Four-Regular Graphs with Rigid Vertices Associated to DNA Recombination

Jonathan Burns, Egor Dolzhenko, Nataša Jonoska[*], Tilahun Muche, Masahico Saito[*]
Department of Mathematics and Statistics
University of South Florida

May 23, 2011

Abstract

Genome rearrangement and homological recombination processes have been modeled by 4-regular spacial graphs with rigid vertices, called assembly graphs [1]. These graphs can also be represented by double occurrence words called assembly words. The rearranged DNA segments are modeled by certain types of paths in the assembly graphs called polygonal paths. The minimum number of polygonal paths visiting all vertices in a graph is called an assembly number for the graph.

In this paper, we give formulas for counting certain types of assembly graphs and assembly words. Some of these formulas produce sequences not previously reported at the Online Encyclopedia of Integer Sequences [11]. We provide a sharp upper bound for the number of polygonal paths in Hamiltonian sets of polygonal paths, and present a family of graphs that achieves this bound. We investigate changes in the assembly numbers as a result of graph compositions. Finally, we introduce polynomial invariants for assembly graphs and show properties of this invariant. These studies provide possible DNA structures during recombination processes.

1 Introduction

Epigenetic genomic rearrangements occur on both evolutionary scale and developmental scale. Massively occurring recombination processes during sexual reproduction in several species of ciliates (unicellular organisms), such as *Oxytricha* and *Stylonychia*, make them ideal model organisms to study gene rearrangements. Here we give only a very short description of the biological phenomenon and refer the reader to [2, 5] and references therein for details. These species of ciliates have two types of nuclei, micronuclear and macronuclear type, possibly in several copies. During mating, only micronuclear genes are exchanged. After the conjugation, the old macronuclei are disintegrated and new macronuclei are developed from one of the newly formed micronuclei. These DNA processing events involve global deletion of 95-98% of the germline DNA, effectively eliminating *all* so-called "junk" DNA, intervening DNA segments (internal eliminated sequences, IESs) that interrupt coding of the genes. Because IESs interrupt coding regions in the micronucleus, each macronuclear gene may appear as several nonconsecutive segments (macronuclear destined sequences, MDSs) in the

[*] Supported in part by NSF grant DMS #0900671.

micronucleus. Moreover, for thousands of genes, the order of these MDS segments in the micronuclei can be permuted or sequences reversed with respect to the micronuclear sequence. Figure 1 shows an example of a typical scrambled gene requiring a variety of recombination events.

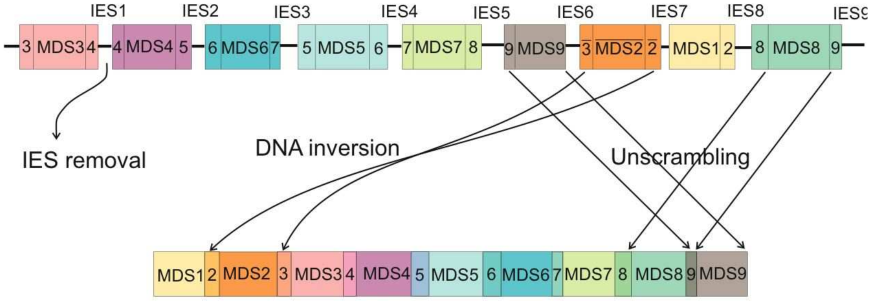


Figure 1: Scrambled Actin I gene in *O. nova* [13] and the required DNA recombination. Micronuclear sequence (top) and macronuclear sequence (bottom).

There are several theoretical models attempting to describe these DNA recombination processes [5, 6, 7, 12]. It has been conjectured that an additional molecule (called a template) takes part in the recombination process [2, 12] and experimental support for this model was obtained in [10]. Based on these observations, a theoretical model with spatial graphs depicting the molecule(s) at the time of recombination was introduced in [2]. This model describes a micronuclear gene(s) as a spatial graph, called an assembly graph, with 1- or 4-valent rigid vertices. Each 4-valent vertex represents the location of the homologous recombination. A single micronuclear gene is modeled as an assembly graph with an Eulerian path in which consecutive edges are not “neighbors” with respect to the common incident vertex. Observe that the sequence containing vertices listed in the order visited by the Eulerian path forms a (double occurrence) word, called an assembly word. The recombination events are modeled by smoothing of the 4-valent vertices. In [1] each macronuclear gene is represented as a so-called polygonal path in the graph which in turn determines the smoothing type at the vertices along these paths. In this paper we study properties of assembly words and graphs as well as polygonal paths within these graphs.

The paper is organized as follows. Section 2 recalls the definitions of assembly graphs, assembly words, polygonal paths as well as some of the relevant properties shown in [1]. In Section 3, we develop formulas for the number of assembly words and graphs with a given number of vertices of various types and classes. We note that assembly words have been studied within other contexts [15, 14] and some of the obtained formulas are known. However, we also show new formulas that generate sequences of numbers not previously reported in the Online Encyclopedia of Integer Sequences [11]. In Section 4, numbers associated with assembly graphs and their polygonal paths are investigated. Each polygonal path determines the type of smoothing associated with the vertices visited by the path and represents a potential macronuclear gene. A Hamiltonian set of polygonal paths represents a possible set of genes encoded by the micronuclear segment corresponding to the graph. We give a bound on the number of Hamiltonian polygonal paths and show that this bound is sharp. A family of graphs that achieves this bound is also presented. In Section 5, assembly numbers (that are the minimal number of Hamiltonian polygonal paths for a graph), and their behavior under graph compositions are studied. The assembly number determines the minimal number of genes encoded by the micronuclear sequence corresponding to the graph. We describe the structure of the graphs that necessarily increase the assembly number via graph compositions. In a manner similar to polynomial knot invariants, we define an assembly polynomial for an assembly graph in Secion 6. We show that the assembly polynomial is an invariant for assembly graphs and show that the all assembly polynomials are products of polynomials corresponding to “indecomposable” (i.e., strongly-irreducible) graphs.

2 Preliminary

Definitions and notations of assembly graphs and related concepts are listed below. For a more detailed explanation of the model and its connection to the genetic processes we refer the reader to [1, 2].

Definitions of assembly words, graphs, and related concepts.

A graph is a pair (V, E) consisting of the set V of vertices and the set E of edges. The endpoints of every edge is either a pair of vertices or a single vertex. In the latter case, the edge is called a *loop*. A *degree* of a vertex v is the number of edges incident to v such that each loop is counted twice. A path in a graph is an alternating sequence of vertices and edges (starting and ending with a vertex) such that consecutive vertices are incident to the edge between them. A single vertex is a path called *singleton*.

A *4-valent rigid vertex* is a vertex of degree 4 for which a cyclic order of edges is specified. For a 4-valent rigid vertex v, if its incident edges appear in order e_1, e_2, e_3, e_4, we say that e_2 and e_4 are *neighbors with respect to* v to e_1 (or e_3). Vice versa, e_1 and e_3 are neighbors to e_2 (or e_4). We note that we allow two of the edges incident to a vertex v to be equal. This does not change the degree of the vertex v, nor the definition of a neighbor. For example, if edges e_1 and e_2 are equal, then this edge is a neighbor to itself, and it is a loop. Similarly if e_1 and e_3 are equal, then this edge is not a neighbor to itself.

