Init search list $L$ of size $L_1$;
2. Init 0-matching of $G$;
3. $currentLevel \leftarrow 0$ and $Improved \leftarrow true$;
4. while $Improved = true$ do
   5. $currentLevel \leftarrow 0$ and $Improved \leftarrow false$;
   6. Insert $G$ into $L[0]$;
   7. while $currentLevel < L_2$ do
      8. $G' \leftarrow$ pop from list $L[currentLevel]$;
      9. if $G'$ improves the median score then
         10. $G \leftarrow G'$;
         11. $Improved \leftarrow true$ and break;
      12. if $currentLevel < L_1$ then
         13. for $x \in \forall$ 0-matching pairs of $G$ do
             14. $G' \leftarrow$ perform DCJ on $G'$ using $x$;
             15. if $num_{-}pair(x) > \delta$ then Insert $G'$ into $L[currentLevel + 1]$;
         else
             16. $G' \leftarrow$ perform DCJ on $G'$ using $x = \text{argmax } x$ $num_{-}pair(x)$;
             17. if $num_{-}pair(x) > \delta$ then Insert $G'$ into $L[currentLevel + 1]$;
         18. $currentLevel \leftarrow currentLevel + 1$;
   19. return 0-matching of $G$;

**Lemma 1.** As long as the irregular vertices do not involve, regular subgraphs are applicable to decompose MBG.

**Proof.** If there are $d$ number of vertices that contain duplicated edges in MBG, we can disambiguate the MBG by generating different subgraphs that contain only one of the duplicate edge. We call these subgraphs disambiguate MBG, ($d$-MBG), and there are $O(\prod_{i < d} \text{deg}(i))$ number of $d$-MBGs. If a regular adequate subgraph exists in the MBG, it must also exists in every $d$-MBG. Based on the $0$-matching solution, we can transform every $d$-MBG into completed $d$-MBG ($cd$-MBG) by constructing the optimal completion [13] between $0$-matching and all the other three types of edges. After this step, the adequate subgraphs in every $d$-MBG still exist in every $cd$-MBG, thus we can use these adequate subgraphs to decompose $cd$-MBG for each median problem without losing accuracy. □

**Search Space Reduction Methods** The performance bottleneck with the median computation is in the exhaustive search step, because for each search level we need to consider $O(|E|^2)$ possible number of edge pairs, which is $O(|E|^{2L_1})$ in total. Unlike the well-studied traveling salesman problem (TSP) where it is cheap to find the best neighbor, here we need to compute the DCJ-Indel-
Exemplar(Matching) problem, NP-hard distance, which makes this step extremely expensive to conclude. Noticing that if we search neighbors on edges that are on the same 0-i color altered connected component (0-i-comp), the DCJ-Indel-Exemplar(Matching) distance for genome 0 and genome i is more likely to reduce [32], thus we can sort each edge pair by how many 0-i-comp they share. Suppose the number of 0-i-comp that an edge pair x share is num_pair(x), when the algorithm is in the exhaustive search step (currentLevel < L1), we set a threshold δ and select the edge pairs that satisfy $\text{num}_\text{pair}(x) > \delta$ to add into the search list. When it comes to the recursive deepening step, we select the edge pair that satisfy $\arg\max_x \text{num}_\text{pair}(x)$ to add into the search list. This strategy has two merits: 1) some of the non-promising neighbor solution is eliminated to reduce the search space; 2) the expensive evaluation step which make a function call to DCJ-Indel-Exemplar(Matching) distance is postponed to the time when a solution is retrieved from the search list.

The LK based median computation algorithm is as Algorithm 2 shows.

4 Experimental Results

We implement our code with python and C++: the python code realized the optimization methods while the C++ code is implemented on a parallel branch-and-bound framework OPTKit. We conduct extensive experiments to evaluate the accuracy and speed of our distance and median algorithms using both simulated and real biological data. Experimental tests ran on a machine with linux operating system configured with 16 Gb of memory and an Intel(R) Xeon(R) CPU E5530 16 core processor, each core has 2.4GHz of speed. All of the experiments ran with a single thread. We choose to use g++-4.8.1 as our compiler.

4.1 Distance Computation

To the best of our knowlege, there is no software package that can handle both duplications and Indels. We compare our DCJ-Indel-Exemplar (Matching) distances with GREDO [25], a software package based on linear programming that can handle duplications.

