

Generalized Hultman Numbers and Cycle Structures of Breakpoint Graphs

Nikita Alexeev,^{*‡} Anna Pologova,[†] and Max A. Alekseyev^{*}

Abstract

Genome rearrangements can be modeled as k -breaks, which break a genome at k positions and glue the resulting fragments in a new order. In particular, reversals, translocations, fusions, and fissions are modeled as 2-breaks, and transpositions are modeled as 3-breaks. While k -break rearrangements for $k > 3$ have not been observed in evolution, they are used in cancer genomics to model chromothripsis, a catastrophic event of multiple breakages happening simultaneously in a genome. It is known that the k -break distance between two genomes (i.e., the minimum number of k -breaks required to transform one genome into the other) can be computed in terms of cycle lengths in the breakpoint graph of these genomes.

In the current work, we address the combinatorial problem of enumerating genomes at a given k -break distance from a fixed unichromosomal genome. More generally, we enumerate genome pairs, whose breakpoint graph has a given distribution of cycle lengths. We further show how our enumeration can be used for uniform sampling of random genomes at a given k -break distance, and describe its connection to various combinatorial objects such as Bell polynomials.

1 Introduction

Genome rearrangements are evolutionary events that change gene order along the genome. The genome rearrangements can be modeled as k -breaks (Alekseyev and Pevzner, 2008), which break a genome at k positions and glue the resulting fragments in a new order. While most frequent genome rearrangements such as *reversals* (which flip segments of a chromosome), *translocations* (which exchange segments of two chromosomes), *fusions* (which merge two chromosomes into one), and *fissions* (which split a single chromosome into two) can be modeled as 2-breaks (also called *Double-Cut-and-Join* or *DCJ* in Yancopoulos et al. 2005), more complex and rare genome rearrangements such as transpositions are modeled as 3-breaks. While k -break rearrangements for $k > 3$ have not been observed in evolution, they are used in cancer genomics to model *chromothripsis*, a catastrophic event of multiple breakages happening simultaneously in the genome (Stephens et al., 2011; Weinreb et al., 2014).

^{*}The George Washington University, Washington, DC, USA

[†]St. Petersburg State University, St. Petersburg, Russia

[‡]Corresponding author. Email: nikita_alexeev@gwu.edu

The *k-break distance* between two genomes is defined as the minimum number of *k*-breaks required to transform one genome into the other. The 2-break (DCJ) distance is often used in phylogenomic studies to estimate the evolutionary remoteness of genomes. The *k*-break distance between two genomes can be expressed in terms of cycles in the breakpoint graph of these genomes. Namely, while the 2-break distance depends only on the number of cycles in this graph, the *k*-break distance in general depends on the distribution of the cycle lengths (Alekseyev and Pevzner, 2008).

In the current work, we address the combinatorial enumeration of genomes at a given *k*-break distance from a fixed unichromosomal genome. More generally, for a fixed unichromosomal genome *P*, we enumerate all genomes *Q* such that the breakpoint graph of *P* and *Q* has a given distribution of cycle lengths. We consider various flavors of this problem, where genes may be arbitrarily oriented or co-oriented along the genomes,¹ while the genomes *Q* may be unichromosomal or multichromosomal. In the multichromosomal case we restrict genomes to contain only circular chromosomes, while in the unichromosomal case we consider both circular and linear genomes.

Previous studies are mostly concerned with 2-break distances between unichromosomal genomes. In particular, unichromosomal genomes with co-oriented genes can be interpreted as permutations, and the number of permutations at a given 2-break distance from the identity permutation is given by Hultman numbers (Hultman, 1999). Doignon and Labarre (2007) gave a closed formula for Hultman numbers, Bóna and Flynn (2009) proved a relation between Hultman numbers and Stirling numbers of the first kind. The case of 2-break distances between genomes with arbitrarily oriented genes was solved by Grusea and Labarre (2013). The asymptotic distribution of 2-break distances was proved to be normal by Alexeev and Zograf (2014). The analog of Hultman numbers for multichromosomal circular genomes was recently studied by Feijão et al. (2014). The current work generalizes all these results.

2 Background

We start our analysis with (multichromosomal) circular genomes and later extend it to unichromosomal linear genomes.

We represent a circular genome consisting of genes $\{1, 2, \dots, n\}$ as a *genome graph*. This graph contains $2n$ vertices: for each gene $i \in \{1, 2, \dots, n\}$, there are the *tail* and *head* vertices i^t and i^h . The graph has n directed *gene edges* of the form (i^t, i^h) encoding n genes, and n undirected *adjacency edges* connecting neighboring head/tail vertices of adjacent genes (Fig. 1a). We remark that for the genomes with co-oriented genes all adjacency edges connect the head of one gene with the tail of another.

Let *P* and *Q* be a pair of circular genomes on the same genes $\{1, 2, \dots, n\}$. We assume that in their genome graphs the adjacency edges of *P* are colored black and the adjacency edges of *Q* are colored gray. The *breakpoint graph* $G(P, Q)$ is defined on the set of vertices $\{i^t, i^h \mid i = 1, \dots, n\}$ with black and gray edges inherited from genome graphs of *P* and *Q* (Fig. 1b). Since each vertex in $G(P, Q)$ has degree 2, the black and gray edges form a

¹The case of unoriented genes is presumably much harder. For example, computing the reversal distance between unichromosomal genomes with unoriented genes is known to be NP-hard (Caprara, 1997).