An *assembly graph* is a finite connected graph where all vertices are rigid vertices of valency 1 or 4. A vertex of valency 1 is called an *endpoint*. Note that the definition of assembly graph implies that the number of endpoints is always even. The number of 4-valent vertices in Γ is called the *size* of Γ and is denoted by $|\Gamma|$. The assembly graph is called *trivial* if $|\Gamma| = 0$. Two assembly graphs are *isomorphic* if they are isomorphic as graphs and the graph isomorphism preserves the cyclic order of the edges incident to a vertex.

Two types of paths are of interest: (a) paths in which consecutive edges are never neighbors with respect to their common incident vertex and (b) paths in which every pair of consecutive edges are neighbors with respect to their common incident vertex. A path of type (a) where no edge is repeated is called a *transverse* path, or simply a *transversal*. A path of type (b) where no vertex is repeated is called a *polygonal* path. Graphs that have an Eulerian transversal are called *simple assembly graphs*. We note that in a simple assembly graph, if a vertex v is an endpoint of a loop e, then e must be a neighbor of itself.

Both transversals and polygonal paths model specific parts of the DNA rearrangement phenomenon. In particular, the transversal represents the micronuclear DNA segment prior to the recombination, the rigid vertices indicate the recombination sites, and polygonal paths are DNA segments after the rearrangements. We are interested to see what types of rearrangements and how many distinct genes can be encoded in a single micronuclear DNA sequence. Therefore,we are interested in graphs with Eulerian transversals, i.e., simple assembly graphs.

Convention: In the rest of our exposition, unless otherwise stated, all graphs are simple assembly graphs with two endpoints.

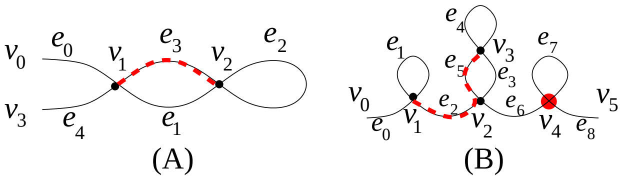


Figure 2: Simple assembly graphs and polygonal paths in dotted lines. In (B) there are two polygonal paths, one of which is a singleton containing the vertex v_4.

Figure 2 depicts two examples of assembly graphs. Eulerian transverse paths for (A) and (B) are respectively given by, given by

$$(v_0, e_0, v_1, e_1, v_2, e_2, v_2, e_3, v_1, e_4, v_3), \quad \text{and}$$
$$(v_0, e_0, v_1, e_1, v_1, e_2, v_2, e_3, v_3, e_4, v_3, e_5, v_2, e_6, v_4, e_7, v_4, e_8, v_5).$$

A path may be given as a sequence of edges only (omitting the vertices) when the sequence of edges uniquely determines the vertices of the path. With this convention the transverse paths for the graphs in (A) and (B) in Figure 2 are $(e_0, e_1, \dots, e_4)$ and $(e_0, e_1, \dots, e_8)$, respectively.

Two transverse paths with endpoints are *equivalent* if they are either identical, or one is the reverse of the other. In [1] it was shown that two simple assembly graphs with two endpoints are isomorphic if and only if their Eulerian transversals are equivalent.

Given a simple assembly graph Γ, designate one of the endpoints as initial (i) and the other endpoint as terminal (t). We call such Γ an *oriented* (or *directed*) simple assembly graph with direction from i to t. We consider the transverse path of a directed simple assembly graph as a path starting at the vertex i and terminating at the vertex t.

An *assembly word* is a word in which every symbol appears exactly twice. We refer to these words also as *double occurrence words*. The *reverse* of a word $w = a_1 \cdots a_k$ is $w^R = a_k \cdots a_1$. If an assembly word w can be written as a product $w = uv$ of two non-empty assembly words u, v, then w is called *reducible*; otherwise, it is called *irreducible*. Two double occurrence words are called *equivalent* if, after renaming the symbols, either they are identical or one is identical to the reverse of the other. For example, $w = 123231$ is equivalent to its reverse $w^R = 132321$ and $w' = 213132$ is w^R after renaming 1 with 2 and 2 with 1. Therefore,all these words are equivalent. Assembly words are related to assembly graphs as follows.

Let Γ be an oriented simple assembly graph with an initial vertex i and a terminal vertex t. Let the set of 4-valent vertices of Γ be $V = \{v_1, \dots, v_n\}$ where $n = |\Gamma|$. Starting from i, write down the sequence of vertices in the order they are encountered along the transversal. This is an assembly word over alphabet V. Thus an oriented assembly graph gives rise to an assembly word, and it is known that equivalence classes of assembly words are in one-to-one correspondence with isomorphism classes of assembly graphs [1]. In particular if w is a double occurrence word, we write Γ_w for the simple assembly graph defined by w.

A *composition* $\Gamma_1 \circ \Gamma_2$ of two oriented simple assembly graphs Γ_1 and Γ_2 is the directed simple assembly graph obtained by identifying the terminal vertex of Γ_1 with the initial vertex of Γ_2. Note that the initial vertex of $\Gamma_1 \circ \Gamma_2$ is the initial vertex of Γ_1 and the terminal vertex of $\Gamma_1 \circ \Gamma_2$ is the terminal vertex of Γ_2. For composition of words and graphs, we have $\Gamma_{w_1} \circ \Gamma_{w_2} = \Gamma_{w_1 w_2}$. In general $\Gamma_1 \circ \Gamma_2$ is not isomorphic to $\Gamma_2 \circ \Gamma_1$. For example, consider Γ_{aa} and Γ_{bbcddc}. A composition $\underbrace{\Gamma \circ \Gamma \circ \cdots \circ \Gamma}_{k}$ is called a *k-fold composition of* Γ and is denoted by Γ^k. If $\Gamma = \Gamma_1 \circ \Gamma_2$ for some non-trivial, oriented assembly graphs Γ_1 and Γ_2, then Γ is called *reducible.* Otherwise it is called *irreducible.* This corresponds to reducible and irreducible assembly words.

Definitions of the assembly number and related concepts.

Two paths are *disjoint* if they do not have a vertex in common. We are interested in disjoint polygonal paths that visit every vertex in an assembly graph. A pairwise disjoint set $\{\gamma_1, \ldots, \gamma_k\}$ of polygonal paths in Γ is called *Hamiltonian* if their union contains all 4-valent vertices of Γ. In particular the set of vertices $V(\Gamma)$ is a Hamiltonian set of singletons. A polygonal path γ with no repeating vertices is called *Hamiltonian* if the set $\{\gamma\}$ is Hamiltonian.

