involve
a journal of mathematics

Distribution of genome rearrangement distance under double cut and join

Jackie Christy, Josh McHugh, Manda Riehl and Noah Williams

msp

2014 vol. 7, no. 4

INVOLVE 7:4 (2014)
dx.doi.org/10.2140/involve.2014.7.491

Distribution of genome rearrangement distance under double cut and join

Jackie Christy, Josh McHugh, Manda Riehl and Noah Williams

(Communicated by Anant Godbole)

Using the double-cut-and-join (DCJ) model for genome rearrangement we use combinatorial techniques to analyze the distribution of genomes under DCJ distance. We present an exponential generating function for the number of genomes that are maximally distant from a given genome and provide a formula for the number of genomes that are any given distance from an arbitrary starting genome.

1. Introduction

Many mathematical models have been developed to aid biologists and bioinformaticians in their study of the genome rearrangement problem, whose goal is to find the optimal sequence of mutations for the transformation of one genome into another. Using the double-cut-and-join (DCJ) model, Bergeron, Mixtacki, and Stoye [Bergeron et al. 2006] found that the distance between two genomes is completely determined by a bipartite graph created from the genomes. We utilize their data structure to find the distribution of genomes that are distance d from a given genome under DCJ. In Section 2, we introduce genome rearrangement, DCJ, and an important result of the same authors. In Section 3, we present a generating function for the number of maximally distant genomes from a given genome, and in Section 4, we obtain the distribution of all genomes by distance from a given genome.

2. Background

A brief history of genome rearrangements. Deoxyribonucleic acid (DNA) contains instructions for the creation of the proteins necessary for the development and

MSC2010: 05E05, 68R15.
Keywords: genome rearrangement, double cut and join, generating function.
The authors received support for this work from UWEC's Office of Research and Sponsored Programs. Williams was also supported by the UWEC Foundation.

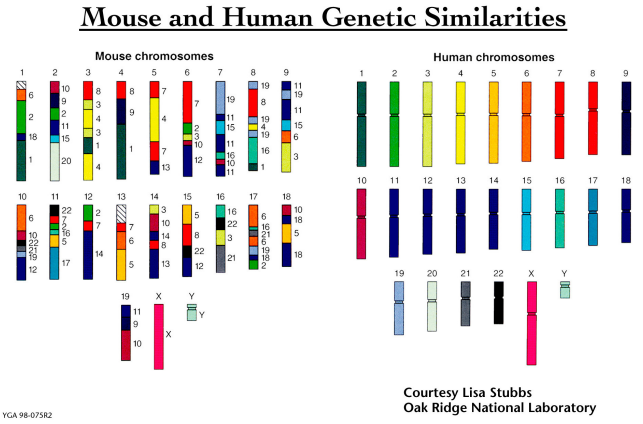

Figure 1. Preserved segments between mouse and human genomes showing long stretches of conserved DNA from their common ancestor. More than ninety percent of the mouse genome consists of shuffled pieces of the human genome [NHGRI 2002].

survival of living organisms. The entire collection of DNA in an organism is called the organism's genome, and this DNA is contained within chromosomes comprised of genes. When DNA is replicated, occasionally something goes awry and a mutation occurs, slightly changing an organism's genetic make-up. A sufficient number of mutations can result in death, disease, or the development of a new species.

In the genome rearrangement problem, the object is to find the optimal sequence of mutations that transforms one genome into another, where both genomes are defined on the same set of genes. The number of mutations in this most efficient scenario is defined to be the *distance* between the two genomes.

In the simplest case, genomes can be modeled by permutations under the assumptions that all genomes share the same set of genes, there are no duplicated genes, and only a single chromosome is considered [Fertin et al. 2009]. Most models now use objects that are more complicated than permutations by removing some or all of these assumptions [Yancopoulos et al. 2005]. For example, signed permutations are utilized to better model that DNA is oriented, and ordered set partitions can be used for multiple chromosomes.

Double cut and join. In the DCJ model, genes are numbered and oriented, as shown in the figures on the next page. Consequently, a gene may be represented as a numbered left or right arrow with labeled ends; for example, $7h$ and $7t$ denote the head and tail of the seventh gene. Chromosomes are collections of arrows that have

been joined head to head, tail to tail, or head to tail, and genomes constitute sets of chromosomes. Alternatively, a genome may be represented as a collection of vertices that correspond to the locations where genes meet. An *internal vertex*, or *adjacency*, occurs where two genes are joined in one of the three fashions mentioned above, and an *external vertex*, or *telomere*, occurs where the head or tail of a gene is not connected to other genes. Note that there are always an even number of telomeres in a genome, and that the number of genes in a genome is equivalent to the sum of the adjacencies and the number of pairs of telomeres present.

DCJ is a broad model that encompasses linear and circular chromosomes and incorporates the following mutations:

- Inversions: reverse the order of a chromosome or part of the genome
- Interchanges: switch two segments of the genome
- Translocations: swap the ends of two chromosomes
- Circularizations and linearizations: convert between linear and circular chromosomes.

A DCJ operation involves making two cuts in a genome and rejoining the pieces in one of the following ways:

i. Two internal vertices $\{a, b\}$ and $\{c, d\}$ can be replaced with two new internal vertices $\{a, d\}$ and $\{b, c\}$ or $\{a, c\}$ and $\{b, d\}$. See Figure 2.