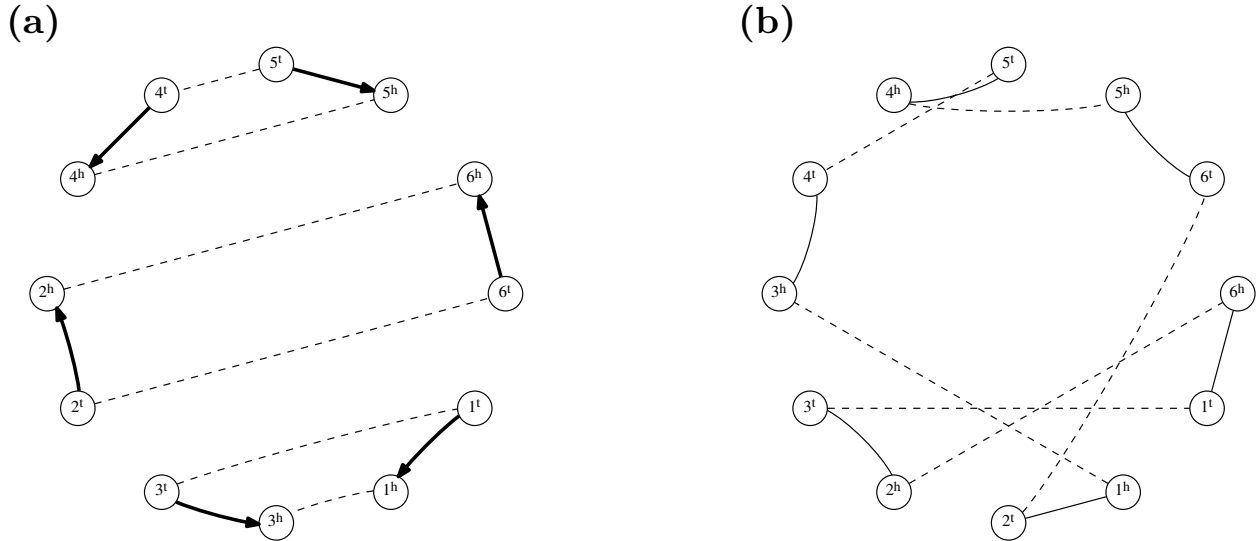


Figure 1: For genomes $P = (1, 2, 3, 4, 5, 6)$ and $Q = (1, -3)(2, -6)(4, -5)$, **(a)** the genome graph of Q ; **(b)** the breakpoint graph $G(P, Q)$, where the adjacency edges of P and Q are colored black (solid) and gray (dashed), respectively. The graph $G(P, Q)$ consists of one 2-cycle and one 4-cycle.

collection of alternating black-gray cycles. We say that a black-gray cycle is an ℓ -cycle if it is composed of ℓ black and ℓ gray edges. Let $c_\ell(P, Q)$ be the number of ℓ -cycles in $G(P, Q)$. Then the total number of black edges in $G(P, Q)$ equals

$$\sum_{\ell \geq 1} \ell \cdot c_\ell(P, Q) = n.$$

A k -break in genome Q corresponds to an operation in its genome graph and the breakpoint graph $G(P, Q)$. Namely, a k -break replaces any k -tuple of gray edges with another k -tuple of gray edges forming a matching on the same set of $2k$ vertices (Fig. 2). A transformation of genome Q into genome P with k -breaks can therefore be viewed as a transformation of the breakpoint graph $G(P, Q)$ into the breakpoint graph $G(P, P)$ with k -breaks on gray edges. The k -break distance $d_k(P, Q)$ between genomes P and Q is the minimum number of k -breaks in such a transformation.

The 2-break distance between genomes P and Q is given by the following formula (Yancopoulos et al., 2005):

$$d_2(P, Q) = n - c(P, Q), \quad (1)$$

where $c(P, Q) = \sum_{\ell \geq 1} c_\ell(P, Q)$ is the total number of cycles in $G(P, Q)$. Formulae for the k -break distance for $k > 2$ are more sophisticated. In particular, $d_3(P, Q)$ and $d_4(P, Q)$ are given by the following formulae (Alekseyev and Pevzner, 2008):

$$d_3(P, Q) = \frac{n - c^{2,1}(P, Q)}{2}, \quad (2)$$

$$d_4(P, Q) = \left\lceil \frac{n - c^{3,1}(P, Q) - \lfloor c^{3,2}(P, Q)/2 \rfloor}{3} \right\rceil, \quad (3)$$

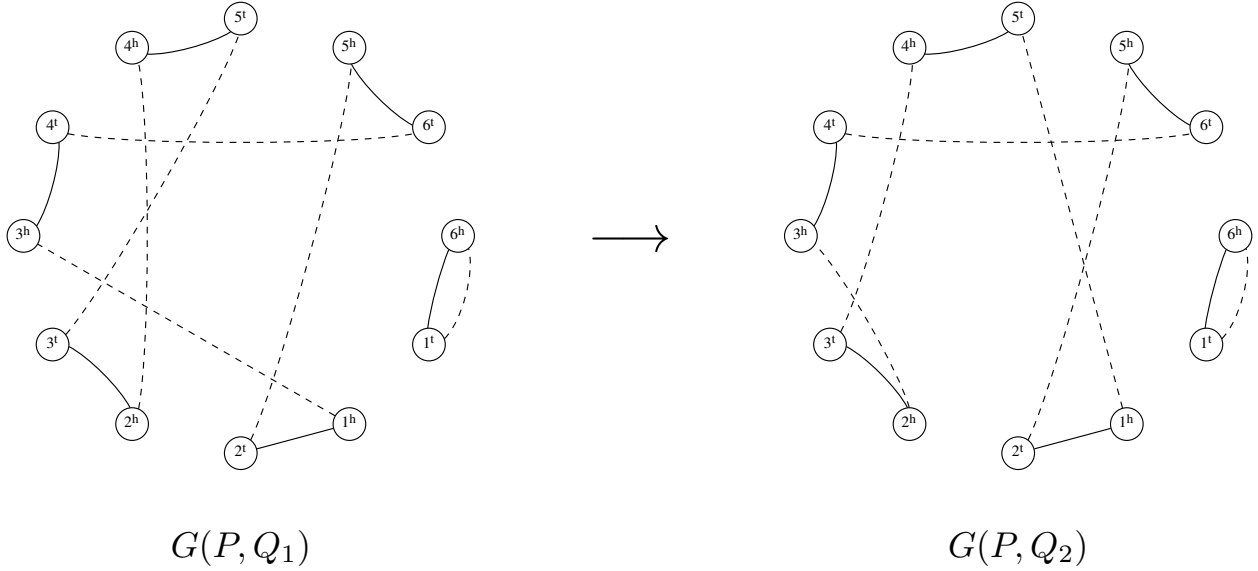


Figure 2: A 3-break transforming genome $Q_1 = (1, -3, 5, 2, -4, 6)$ into genome $Q_2 = (1, 5, 2, -3, -4, 6)$ corresponds to a transformation of the breakpoint graph $G(P, Q_1)$ into $G(P, Q_2)$ by replacing the gray edges $\{1^h, 3^h\}$, $\{2^h, 4^h\}$, and $\{3^t, 5^t\}$ with the gray edges $\{1^h, 5^t\}$, $\{2^h, 3^h\}$, and $\{3^t, 4^h\}$.

where

$$c^{m,i}(P, Q) = \sum_{\ell \equiv i \pmod{m}} c_\ell(P, Q).$$

For a fixed unichromosomal genome P with n genes and a given vector $(c_1, c_2, c_3, \dots, c_n)$ of nonnegative integers such that $\sum_{\ell=1}^n \ell \cdot c_\ell = n$, we will compute the number of genomes Q such that $G(P, Q)$ consists of c_ℓ ℓ -cycles (for each $\ell \in \{1, 2, \dots\}$). As an application, this enumeration will allow us to find the distribution of k -break distances from various genomes Q to a fixed genome P for any $k \geq 2$.

3 Genomes With A Fixed Breakpoint Graph

Let $\mathbf{c} = (c_1, c_2, c_3, \dots)$ be a sequence of nonnegative integers with a finite number of nonzero (i.e., strictly positive) terms. Then $L(\mathbf{c}) = \sum_{\ell \geq 1} \ell \cdot c_\ell$ is a finite integer. We say that a breakpoint graph has *cycle structure* \mathbf{c} if for every positive integer ℓ , the number of ℓ -cycles in this graph equals c_ℓ .