Simulated Data The simulated data sets are generated with genomes containing 1000 genes. The Indels rate is set ($\gamma$) as 5%, inline with the duplication rate ($\phi$) as 10%. Considering GREDO can not process Indel data, all Indels for GREDO are removed. We compare the change of distance estimation with the variation of mutation rate ($\theta$), which grows from 10% to 100%. The experimental results for simulated data are displayed in Figure 5.

1. For computational time, since the results of time spans over a range of thousands of seconds, we display the time with log scale to construe results clearly.
(a) Time result for simulated data.  
(b) Distance result for simulated data.

Fig. 5. Experimental results for distance computation using simulated data.

(a) $\gamma = \phi = 0\%$ and $\theta$ varies from 10% to 100%.  
(b) $\gamma = \phi = 5\%$ and $\theta$ varies from 10% to 60%.

Fig. 6. Experimental results for median computation applying DCJ-Indel-Exemplar distance.

When the mutation rate is less than 50%, all three methods perform similarly, with the fact that GREDO is faster than both of our branch-and-bound methods. However, GREDO slows down dramatically when the mutation
| Data               | Distance Results | Time Results |
|-------------------|------------------|--------------|
|                   | GREDO | Exem | Matc | GREDO  | Exem | Matc |
| brownrat_chicken  | 1678  | 24546| 24704| 3604.28| 172.73| 7.45 |
| brownrat_gorilla  | 1274  | 17922| 17966| 5707.13| 12.64 | 12.10 |
| brownrat_human    | 1083  | 17858| 17900| 3725.76| 12.14 | 12.19 |
| brownrat_mouse    | 790   | 15433| 15445| 3725.66| 14.51 | 15.06 |
| chicken_gorilla   | 1491  | 16379| 16421| 3725.62| 7.54  | 7.57 |
| chicken_human     | 1521  | 16231| 16276| 3725.65| 7.74  | 7.47 |
| chicken_mouse     | 1528  | 15712| 15745| 3726.03| 9.82  | 8.16 |
| gorilla_human     | 486   | 17798| 17798| 3607.63| 13.94 | 13.81 |
| gorilla_mouse     | 860   | 18914| 18935| 4816.31| 12.60 | 12.13 |
| human_mouse       | 749   | 18126| 18144| 94.64  | 12.45 | 12.61 |

Table 1. Experimental results for distance computation with real data set.

Fig. 7. Experimental results for median computation applying DCJ-Indel-Matching distance.

(a) $\gamma = \phi = 5\%$ and $\theta$ varies from 10\% to 100\%. (b) $\gamma = \phi = 10\%$ and $\theta$ varies from 10\% to 100\%.

rate is increased, while our branch-and-bound based method takes less increased time to finish.
2. For computational accuracy, we show the distance results corrected by EDE approach which is one of the best true distance estimator. As for simulated data, we can see that when the mutation rate is small (< 50%) GREDO under estimate the distance as opposed to our two branch-and-bound methods; but it will over estimate the distance with the growth of mutation rate.

**Real data** We prepare the real data sets using genomes downloaded from Ensembl and processed them following the instructions in [25]. The real data set contains 5 species: brown-rat, chicken, human, mouse and gorilla. For DCJ-Indel-Exemplar (Matching) distance, we only convert the Ensembl format to adapt the data to our program. Meanwhile, just as the simulated data, all Indels in real data set for GREDO are removed. The results for real data are shown in Table 1.

1. For computational time, the branch-and-bound method shows orders of magnitudes of speed up compared with GREDO. We analyze the data, the reason can be construed as the existence of multiple connected component in BPG. So that our method can divide the graph into much smaller size, versus GREDO which doesn’t have this mechanism.

2. For computational accuracy, the distance results of the real data gives us a taste of how frequently Indels happen in the genome evolution. We can see orders of magnitude of difference between our distance results and GREDO, which is mainly due to the large amount of Indels in the real data set. Note that we did not change the way GREDO compute its distance as in paper [25], in the real distance computation, we should consider Indels in alignment with duplications.

4.2 Median Computation

**Median Computation** We simulate the median data of three genomes using the same strategy as in the distance simulation. In our experiments, each genome is “evolved” from a seed genome, which is identity, and they all evolve with the same evolution rate ($\theta$, $\gamma$, and $\phi$). The sequence length in the median experiments are reduced to 50, due to performance issues.

**DCJ-Indel-Exemplar median** We analyze the result of using LK algorithm with $L1 = 2$ and $L2 = 3$, and the K-OPT algorithm of $K = 2$. Search space reduction methods are used, with $\delta = 2$ and $\delta = 3$ respectively.