Let Γ be a non-trivial assembly graph. The *assembly number* of Γ that is denoted by $\mathrm{An}(\Gamma)$, is defined by $\mathrm{An}(\Gamma) = \min\{\ k \mid$ there exists a Hamiltonian set of polygonal paths $\{\gamma_1, \ldots, \gamma_k\}$ in $\Gamma\}$. Graphs Γ with $\mathrm{An}(\Gamma) = 1$ are called *realizable.* Otherwise they are called *unrealizable.* These names reflect whether a given assembly graph corresponds to a single scrambled gene or not. In particular the assembly number of a graph gives the minimal number of genes that can be encoded by a corresponding DNA sequence.

For a positive integer n, we define the *minimum realization number for* n to be $R_{\min}(n) = \min\{|\Gamma| : \mathrm{An}(\Gamma) = n\}$, where $|\Gamma|$ is the number of 4-valent vertices in Γ. A graph Γ that gives the minimum for $R_{\min}(n)$ is called *a realization of assembly number* n and gives a structure with minimal number of vertices representing a DNA sequence that encodes at least n genes.

A Hamiltonian polygonal path for the assembly graph depicted in Figure 2 (A) is (v_1, e_3, v_2), depicted by a thick dotted curve. Note that for the graph Γ depicted in (B), there is no Hamiltonian polygonal path. Therefore,Γ is unrealizable with $\mathrm{An}(\Gamma) = 2$.

The following properties were shown in [1].

- For each pair of directed simple assembly graphs Γ_1 and Γ_2, one of the following equalities hold: $\mathrm{An}(\Gamma_1 \circ \Gamma_2) = \mathrm{An}(\Gamma_1) + \mathrm{An}(\Gamma_2)$ or $\mathrm{An}(\Gamma_1 \circ \Gamma_2) = \mathrm{An}(\Gamma_1) + \mathrm{An}(\Gamma_2) - 1$.
- For any positive integer n, there exists
  (i) a reducible assembly graph Γ such that $\mathrm{An}(\Gamma) = n$,
  (ii) an irreducible assembly graph Γ such that $\mathrm{An}(\Gamma) = n$, and
  (iii) an assembly graph Γ with no endpoints such that $\mathrm{An}(\Gamma) = n$.
- The following properties hold for $R_{\min}$.
  (i) For every positive integer n, $R_{\min}(n) < R_{\min}(n+1)$.
  (ii) If $R_{\min}(n) = k$ for some n and k, then for every $s \geq k$ there is an assembly graph Γ with s 4-valent vertices such that $\mathrm{An}(\Gamma) = n$.
  (iii) $R_{\min}(n) \leq 3(n-1) + 1$ for every positive integer n.

A case-by-case inspection shows that $\mathrm{An}(\Gamma) = 1$ for all assembly graphs Γ with $|\Gamma| \leq 3$. The graph Γ in Figure 2 (B) has four 4-valent vertices and $\mathrm{An}(\Gamma) = 2$. Therefore, $R_{\min}(1) = 1$ and $R_{\min}(2) = 4$ (see Section 5).

3 Number of Isomorphism Classes of Assembly Graphs

We start this section with some classifications of assembly graphs having small numbers of rigid vertices. We use natural numbers as the symbols for the double-occurence words representing the assembly graphs.

Consider an assembly graph with designated initial and terminal vertices. For convenience we choose 1 as the first letter assigned to the first 4-valent rigid vertex that is encountered along the transversal after the initial vertex. Choose 2 as the second letter, unless the second letter is also 1 (in this case a loop brings us back to the first vertex) and so on. A double occurrence word written according to this convention, where the first appearance of a symbol must be one greater than the largest of all preceding symbols, is an assembly word *in ascending order.* As mentioned in Section 2, the word corresponding to the reverse orientation of an assembly graph is equivalent to the word corresponding to the original orientation of the graph. Therefore, there are at most two words in ascending order in every equivalence class of double occurrence words [1]. For example, 1212 is an assembly word in ascending order. Its reverse is 2121 which is not in ascending order. By renaming the letters, we obtain 1212 in ascending order, which is the same as the original. In the rest of the exposition we always use words in an ascending order.

There is only one assembly word with one letter $w = 11$. Hence there is one corresponding assembly graph, denoted $\Gamma_{(1)}$, which is the loop shown in Figure 3(A). The irreducible assembly words with 2 letters are 1212 and 1221 whose corresponding assembly graphs are depicted in Figure 3(B) and (C), respectively. The assembly graph corresponding to the reducible word 1122 is depicted in Figure 3(D). All these have the assembly number 1.

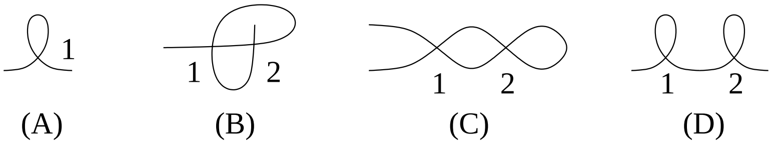


Figure 3: List of assembly graphs of sizes 1 and 2.

For irreducible assembly words with three distinct symbols we have the following classification.

Lemma 3.1 *There are* 8 *equivalence classes of irreducible assembly words of* 3 *letters. They are represented by* 121323*,* 121332*,* 122331*,* 123123*,* 123132*,* 123231*,* 123312*,* 123321*.*

Proof. We may assume that a word starts with 12, since 11 would give a reducible word. If the word starts with 121, then the irreducibility implies that the next symbol must be 3, so it starts with 1213. There are two possibilities in this case: 121323 and 121332. If the word starts with 122, again, the next symbol must be 3, such as 1223. The word 122313 is equivalent to 121332. Therefore,the only possibility is 122331. If the word starts with 123, among 6 possibilities for the remaining three letters, 123213 is equivalent to 123132, and all others are distinct. □

Isomorphism classes for graphs corresponding to the words in Lemma 3.1 are depicted in Figure 4. One can check directly from the diagrams that all these graphs have an assembly number 1.

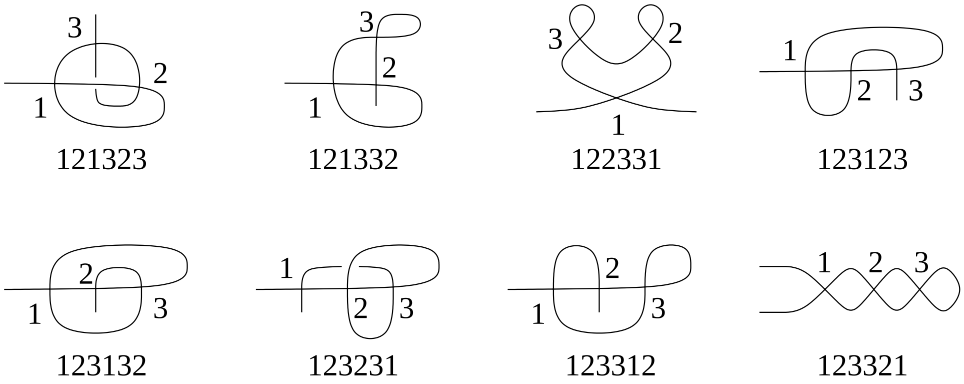


Figure 4: Representatives of isomorphism classes of irreducible graphs of size 3.