Let P be a fixed unichromosomal genome with n genes and $\mathcal{Q}_n(h; \mathbf{c})$ be the set of h -chromosomal genomes Q on the same n genes such that $G(P, Q)$ has cycle structure \mathbf{c} .² Let $M_n(h; \mathbf{c})$ be the cardinality of $\mathcal{Q}_n(h; \mathbf{c})$, i.e., $M_n(h; \mathbf{c}) = |\mathcal{Q}_n(h; \mathbf{c})|$. Clearly, we have

²We remark that the genome P essentially corresponds to a cyclic order on the genes of genomes Q . Hence, $\mathcal{Q}_n(h; \mathbf{c})$ is well-defined as soon as we are given a cyclic order on the genes. Without loss of generality, we may assume that the genes are labeled by numbers from 1 to n (up to a cyclic rotation).

$M_n(h; \mathbf{c}) = 0$ unless $L(\mathbf{c}) = n$.³ We remark that $M_n(h; \mathbf{c})$ does not depend on the order of genes in P but only on their quantity.

The generating function of numbers $M_n(h; \mathbf{c})$ is defined by

$$\begin{aligned} F(x; u; s_1, s_2, \dots) &= \sum_{\mathbf{c}} x^{L(\mathbf{c})-1} \sum_{h=1}^{\infty} M_{L(\mathbf{c})}(h; \mathbf{c}) u^{h-1} \prod_{i=1}^{\infty} s_i^{c_i} \\ &= \sum_{n=1}^{\infty} x^{n-1} \sum_{\mathbf{c}: L(\mathbf{c})=n} \sum_{h=1}^{\infty} M_n(h; \mathbf{c}) u^{h-1} \prod_{i=1}^{\infty} s_i^{c_i}. \end{aligned}$$

We remark that $F(x; u; s_1, s_2, \dots)$ at $x = 0$ equals s_1 (which corresponds to $G(P, Q)$, where $Q = P$ consists of $n = 1$ gene), while at $u = 0$ it enumerates breakpoint graphs $G(P, Q)$ for unichromosomal genomes Q .

Theorem 3.1. *The following equation, together with the initial condition $F(0; u; s_1, s_2, \dots) = s_1$, uniquely determines the generating function $F(x; u; s_1, s_2, \dots)$.*

$$\begin{aligned} \frac{\partial F}{\partial x} &= \sum_{i=2}^{\infty} \sum_{j=1}^{i-1} (i-1) s_j s_{i-j} \frac{\partial F}{\partial s_{i-1}} + \sum_{i=2}^{\infty} (i-1)^2 s_i \frac{\partial F}{\partial s_{i-1}} \\ &+ 2 \sum_{i=2}^{\infty} \sum_{j=1}^{i-1} j(i-j) s_{i+1} \frac{\partial^2 F}{\partial s_j \partial s_{i-j}} + u \sum_{i=1}^{\infty} i s_{i+1} \frac{\partial F}{\partial s_i}. \end{aligned} \tag{4}$$

Proof. The theorem statement follows from Lemma 3.2 below, which essentially restates the equation (4) as equalities of the coefficients of $x^{n-2} u^{h-1} \prod_{i \geq 1} s_i^{c_i}$ in the left- and right-hand sides of (4). Furthermore, these equalities uniquely determine the values of all $M_n(h; \mathbf{c})$ by induction on n , thus determining $F(x; u; s_1, s_2, \dots)$. \square

Lemma 3.2. *For any positive integers n, h , we have (the initial condition)*

$$M_1(h; \mathbf{c}) = \begin{cases} 1, & \text{if } h = 1 \text{ and } \mathbf{c} = \mathbf{e}_1; \\ 0, & \text{otherwise;} \end{cases}$$

³In fact, everywhere below in $M_n(h; \mathbf{c})$ we have $n = L(\mathbf{c})$, making the index n redundant. However, we find beneficial to have it as a ‘‘checksum’’ for \mathbf{c} .

and for all $n > 1$,

$$(n-1)M_n(h; \mathbf{c}) = \sum_{i=2}^{\infty} \sum_{j=1}^{i-1} (i-1)(c_{i-1} + 1 - \delta_{j,1} - \delta_{j,i-1})M_{n-1}(h; \mathbf{c} + \mathbf{e}_{i-1} - \mathbf{e}_j - \mathbf{e}_{i-j}) \quad (5)$$

$$+ \sum_{i=2}^{\infty} (i-1)^2(c_{i-1} + 1)M_{n-1}(h; \mathbf{c} + \mathbf{e}_{i-1} - \mathbf{e}_i) \quad (6)$$

$$+ 2 \sum_{i=2}^{\infty} \sum_{j=1}^{i-1} j(i-j)(c_j + 1)(c_{i-j} + 1 + \delta_{j,i-j})M_{n-1}(h; \mathbf{c} - \mathbf{e}_{i+1} + \mathbf{e}_j + \mathbf{e}_{i-j}) \quad (7)$$

$$+ \sum_{i=2}^{\infty} (i-1)(c_{i-1} + 1)M_{n-1}(h-1; \mathbf{c} + \mathbf{e}_{i-1} - \mathbf{e}_i), \quad (8)$$

where $\delta_{i,j}$ is the Kronecker delta, and $\mathbf{e}_i = (\delta_{i,1}, \delta_{i,2}, \dots)$ is a unit vector (where all coordinates except the i -th are zero).⁴

Proof. We prove the lemma statement using double counting.⁵

Let $n > 1$, $Q \in \mathcal{Q}_n(h; \mathbf{c})$, and $l \in \{1, \dots, n-1\}$. We remove gene l from both genomes P and Q to obtain new genomes P' and Q' on $n-1$ genes. Then the breakpoint graph $G(P', Q')$ can be obtained from $G(P, Q)$ by removal of vertices l^t and l^h and incident gray edges $\{l^t, a\}$, $\{l^h, c\}$ and black edges $\{l^t, b\}$, $\{l^h, d\}$, and addition of a new gray edge $\{a, c\}$ (unless $a = l^h$ and $b = l^t$) and a new black edge $\{b, d\}$ (Fig. 3). Clearly, in $G(P, Q)$ vertices a, b belong to the same black-gray cycle and so do vertices c, d . Similarly, in $G(P', Q')$ vertices b, d belong to the same black-gray cycle and so do vertices a, c (if present).

Below we analyze how the cycle structure of $G(P', Q')$ may differ from the cycle structure of $G(P, Q)$. There are four cases to consider:

Case 1. Vertices a, b belong to a different cycle in $G(P, Q)$ than vertices c and d . If these cycles are a j -cycle and a $(i-j)$ -cycle ($i > j \geq 1$), respectively, then $Q' \in \mathcal{Q}_{n-1}(h; \mathbf{c} + \mathbf{e}_{i-1} - \mathbf{e}_j - \mathbf{e}_{i-j})$ and vertices a, b, c, d belong to the same $(i-1)$ -cycle in $G(P', Q')$.