1. To begin with, we compare our result along with equal content data, since there are already benchmark programs to help us performing analysis. We run the exact DCJ median solver (we use the one in [32]) to compare our heuristic with the exact median results. In Fig 6(a) it shows the accuracy of our heuristic versus the exact result. It is shown that when $\theta \leq 60\%$, all results of the LK and K-OPT methods are quite close to the exact solver.
For parameter of $\delta = 2$, both $LK$ and $K$-OPT methods can generate exactly the same results for most of the cases.

2. As for the median results for unequal contents, we set both $\gamma$ and $\phi$ to 5\% and increase the mutation (inversion) rate $\theta$ from 10\% to 60\%. We compare our results with the accumulated distance of the three genomes to their simulation seed. Although it can not show the accuracy of our method (since we do not have an exact solver), it can be used as an indicator of how close that our method is to the real evolution. Fig 6(b) shows that when $\delta = 3$, both the $LK$ and $K$-OPT algorithms get results quite close to the real evolutionary distance.

**DCJ-Indel-Matching median** Since DCJ-Indel-Exemplar median has already given us the result of how $LK$ performs against exact solver, and how different parameters of $LK$ performs. With these things in mind, we choose to use $LK$ with $L_1 = 2$ and $L_2 = 3$ having $\delta = 2$ as the configuration for our DCJ-Indel-Matching median solver. We use the same data as in the previous experiments, and the experimental results are shown in Figure 7(a) and Figure 7(b). We can see that in general, the new implementation is quite close to the real result when $\gamma = 5\%$ and $\phi = 5\%$ and slightly worse than real result when $\gamma = 10\%$ and $\phi = 10\%$.

5 Conclusion

In this paper, we proposed a new way to compute the distance and median between genomes with unequal contents (with Indels and duplications). Our distance method can handle Indels which is ubiquitous in the real data set, and is proved to be more efficient as opposed to GREDO. We designed a Lin-Kernighan based method to compute median, which can get close to optimal results in alignment with the exact median solver, and our methods can handle duplications and Indels as well.

6 Acknowledgements

This Research was sponsored in part by the NSF OCI-0904461 (Bader), OCI-0904179, IIS-1161586 (Tang) and OCI-0904166 (Schaeffer).

References

1. Angibaud, S., Fertin, G., Rusu, I., Thévenin, A., Vialette, S.: On the approximability of comparing genomes with duplicates. J. Graph Algorithms Appl. 13(1), 19–53 (2009)
2. Bader, D.A., Moret, B.M.E., Yan, M.: A linear-time algorithm for computing inversion distance between signed permutations with an experimental study. Journal of Computational Biology 8, 483–491 (2001)

3. Bafna, V., Pevzner, P.A.: Sorting by transpositions. SIAM J. Discrete Math. 11(2), 224–240 (1998)

4. Bergeron, A., Mixtacki, J., Stoye, J.: On sorting by translocations. In: Journal of Computational Biology. pp. 615–629. Springer (2005)

5. Blin, G., Chauve, C., Fertin, G.: The breakpoint distance for signed sequences. In: Proc. CompBioNets 2004. vol. Text in Algorithms, Volume 3, pp. 3–16. King’s College London (2004)

6. Bourque, G., Pevzner, P.A.: Genome-Scale Evolution: Reconstructing Gene Orders in the Ancestral Species. Genome Res. 12(1), 26–36 (2002)

7. Braga, M.D.V., Willing, E., Stoye, J.: Genomic distance with dcj and indels. In: Proceedings of the 10th international conference on Algorithms in bioinformatics. pp. 90–101. WABI’10, Springer-Verlag, Berlin, Heidelberg (2010)

8. Brewer, C., Holloway, S., Zawalnyski, P., Schinzel, A., FitzPatrick, D.: A chromosomal duplication map of malformations: Regions of suspected haplo and triplo-lethality and tolerance of segmental aneuploidy in humans. The American Journal of Human Genetics 64(6), 1702 – 1708 (1999)

9. Caprara, A.: The Reversal Median Problem. INFORMS Journal on Computing 15(1), 93–113 (2003)

10. Chen, X., Zheng, J., Fu, Z., Nan, P., Zhong, Y., Lonardi, S., Jiang, T.: Assignment of orthologous genes via genome rearrangement. IEEE/ACM Trans. Comput. Biology Bioinform. 2(4), 302–315 (2005)