The rest of this section contains general formulae for the number of equivalence classes of assembly words and corresponding assembly graph isomorphism classes. For a positive integer n, we use the notation $[n] = \{1, 2, \dots, n\}$.

The following has been previously observed over 50 years ago in different context [15] and is part of the folklore, but we provide a constructive proof for completeness.

Lemma 3.2 *There is a one-to-one correspondence between the set of assembly words on n letters and the set $[2n-1] \times \cdots \times [3] \times [1]$. The cardinality W_n of this set is $(2n-1)!!$.*

Proof. The proof follows by induction on n. As noted above, there is only one double occurrence word with one symbol.

Let w_n be an arbitrary assembly word with n distinct letters. The second occurrence of the letter 1 can appear without restriction in any of the remaining $2n-1$ positions. If $n \geq 2$, we may remove both 1's to form a word w_{n-1} with $n-1$ letters and decrease by 1 the remaining symbols. Note that w_{n-1} is also a double occurrence word in ascending order. By the inductive hypothesis there are $1 \cdot 3 \cdots (2n-3) = (2n-3)!!$ possible choices for w_{n-1}. Since there are $(2n-1)$ choices to place the 1's to obtain w_n, the number of possible double occurrence words of n symbols is $(2n-3)!! \cdot (2n-1) = (2n-1)!!$. □

Definition 3.3 An assembly word is a *palindrome* if it is equal to its reverse written in ascending order. The graph Γ_w is said to be *palindromic* if w is a palindrome.

By definition of a palindrome it follows that there is only one word in the equivalence class of a palindrome. Note that there is a unique palindrome of n letters whose reverse is identical to itself:

$$123 \cdots (n-1)nn(n-1) \cdots 321.$$

The assembly word $w = 121323$ is also a palindrome and its reverse starts with 3, and is not identical to the original. But w^R written in ascending order is identical to w.

The next lemma gives a closed formula for the number of palindromes in n letters.

Lemma 3.4 *The number of palindromes with n letters, for any positive integer n is*

$$P_n = \sum_{k=\lfloor \frac{n}{2} \rfloor}^{n} \binom{k}{n-k} \frac{n!}{k!}.$$

Proof. Let P_n denote the number of palindromes with n letters. Then we show that P_n satisfies the recursive formula

$$P_n = P_{n-1} + 2(n-1)P_{n-2} \text{ for } n > 1,\ P_0 = 1 \text{ and } P_1 = 1.$$

Observe that $P_1 = 1$ since there is a unique one letter palindrome. And $P_2 = 3$ since 1122, 1212, and 1221 are all two letter palindromes.

A word w with n letters that starts and ends with 1 is a palindrome if and only if a word obtained from w by removing both occurrences of 1 is a palindrome with $n-1$ letters. Hence there are P_{n-1} palindromes with n letters that start and end with 1.

Consider a word w with $n > 2$ letters where the second symbol 1 is at the position $j \neq 2n$. Then the word w is a palindrome if and only if w contains the same symbol s at both positions $2n$ and $2n-j+1$, and removing symbols 1 and s from w produces a palindrome of length $n-2$. Hence there are P_{n-2} palindromes that have a symbol 1 at the j-th position for $1 < j < 2n$ ($2n-2$ choices).

According to the above argument,

$$P_n = P_{n-1} + 2(n-1)P_{n-2} \text{ for } n > 2,\ P_1 = 1 \text{ and } P_2 = 3.$$

is a recurrence relation for P_n. Note that the empty word with no symbols is palindrome by default, so $P_0 = 1$. It is known [11] that the closed formula for this recursive relation is $\sum_{k=\lfloor \frac{n}{2} \rfloor}^{n} \binom{k}{n-k} \frac{n!}{k!}$.

□

Considering that each isomorphism class of assembly graphs with two endpoints corresponds to one palindrome or two non-palindromic assembly words, we have the following formula for the number of isomorphism classes of assembly graphs.

Proposition 3.5 *The number of isomorphism classes of simple assembly graphs of size n is*

$$G_n = \frac{1}{2}\left[(2n-1)!! + \sum_{k=\lfloor \frac{n}{2} \rfloor}^{n} \binom{k}{n-k} \frac{n!}{k!}\right] = \frac{1}{2}(W_n + P_n). \tag{1}$$

Proof. Each assembly word corresponds to one orientation of a transverse path of a simple assembly graph. It is known that for two words in ascending order w and w', Γ_w is isomorphic to $\Gamma_{w'}$ if and only if $w = w'$ or $w' = w^R$ [1]. Assembly words that are palindromic correspond to simple assembly graphs in which both orientations of the transverse path yield the same word. Thus the total number $(2n-1)!!$ of assembly words in Lemma 3.2 counts the isomorphism classes that correspond to non-palindromic words twice while counting the isomorphism classes that correspond to palindromes only once. By Lemma 3.2, the second term in (1) adds one count of each palindrome.

Therefore,(1) computes the number of assembly graph isomorphism classes. □

Lemma 3.6 *The number I_n of irreducible assembly words with n letters is given by $I_1 = 1$ and*

$$I_n = W_n - \sum_{k=1}^{n-1} W_k \, I_{n-k} \quad \textit{for } n > 1,$$

where $W_n = (2n-1)!!$ is the total number of assembly words.

Proof. We shall count the number of irreducible assembly words by subtracting the number of reducible assembly words from the total number W_n of assembly words with n letters.

Without loss of generality, let $w = uv$ be a reducible assembly word with n letters such that u is an irreducible assembly word. If the length of v is $2k$, for some $1 \leq k \leq n-1$, then the length of u is $2(n-k)$. By construction, u is irreducible and is counted among I_{n-k}, and v is counted among W_k possible assembly words of length $2k$.

Summing over the possible lengths of v yields the desired count. Since u is irreducible and v is non-empty, this ensures that each reducible assembly word w is counted exactly once. □

Lemma 3.7 *The number J_n of irreducible palindromes with n letters is given by $J_1 = 1$ and*

$$J_n = P_n - \sum_{k=1}^{\lfloor n/2 \rfloor} W_k \, J_{n-2k} \quad \textit{for } n > 1,$$

where W_n, P_n are the total numbers of assembly words and palindromes with n letters, respectively.

Proof. Similarly to the proof of Lemma 3.6, we first count the reducible palindromes. Let $w = uvu'$ where u is a non-empty arbitrary assembly word with k letters ($1 \leq k \leq \lfloor n/2 \rfloor$), u' belongs to the same isomorphism class as the reverse of u, and v is an irreducible palindrome with $(n-2k)$ letters.

There are W_k possible values of u, J_{n-2k} possible values for v and u' is dependent on u. Summing over all the valid lengths of u yields the desired count. □

A non-empty word u is a *subword* of w if $w = suv$ for some words s and v. The word u is called a *proper subword* if at least one of s or v is non-empty.

Definition 3.8 An assembly word is *strongly-irreducible* if it does not contain a proper subword that is an assembly word. An assembly graph that corresponds to a strongly-irreducible word is said to be *strongly-irreducible.*

Note that a strongly-irreducible assembly word is irreducible, but the converse is not true: 1221 is irreducible but not strongly-irreducible.