Case 2. Vertices a, b, c, d belong to the same $(i+1)$ -cycle ($i \geq 2$) in $G(P, Q)$ and their order is $(a, l^t, b, \dots, c, l^h, d, \dots)$. In this case, $Q' \in \mathcal{Q}_{n-1}(h; \mathbf{c} + \mathbf{e}_i - \mathbf{e}_{i+1})$ and vertices a, b, c, d belong to the same i -cycle in $G(P, Q)$.

Case 3. Vertices a, b, c, d belong to the same $(i+1)$ -cycle ($i \geq 2$) in $G(P, Q)$ and their order is $(a, l^t, b, \dots, d, l^h, c, \dots)$. In this case, $Q' \in \mathcal{Q}_{n-1}(h; \mathbf{c} - \mathbf{e}_{i+1} + \mathbf{e}_j + \mathbf{e}_{i-j})$, edge $\{a, c\}$ belongs to some j -cycle ($1 \leq j < i$) in $G(P', Q')$, and edge $\{b, d\}$ belongs to some $(i-j)$ -cycle in $G(P', Q')$.

⁴We remark that in (5)-(8) all indices of M are in agreement with the corresponding cycle structure, i.e., $L(\mathbf{c}) = n$ and each of $L(\mathbf{c} + \mathbf{e}_{i-1} - \mathbf{e}_j - \mathbf{e}_{i-j})$, $L(\mathbf{c} + \mathbf{e}_{i-1} - \mathbf{e}_i)$, $L(\mathbf{c} - \mathbf{e}_{i+1} + \mathbf{e}_j + \mathbf{e}_{i-j})$ equals $n-1$.

⁵A similar technique for a different enumeration problem was used in Alexeev et al. (2016).

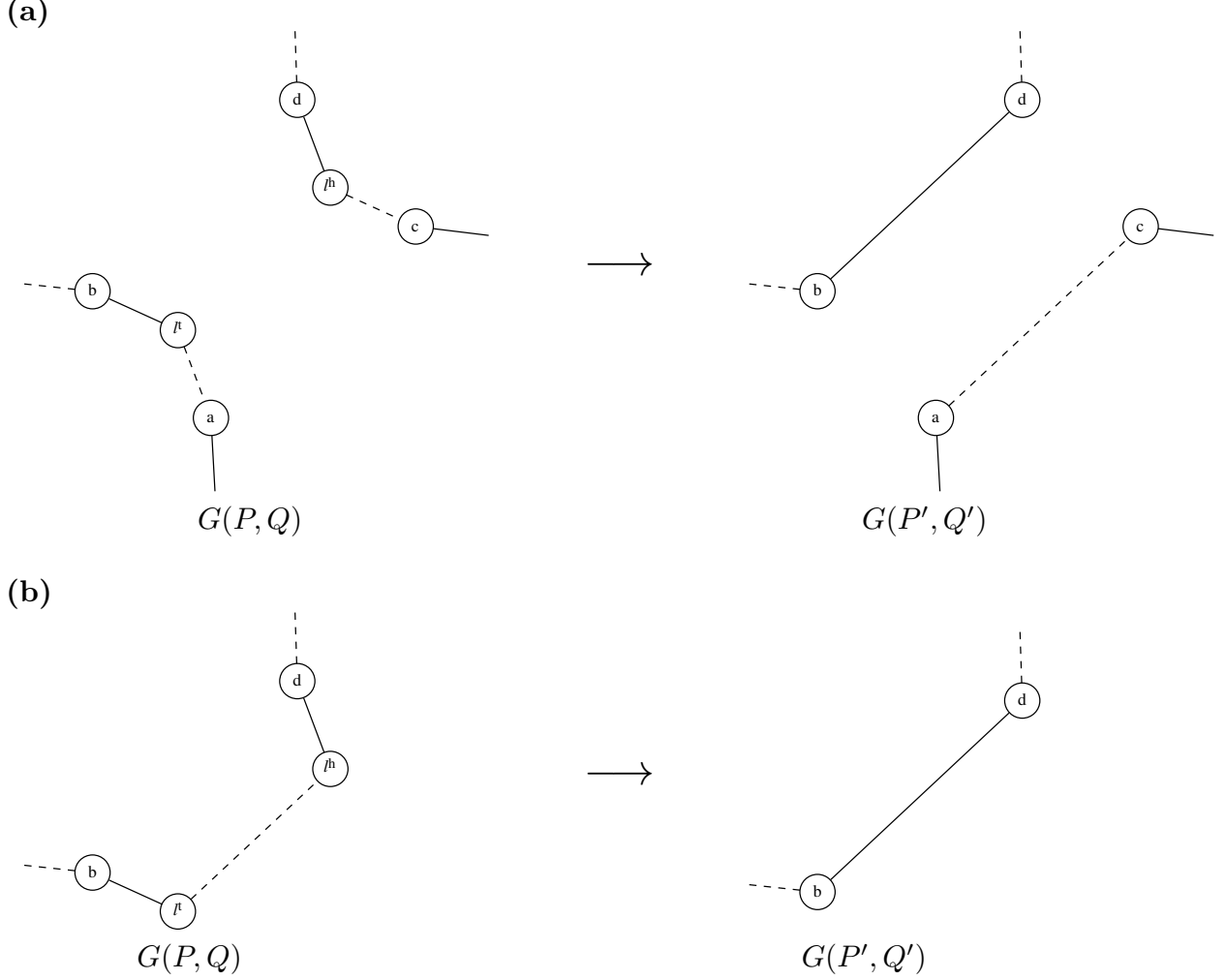


Figure 3: A transformation of breakpoint graphs corresponding to removal of gene l from genomes P and Q resulting in genomes P' and Q' . **(a)** The graph $G(P, Q)$ has no gray edge $\{l^t, l^h\}$, i.e., $a \neq l^h$ and $c \neq l^t$. **(b)** The graph $G(P, Q)$ contains the gray edge $\{l^t, l^h\}$, i.e., $a = l^h$ and $c = l^t$.

Case 4. Vertices a, c coincide with l^h, l^t , i.e., $a = l^h$ and $c = l^t$.⁶ This means that gene l forms its own chromosome in Q and this chromosome is removed in Q' (Fig. 3b). In this case, vertices a, b, c, d belong to the same i -cycle in $G(P, Q)$ for some $i \geq 2$, and b, d belong to an $(i - 1)$ -cycle in $G(P', Q')$. Hence, $Q' \in \mathcal{Q}_{n-1}(h - 1; \mathbf{c} + \mathbf{e}_{i-1} - \mathbf{e}_i)$.