11. Chen, Z., Fu, B., Zhu, B.: Erratum: The approximability of the exemplar breakpoint distance problem. In: FAW-AAIM. p. 368 (2012)

12. Compeau, P.E.C.: A simplified view of dcj-indel distance. In: Proceedings of the 12th international conference on Algorithms in Bioinformatics. pp. 365–377. WABI’12, Springer-Verlag, Berlin, Heidelberg (2012)

13. Gao, N., Yang, N., Tang, J.: Ancestral genome inference using a genetic algorithm approach. PLoS ONE 8(5) (2013)

14. Hu, F., Zhou, J., Zhou, L., Tang, J.: Probabilistic reconstruction of ancestral gene orders with insertions and deletions. IEEE/ACM Trans. Comput. Biology Bioinform. 11(4), 667–672 (2014), http://doi.ieeecomputersociety.org/10.1109/TCBB.2014.2309602

15. Lenne, R., Solnon, C., Stutzle, T., Tannier, E., Birattari, M.: Reactive Stochastic Local Search Algorithms for the Genomic Median Problem. In: Carlos Cotta, J.v.H. (ed.) Eighth European Conference on Evolutionary Computation in Combinatorial Optimisation (EvoCOP). pp. 266–276. LNCS, Springer (Mar 2008)

16. Mabrouk, N.E.: Sorting Signed Permutations by Reversals and Insertions/Deletions of Contiguous Segments. Journal of Discrete Algorithms 1(1), 105–122 (2001)

17. Moret, B.M.E., Tang, J., san Wang, L., Warnow, Y.: Steps toward accurate reconstructions of phylogenies from gene-order data. J. Comput. Syst. Sci 65, 508–525 (2002)
19. Moret, B.M.E., Wang, L.S., Warnow, T., Wyman, S.K.: New approaches for reconstructing phylogenies from gene order data. In: ISMB (Supplement of Bioinformatics). pp. 165–173 (2001)

20. Nguyen, C.T., Tay, Y.C., Zhang, L.: Divide-and-conquer approach for the exemplar breakpoint distance. Bioinformatics 21(10), 2171–2176 (May 2005)

21. Pe’er, I., Shamir, R.: The median problems for breakpoints are np-complete. Elec. Colloq. on Comput. Complexity 71 (1998)

22. Rajan, V., Xu, A.W., Lin, Y., Swenson, K.M., Moret, B.M.E.: Heuristics for the inversion median problem. BMC Bioinformatics 11(S-1), 30 (2010)

23. Sankoff, D.: Genome rearrangement with gene families. Bioinformatics 15(11), 909–917 (1999)

24. Shao, M., Lin, Y.: Approximating the edit distance for genomes with duplicate genes under dcj, insertion and deletion. BMC Bioinformatics 13(S-19), S13 (2012)

25. Shao, M., Lin, Y., Moret, B.M.E.: An exact algorithm to compute the dcj distance for genomes with duplicate genes. In: RECOMB. pp. 280–292 (2014)

26. Xu, A.W.: Dcj median problems on linear multichromosomal genomes: Graph representation and fast exact solutions. In: RECOMB-CG, pp. 70–83 (2009)

27. Xu, A.W.: A fast and exact algorithm for the median of three problem: A graph decomposition approach. Journal of Computational Biology 16(10), 1369–1381 (2009)

28. Xu, A.W., Moret, B.M.E.: Gasts: Parsimony scoring under rearrangements. In: WABI. pp. 351–363 (2011)

29. Xu, A.W., Sankoff, D.: Decompositions of multiple breakpoint graphs and rapid exact solutions to the median problem. In: Proceedings of the 8th international workshop on Algorithms in Bioinformatics. pp. 25–37. WABI ’08, Springer-Verlag, Berlin, Heidelberg (2008)

30. Yancopoulos, S., Attie, O., Friedberg, R.: Efficient sorting of genomic permutations by translocation, inversion and block interchange. Bioinformatics 21(16), 3340–3346 (2005)

31. Yancopoulos, S., Friedberg, R.: Sorting genomes with insertions, deletions and duplications by dcj. In: Nelson, C.E., Vialette, S. (eds.) RECOMB-CG. Lecture Notes in Computer Science, vol. 5267, pp. 170–183. Springer (2008)

32. Yin, Z., Tang, J., Schaeffer, S.W., Bader, D.A.: Streaming breakpoint graph analytics for accelerating and parallelizing the computation of dcj median of three genomes. In: ICCS. pp. 561–570 (2013)