Remark 3.9 In a simple assembly graph without endpoints an Eulerian transversal is a cycle.

Hence for every assembly word corresponding to the graph any of its cyclic permutations of symbols represents the same assembly graph. Therefore, in the case of simple assembly graphs without endpoints the notions of irreducible and strongly-irreducible coincide. This is used in Section 6.

The numbers of strongly-irreducible words and strongly-irreducible palindromes have been previously studied within different contexts (see, for example, [14]). In [9] strongly-irreducible palindromes are called “connected linked diagrams” and the following formula is obtained.

Proposition 3.10 [9] *Let S_n be the number of strongly-irreducible assembly words and T_n be the number of strongly-irreducible palindromes. Then*

$$S_n = (n-1)\sum_{i=1}^{n-1} S_i S_{n-i}$$

for $n > 1$ when $S_1 = 1$ and

$$T_n = \sum_{i=1}^{n-2} T_i T_{n-i} + \sum_{i=1}^{\lfloor n/2 \rfloor} (2n-4i-1) S_i T_{n-2i}$$

for $n > 1$ when $T_0 = -1$ and $T_1 = 1$.

Symbols	All Words W_n	Palindromes P_n	Irreducible I_n	Strongly-Irred. S_n
1	1	1	1	1
2	3	3	2	1
3	15	7	10	4
4	105	25	74	27
5	945	81	706	248
6	10395	331	8162	2830
7	135135	1303	110410	38232
8	2027025	5937	1708394	593859
9	34459425	26785	29752066	10401712
10	654729075	133651	576037442	202601898
11	13749310575	669351	12277827850	4342263000
12	316234143225	3609673	285764591114	101551822350
13	7905853580625	19674097	7213364729026	2573779506192
14	213458046676875	113525595	196316804255522	70282204726396
15	6190283353629375	664400311	5731249477826890	2057490936366320
OEIS	A001147	A047974	A000698	A000699

Table 1: Sequences of numbers of distinct equivalence classes for various types of assembly words.

Table 1 shows numbers of various types of assembly words, and Table 2 shows the numbers of assembly graph isomorphism classes for each type of assembly words. In the last row, we give a reference to the sequence in the Online Encyclopedia of Integer Sequences [11], whenever possible. The top reference of the last row in Table 2 corresponds to the sequence in OEIS providing the total number of words of the given type, and the bottom reference corresponds to the sequence providing the number of palindromes of the given type. The combination of the two sequences as in formula (1) of Proposition 3.5 gives the numbers that are listed in the tables. Some of these sequences have no OEIS reference as seen in Table 2. Hence those sequences correspond to previously unreferenced sequences in [11] and we believe they may be of interest. More computational results are posted

Rigid Vertices	All Classes G_n	An$(\Gamma) = 1$	Irreducible	Strongly-Irred.
1	1	1	1	1
2	3	3	2	1
3	11	11	8	3
4	65	64	47	17
5	513	504	389	135
6	5363	5241	4226	1463
7	68219	66515	55804	19306
OEIS	W_n=A001147 P_n=A047974	—— ——	I_n=A000698 ——	S_n=A000699, T_n=A004300

Table 2: Isomorphism classes of assembly graphs. The bars indicate that these sequences are not part of the known sequences listed in [11].

on the website [17]. The *Mathematica* program that generates the tables is posted online [16]. We do not know closed form formulas for I_n, J_n, S_n, and T_n.

4 Sets of Hamiltonian Polygonal Paths

Recall that a polygonal path in an assembly graph is a path in which every two consecutive edges are neighbors with respect to their common vertex. This can be seen as if the path makes a "90-degree" turn at every rigid vertex. A set of paths $\gamma = \{\gamma_1, \dots, \gamma_s\}$ is called Hamiltonian if every vertex is visited by exactly one γ_i for some $i = 1, \dots, s$. In this section, we investigate the number (of sets) of Hamiltonian polygonal paths in assembly graphs. First we consider the number of all sets of such paths.

Theorem 4.1 *If Γ is a simple assembly graph with $|\Gamma| = n$ and $\mathcal{C}$ is the collection of all sets of Hamiltonian polygonal paths of Γ, then*

$$|\mathcal{C}| \leq F_{2n+1} - 1,$$

where F_n is the n^{th} Fibonacci number.

Proof. Orient Γ and label the edges between the 4-valent vertices successively along the transversal as $\{1, \dots, 2n-1\}$ (the edges incident to the endpoints are not enumerated). For a set $\gamma = \{\gamma_1, \dots, \gamma_s\}$ of Hamiltonian polygonal paths, let $||\gamma||$ be the number of edges visited by paths in γ. Vertices not incident to edges in γ are considered to be singleton paths. Let $\mathcal{C}_k$ be the collection of all γ such that $||\gamma|| = k$. Clearly, $||\gamma|| \leq n-1$ for all γ and $|\mathcal{C}| = |\mathcal{C}_{n-1}| + \cdots + |\mathcal{C}_1| + |\mathcal{C}_0|$.

To obtain $|\mathcal{C}_k|$, note that each $\gamma \in \mathcal{C}_k$ corresponds to a unique subset $S_\gamma \in \binom{[2n-1]}{k}$ formed by the labeled edges belonging to γ. Since γ is a set of polygonal paths, S_γ contains no consecutive integers, and we can count the elements in $\mathcal{C}_k$ by the classic *stars and bars* arguement.

Take any S_γ derived from $\gamma \in \mathcal{C}_k$ and represent the elements of S_γ by stars "$\star$" and the elements of the complement of S_γ within set $[2n-1]$ with $2n-1-k$ vertical bars "|". Note that there is at least one bar between any two stars. Every position before, between, and after the vertical bars represents a possible slot for k different stars "$\star$" for a total of $2n-k$ positions. Therefore,there are a total of $\binom{2n-k}{k}$ arrangements of stars and bars such that no arrangement contains two consecutive stars. Thus, for each k such that $0 \leq k \leq n-1$, $|\mathcal{C}_k| \leq \binom{2n-k}{k}$ and

$$|\mathcal{C}| = \sum_{k=0}^{n-1} |\mathcal{C}_k| \leq \sum_{k=0}^{n-1} \binom{2n-k}{k} = F_{2n+1} - 1.$$

The last equality is straightforward by induction (see also [11]). □

Definition 4.2 The *tangled cord* TC_n of size n, for a positive integer n, is an assembly graph with assembly word

$$1213243 \cdots (n-1)(n-2)n(n-1)n.$$

Specifically, $\mathrm{TC}_1 = 11$, $\mathrm{TC}_2 = 1212$, $\mathrm{TC}_3 = 121323$, and TC_n is obtained from TC_{n-1} by replacing the last letter $(n-1)$ by the subword $n(n-1)n$. Figure 5(a) shows the structure of the tangled cord.