We define a function Γ_l , which maps a genome Q to a pair $(Q', (a, c))$ (Cases 1-3) or a genome Q' (Case 4), where (a, c) is an ordered pair of vertices corresponding to a gray edge in $G(P', Q')$. For any integers $n > 1$ and $l \in \{1, \dots, n - 1\}$, we will prove that Γ_l is a bijection between (i) the h -chromosomal genomes Q on n genes; and (ii) the union of the h -chromosomal genomes Q' on $n - 1$ genes with a marked gray edge in $G(P', Q')$ and the

⁶We remark that a similarly looking case $b = l^h$ and $d = l^t$ is not possible, since P is a unichromosomal genome with $n > 1$ genes.

$(h - 1)$ -chromosomal genomes on $n - 1$ genes. Namely, we will show that Γ_l is invertible. Indeed, given an h -chromosomal genome Q' on genes $\{1, 2, \dots, n - 1\}$ and a pair (a, c) , we relabel the genes consecutively into $\{1, 2, \dots, l - 1, l + 1, \dots, n\}$. To reconstruct a genome Q from Q' , we insert gene l in between of the genes corresponding to vertices a and c (in the direction from a to c). Similarly, given an $(h - 1)$ -chromosomal genome Q' on genes $\{1, 2, \dots, n - 1\}$, we relabel its genes and construct genome Q from Q' by adding a new chromosome consisting of a single gene l .

To obtain a formula $M_n(h; \mathbf{c})$ for given integer $h \geq 1$ and cycle structure \mathbf{c} (with $n = L(\mathbf{c})$), we restrict functions Γ_l to the genomes $Q \in \mathcal{Q}_n(h; \mathbf{c})$. Since there are $n - 1$ values of l , the total number of pairs $(Q, \Gamma_l(Q))$ equals $(n - 1)M_n(h; \mathbf{c})$. Since each Γ_l is a bijection, this amount also equals the sum of

- number of pairs $(\Gamma_l^{-1}((Q', (a, c))), (Q', (a, c)))$, where $\Gamma_l^{-1}((Q', (a, c))) \in \mathcal{Q}_n(h; \mathbf{c})$ and Q' belongs to $\mathcal{Q}_{n-1}(h; \mathbf{c} + \mathbf{e}_{i-1} - \mathbf{e}_j - \mathbf{e}_{i-j})$, $\mathcal{Q}_{n-1}(h; \mathbf{c} + \mathbf{e}_i - \mathbf{e}_{i+1})$, or $\mathcal{Q}_{n-1}(h; \mathbf{c} - \mathbf{e}_{i+1} + \mathbf{e}_j + \mathbf{e}_{i-j})$ for some $i > j \geq 1$ (Cases 1,2,3, respectively); and
- number of pairs $(\Gamma_l^{-1}(Q'), Q')$, where $\Gamma_l^{-1}(Q') \in \mathcal{Q}_n(h; \mathbf{c})$ and $Q' \in \mathcal{Q}_{n-1}(h - 1; \mathbf{c} + \mathbf{e}_{i-1} - \mathbf{e}_i)$ (Case 4).

We consider Cases 1 and 3 in details.

In Case 1, for any given integers $i > j \geq 1$, we consider a genome $Q' \in \mathcal{Q}_{n-1}(h; \mathbf{c} + \mathbf{e}_{i-1} - \mathbf{e}_j - \mathbf{e}_{i-j})$ composed of genes $\{1, 2, \dots, n - 1\}$ and enumerate the ways to reconstruct some genome $Q \in \mathcal{Q}_n(h; \mathbf{c})$ from Q' . First, we choose an $(i - 1)$ -cycle C in $G(P', Q')$, which can be done in $c_{i-1} + 1 - \delta_{j,1} - \delta_{j,i-1}$ ways. Then we choose an integer l such that the black edge $\{(l - 1)^h, l^t\}$ belongs to C , which can be done in $i - 1$ ways. Then the cycle C has the form $((l - 1)^h, l^t, \dots, c, a, \dots)$, where there are $2i - 2j$ edges between vertices l^t and c (and thus $\{a, c\}$ represents a gray edge in C). Then we reconstruct a genome Q as $Q = \Gamma_l^{-1}((Q', (a, c)))$. Summing over the values of i, j gives the term (5) for the total number of such genomes Q .

In Case 3, for any given integers $i > j \geq 1$, we consider a genome $Q' \in \mathcal{Q}_{n-1}(h; \mathbf{c} - \mathbf{e}_{i+1} + \mathbf{e}_j + \mathbf{e}_{i-j})$ composed of genes $\{1, 2, \dots, n - 1\}$ and enumerate the number of ways to reconstruct some genome $Q \in \mathcal{Q}_n(h; \mathbf{c})$ from Q' . First, we choose a j -cycle and an $(i - j)$ -cycle in $G(P', Q')$, which can be done in $(c_j + 1)(c_{i-j} + 1 + \delta_{j,i-j})$ ways. Then we choose a gray edge $\{u, v\}$ in the j -cycle (in j ways) and choose an integer l such that the black edge $\{(l - 1)^h, l^t\}$ is in the $(i - j)$ -cycle (in $i - j$ ways). Then we reconstruct a genome Q in two ways: $Q = \Gamma_l^{-1}((Q', (u, v)))$ and $Q = \Gamma_l^{-1}((Q', (v, u)))$, which gives factor 2. Summing over the values of i, j gives the term (7) for the total number of such genomes Q .

Cases 2 and 4 follow similarly and deliver the terms (6) and (8), respectively. \square

4 Applications

4.1 Hultman Numbers

Let P be a fixed linear unichromosomal genome on n co-oriented genes and $H(n, n + 1 - d)$ be the number of linear unichromosomal genomes Q on the same co-oriented genes such that

the 2-break distance between P and Q is d . The numbers $H(n, m)$ are called *Hultman numbers* (Doignon and Labarre, 2007; Bóna and Flynn, 2009; Alexeev and Zograf, 2014) and present in the OEIS (The OEIS Foundation, 2016) as the sequence **A164652**. The problem of enumerating linear unichromosomal genomes can be reduced to enumerating circular genomes as follows. One can add a virtual gene 0 to the genomes P and Q in between of the first and last genes on their chromosomes, making them circular. Then the 2-break distance between P and Q equals $n + 1 - m$, where m is the number of cycles in the (modified) breakpoint graph $G(P, Q)$.

The Hultman numbers can be obtained from a modification of Theorem 3.1. Namely, let P be a fixed unichromosomal circular genome with genes $\{1, 2, \dots, n\}$ and let $\mathcal{Q}_n^+(h; \mathbf{c})$ be the set of h -chromosomal circular genomes Q on the same co-oriented n genes such that $G(P, Q)$ has cycle structure \mathbf{c} . Denote the cardinality of $\mathcal{Q}_n^+(h; \mathbf{c})$ by $M_n^+(h; \mathbf{c})$.

