GENOME HALVING WITH DOUBLE CUT AND JOIN

ROBERT WARREN AND DAVID SANKOFF
University of Ottawa

The genome halving problem, previously solved by El-Mabrouk for inversions and reciprocal translocations, is here solved in a more general context allowing transpositions and block interchange as well, for genomes including multiple linear and circular chromosomes. We apply this to several data sets and compare the results to the previous algorithm.

1. Introduction

In this paper we discuss a generalization of the genome halving process studied by El-Mabrouk.[3] Before stating and solving the problem formally in the ensuing sections, we first give some motivation for the generalization.

Models of genome rearrangement processes have permitted different repertoires of operations. Certainly, realistic models must account for inversion. Likewise, reciprocal translocations, Robertsonian translocations and other processes of chromosome fusion and fission, all of which involve transferring an entire telometric (i.e., suffix or prefix) region of at least one chromosome, are widespread across all eukaryotic domains.

Other movements of chromosomal fragments, usually not involving telomeres, are widely attested, and grouped together under the label of transpositions. They are produced by a variety of processes, such as gene duplication followed by the loss of the original copy, or retrotransposition, or recombination errors.

Of the three true movement rearrangements,[a] inversion, translocation and transposition, only the first two, separately or in combination, have proved very amenable to mathematical modeling, as exemplified by the Hannenhalli-Pevzner formula for the edit distance between two genomes, i.e., the minimum number of operations required to transform one genome into another, and the efficient algorithm for producing such a series of operations. No formula or efficient algorithm exists for transposition, either by itself or in combination with the other two operations.

Recently, Yancopoulos *et al.*[6] introduced the "double cut and join" (DCJ) operation as the basis for generating all the movement rearrangements. This allowed for the inclusion of transposition with inversion and translocation in a single model

[a]Duplications of genes or of chromosomal segments, as well as deletions and insertions are often considered as aspects of genome rearrangement, but they are not really of the same biological nature as the movements inherent in inversion, translocation and transposition, and mathematical models of rearrangement are not easily extended to encompass them.

and resulted in a simpler formula for the edit distance and a simpler algorithm for recovering a corresponding series of operations. A double cut and join operation simply cuts the chromosome in two places and joins the four ends of the cut in a new way.

The DCJ model, however, allows for the generation of a new kind of movement operation, a generalized transposition called block interchange, which is not represented in the biological genome rearrangement literature, though it has long been studied in the mathematical literature on rearrangement. Both transposition and block interchange can be thought of as the excision of a fragment, its circularization, together counting as one DCJ operation, followed by a second set of cuts, where the circle is not necessarily cut in the same place it was originally created through a join, and then reincorporated at a new site in the chromosome. Transpositions and block interchanges thus count as two DCJ operations whereas inversions and translocations each count as one.

The question arises, what is the biological significance of these chromosomal circles? On the evolutionary level, very little is known, but circular DNA structures abound in all sorts of organisms, even eukaryotes. Circular chromosomes are well-known in clinical studies[4] and the process of excision, circularization, linearization and reincorporation is exactly what happens in the configuration of the immune response in higher animals. Because the evolutionary consequences of block interchange could have come about in other ways, there has been no reason to look for evidence of this process or even to notice it. The question of its existence or importance remains open.

Yancopoulos *et al.*'s original publication[6] pointed out that the running time of their algorithm could be reduced to linear if circles were not constrained to be reincorporated into linear chromosomes as as soon as they were generated. Bergeron *et al.*[2] recently restated the DCJ model and produced a simplified (linear) algorithm ignoring the reincorporation constraint and, as in the mathematical justification of DCJ in Ref. 6, without any explicit mention of the particular operations of inversion, translocation, transposition, interchange, fusion and fission. It is thus the most general existing algorithm for movement rearrangements. As it has a form which lends itself well to constraints on the operations allowed, it can largely emulate other algorithms, e.g., the Hannenhalli-Pevzner algorithm (but without taking into account "hurdles" and "knots") or the Yancoupoulos-Attie-Friedberg algorithm (at the cost of losing its computational efficiency).

It is with this background that we ask how generalizations of the genomic distance problem, such as genome halving or rearrangement median (considered elsewhere), behave under the DCJ context.