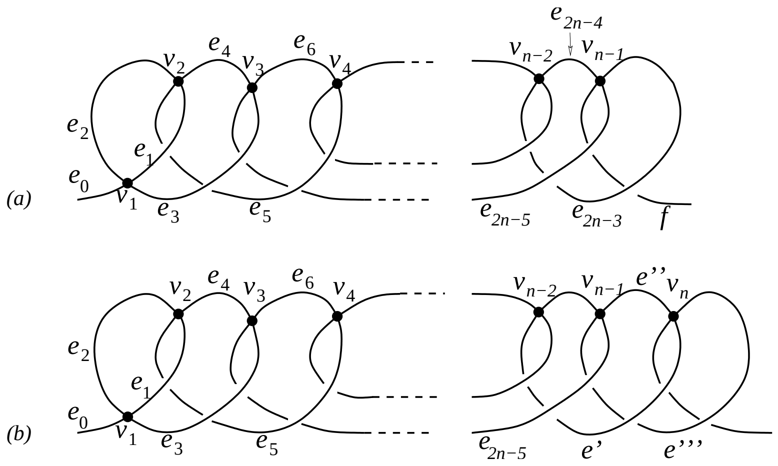


Figure 5: Addition of a new vertex to TC_{n-1} to obtain TC_n.

Theorem 4.3 *The tangled cord TC_n has $\binom{n+1}{2}$ distinct Hamiltonian polygonal paths.*

Proof. Observe that $\binom{n+1}{2} = \binom{2n-n+1}{n-1}$ which is precisely the number of ways one can chose $n-1$ non-consecutive numbers from $1, 2, \ldots, 2n-1$. We prove the theorem by induction on n.

Trivially, the singleton vertex of TC_1 constitutes a Hamiltonian polygonal path. Further, all three edges between the two 4-valent rigid vertices of TC_2 constitute Hamiltonian polygonal paths (see Figure 3(B)). For $n = 3$, TC_3 contains 7 edges labeled $e_0, e_1, \ldots, e_5, f$. Out of these, the following six form polygonal paths: e_1e_3, e_1e_4, e_1e_5, e_2e_4, e_2e_5 and e_3e_5. These paths are indicated in Figure 6. Hence, the claim holds for $n = 1, 2, 3$. Let $n \geq 4$ and consider TC_{n-1}.

Label the 4-valent vertices of TC_{n-1} and edges between them sequentially along a transversal as $v_1, v_2, \ldots, v_{n-1}$ and $e_0, e_1, \ldots, e_{2n-3}, f$ respectively (see Figure 5(a)). In this ordering, the edges

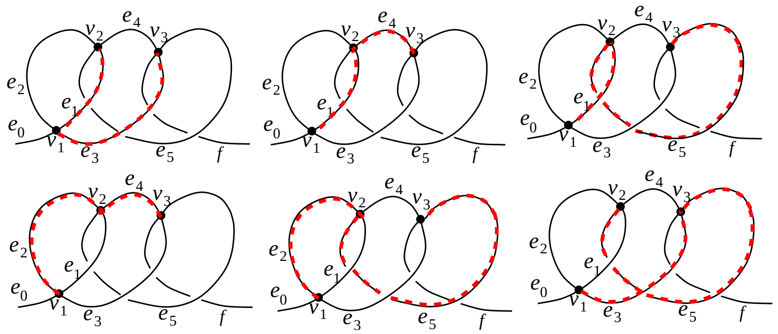


Figure 6: Six polygonal paths for TC_3.

e_0 and f are incident to the endpoints. Suppose that every choice of $n-2$ non-consecutive edges from $e_1, e_2, \ldots, e_{2n-3}$ forms a Hamiltonian polygonal path. The number of such paths is $\binom{n}{2}$.

Observe that TC_n is formed from TC_{n-1} by extending f and intersecting it with edge e_{2n-3}. Such a construction makes the following changes to TC_{n-1} (see Figure 5(b)):

1. The intersection of f with e_{2n-3} becomes a new vertex which we call v_n,
2. The edge e_{2n-3} is split into two edges by v_n which we call e' and e'' in the direction of the original transversal, and
3. Another edge is created between v_{n-1} and v_n following e'' which we call e'''.

By the induction hypothesis, every choice of $n-2$ non-consecutive edges in TC_{n-1} forms a polygonal path and there are $\binom{n}{2}$ distinct Hamiltonian polygonal paths of TC_{n-1}. Let γ be a Hamiltonian polygonal path in TC_{n-1} that contains v_{n-1}. We have the following possibilities for reaching v_{n-1} with a Hamiltonian polygonal path in TC_n.

Case 1. Suppose $e_{2n-3} \in \gamma$. Let g_1 be the number of such paths in TC_{n-1}. Note that v_{n-1} and v_{n-2} are ends of e_{2n-3}, and the neighbors of e_{2n-3} are e_{2n-4} and e_{2n-5}. The path $e'e'''$ also ends at v_{n-1} and v_{n-2}, and the neighboring edges are e_{2n-4} and e_{2n-5}. Therefore, γ may be extended to a Hamiltonian polygonal path $\gamma' = (\gamma \setminus e_{2n-3}) \cup \{e', e'''\}$ of TC_n. Hence with this extension, the number of paths in TC_n that contain both e' and e''' is the same as the number of paths in TC_{n-1} that contain e_{2n-3}. We have g_1 such paths in TC_n.

Case 2. Now suppose $e_{2n-3} \notin \gamma$, but $e_{2n-5} \in \gamma$. Let g_2 be the number of such paths in TC_{n-1}. Then v_{n-1} is an end of γ. Since e_{2n-5} is in γ, the vertex v_{n-2} must be visited by e_{2n-7}, so v_{n-2} must be an end of γ as well. Hence, γ can extend to a Hamiltonian polygonal path in TC_n in three possible ways: $\gamma_1 = \gamma \cup \{e''\}$, or $\gamma_2 = \gamma \cup \{e'''\}$ or $\gamma_3 = \gamma \cup \{e'\}$. Thus there are $3g_2$ such paths in TC_n. Note that there is only one path in TC_{n-1} that contains e_{2n-5} but not e_{2n-3} (the path with $n-2$ edges: $e_1 e_3 \cdots e_{2n-5}$). Hence $g_2 = 1$.

Case 3. Suppose $e_{2n-4} \in \gamma$. Let g_3 be the number of such paths in TC_{n-1}. Since γ is polygonal and Hamiltonian, it visits every vertex in TC_{n-1} and neither e_{2n-5} nor e_{2n-3} are in γ. In this case γ ends at vertex v_{n-1}. Hence, γ may be extended to one of two different Hamiltonian polygonal paths $\gamma_1' = \gamma \cup \{e''\}$ or $\gamma_2' = \gamma \cup \{e'''\}$ in TC_n. Thus there are $2g_3$ such paths in TC_n. Now γ has $n-2$ edges, and if the last edge is fixed to be e_{2n-4} then the rest of $n-3$ non-consecutive edges are chosen from $e_1, e_2, \ldots, e_{2n-6}$. Note that there are $\binom{(2n-6)-(n-3)+1}{n-3} = \binom{n-2}{n-3} = n-2$ such paths, i.e., $g_3 = n-2$.