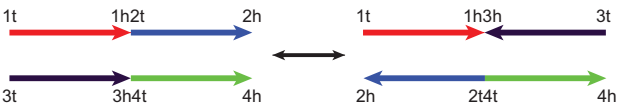

Figure 2. An illustration of the first type of mutation allowed under DCJ. The genome at right is obtained by replacing internal vertices $\{1h, 2t\}$ and $\{3h, 4t\}$ in the leftmost genome with internal vertices $\{1h, 3h\}$ and $\{2t, 4t\}$.

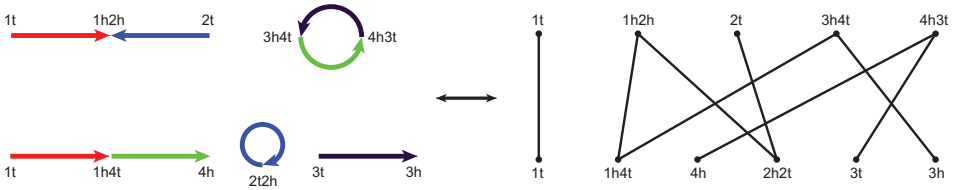

Figure 3. The bipartite adjacency graph constructed from two multi-chromosomal genomes.

ii. An internal vertex $\{a, b\}$ and an external vertex $\{c\}$ can be replaced with new internal and external vertices $\{a, c\}$ and $\{b\}$ or $\{b, c\}$ and $\{a\}$.

iii. Two external vertices $\{a\}$ and $\{b\}$ can be replaced by an internal vertex $\{a, b\}$.

iv. An internal vertex $\{a, b\}$ can be replaced by two external vertices $\{a\}$ and $\{b\}$.

Any genome can be represented by a distinct arrangement of adjacencies and telomeres. Bergeron, Mixtacki, and Stoye found that the DCJ distance between two genomes is completely determined by a bipartite graph whose vertices correspond to the sets of adjacencies and telomeres of the two genomes. In this graph, two vertices are connected with an edge for every head or tail that they share (see Figure 3). The DCJ distance between the genomes can be determined based on the number of cycles and odd-length paths in this graph.

Theorem 1 [Bergeron et al. 2006]. *The DCJ distance between two genomes, A and B, defined on the same set of N genes, is given by*

$$d_{DCJ}(A, B) = N - (C + I/2),$$

where C is the number of cycles and I is the number of odd-length paths in the adjacency graph of A and B.

Consider Figure 3, which depicts two genomes and their adjacency graph. Notice that the adjacency $\{3h, 4t\}$ in the first genome is connected to the adjacency $\{1h, 4t\}$ and the telomere $\{3h\}$ in the second genome. Using Theorem 1, we can calculate that the DCJ distance is $4 - (0 + 2/2) = 3$ because there are no cycles and two odd-length paths in the adjacency graph. The following sequence of three DCJ operations demonstrates one way that the first genome may be transformed into the second genome using the fewest number of mutations.

Operation 1: Replace internal vertices $\{1h, 2h\}$ and $\{3h, 4t\}$ with internal vertices $\{1h, 4t\}$ and $\{2h, 3h\}$; see DCJ operation i. This is a linearization and an insertion.

Operation 2: Exchange internal vertex $\{2h, 3h\}$ and external vertex $\{2t\}$ for internal vertex $\{2h, 2t\}$ and external vertex $\{3h\}$; see DCJ operation ii. This constitutes a translocation and a circularization.

Operation 3: Replace internal vertex $\{4h, 3t\}$ with external vertices $\{4h\}$ and $\{3t\}$; see DCJ operation iv. This models a translocation.

3. Counting maximally distant genomes

Building on the result of Theorem 1, we observed that the maximum distance between two genomes defined on N genes is N and occurs when $C + I/2 = 0$. This means that there are no cycles and no odd-length paths in the adjacency graph of two maximally distant genomes. We established the following result by considering

an arbitrary starting genome defined on N genes and counting the number of distinct adjacency graphs that could be created from it, where each adjacency graph contained only even-length paths.

Theorem 2. *The number of genomes that are the maximum DCJ distance away from a genome containing* $2m$ *telomeres and* n *adjacencies is given by*

$$G_{\max}(m,n) = (2m-1)!! \sum_{k=0}^{n} \binom{m+n-1}{k} \binom{n}{k} k!\, 2^k.$$

Proof. We count the number of distinct adjacency graphs that contain exclusively even-length paths, where the upper genome contains m pairs of telomeres and n adjacencies. In this proof and in subsequent proofs, we refer to upper and lower vertices as those adjacencies and telomeres located in the upper and lower genomes, respectively. Consider the following procedure:

1. Sum over the number of even-length paths j. The minimum number is m since each even-length path may contain no more than one pair of upper telomeres. The maximum number of even-length paths is $m+n$ since each pair of upper telomeres and each upper adjacency may be in a path of length two. Note that if we let $k = m+n-j$, then the sum from $j = m$ to $m+n$ becomes the sum from $k = 0$ to n.
2. Place the $2m$ upper telomeres into pairs. Each pair will define the endpoints of an even-length path. There are $(2m-1)(2m-3)\cdots 3\cdot 1 = (2m-1)!!$ ways to accomplish this.
3. Arrange the upper adjacencies into the order that they will appear in the even-length paths. This can be done in $n!$ ways. The next step will involve partitioning these adjacencies into paths.
4. The even-length paths are constructed in the following way. We begin by partitioning the n upper adjacencies into j even-length paths. Since each path must be nonempty, one adjacency must be placed into each of the $j-m$ paths without upper telomeres. The number of ways to do this is

$$\binom{(n-j+m)+(j-1)}{j-1} = \binom{m+n-1}{j-1}.$$

   In addition, once all of the upper vertices have been arranged and assigned to paths, there are two choices for how to connect each upper adjacency with its neighbors on its path. (For instance, if a path contains ordered upper vertices $\{1t\}$, $\{1h, 2t\}$, $\{2h\}$, there are two possibilities for the lower adjacencies in between: $\{1t, 1h\}$, $\{2t, 2h\}$ and $\{1t, 2t\}$, $\{1h, 2h\}$.) To accomplish this, we multiply by two for every upper adjacency except for those in paths of length two.

We have overcounted since the even-length paths we are creating are non-directed and the upper adjacencies in each non-upper-telomere-containing path can be in right-to-left or left-to right-order (except of course for paths of length two). For the upper-telomere-containing paths, this ordering of the upper adjacencies is significant because it determines which telomere is adjacent to which adjacency along the path. Hence, we divide by two for every non-upper-telomere-containing even-length path of length greater than two. Since the number of non-upper-telomere-containing paths of length two is equivalent to the number of upper adjacencies in even-length paths of length two, we multiply by 2^{m+n-j}.

We have overcounted further since the paths that do not contain upper telomeres are not distinct. To resolve this situation, we divide by $(j-m)!$.

Combining these four steps yields

$$G_{\max}(m,n) = \sum_{j=m}^{m+n} (2m-1)!!\, n! \binom{m+n-1}{j-1} \frac{2^{m+n-j}}{(j-m)!}. \tag{1}$$

By defining $k = m+n-j$, and rearranging the summation above, we obtain

$$G_{\max}(m,n) = (2m-1)!! \sum_{k=0}^{n} \binom{m+n-1}{k} \binom{n}{k} k!\, 2^k.$$

Alternatively, if we define $k = j-m$ in (1), we have

$$G_{\max}(m,n) = (2m-1)!! \sum_{k=0}^{n} \binom{m+n-1}{n-k} \frac{n!}{k!} 2^{n-k}, \tag{2}$$

which is useful in simplifying the formula of Theorem 7 below. □

Next, fix m, and consider the collection of genomes having a $2m$ telomeres and a variable number of adjacencies n. For such a collection, we obtain an infinite sequence $\{g_m^n\}$ over n, where each term represents the number of maximally distant genomes from a genome having $2m$ telomeres and n adjacencies. For example, the sequence associated with a genome containing two pairs of external vertices is

$$3, 15, 111, 1083, 13083, \ldots, g_2^n, \ldots$$

where g_2^n is given by $G_{\max}(2,n)$ and represents the number of maximally distant genomes from a starting genome with two pairs of external vertices and n internal vertices. We now find the exponential generating function for this sequence.

Lemma 3.

$$\left(\frac{1}{1-2x}\right)^{m+n} = \sum_{j=0}^{\infty} \binom{m+n+j-1}{j} (2x)^j.$$

Proof. We have $\frac{1}{1-2x} = \sum_{i=0}^{\infty} (2x)^i$ and hence

$$\left(\frac{1}{1-2x}\right)^{m+n} = \left(\sum_{i=0}^{\infty} (2x)^i\right)^{m+n}$$
$$= \underbrace{(1+2x+\cdots+(2x)^i+\cdots)\cdots(1+2x+\cdots+(2x)^i+\cdots)}_{m+n \text{ terms}}.$$

Next, consider this multiplication in a combinatorial sense where the resulting product, an infinite series, is formed term by term and where each term is the product of $m+n$ elements, one coming from each initial series. When adding these terms, consider the coefficient of x^j. Using a bijection to a familiar problem of placing j balls into $m+n$ bins, one can count the number of terms having degree j, and then, 2^j can be factored from each term. Thus, the coefficient of $(2x)^j$ is simply the number of terms having degree j and is expressed by

$$\binom{m+n+j-1}{j}.$$

Furthermore, the sum of all x^j and their coefficients is equivalent to the product of the $m+n$ series. That is,

$$\underbrace{(1+2x+\cdots+(2x)^i+\cdots)\cdots(1+2x+\cdots+(2x)^i+\cdots)}_{m+n \text{ terms}}$$
$$= \sum_{j=0}^{\infty} \binom{m+n+j-1}{j} (2x)^j.$$

Thus,

$$\left(\frac{1}{1-2x}\right)^{m+n} = \sum_{j=0}^{\infty} \binom{m+n+j-1}{j} (2x)^j. \qquad \square$$

Theorem 4. *The exponential generating function for the sequence* $\left\{g_m^n\right\}$ *is*

$$g_m(x) = \left(\frac{(2m-1)!}{2^{m-1}(m-1)!}\right) \frac{e^{\frac{x}{1-2x}}}{(1-2x)^m},$$

where the n-th term of the sequence $\left\{g_m^n\right\}$*, or* $G_{\max}(m,n)$*, is given by* $g_m^{(n)}(0)$.