2. Background

In this section we introduce our notation for genomes. A gene a represents an oriented sequence of DNA whose two *extremities* are its *tail* $\overrightarrow{a}$ and its *head* $\overrightarrow{a}$.

The *adjacency* of two consecutive genes a and b is denoted by an unordered set, either $\{\overrightarrow{a}, \overleftarrow{b}\}$, $(= \{\overleftarrow{b}, \overrightarrow{a}\})$, $\{\overrightarrow{a}, \overrightarrow{b}\}$, $\{\overleftarrow{a}, \overleftarrow{b}\}$, $\{\overleftarrow{a}, \overrightarrow{b}\}$, depending on the order and orientation of a and b. An extremity that is not adjacent to any other extremity is called a *telomere* and is represented by a singleton set $\{\overrightarrow{a}\}$ or $\{\overleftarrow{a}\}$. A *genome* is represented by an unordered set of adjacencies and telomeres such that the head and tail of each gene appear exactly once.

A *duplicated genome* is a genome with two copies of each gene such that the head and the tail of every gene appear exactly twice. To differentiate the two genes we arbitrarily assign each gene a subscript. Thus, we say that gene a is a *unique gene* with *paralogs* a_1 and a_2 with corresponding *paralogous extremities* $\overleftarrow{a_1}$ and $\overleftarrow{a_2}$, and $\overrightarrow{a_1}$ and $\overrightarrow{a_2}$. To denote paralogous extremities, we have a special notation: if p is an extremity then $\overline{p}$ is its corresponding paralogous extremity. Thus, if $p = \overleftarrow{a_1}$ then $\overline{p} = \overleftarrow{a_2}$. Given $V \subseteq A$, where A is a duplicated genome, we can retrieve the set of all extremities using the function $\pi(V) = \bigcup_{v \in V} v$. The set of paralogous extremities in V can be retrieved using the function $\varphi(V)$ defined as follows: if p and $\overline{p}$ are in $\pi(V)$ then p is in $\varphi(V)$.

Definition 1. Let A be a duplicated genome. A is *valid* if and only if:

- If $\{u, v\} \in A$ then $\{\overline{u}, \overline{v}\} \in A$
- If $\{u\} \in A$ then $\{\overline{u}\} \in A$

A duplicated genome that is valid is a *perfectly duplicated genome*. Similarly, an invalid duplicated genome is called a *rearranged duplicated genome*.

Observant readers may notice that the above definition of validity is very general and will allow many genomes with some questionable halvings. This is intentional. One of the advantages of double cut and join is the ease with which it handles circular chromosomes. However, what is considered a valid halving in a linear multichromosonal genomes and what is considered valid for circular unichromosonal genomes are very different. For the case of an input consisting of multiple chromosomes that can be either linear of circular neither definition suffices. Our definition of validity combines both definitions of validity but, because we do not try and conserve chromosomes, it can result in some surprising results. However, typically these results are not desirable but adding additional constraints to prevent may occasionally increase the cost. For a better treatment of validity for linear multichromosonal genomes consult Ref. 3. For a better treatment of validity for circular unichromosonal genomes consult Ref. 1.

We can now define the problem:

Definition 2. The *genome halving problem* is defined as follows: given a rearranged duplicated genome A find a perfectly duplicated genome B such that the distance between A and B is minimal with respect to some distance metric.

As mentioned in the introduction, in this paper the distance metric we will use

is the *double cut and join distance*. To understand the DCJ distance we must first introduce the following data structure:

Definition 3. The *adjacency graph* $AG(A, B)$ is a graph whose set of vertices are the adjacencies and telomeres of A and B. For each $u \in A$ and $v \in B$ there are $|u \cap v|$ edges between u and v.

Since every vertex in an adjacency graph has a degree of two, there are only two types of components: cycles and paths. Since the graph is bipartite, all the cycles have an even number of edges. Paths may have an odd or even number of edges. We refer to paths with an odd number of edges as *odd paths* and paths with an even number of edges as *even paths*. The difference between odd and even paths is important, thus, overall, there are three types of components to consider. Since an adjacency graph is bipartite, we can deduce the following useful lemma:

Lemma 1. *An adjacency graph $AG(A, B)$ contains a path with an odd number of edges if and only if telomere $\{u\} \in A$ and telomere $\{v\} \in B$ are endpoints of a path.*

Since double cut and join is defined for non-duplicated genomes, for the purposes of measuring distance we consider each paralog to be a different gene.