We observe that all Hamiltonian polygonal paths in TC_n are obtained as described in the three cases above. Let β be a Hamiltonian polygonal path of TC_n. Then β contains $(n-1)$ edges. Since β is Hamiltonian, for the edges e', e'', e''' we have four mutually exclusive cases: (a) only $e' \in \beta$, (b) only $e'' \in \beta$, (c) only $e''' \in \beta$, or (d) $e', e''' \in \beta$, but $e'' \notin \beta$. In the first three cases (a)–(c), $(n-2)$ edges of β correspond to non-consecutive edges in TC_{n-1}, and these cases correspond to extensions of Hamiltonian polygonal paths in TC_{n-1}. Hence, those are the cases when the polygonal paths in TC_{n-1} end at vertices v_{n-1} or v_{n-2}. Such paths are described in Case 2 and Case 3. In the case of (d), we observe that for every path in TC_n that contains e' and e''' the neighboring edge of e''', the edge e_{2n-4} cannot be part of β. This means that the rest of $n-3$ edges of β are non-consecutive edges in TC_{n-1} chosen from $e_1, \ldots, e_{2n-5}$. Hence we can obtain β from a Hamiltonian polygonal path γ in TC_{n-1} containing e_{2n-3} as described in Case 1.

Now we observe that there are exactly $\binom{n+1}{2}$ Hamiltonian paths in TC_n. From Case 1 we have g_1 paths in TC_n, from Case 2 there are $3g_2$ paths in TC_n and from Case 3 there are $2g_3$ paths in TC_n. Considering that $g_1 + g_2 + g_3$ is the number of Hamiltonian polygonal paths in TC_{n-1} and the fact that $g_2 = 1$ and $g_3 = n-2$ we have total of

$$g_1 + 3g_2 + 2g_3 = \binom{n}{2} + 2g_2 + g_3 = \binom{n}{2} + 2 + n - 2 = \binom{n+1}{2}$$

Hamiltonian polygonal paths of TC_n. □

Corollary 4.4 The upper bound in Theorem 4.1 is achieved for every positive integer n, and it is achieved by a tangled cord TC_n.

Proof. We observe that the only cycle that can be formed by non-consecutive edges in TC_n is the cycle that contains all odd numbered edges. Hence any set of k non-consecutive edges in TC_n where $k < n$ does not form a cycle. This implies that every subset of k non-consecutive edges corresponds to a Hamiltonian set of polygonal paths (the vertices not visited by these k edges form paths that are singletons). Hence the inequality in Theorem 4.1 is tight. □

Conjecture 4.5 The upper bound in Theorem 4.1 is achieved for every positive integer n *only* by the tangled cord TC_n.

5 Assembly Numbers

We recall that the assembly number of an assembly graph is the minimum number of scrambled genes that can be represented by the graph. In particular, we say that an assembly graph is realizable if it can encode a single gene, in other words, if the graph has a Hamiltonian polygonal path. In this section, we present properties of assembly numbers.

Table 3 shows the assembly numbers for assembly graph isomorphism classes corresponding to graphs with a small number of rigid vertices. These numbers are obtained by computer calculations.

In particular, we notice that the assembly numbers 2 and 3 appear for the first time for graphs with 4 and 7 vertices, respectively. This implies that the minimum realization number $R_{\min}(n) = \min\{|\Gamma| : \mathrm{An}(\Gamma) = n\}$ for $n = 2$ and $n = 3$ to be 4 and 7, respectively. In the case of assembly number 2 there is only one realization graph with 4 vertices out of the total 65 isomorphism classes, and in the case of assembly number 3 there are 3 realization graphs (with 7 vertices) among over 67,000 isomorphism classes.

Assembly number \ # of rigid vertices	2	3	4	5	6	7
1	3	11	64	504	5241	66515
2	0	0	1	9	122	1701
3	0	0	0	0	0	3

Table 3: Assembly numbers.

We observe that the assembly number increases if the polygonal paths are “forced” to travel along certain edges. This enforcement can be obtained by introducing loops on the edges, therefore introducing the necessity for the polygonal paths to visit vertices of the loops. Hence, here we study the assembly number of graphs obtained by adding loops on edges, and investigate possible lengths of polygonal paths for such graphs.

We recall that $\Gamma_{(1)}$ denotes a loop, a simple assembly graph corresponding to 11, Figure 3(A).

Definition 5.1 The assembly graph $\tilde{\Gamma}$ obtained from a given assembly graph Γ by substituting every edge by a loop $\Gamma_{(1)}$ is called the *assembly graph obtained from* Γ *by loop-saturation*, or a *loop-saturated graph.*

The assembly graph $\tilde{\Gamma}^{\circ}$ obtained from a given assembly graph Γ by substituting every edge by a loop $\Gamma_{(1)}$, except the edges incident to the endpoints, is called the *assembly graph obtained from* Γ *by interior loop-saturation*, or an *interior loop-saturated graph.*

Loop-saturated and interior loop-saturated graphs obtained from graphs with 1 and 2 vertices are depicted in Figure 7. If $\tilde{\Gamma}$ is obtained from Γ by loop-saturation, then all vertices of Γ are vertices of $\tilde{\Gamma}$. In fact, the vertices of $\tilde{\Gamma}$ consist of those from Γ and vertices incident to loops added by loop-saturation. Moreover, any loop in $\tilde{\Gamma}$ is incident to a vertex that is not in Γ. The loop $\Gamma_{(1)}$ is a loop-saturation of the trivial graph with two endpoints and no 4-valent vertices.

Definition 5.2 For an integer $n \geq 0$, the set of assembly graphs obtained by interior loop-saturation from assembly graphs of size n is denoted by $\mathcal{G}_n$.

For example, $\mathcal{G}_0$ consists of only the trivial graph, $\mathcal{G}_1$ consists of only one graph corresponding to the assembly word 1221 depicted in Figure 7(1) bottom right. For $n = 2$, $\mathcal{G}_2$ is the set of assembly graphs corresponding to the assembly words 1221334554, 1223443551 and 1223441553. These graphs correspond to the interior loop-saturated graphs of graphs of size 2 in Figure 3 and are bottom right of Figure 7(2A), (2B) and (2C), respectively. Note that the assembly number of these three graphs is 1.

Definition 5.3 The *length* of a polygonal path γ, denoted $|\gamma|$, is the number of 4-valent vertices that the path contains.