Proof. We have

$$\frac{e^{\frac{x}{1-2x}}}{(1-2x)^m} = \sum_{n=0}^{\infty} \frac{\left(\frac{x}{1-2x}\right)^n}{n!\,(1-2x)^m} = \sum_{n=0}^{\infty} \frac{x^n}{n!} \left(\frac{1}{1-2x}\right)^{m+n}.$$

Using Lemma 3, we obtain

$$\frac{e^{\frac{x}{1-2x}}}{(1-2x)^m} = \sum_{n=0}^{\infty}\left(\frac{x^n}{n!}\sum_{j=0}^{\infty}\binom{m+n+j-1}{j}(2x)^j\right).$$

Expanding this series yields

$$\begin{aligned}
&\frac{e^{\frac{x}{1-2x}}}{(1-2x)^m}\\
&\quad=\binom{m-1}{0}+\binom{m}{1}(2x)+\cdots+\binom{m+j-1}{j}(2x)^j+\cdots\\
&\qquad+\binom{m}{0}x+\binom{m+1}{1}(2x)x+\cdots+\binom{m+j}{j}(2x)^j x+\cdots\\
&\qquad+\cdots\\
&\qquad+\binom{m+n-1}{0}\frac{x^n}{n!}+\binom{m+n}{1}(2x)\frac{x^n}{n!}+\cdots+\binom{m+n+j-1}{j}(2x)^j\frac{x^n}{n!}+\cdots.
\end{aligned}$$

By looking at the coefficient of each power of x, the following infinite series is created (consider a diagonal argument).

$$\begin{aligned}
\frac{e^{\frac{x}{1-2x}}}{(1-2x)^m} &= \sum_{n=0}^{\infty}\left(x^n\sum_{k=0}^{n}\binom{m+n-1}{k}\frac{2^k}{(n-k)!}\right)\\
&= \sum_{n=0}^{\infty}\left(x^n\frac{n!}{n!}\sum_{k=0}^{n}\binom{m+n-1}{k}\frac{k!}{k!}\frac{2^k}{(n-k)!}\right)\\
&= \sum_{n=0}^{\infty}\left(x^n\frac{1}{n!}\sum_{k=0}^{n}\binom{m+n-1}{k}\frac{n!\,k!\,2^k}{k!\,(n-k)!}\right)\\
&= \sum_{n=0}^{\infty}\left(\frac{x^n}{n!}\sum_{k=0}^{n}\binom{m+n-1}{k}\binom{n}{k}k!2^k\right).
\end{aligned}$$

Thus,

$$\begin{aligned}
&\left(\frac{(2m-1)!}{2^{m-1}(m-1)!}\right)\frac{e^{\frac{x}{1-2x}}}{(1-2x)^m}\\
&\qquad=\sum_{n=0}^{\infty}\left(\frac{x^n}{n!}\left(\frac{(2m-1)!}{2^{m-1}(m-1)!}\right)\sum_{k=0}^{n}\binom{m+n-1}{k}\binom{n}{k}k!\,2^k\right). \quad \square
\end{aligned}$$

4. Distribution of DCJ distance

To understand the way in which distance from a given genome is distributed across all genomes, we count the total number of genomes that are each distance away from an arbitrary genome. Employing Theorem 1, we recognize that a destination

genome is distance d away from a starting genome precisely when there are a total of $N - d$ cycles and pairs of odd-length paths in the adjacency graph between the two genomes. Consequently, we count the number of distinct adjacency graphs we can construct from a given starting genome that include exactly $N - d$ cycles and pairs of odd-length paths.

Lemma 5. *The number of ways to arrange $2p$ upper telomeres and k upper adjacencies into distinct adjacency graphs that contain exclusively odd-length paths is*

$$k! \binom{2p+k-1}{k} 2^k.$$

Proof. Consider the following counting procedure.

1. We begin by arranging the k upper adjacencies according to the order that they will appear in the odd-length paths. This can be accomplished in $k!$ ways.
2. Next, partition the k upper adjacencies into $2p$ odd-length paths. Each of these paths is distinct because it contains a distinct upper telomere. The number of ways to do this is
$$\binom{2p+k-1}{k}.$$
3. Once all of the upper vertices have been assigned to paths and have been arranged, there are two choices for how to connect each upper adjacency with its neighbors on its path. (For instance, if an odd-length path contains ordered upper vertices $\{1t\}$ and $\{1h, 2t\}$, there are two possibilities for the lower adjacencies in between: $\{1t, 1h\}$, $\{2t\}$ and $\{1t, 2t\}$, $\{1h\}$). To accomplish this, we multiply by 2^k.

Multiplying these terms yields

$$k! \binom{2p+k-1}{k} 2^k. \qquad \square$$

Lemma 6. *The number of ways to arrange i upper internal vertices into distinct adjacency graphs that contain q cycles and no even-length or odd-length paths is*

$$s(i, q)2^{i-q},$$

where $s(a, b)$ are the unsigned Stirling numbers of the first kind.