Theorem 1. ***(Ref. 2)*** *Let A and B be two genomes defined on the same set of n genes, then we have*

$$d(A, B) = n - c - \frac{i}{2}$$

where c is the number of cycles and i the number of odd paths in $AG(A, B)$.

For simplicity, throughout the rest of this paper we will use the symbol A to represent a rearranged genome and genome B to represent a perfectly duplicated genome. A and B have n unique genes, thus, they each have $2n$ paralogs, $4n$ extremities and $2n$ paralogous extremities. Thus, $d(A, B) = 2n - c - i/2$.

3. Natural Graphs

From the definition of validity, it is clear that the relationships between paralogous extremities is important. Following El-Mabrouk,[3] we define a data structure, a natural graph, to capture that relationship.

Definition 4. For each $u \in A$ we define the set V_u recursively as follows:

- Basis Step: $u \in V_u$
- Recursive Step: If $u \in V_u$ and u is an adjacency $\{p, q\}$ then $v, w \in A$ where $\overline{p} \in v$ and $\overline{q} \in w$ are also in V_u. If $u \in V_u$ and u is a telomere $\{p\}$ then $v \in A$ where $\overline{p} \in v$ is also in V_u

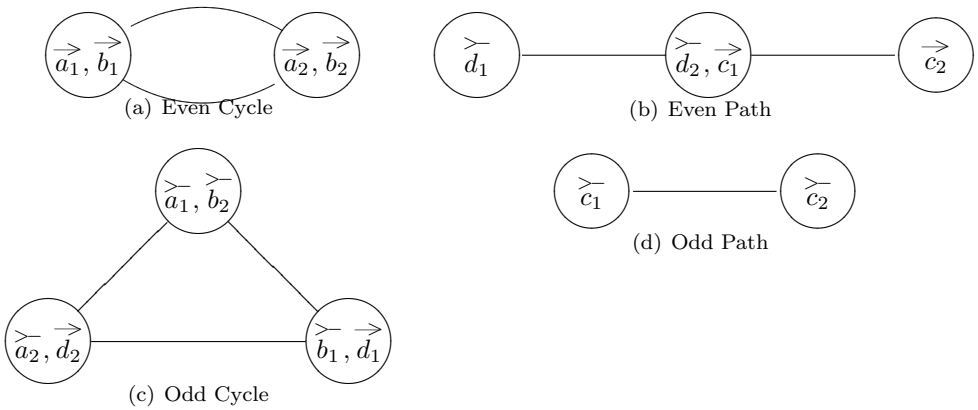

Figure 1. The natural graphs for $A = \{\{\overleftarrow{c_1}\}, \{\overrightarrow{c_1}, \overleftarrow{d_2}\}, \{\overrightarrow{d_2}, \overleftarrow{a_2}\}, \{\overrightarrow{a_2}, \overrightarrow{b_2}\}, \{\overleftarrow{b_2}, \overleftarrow{a_1}\}, \{\overrightarrow{a_1}, \overrightarrow{b_1}\}, \{\overleftarrow{b_1}, \overrightarrow{d_1}\}, \{\overleftarrow{d_1}\}, \{\overleftarrow{c_2}\}, \{\overrightarrow{c_2}\}\}$.

We define the set $E_u = \{(v, w) \in V_u | p \in v \wedge \overline{p} \in w\}$. We say that $G(V_u, E_u)$ is a *natural graph*, $G(V, E)$, generated by u. Note that if there exists an $G(V, E)$ and an $G'(V', E')$ such that $v \in V$ and $v \in V'$ then $G = G'$.

Let NG be the set of all natural graphs defined on A.

Like adjacency graphs, every vertex in a natural graph has degree of at most two. Therefore we can classify natural graphs in one of four ways:

Definition 5. Let $G(V, E) \in NG$. NG consists of four mutually exclusive subsets:

(1) If G is a cycle and $|E|$ is even then it is called an *even cycle*. We call the set of all natural graphs that are even cycles EC
(2) If G is a path and $|E|$ is odd then it is called an *odd path*. We call the set of all natural graphs that are odd paths OP
(3) If G is a cycle and $|E|$ is odd then it is called an *odd cycle*. We call the set of all natural graphs that are odd cycles OC
(4) If G is a path and $|E|$ is even then it is called an *even path*. We call the set of all natural graphs that are even paths EP

Using the properties of natural graphs we can derive some useful lemmas, the proofs of which will be included in a full version of this paper.