The generating functions of numbers $M_n^+(h; \mathbf{c})$ is defined by

$$G(x; u; s_1, s_2, \dots) = \sum_{n=1}^{\infty} x^{n-1} \sum_{h=1}^{\infty} u^{h-1} \sum_{\mathbf{c}:L(\mathbf{c})=n} M_n^+(h; \mathbf{c}) \prod_{i=1}^{\infty} s_i^{c_i}.$$

Theorem 4.1. *The following equation, together with the initial condition $G(0; u; s_1, s_2, \dots) = s_1$, uniquely determines the generating function $G(x; u; s_1, s_2, \dots)$.*

$$\begin{aligned} \frac{\partial G}{\partial x} &= \sum_{i=2}^{\infty} \sum_{j=1}^{i-1} (i-1) s_j s_{i-j} \frac{\partial G}{\partial s_{i-1}} \\ &+ \sum_{i=2}^{\infty} \sum_{j=1}^{i-1} j(i-j) s_{i+1} \frac{\partial^2 G}{\partial s_j \partial s_{i-j}} \\ &+ u \sum_{i=1}^{\infty} i s_{i+1} \frac{\partial G}{\partial s_i}. \end{aligned}$$

Proof. The proof is similar to the proof of Theorem 3.1 and Lemma 3.2, except that genome Q here has to have co-oriented genes and thus there is no Case 2 and there is no factor 2 for Case 3. \square

Let $F_n(u; s_1, s_2, \dots)$ and $G_n(u; s_1, s_2, \dots)$ be the coefficients of x^{n-1} in $F(x; u; s_1, s_2, \dots)$ and $G(x; u; s_1, s_2, \dots)$, respectively. The first few values⁷ of $F_n(0; s_1, s_2, \dots)$ and $G_n(0; s_1, s_2, \dots)$ corresponding to unichromosomal genomes are listed below:

$$\begin{aligned} F_1(0; s_1, s_1, \dots) &= s_1, \\ F_2(0; s_1, s_2, \dots) &= s_1^2 + s_2, \\ F_3(0; s_1, s_2, \dots) &= s_1^3 + 3s_1s_2 + 4s_3, \\ F_4(0; s_1, s_2, \dots) &= s_1^4 + 6s_1^2s_2 + (5s_2^2 + 16s_1s_3) + 20s_4, \\ F_5(0; s_1, s_2, \dots) &= s_1^5 + 10s_1^3s_2 + (40s_1^2s_3 + 25s_1s_2^2) + (100s_1s_4 + 60s_2s_3) + 148s_5. \end{aligned}$$

⁷These values are computed with Mathematica code given in Appendix.

$$\begin{aligned}
G_1(0; s_1, s_1, \dots) &= s_1, \\
G_2(0; s_1, s_2, \dots) &= s_1^2, \\
G_3(0; s_1, s_2, \dots) &= s_1^3 + s_3, \\
G_4(0; s_1, s_2, \dots) &= s_1^4 + (4s_1s_3 + s_2^2), \\
G_5(0; s_1, s_2, \dots) &= s_1^5 + (10s_1^2s_3 + 5s_1s_2^2) + 8s_5, \\
G_6(0; s_1, s_2, \dots) &= s_1^6 + (20s_1^3s_3 + 15s_1^2s_2^2) + (48s_1s_5 + 12s_3^2 + 24s_2s_4).
\end{aligned}$$

Taking $s_i = s$ for all $i = 1, 2, \dots$, we get

$$G_n(0; s, s, \dots) = \sum_{m=1}^{n+1} H(n-1, m) s^m.$$

In particular, we obtain the following formula for Hultman numbers:

$$H(n-1, m) = \sum_{\mathbf{c} \in \mathcal{C}_{n,m}} M_n^+(1; \mathbf{c}),$$

where $\mathcal{C}_{n,m} = \{\mathbf{c} : L(\mathbf{c}) = n \text{ and } \sum_{i=1}^n c_i = m\}$.

Grusea and Labarre (2013) introduced the problem of enumerating linear unichromosomal genomes, where genes may be arbitrarily oriented. The corresponding *signed Hultman numbers* $H^\pm(n, m)$ form the sequence A189507 in the OEIS. Theorem 3.1 allows us to compute these numbers as follows:

$$H^\pm(n-1, m) = \sum_{\mathbf{c} \in \mathcal{C}_{n,m}} M_n(1; \mathbf{c}). \quad (9)$$

The first few numbers $H(n, m)$ and $H^\pm(n, m)$ are listed in Table 1.

Table 1: Values of Hultman numbers.

(a) Values of $H(n, m)$.							(b) Values of $H^\pm(n, m)$.						
$n \backslash m$	1	2	3	4	5	6	$n \backslash m$	1	2	3	4	5	6
0	1						0	1					
1	0	1					1	1	1				
2	1	0	1				2	4	3	1			
3	0	5	0	1			3	20	21	6	1		
4	8	0	15	0	1		4	148	160	65	10	1	
5	0	84	0	35	0	1	5	1348	1620	701	155	15	1

4.2 Bell Polynomials

The numbers $M_n(h; \mathbf{c})$ have multiple connections to well-known combinatorial objects. Some of these connections are straightforward, and some appear to be new.

It is easy to see that in the unichromosomal case, $G_n(0; 1, 1, \dots)$ enumerates permutations of order $n - 1$, and so

$$G_n(0; 1, 1, \dots) = (n - 1)!.$$

Similarly, $F_n(0; 1, 1, \dots)$ enumerates signed permutations of order $n - 1$, and so

$$F_n(0; 1, 1, \dots) = 2^{n-1}(n - 1)!.$$

In the multichromosomal case, we get more general formulae:

$$G_n(u; 1, 1, \dots) = \sum_{h=1}^n \begin{bmatrix} n \\ h \end{bmatrix} u^{h-1}$$

and

$$F_n(u; 1, 1, \dots) = \sum_{h=1}^n 2^{n-h} \begin{bmatrix} n \\ h \end{bmatrix} u^{h-1},$$

where $\begin{bmatrix} n \\ h \end{bmatrix}$ are unsigned Stirling numbers of the first kind (A094638 in the OEIS). Moreover, for $u = 1$, we have

$$G_n(1; s_1, s_2, \dots) = \sum_{\mathbf{c}: L(\mathbf{c})=n} \frac{n!}{\prod_{i=1}^n c_i!} \prod_{i=1}^n \left(\frac{s_i}{i}\right)^{c_i}. \quad (10)$$