Proof. The unsigned Stirling numbers of the first kind $s(i, q)$ count the number of permutations of i elements (the upper adjacencies) into q disjoint cycles. Note that for this Stirling sequence, the clockwise or counterclockwise orientation of each cycle that contains more than two upper adjacencies is distinct. If we impose a lexicographic ordering of the upper adjacencies in each cycle, the clockwise or counterclockwise orientation of these adjacencies can represent the two ways

in which the adjacency with the smallest value in the lexicographic ordering can connect with its neighbors in the cycle. (Suppose the upper adjacency $\{1h, 2t\}$ has the smallest value in its cycle with respect to the lexicographic ordering. $\{1h, 2t\}$ can connect to its left neighbor through a lower adjacency that contains the end $1h$ or through a lower adjacency that contains the end $2t$. The end that $\{1h, 2t\}$ contributes to the lower adjacency to its right is determined by this choice.)

Once all of the upper adjacencies have been arranged into cycles, there are two choices for how to connect each upper adjacency with its neighbors on its path. (If a path contains ordered upper adjacencies $\{1h, 2t\}$, $\{2h, 1t\}$, $\{3t, 3h\}$, there are eight possibilities for the lower adjacencies in between. These include $\{2t, 2h\}$, $\{1t, 3t\}$, $\{3h, 1h\}$ and $\{2t, 1t\}$, $\{2h, 3t\}$, $\{3h, 1h\}$ for example.) We have already connected one upper adjacency to its neighbors in each cycle that has more than two upper adjacencies. To connect the others in these cycles, we multiply by two for each additional upper adjacency. For cycles containing exactly one upper adjacency, there is only one way to create the lower adjacency in the cycle. For cycles containing exactly two upper adjacencies, there are two ways to form the two lower adjacencies. (For $\{1h, 3t\}$ and $\{2t, 2h\}$ the lower adjacencies could be $\{1h, 2t\}$, $\{3t, 2h\}$ or $\{1h, 2h\}$, $\{3t, 2t\}$.)

Hence, we multiply by two for every upper adjacency in a cycle beyond the first upper adjacency in that cycle. Since there are q cycles, we multiply by 2^{i-q}.

Collecting everything together yields $s(i, q)2^{i-q}$. □

Combining Lemmas 6 and 5 and Theorem 2, we establish the following result that classifies all genomes according to their distance from a given genome.

Theorem 7. *The number of genomes that are a distance d away from a starting genome having $2m$ telomeres and n adjacencies is*

$$G(m,n,d) = \sum_{c=\max\{0,n-d\}}^{\min\{n,m+n-d\}} \sum_{i=c}^{n} \sum_{j=0}^{n-i} \sum_{k=0}^{n-i-j} \frac{s(i,c)n!\,(2(d+c-n)-1)!!}{i!\,k!\,2^{c+k-n}} \binom{2m}{2(d+c-n)} \times \binom{2(m+n-d-c)+j-1}{j} \binom{d+c-i-j-1}{n-i-j-k},$$

where $s(a, b)$ are the unsigned Stirling numbers of the first kind.

Proof. We count the number of distinct adjacency graphs from a genome with $2m$ telomeres and n adjacencies, where each graph contains a total of $N - d$ cycles and pairs of odd-length paths (the remaining paths are of even length). Let i and j represent the number of upper adjacencies in cycles and in odd-length paths, respectively, and define c to be the number of cycles in the adjacency graph. Consider the following counting procedure.

1. We begin by summing over the number of cycles c in the adjacency graph. The minimum number is $\max\{0, n-d\}$ because the number of cycles must be nonnegative, and we are restricted by the maximum number of odd-length paths that can be formed (recall that the adjacency graph must have $N-d$ cycles and pairs of odd-length paths). Since each odd-length path must contain an upper telomere, the number of odd-length paths that can be formed is at most $2m$. Recall that $N-d = I/2+c$, where I is the number of odd-length paths in the adjacency graph (Theorem 1). It follows that $2m$ is the maximum value for I, and in this case, $N-d = m+c$. Substituting $m+n$ for N and simplifying yields $c = n-d$.

   The maximum number of cycles that can be in the adjacency graph is $\min\{n, m+n-d\}$. We are restricted by the number of upper adjacencies n since each cycle contains exclusively adjacencies. We are also restricted by $N-d$ since the total number of cycles and pairs of odd-length paths must not exceed $N-d = m+n-d$.

2. Next, we sum over i, the number of upper adjacencies that are in cycles. This is at least c and at most n.

3. We then sum over j, the number of upper adjacencies that are in odd-length paths. j can be 0, but it must not exceed $n-i$ since $n-i$ is the number of upper adjacencies that remain after the first two steps.

4. Now, we choose $2(m+n-d-c)$ upper telomeres to be in odd-length paths. Notice that after we have decided on the number of cycles c, we know from Theorem 1 that there are $N-d-c = m+n-d-c$ pairs of odd-length paths in the adjacency graph. Thus, we multiply by

$$\binom{2m}{2(m+n-d-c)} = \binom{2m}{2(d+c-m)}.$$

5. Next, we pick the upper adjacencies that are in cycles and those that are in odd-length paths. This can be done in

$$\binom{n}{i}\binom{n-i}{j}$$

   ways.