Lemma 2. *Let $G(V, E) \in NG$. There is no subgraph G' of G, such that G' is perfectly duplicated.*

Lemma 3. *Let $G(V, E) \in NG$ and $\{u, v\} \in V$. Let B and B' be two identical perfectly duplicated genomes except $\{u, v\}, \{\overline{u}, \overline{v}\} \in B$ and $\{u\}, \{\overline{u}\} \in B'$. $d(A, B) \leq d(A, B')$.*

Lemma 4. *Let $G(V, E) \in NG$. If $|\varphi(V)|$ is even then there exists an $H \subseteq B$ with $\pi(H) = \pi(V)$ such that $d(V, H)$ is minimal.*

From Lemma 2 and Lemma 4 we can now effectively redefine the genome halving problem to: for each natural graph, construct a subset of B such that the distance between the natural graph and its corresponding subset is minimal.

From Lemma 4 we can conclude that all the natural graphs should contain an even number of paralogous extremities. Observe that this is the case for the natural graphs in the sets EC and EP. For the remaining graphs we observe that since there are an even number of paralogous extremities in A it must be the case that $|OP| + |OC|$ is even.

4. Lower Bounds

Before we can derive lower bounds on the distance, we once again need a new data structure:

Definition 6. Let $G(V, E) \in SN$. Let $H \subseteq B$ such that $\pi(H) = \pi(V)$. Let C be a component of $AG(V, H)$. We define the *signature* S_C as follows:

(1) If $u \in \pi(C \cap V)$ then u is in S_C unless $\overline{u}$ is already in S_C;
(2) If $\{u, v\}$ is an adjacency in $C \cap V$ and u is in S_C then neither v nor $\overline{v}$ is in S_C;

A *maximal signature* is a signature which includes as many extremities as possible. Let S be a set of maximal signatures for all components in $AG(V, H)$. We define a *signature graph* $SG(S, F)$ where S is the vertices and F is the set of edges. F is defined as follows: for all $S_1, S_2 \in S$ there exists an edge $\{S_1, S_2\}$ if and only if there exists an extremity x such that $x \in S_1$ and $\overline{x} \in S_2$. We also define the function $\delta(S_C)$, where $S_C \in S$, which denotes the degree of S_C.

From the fact that $|\varphi(V)|$ is odd and from the definition of a maximal signature we get the following lemma:

Lemma 5. *Let $G(V, E) \in EP \cup OC$ and let $SG(S, F)$ be a signature graph defined on G. $\sum_{S_C \in S} |S_C| \leq |V| - 1$*

We now have enough information to establish lower bounds on the distance:

Theorem 2. *Let $G(V, E) \in SN$ with $2n$ extremities where $n \geq 1$. Let $H \subseteq B$ with $\pi(H) = \pi(V)$. The following statements are all true:*

(1) If $G \in EC$ then $d(V, H) \geq n - 1$;
(2) If $G \in OP$ then $d(V, H) \geq n - 2$;
(3) If $G \in OC$ then $d(V, H) \geq n - 1$;
(4) If $G \in EP$ then $d(V, H) \geq n$;

Proof. From the definition of a signature graph, we know that $\sum_{S_C \in S} \delta(S_C) \leq \sum_{S_C \in S} |S_C| \leq |V|$. Therefore, $|F| = \frac{1}{2} \sum_{S_C \in S} \delta(S_C) \leq \frac{1}{2} \sum_{S_C \in S} |S_C| \leq \frac{1}{2}|V|$. It follows from Lemma 2 that all signature graphs are connected, therefore, $|S| \leq |F| + 1 \leq \frac{1}{2}|V| + 1$. Thus, $AG(V, H)$ contains $\leq \frac{1}{2}|V| + 1$ components.

When $G \in EC$, $|V| = 2n$ and thus $AG(V, H)$ contains $\leq n + 1$ components.

When $G \in OP$, $|V| = 2n$ and thus $AG(V, H)$ contains $\leq n + 1$ components.