The numbers $L(\mathbf{c})!/(c_1!1^{c_1}c_2!2^{c_2}\dots)$ enumerate permutations with the cycle structure \mathbf{c} and form the sequence A124795 in the OEIS. The functions $G_n(1; s_1, s_2, \dots)$ are closely related to the complete exponential Bell polynomials (Comtet, 1974, Section 3.3)

$$Y_n(x_1, x_2, \dots) = \sum_{\mathbf{c}: L(\mathbf{c})=n} \frac{n!}{\prod_{i=1}^n c_i!} \prod_{i=1}^n \left(\frac{x_i}{i!}\right)^{c_i}. \quad (11)$$

Namely, from (10) and (11) it follows that

$$G_n\left(1; \frac{s_1}{0!}, \frac{s_2}{1!}, \frac{s_3}{2!}, \dots, \frac{s_k}{(k-1)!}, \dots\right) = Y_n(s_1, s_2, \dots).$$

Hence, Theorem 4.1 implies the following (apparently new) differential equation for Bell polynomials:

$$\begin{aligned} (n-1)Y_n(x_1, x_2, \dots) &= \sum_{i=2}^{\infty} \sum_{j=1}^{i-1} (i-1) \binom{i-2}{j-1} x_j x_{i-j} \frac{\partial Y_{n-1}}{\partial x_{i-1}} \\ &+ \sum_{i=2}^{\infty} \sum_{j=1}^{i-1} \frac{x_{i+1}}{\binom{i}{j}} \frac{\partial^2 Y_{n-1}}{\partial x_j \partial x_{i-j}} \\ &+ \sum_{i=2}^{\infty} x_i \frac{\partial Y_{n-1}}{\partial x_{i-1}}. \end{aligned}$$

4.3 Distribution Of k -Break Distances

Let $H_k^h(n, d)$ be the number of h -chromosomal circular genomes with n genes at the k -break distance d from a fixed unichromosomal circular genome. For $k = 2$ and $h = 1$, these numbers represent signed Hultman numbers: $H_2^1(n, d) = H^\pm(n - 1, n - d)$.

Using formulae (2) and (3), we can further obtain $H_3^h(n, d)$ and $H_4^h(n, d)$. The first few numbers $H_3^1(n, d)$, $H_3^2(n, d)$, $H_4^1(n, d)$, and $H_4^2(n, d)$ are listed in Table 2.

Table 2: Values of generalized Hultman numbers.

(a) Values of $H_3^1(n, d)$.

$n \setminus d$	0	1	2	3
1	1	0	0	0
2	1	1	0	0
3	1	7	0	0
4	1	22	25	0
5	1	50	333	0
6	1	95	1851	1893
7	1	161	6839	39079

(b) Values of $H_3^2(n, d)$.

$n \setminus d$	1	2	3
1	0	0	0
2	1	0	0
3	6	0	0
4	18	26	0
5	40	360	0
6	75	2034	2275
7	126	7588	48734

(c) Values of $H_4^1(n, d)$.

$n \setminus d$	0	1	2	3
1	1	0	0	0
2	1	1	0	0
3	1	7	0	0
4	1	47	0	0
5	1	175	208	0
6	1	470	3369	0
7	1	1036	45043	0
8	1	2002	315213	327904

(d) Values of $H_4^2(n, d)$.

$n \setminus d$	1	2	3
1	0	0	0
2	1	0	0
3	6	0	0
4	44	0	0
5	170	230	0
6	465	3919	0
7	1036	55412	0
8	2016	396764	437572

4.4 Sampling Of Random Genomes

Theorem 3.1 and Lemma 3.2 allow us to sample a (uniformly) random genome Q with given number of genes n , number of chromosomes h , and cycle structure \mathbf{c} of the breakpoint graph $G(P, Q)$. Namely, we define a Markov chain \mathcal{M} as follows:

- the states of \mathcal{M} are genome classes $\mathcal{Q}_n(h; \mathbf{c})$;
- the probability of transition between $\mathcal{Q}_n(h; \mathbf{c})$ and $\mathcal{Q}_{n-1}(h; \mathbf{c} + \mathbf{e}_{i-1} - \mathbf{e}_j - \mathbf{e}_{i-j})$ (for any $i \geq 2$ and $1 \leq j < i$) is

$$\frac{(i-1)(c_{i-1} + 1 - \delta_{j,1} - \delta_{j,i-1})M_{n-1}(h; \mathbf{c} + \mathbf{e}_{i-1} - \mathbf{e}_j - \mathbf{e}_{i-j})}{(n-1)M_n(h; \mathbf{c})};$$

- the probability of transition between $\mathcal{Q}_n(h; \mathbf{c})$ and $\mathcal{Q}_{n-1}(h; \mathbf{c} + \mathbf{e}_{i-1} - \mathbf{e}_i)$ (for any $i \geq 2$) is

$$\frac{(i-1)^2(c_{i-1}+1)M_{n-1}(h; \mathbf{c} + \mathbf{e}_{i-1} - \mathbf{e}_i)}{(n-1)M_n(h; \mathbf{c})};$$

- the probability of transition between $\mathcal{Q}_n(h; \mathbf{c})$ and $\mathcal{Q}_{n-1}(h; \mathbf{c} - \mathbf{e}_{i+1} + \mathbf{e}_j + \mathbf{e}_{i-j})$ (for any $i \geq 2$ and $1 \leq j < i$) is

$$\frac{2j(i-j)(c_j+1)(c_{i-j}+1+\delta_{j,i-j})M_{n-1}(h; \mathbf{c} - \mathbf{e}_{i+1} + \mathbf{e}_j + \mathbf{e}_{i-j})}{(n-1)M_n(h; \mathbf{c})};$$

- the probability of transition between $\mathcal{Q}_n(h; \mathbf{c})$ and $\mathcal{Q}_{n-1}(h-1; \mathbf{c} + \mathbf{e}_{i-1} - \mathbf{e}_i)$ (for any $i \geq 2$) is

$$\frac{(i-1)(c_{i-1}+1)M_{n-1}(h-1; \mathbf{c} + \mathbf{e}_{i-1} - \mathbf{e}_i)}{(n-1)M_n(h; \mathbf{c})};$$

- the probability of transition between $\mathcal{Q}_1(1; \mathbf{e}_1)$ and itself is equal to 1;
- in the other cases, the transition probability equals to 0.

Lemma 3.2 implies that the Markov chain \mathcal{M} is well-defined. For any initial state $\mathcal{Q}_n(h; \mathbf{c})$, the process after $n-1$ steps comes into the terminal state $\mathcal{Q}_1(1; \mathbf{e}_1)$, which consists of a single genome.

To sample a random genome $Q \in \mathcal{Q}_n(h; \mathbf{c})$, we first sample a random path $(\mathcal{Q}_n, \mathcal{Q}_{n-1}, \dots, \mathcal{Q}_1)$ starting at $\mathcal{Q}_n = \mathcal{Q}_n(h; \mathbf{c})$ and ending at the termination state $\mathcal{Q}_1 = \mathcal{Q}_1(1; \mathbf{e}_1)$. We start with $Q \in \mathcal{Q}_1$ (i.e., Q is a genome with a single gene) and for every j from 1 to $n-1$, we randomly add a gene into Q such that the resulting genome belongs to \mathcal{Q}_{j+1} . By construction, at the end of this process the genome Q represents a uniformly random element of $\mathcal{Q}_n(h; \mathbf{c})$.