6. We now arrange into odd-length paths the $n+m-d-c$ pairs of upper telomeres and j upper adjacencies that we have selected to be in odd-length paths. From Lemma 5, the number of ways to do this is

$$j!\binom{2(n+m-d-c)+j-1}{j}2^j.$$

7. We proceed by arranging into c cycles, the i upper adjacencies that we have selected for this purpose. Lemma 6 establishes that there are

$$s(i,c)2^{i-c}$$

ways to accomplish this, where $s(a,b)$ are the unsigned Stirling numbers of the first kind.

8. The remaining $2(d+c-n)$ upper telomeres and $n-i-j$ upper adjacencies are placed into paths of even length. There are

$$(2(d+c-n)-1)!!\sum_{k=0}^{n-i-j}\binom{(d+c-n)+(n-i-j)-1}{n-i-j-k}\frac{(n-i-j)!}{k!}2^{n-i-j-k}$$

ways to do this by (see Equation (2) in the proof of Theorem 2).

We now combine these eight steps and place the sums together, using the abbreviation

$$\sum{}' = \sum_{c=\max\{0,n-d\}}^{\min\{n,m+n-d\}}\sum_{i=c}^{n}\sum_{j=0}^{n-i}\sum_{k=0}^{n-i-j}$$

for simplicity. We obtain

$$\begin{aligned}
&G(m,n,d)\\
&=\sum{}'\binom{2m}{2(d+c-m)}\binom{n}{i}\binom{n-i}{j}\\
&\qquad\times j!\binom{2(n+m-d-c)+j-1}{j}2^j s(i,c)2^{i-c}\\
&\qquad\times(2(d+c-n)-1)!!\binom{d+c-i-j-1}{n-i-j-k}\frac{(n-i-j)!}{k!}2^{n-i-j-k}\\
&=\sum{}'\frac{s(i,c)\,(2(d+c-n)-1)!!\,j!\,(n-i-j)!}{k!\,2^{c+k-n}}\binom{n}{i}\binom{n-i}{j}\\
&\qquad\times\binom{2m}{2(d+c-n)}\binom{2(m+n-d-c)+j-1}{j}\binom{d+c-i-j-1}{n-i-j-k}\\
&=\sum{}'\frac{s(i,c)\,(2(d+c-n)-1)!!\,j!\,(n-i-j)!\,n!\,(n-i)!}{k!\,2^{c+k-n}i!\,(n-i)!\,(n-i-j)!\,j!}\\
&\qquad\times\binom{2m}{2(d+c-n)}\binom{2(m+n-d-c)+j-1}{j}\binom{d+c-i-j-1}{n-i-j-k}\\
&=\sum{}'\frac{s(i,c)n!\,(2(d+c-n)-1)!!}{i!\,k!\,2^{c+k-n}}\binom{2m}{2(d+c-n)}\\
&\qquad\times\binom{2(m+n-d-c)+j-1}{j}\binom{d+c-i-j-1}{n-i-j-k}.\quad\square
\end{aligned}$$

Remark. Let $G(m, n, d)$ be the number of genomes that are a distance d away from a starting genome having $2m$ telomeres and n adjacencies, and let $G_{\max}(m, n)$ be the number of genomes that are the maximum DCJ distance away from the same genome. Then $G_{\max}(m, n) = G(m, n, m+n)$.

In the case where $d = m+n$ the inner two sums in $G(m, n, m+n)$ collapse with $c = i = 0$, and we have,

$$G(m, n, m+n) = \sum_{j=0}^{n} \sum_{k=0}^{n-j} \frac{s(0,0)n!\,(2m-1)!!}{0!\,k!\,2^{k-n}} \binom{2m}{2m} \binom{j-1}{j} \binom{m+n-j-1}{n-j-k}.$$

Since $\binom{j-1}{j} = 0$ unless $j = 0$, the outer sum collapses, and we obtain

$$\begin{aligned} G(m, n, m+n) &= \sum_{k=0}^{n} \frac{n!\,(2m-1)!!}{k!} 2^{n-k} \binom{m+n-1}{n-k} \\ &= (2m-1)!! \sum_{k=0}^{n} \binom{m+n-1}{n-k} \frac{n!}{k!} 2^{n-k} \\ &= G_{\max}(m, n). \end{aligned}$$

Theorem 8. *Let $G(m, n, d)$ be the number of genomes that are a distance d away from a starting genome having $2m$ telomeres and n adjacencies. Then, $G(m, 0, d) = G(m-1, 1, d)$.*

Proof. From Theorem 7 we have

$$\begin{aligned} &G(m, 0, d) \\ &= \sum_{c=0}^{0} \sum_{i=0}^{0} \sum_{j=0}^{0} \sum_{k=0}^{0} \frac{s(0,0)0!\,(2(d+0)-1)!!}{0!\,0!\,2^0} \binom{2m}{2(d+0)} \binom{2(m+0)+0-1}{0} \binom{d+0-1}{0} \\ &= (2d-1)!! \binom{2m}{2d}, \end{aligned}$$

$$\begin{aligned} &G(m-1, 1, d) \\ &= \sum_{c=\max\{0,1-d\}}^{\min\{1,m+1-d\}} \sum_{i=c}^{1} \sum_{j=0}^{1-i} \sum_{k=0}^{1-i-j} \frac{s(i,c)1!\,(2(d+c-1)-1)!!}{i!\,k!\,2^{c+k-1}} \binom{2(m-1)}{2(d+c-1)} \\ &\qquad \times \binom{2(m-d-c)+j-1}{j} \binom{d+c-i-j-1}{1-i-j-k}. \end{aligned}$$