When $G \in EP$, $|V|-1 = 2n$ and, from Lemma 5, only $|V|-1$ vertices contribute towards the signature graph. Thus, $AG(V, H)$ contains $\leq n + 1$ components.

When $G \in OC$, $|V| - 1 = 2n - 2$ and, from Lemma 5, only $|V| - 1$ vertices contribute towards the signature graph. Thus, $AG(V, H)$ contains $\leq n$ components.

From Theorem 1 we can observe that $d(V, S_G \cup \overline{S_G}) = |\varphi(V)| - c - i$ where c and i are the number of cycles and odd paths respectively in $AG(V, S_G \cup \overline{S_G})$. When $G \in EC \cap EP$ we know that $|\varphi(V)| = 2n$. For the remaining two cases $|\varphi(V)| = 2n-1$. We know from the above that the maximum number of components each type of natural graph contains, to establish lower bounds on the distance we need only determine which of those components are cycles and which are odd paths.

It can be proven (proof omitted in this abstract) that some of components must be paths. In the worst case, graphs in OP contains 2 odd paths, graphs in EP contain 1 even path. For the purposes of establishing a lower bound we can safely assume that the remaining components are cycles. □

5. Upper Bounds

Using the structure of a natural graph, we can define an ordering of the vertices and extremities that will simplify ensuing developments. We define the ordering as follows:

Definition 7. Let $G(V, E) \in NG$. We relabel the extremities in V to define a *suitable order* of the vertices.

(1) $G \in EC$. $|V| = |\varphi(V)| = 2n$, for all $n \geq 1$. Let $V = \{v'_1, v_1, v'_2, v_2, \ldots, v'_n, v_n\}$ such that the following hold:
- $v'_1 = \{\overline{x_{2n}}, \overline{x_1}\}$ and $v_1 = \{x_1, x_2\}$;
- For each $i, 1 < i \leq n$, $v'_i = \{\overline{x_{2i-2}}, \overline{x_{2i-1}}\}$ and $v_i = \{x_{2i-1}, x_{2i}\}$;

(2) $G \in OP$. $|\varphi(V)| = 2n - 1$ and $|V| = 2n$, for all $n \geq 1$. Let $V = \{v'_1, v_1, v'_2, v_2, \ldots, v'_n, v_n\}$ such that the following hold:
- $v'_1 = \{\overline{x_1}\}$ and $v_1 = \{x_1, x_2\}$;
- For each $i, 1 < i < n$, $v'_i = \{\overline{x_{2i-2}}, \overline{x_{2i-1}}\}$ and $v_i = \{x_{2i-1}, x_{2i}\}$;
- $v'_n = \{\overline{x_{2n-2}}, \overline{x_{2n-1}}\}$ and $v_n = \{x_{2n-1}\}$;

(3) $G \in OC$. $|V| = |\varphi(V)| = 2n - 1$, for all $n \geq 1$. Let $V = \{v'_1, v_1, v'_2, v_2, \ldots, v'_{n-1}, v_{n-1}, v'_n\}$ such that the following hold:
- $v'_1 = \{x_{2n-1}, \overline{x_1}\}$;

- For each $i, 1 < i \leq n-1$, $v_{i-1} = \{x_{2i-3}, x_{2i-2}\}$ and $v'_i = \{\overline{x_{2i-2}}, \overline{x_{2i-1}}\}$;

(4) $G \in EP$. $|\varphi(V)| = 2n$ and $|V| = 2n+1$, for all $n \geq 1$. Let $V = \{v'_1, v_1, v'_2, v_2, \ldots, v'_n, v_n, v'_{n+1}\}$ such that the following hold:

- $v'_1 = \{\overline{x_1}\}$ and $v_1 = \{x_1, x_2\}$;
- For each $i, 1 < i \leq n$, $v'_i = \{\overline{x_{2i-2}}, \overline{x_{2i-1}}\}$ and $v_i = \{x_{2i-1}, x_{2i}\}$;
- $v'_{n+1} = \{\overline{x_{2n}}\}$;

From the definition of suitable order we can derive the sets S and $\overline{S}$ for each natural graph. Let $G(V, E) \in NG$ where V is suitably ordered. As noted before, $\varphi(V)$ is $2n$ when $G \in EC \cup EP$, $2n-1$ when $G \in OP \cup OC$, for all $n \geq 1$. If $G \notin OC$ then let $S_G = \{v_1, v_2, \ldots, v_n\} \in V$. We define $\overline{S_G}$ as follows: if the adjacency $\{x, y\} \in S_G$ then $\{\overline{x}, \overline{y}\} \in \overline{S_G}$, if the telomere $\{x\} \in S_G$ then $\{\overline{x}\} \in \overline{S_G}$. We call the set S_G the set of *selected vertices*. Set $\overline{S_G}$ is the set of *derived vertices*.