5 Discussion

In the current work, we address the problem of enumeration of genomes with n genes that are at a given k -break distance from a fixed unichromosomal genome. It is known that the k -break distance between two genomes can be computed in terms of cycle lengths in the breakpoint graph of these genomes (Alekseyev and Pevzner, 2008).

Our main result is the recurrent formula for the numbers $M_n(h; \mathbf{c})$ (and their generating function) of breakpoint graphs with the cycle structure \mathbf{c} of h -chromosomal genomes with n genes. We show connection between these numbers and various combinatorial objects (such as Bell polynomials) and further compute numbers $H_k^h(n, d)$ of h -chromosomal genomes with n genes at the k -break distance d from a fixed unichromosomal genome, which generalize Hultman numbers (Hultman, 1999; Doignon and Labarre, 2007; Bóna and Flynn, 2009; Alexeev and Zograf, 2014; Grusea and Labarre, 2013).

We believe that our approach can further lead to finding a formula for the numbers $H_k^h(n, d)$ and then to evaluating the asymptotic distribution of the k -break distances for a general k . Other open questions of interest include enumeration of genomes Q at a given k -break distance from a fixed genome P , where (i) P is unichromosomal and Q is linear

multichromosomal (the case $k = 2$ was addressed by Feijão et al. (2014)); or (ii) P and Q are both multichromosomal. Both questions may be addressed under the assumption of co-oriented or arbitrarily oriented genes. Defining *proper* k -breaks as those that are not $(k - 1)$ -breaks, we may ask similar questions for the *graded* $(2, 3, \dots, k)$ -break distance specifying the number of proper i -breaks for each $i = 2, 3, \dots, k$. Further assuming that proper k -breaks for different k have different rates in the course of evolution, we may be able to estimate these rates from given (extant) genomes, using the technique proposed by Alexeev et al. (2015) for $k = 2, 3$.

Acknowledgements

The work is supported by the National Science Foundation under the grant No. IIS-1462107.

References

- Alekseyev, M. and Pevzner, P. (2008). Multi-break rearrangements and chromosomal evolution. *Theoretical Computer Science*, 395(2):193–202.
- Alexeev, N., Aidagulov, R., and Alekseyev, M. A. (2015). A computational method for the rate estimation of evolutionary transpositions. In Ortuño, F. and Rojas, I., editors, *Proceedings of the 3rd International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO)*, volume 9043 of *Lecture Notes in Computer Science*, pages 471–480.
- Alexeev, N., Andersen, J., Penner, R., and Zograf, P. (2016). Enumeration of chord diagrams on many intervals and their non-orientable analogs. *Advances in Mathematics*, 289:1056 – 1081.
- Alexeev, N. and Zograf, P. (2014). Random matrix approach to the distribution of genomic distance. *Journal of Computational Biology*, 21(8):622–631.
- Bóna, M. and Flynn, R. (2009). The average number of block interchanges needed to sort a permutation and a recent result of stanley. *Information Processing Letters*, 109(16):927–931.
- Caprara, A. (1997). Sorting by reversals is difficult. In *Proceedings of the first annual international conference on Computational molecular biology (RECOMB)*, pages 75–83.
- Comtet, L. (1974). *Advanced Combinatorics*. D. Reidel Publishing Company, Dordrecht, Holland.
- Doignon, J.-P. and Labarre, A. (2007). On Hultman numbers. *Journal of Integer Sequences*, 10(6):Article 07.6.2.
- Feijão, P., Martinez, F. V., and Thévenin, A. (2014). On the Multichromosomal Hultman Number. In Campos, S., editor, *Proceedings of the 9th Brazilian Symposium on Bioinformatics (BSB)*, volume 8826 of *Lecture Notes in Computer Science*, pages 9–16.

- Grusea, S. and Labarre, A. (2013). The distribution of cycles in breakpoint graphs of signed permutations. *Discrete Applied Mathematics*, 161(10):1448–1466.
- Hultman, A. (1999). Toric permutations. *Master’s thesis, Dept. of Mathematics, KTH, Stockholm, Sweden.*
- Stephens, P. J., Greenman, C. D., Fu, B., Yang, F., Bignell, G. R., Mudie, L. J., Pleasance, E. D., Lau, K. W., Beare, D., Stebbings, L. A., et al. (2011). Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell*, 144(1):27–40.
- The OEIS Foundation (2016). *The On-Line Encyclopedia of Integer Sequences*. Published electronically at <http://oeis.org>.
- Weinreb, C., Oesper, L., and Raphael, B. J. (2014). Open adjacencies and k -breaks: detecting simultaneous rearrangements in cancer genomes. *BMC Genomics*, 15(Suppl 6):S4.
- Yancopoulos, S., Attie, O., and Friedberg, R. (2005). Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics*, 21(16):3340–3346.

Appendix. Mathematica Code

Here we provide Wolfram Mathematica code for computing the functions $G_0(u; s_1, s_2, \dots)$, $G_n(u; s_1, s_2, \dots)$, $F_n(0; s_1, s_2, \dots)$, $F_n(u; s_1, s_2, \dots)$:

```
(*Implementation of the summands in the formula in Theorem 3.1*)
L0[f_, n_] :=
  Sum[Sum[(i - 1)*s[j]*s[i - j]*D[f, s[i - 1]], {j, 1, i - 1}], {i, 2,
  n}]
L1[f_, n_] := Sum[(i - 1)^2*s[i]*D[f, s[i - 1]], {i, 2, n}]
L2[f_, n_] :=
  Sum[s[i + 1]*Sum[j*(i - j)*D[f, s[j], s[i - j]], {j, 1, i - 1}], {i,
  2, n}];
Ln[f_, n_] := Sum[(i - 1)*u*s[i]*D[f, s[i - 1]], {i, 2, n}];
FG[n_, orient_, multichr_] := {ff := {s[1]}; Do[g := Last[ff];
  f := 1/k*(L0[g, n] + orient*L1[g, n] + (1 + orient)*L2[g, n]
  + multichr*Ln[g, n]);
  AppendTo[ff, Simplify[f]], {k, n}];
ff[[n]]}
```

```
(*Implementation of function G_n(0;s1,s2,...)*)
G0[n_] := FG[n, 0, 0]
(*Implementation of function G_n(u;s1,s2,...)*)
Gu[n_] := FG[n, 0, 1]
(*Implementation of function F_n(0;s1,s2,...)*)
F0[n_] := FG[n, 1, 0]
(*Implementation of function F_n(u;s1,s2,...)*)
Fu[n_] := FG[n, 1, 1]
```