In this last equation, suppose $m - d \neq 0$ and $d \neq 0$. Then, we have

$$G(m-1,1,d)=\sum_{c=0}^{1}\sum_{i=c}^{1}\sum_{j=0}^{1-i}\sum_{k=0}^{1-i-j}\frac{s(i,c)\,(2(d+c-1)-1)!!}{i!\,k!\,2^{c+k-1}}\binom{2(m-1)}{2(d+c-1)}\times\binom{2(m-d-c)+j-1}{j}\binom{d+c-i-j-1}{1-i-j-k}.$$

This sum becomes

$$\begin{aligned}
&G(m-1,1,d)\\
&\quad=(2d-1)!!\binom{2m-2}{2d}+0+2(2d-3)!!\binom{2m-2}{2d-d}2(m-d)\\
&\qquad\qquad+(2d-3)!!\binom{2m-2}{2d-2}+2(2d-3)!!\binom{2m-2}{2d-2}(d-1)\\
&\quad=(2d-1)!!\binom{2m}{2d}\frac{(2m-2d)(2m-2d-1)}{(2m)(2m-1)}\\
&\qquad\qquad+(2d-3)!!\binom{2m}{2d}\frac{(2d)(2d-1)}{(2m)(2m-1)}\,(4(m-d)+1+2(d-1))\\
&\quad=(2d-1)!!\binom{2m}{2d}\left(\frac{(2m-2d)(2m-2d-1)}{(2m)(2m-1)}+\frac{2d(4m-4d+1+2d-2)}{(2m)(2m-1)}\right).
\end{aligned}$$

Obtaining a common denominator and simplifying yields

$$G(m-1,1,d)=(2d-1)!!\binom{2m}{2d}\left(\frac{4m^2-2m}{4m^2-2m}\right)=(2d-1)!!\binom{2m}{2d}.$$

Hence, $G(m,0,d)=G(m-1,1,d)$ when $m-d\neq 0$ and $d\neq 0$. A similar argument shows that $G(m,0,d)=G(m-1,1,d)$ when $m=d$. Now, if $d=0$, we have $G(m,0,0)=G(m-1,1,0)$, since there is only one genome that is a distance 0 away from a starting genome regardless of the starting genome. Thus, in all cases, we have established that

$$G(m,0,d)=G(m-1,1,d)=(2d-1)!!\binom{2m}{2d}.\qquad\square$$

Definition 9. Consider all DCJ genomes defined the same set of N genes. Observe that for each of these genomes, $N=m+n$, where $2m$ is the number of telomeres and n is the number of adjacencies in the genome. We define the *distance distribution* on N genes with respect to n to be the distribution of genomes according to their distance from a given genome containing n adjacencies and $2(N-n)$ telomeres.

Figure 4 depicts the distance distribution on five and on ten genes for all possiblities of n adjacencies. These results contribute to the understanding of how DCJ distance is distributed over all genomes. The figure displays one property that

we have observed for every distance distribution we have considered thusfar. The following conjecture summarizes this feature.

Conjecture 1. *The distance distribution on N genes with respect to n is unimodal for $n = 0, 1, \ldots, N-1$.*