The case of $G \in OC$ is a special case. We $S_G = \{v_1, v_2, \ldots, v_{n-1}\} \in V$. If we define $\overline{S_G}$ as normal we end up missing the extremities x_{2n-1} and $\overline{x_{2n-1}}$. Thus, for $G \in OC$ we define $\overline{S_G}$ as $\overline{S_G}' \cup \{x_{2n-1}, \overline{x_{2n-1}}\}$ where $\overline{S_G}'$ is the same as the definition for $\overline{S_G}$ when $G \notin OC$. Note that this definition for $\overline{S_G}$ has a tendency to produce circular chromosomes which may not be desirable. Alternative definitions which avoid circles do exist but they produce a worse distance.

We can now derive a solution for B:

$$B = \bigcup_{G \in SN} S_G \cup \overline{S_G}$$

For the ensuing proofs, it is useful to keep track of the *unselected vertices* in V. Let $U_G = V \setminus S$ which is $\{v'_1, v'_2, \ldots, v'_n\}$ when $G \notin EP$ or $\{v'_1, v'_2, \ldots, v_n, v'_{n+1}\}$ when $G \in EP$.

The following useful observation describes the motivation for constructing set S in this manner:

Observation 1. Let $G(V, E) \in NG$. Each adjacency $\{x_i, x_j\} \in S_G$ corresponds to a cycle in $AG(V, S_G \cup \overline{S_G})$ and each telomere $\{x_k\} \in S_G$ corresponds to an odd path in $AG(V, S_G \cup \overline{S_G})$.

In order for B to be valid, the set of derived vertices must be constructed as above. Observe that, for $G(V, E) \in NG$, $U_G \subset V$ and that $\pi(U_G) = \pi(\overline{S_G})$. Thus, $AG(U_G, \overline{S_G}) \subseteq AG(A, B)$ which has the following properties:

Lemma 6. *The following statements are all true:*

(1) If $G \in EC$ then $AG(U_G, \overline{S_G})$ contains exactly one cycle and no paths.
(2) If $G \in OP$ then $AG(U_G, \overline{S_G})$ contains exactly one odd path and no cycles.
(3) If $G \in OC$ then $AG(U_G, \overline{S_G})$ contains exactly one cycle and no paths.
(4) If $G \in EP$ then $AG(U_G, \overline{S_G})$ contains no cycles or odd paths.

Proof. This lemma follows from the definitions of U_G and $\overline{S_G}$ as well as the definition of an adjacency graph. □

We can now define the distance between any natural graph $G(V, E)$ and $S_G \cup \overline{S_G}$:

Theorem 3. *The following statements are all true:*

(1) If $G \in EC$ *then* $d(V, S_G \cup \overline{S_G}) \leq n - 1$
(2) If $G \in OP$ *then* $d(V, S_G \cup \overline{S_G}) \leq n - 2$
(3) If $G \in OC$ *then* $d(V, S_G \cup \overline{S_G}) \leq n - 1$
(4) If $G \in EP$ *then* $d(V, S_G \cup \overline{S_G}) \leq n$

Proof. From Theorem 1 we can observe that $d(V, S_G \cup \overline{S_G}) = |\varphi(V)| - c - i$ where c and i are the number of cycles and odd paths respectively in $AG(V, S_G \cup \overline{S_G})$.

From Observation 1 and Lemma 6 we can immediately conclude that $c = n$ in all cases except when $G \in EC$ in which case $c = n + 1$ and that $i = 0$ in all cases except when $G \in OP$ in which case $i = 2$. When $G \in EC \cap EP$ we know that $|\varphi(V)| = 2n$. For the remaining two cases $|\varphi(V)| = 2n - 1$. □

Theorem 4. *Let* A *be a rearranged duplicated genome and* B *be a perfectly duplicated genome with* $2n$ *genes where* n *is the number of unique genes and* $n \geq 1$ *then the minimum distance between* A *and* B *is:*

$$\begin{aligned} d(A, B) &= n - |EC| - \left\lfloor \frac{2 \cdot |OP| - |OC|}{2} \right\rfloor \\ &= n - |EC| - |OP| - \left\lfloor \frac{|OC|}{2} \right\rfloor \end{aligned}$$

Proof. This theorem follows immediatly from Theorem 2 and Theorem 3 and the fact that $|OP|$ and $|OC|$ are odd. □

6. Experiments

We have implemented the DCJ halving algorithm so that it runs in (provably minimum) linear time. We applied it to data sets on three present-day genomes that are descended from a genome doubling event: *Zea mays*,[7] with two copies of 34 markers, *Saccharomyces cerevisiae*, with two copies of 300 markers, and *Candida glabrata*, with two copies of 300 markers.[5] The number of operations from the doubling event to the present-day genome was 27, 193 and 249 respectively. The El-Mabrouk algorithm gave a result of 250 for *Candida glabrata* but otherwise the results were exactly the same. Applying the algorithm as written in the paper produced circular chromosomes in each case. However, borrowing the look-ahead routine (which prevents the formation of circular chromosomes) from the El-Mabrouk paper[3] we got the same result as El-Mabrouk with no circular chromosomes and no asymptotic increase in complexity.

7. Conclusion

We have shown that the main ideas of the El-Mabrouk algorithm carry over to the DCJ context, although the case analysis here involves both cycles and paths, instead of just cycles in the breakpoint graph. In one respect, however, the algorithm is much simpler, due to the simplifications inherent in the DCJ approach. Where El-Mabrouk had to attend to the complex components of the breakpoint graph known as hurdles and knots, the DCJ formulation avoids this completely.

Since the repertoire of movement rearrangements in the DCJ formulation is complete, the results of applying our algorithm will always be a lower bound on any result using a constrained set of operations. At the same time, constraining the DCJ operations may not yield an optimal result, since these constraints are ad hoc and may not yield the minimum number of operations. Thus the method yields both a lower bound (using unconstrained operations) and an upper bound (using the constraints) on the results of algorithms yielding optimal answers for a specific set of constraints.

8. Acknowledgments

We thank Julia Mixtacki, Jens Stoye, Chunfang Zheng and Nadia El-Mabrouk for helpful discussion. Research supported in part by a grant to David Sankoff from the Natural Sciences and Engineering Research Council of Canada (NSERC). David Sankoff holds the Canada Research Chair in Mathematical Genomics and is a Fellow of the Royal Society of Canada.

References

1. Max A. Alekseyev and Pavel A. Pevzner. Whole genome duplications and contracted breakpoint graphs. *SIAM Journal on Computing*, 36(6):1748–1763, 2007.
2. Anne Bergeron, Julia Mixtacki, and Jens Stoye. A unifying view of genome rearrangements. In Philipp Bücher and Bernard M.E. Moret, editors, *Algorithms in Bioinformatics: 6th International Workshop*, volume 4175 of *Lecture Notes in Computer Science*, pages 163–173. Berlin, Heidelberg: Springer-Verlag, 2006.
3. Nadia El-Mabrouk and David Sankoff. The reconstruction of doubled genomes. *SIAM Journal on Computing*, 32:754–792, 2003.
4. Fabien Kuttler and Sabine Mai. Formation of non-random extrachromosomal elements during development, differentiation and oncogenesis. *Seminars in Cancer Biology*, 17(1):56–64, 2007.
5. David Sankoff, Chunfang Zheng, and Qian Zhu. Polyploids, genome halving and phylogeny. *Bioinformatics*, 23:i431 – i439, 2007.
6. Sophia Yancopoulos, Oliver Attie, and Richard Friedberg. Efficient sorting of genomic permutations by translocation, inversion, and block interchange. *Bioinformatics*, 21(16):3340–3346, 2005.
7. Chunfang Zheng, Qian Zhu, and David Sankoff. Parts of the problem of polyploids in rearrangement phylogeny. In Glenn Tesler and Dannie Durand, editors, *Proceedings of the RECOMB 2007 Workshop on Comparative Genomics*, volume 4751 of *Lecture Notes in Computer Science*, pages 162–176. Berlin, Heidelberg: Springer-Verlag, 2007.