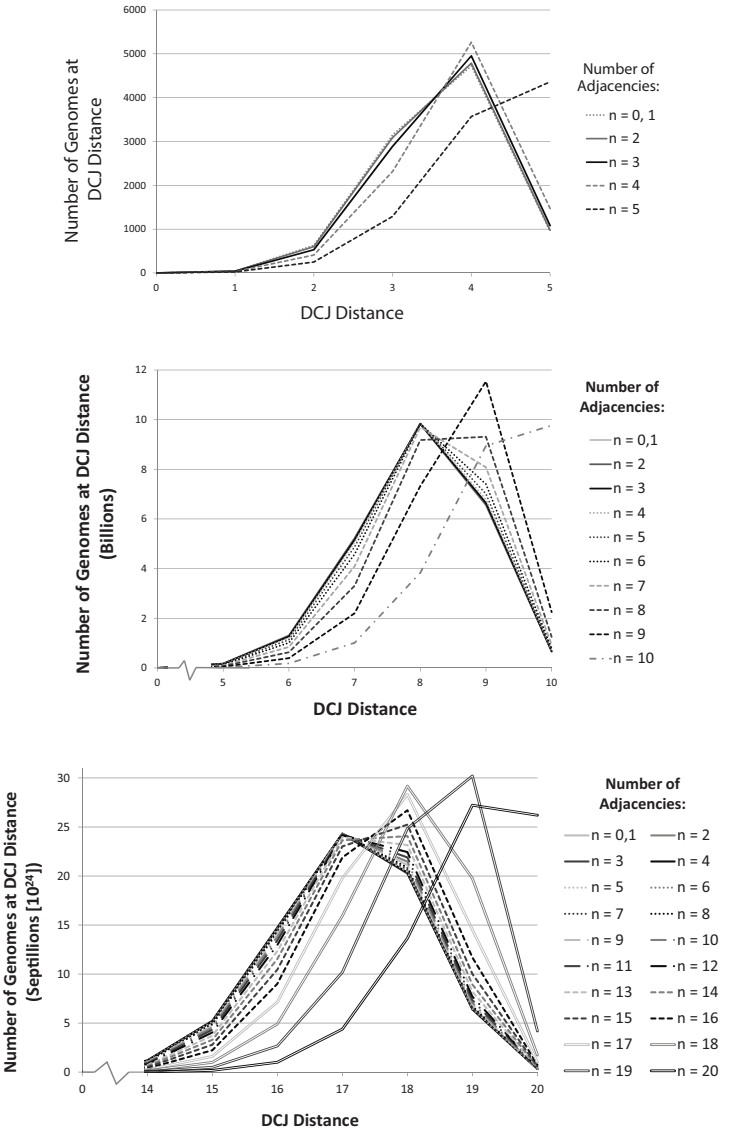

Figure 4. The distance distribution on n genes with respect to $0, 1, \ldots, n$ adjacencies, for $n = 5$ (top), $n = 10$ (middle) and $n = 20$ (bottom).

Although we have yet to prove this claim, it has been verified for all distance distributions on N genes where $1 \leq N \leq 10$, and for the cases where $N = 16$ and $N = 20$.

5. Concluding remarks

We would like to extend our results to an unsigned version of the DCJ model. Using a computer program that we created to simulate DCJ operations on unsigned genomes, we collected information about maximally distant genomes. For small values of N, we counted the total number of genomes that can be defined on a fixed number of genes.

In addition to examining maximally distant genomes, we investigated properties of the maximum distance graph M, whose vertices constitute all possible genomes of length N and whose edges link two vertices a and b whenever genome a is maximally distant from genome b. In Figure 5, we show such a graph for all genomes on three genes.

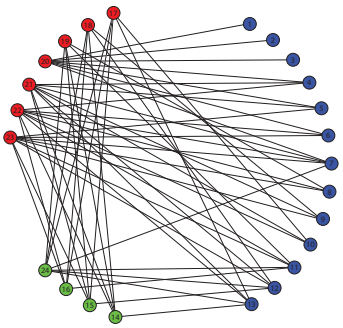

Figure 5. Maximum distance graph M for all genomes on 3 genes, with genomes labeled in lex order. It is a tripartite graph with maximally independent subsets of sizes 4, 7, and 11.

Ultimately, we would like to develop a formula for the distance between unsigned genomes and one that counts all of the possible unsigned genomes defined on a fixed set of genes. We could then extend our results from Sections 3 and 4 to the unsigned DCJ model.

References

[Bergeron et al. 2006] A. Bergeron, J. Mixtacki, and J. Stoye, "A unifying view of genome rearrangements", pp. 163–173 in *Algorithms in Bioinformatics*, edited by P. Bücher and B. M. E. Moret, Lecture Notes in Computer Science **4175**, 2006.

[Fertin et al. 2009] G. Fertin, A. Labarre, I. Rusu, É. Tannier, and S. Vialette, *Combinatorics of genome rearrangements*, MIT Press, Cambridge, MA, 2009. MR 2010e:92083

[NHGRI 2002] National Human Genome Research Institute, "The mouse genome and the measure of man", technical report (NIH News Advisory), 2002.

[Yancopoulos et al. 2005] S. Yancopoulos, O. Attie, and R. Friedberg, "Efficient sorting of genomic permutations by translocation, inversion and block interchange", *Bioinformatics* **21** (2005), 3340–3346.

Received: 2012-12-14 Revised: 2013-03-30 Accepted: 2013-04-01

christyj@uwec.edu *Department of Mathematics, University of Wisconsin, Eau Claire, WI 54702-4004, United States*

joshua.mchugh.129@gmail.com *Department of Mathematics, University of Wisconsin, Eau Claire, WI 54702-4004, United States*

riehlar@uwec.edu *Department of Mathematics, University of Wisconsin, Eau Claire, WI 54702-4004, United States*

noah.williams@colorado.edu *Department of Mathematics 340, University of Colorado, Campus Box 395, Boulder, CO 80309-0395, United States*

involve

msp.org/involve

See inside back cover or msp.org/involve for submission instructions. The subscription price for 2014 is US $120/year for the electronic version, and $165/year (+$35, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues from the last three years and changes of subscribers address should be sent to MSP.

Involve (ISSN 1944-4184 electronic, 1944-4176 printed) at Mathematical Sciences Publishers, 798 Evans Hall #3840, c/o University of California, Berkeley, CA 94720-3840, is published continuously online. Periodical rate postage paid at Berkeley, CA 94704, and additional mailing offices.

Involve peer review and production are managed by EditFLOW® from Mathematical Sciences Publishers.

PUBLISHED BY

mathematical sciences publishers

nonprofit scientific publishing

http://msp.org/

involve
2014 vol. 7 no. 4

1944-4176(2014)7:4;1-5