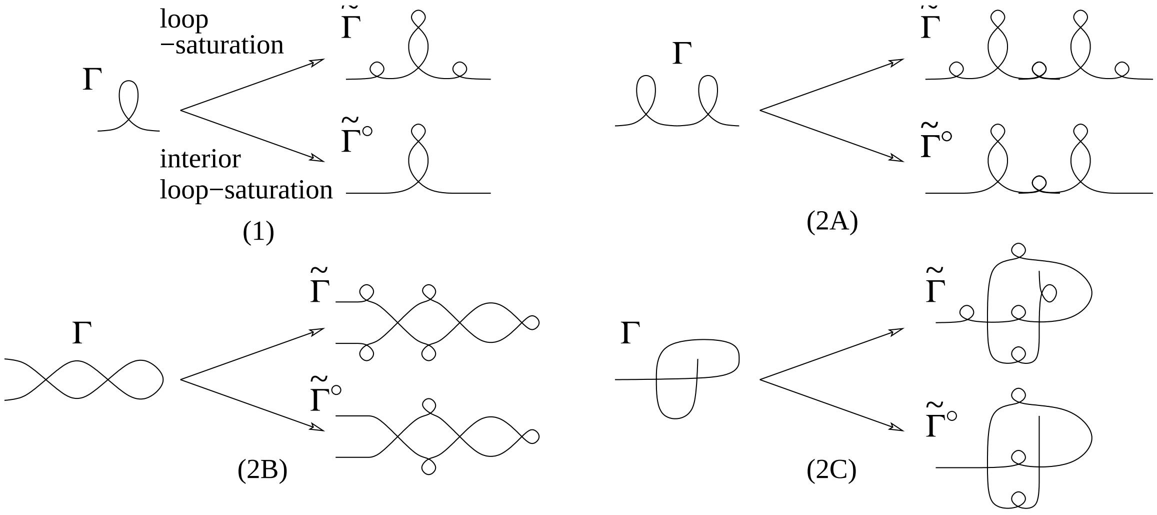


Figure 7: Loop-saturated and interior loop-saturated graphs.

Let $\gamma = \{\gamma_1, \dots, \gamma_k\}$ be a set of polygonal paths in an assembly graph. Without loss of generality assume that $|\gamma_1| \geq |\gamma_2| \geq \cdots \geq |\gamma_k|$. The *height* of γ, denoted by $\mathrm{Ht}(\gamma)$, is a sequence of positive integers $(|\gamma_1|, \dots, |\gamma_k|)$.

Lemma 5.4 *Let H be a loop-saturated or interior loop-saturated assembly graph with $|H| > 2$, and $\gamma = \{\gamma_1, \gamma_2, \dots, \gamma_k\}$ be a minimal Hamiltonian set of polygonal paths. Then $|\gamma_i|$ is odd for each $i = 1, 2, \dots, k$.*

Proof. Let H be a loop-saturated or interior loop-saturated graph obtained from Γ and let $L = V(H) \setminus V(\Gamma)$. Therefore,$L$ is the set of all new vertices obtained by loop saturation of Γ. Note that the vertices in every polygonal path alternate between vertices in $V(\Gamma)$ and those in L. We show that each polygonal path γ_i in a minimal Hamiltonian set of polygonal paths starts and ends at a loop. This implies that $|\gamma_i|$ is odd.

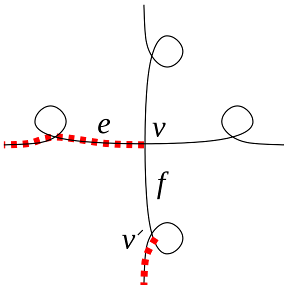


Figure 8: A polygonal path that doesn't end at a loop.

Suppose γ_i is a path in a Hamiltonian set of polygonal paths γ that ends at a vertex $v \in V(\Gamma)$ (see Figure 8). There are two cases; either v is incident to an endpoint and $H = \tilde{\Gamma}^\circ$ is interior loop-saturated, or v is not incident to an endpoint. In the latter case, consider the end edge of γ_i

denoted by e. Both edges that are neighbors of e with respect to v are incident to vertices in L (by construction). Let v' be one of these vertices and f the edge with endpoints v and v'. There exists a polygonal path γ_j that visits v' so v' must be an end-vertex of γ_j. Then substitute the two polygonal paths γ_i, γ_j in the Hamiltonian set of polygonal paths γ with a single path $\gamma_i f \gamma_j$. The new set of paths is Hamiltonian and contains less number of paths than γ. Hence, γ is not a minimal Hamiltonian set of polygonal paths.

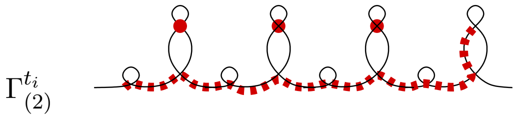
The case when $H = \tilde{\Gamma}^\circ$ is interior loop-saturated and v is adjacent to an endpoint follows similarly. In this case, because H has more than two vertices, the vertex v is adjacent to at least two and at most three loops, one of which must be an endpoint of a neighbor of the last edge e of the polygonal path γ_i ending at v. □

In the case when $|H| = 2$, H must be the interior loop-saturated graph of $\Gamma_{(1)}$, and it has a polygonal path of length 2.

Proposition 5.5 *For any positive integer k, every non-increasing sequence of positive odd integers $(d_1, \ldots, d_k)$ satisfying $\sum_{i=1}^{k} d_i = 3k-2$ is the height of a Hamiltonian set of polygonal paths for some assembly graph Γ of size $3k-2$ and $\mathrm{An}(\Gamma) = k$.*

Proof. Let $\Gamma_{(1)}$ and $\Gamma_{(2)}$ be the assembly graphs corresponding to the assembly words 11 and 112332, respectively. We construct a graph $\Gamma_{(k)}$ using the $(k-1)$-fold composition ${\Gamma_{(2)}}^{k-1}$ (see Figure 9 for an example ${\Gamma_{(2)}}^4$) as $\Gamma_{(k)} = {\Gamma_{(2)}}^{k-1} \circ \Gamma_{(1)}$. We show that this graph has the stated property. By construction $\Gamma_{(k)}$ has $3k-2$ vertices and in [1] it was proven that $\mathrm{An}(\Gamma_{(k)}) = k$.

Note that the condition $\sum_{i=1}^{k} d_i = 3k-2$ implies that any height of Hamiltonian polygonal paths must end with 1 and must start with a number greater than 1. Consider the sequence $D_0 = (d_1, \ldots, d_k) = (3, \ldots, 3, 1)$. A sequence $(5, 3, \ldots, 3, 1, 1)$ is obtained from D_0 by decreasing d_{k-1} by 2 and increasing d_1 by 2. We say that we have "moved 2 from d_{k-1} to d_1" (subtracted 2 from d_{k-1} and added 2 to d_1). Any given sequence $D = (d_1, \ldots, d_k)$ of odd integers satisfying $d_1 \geq d_2 \geq \cdots \geq d_k$ and $\sum_{i=1}^{k} d_i = 3k-2$ is obtained from D_0 by a sequence of moves of 2 from the last position that is not 1 to one of the earlier positions that are not 1. Suppose $d_i \geq 3$ for $i = 1, \ldots, \ell$ and let t_i be such that $d_i = 2t_i + 1$. For each i associate $t_i - 1 = (d_i - 3)/2$ copies of 1s to d_i. Note that the last 1 in the sequence is not associated to any of the d_i's. Then one computes $\sum_{j=1}^{k} d_j = (k-\ell) + \sum_{i=1}^{\ell} d_i = 3k-2$ so that $\sum_{i=1}^{\ell} d_i = 2k+\ell-2$. This implies that



$$d_i = 9, \quad t_i = \frac{d_i - 1}{2} = 4$$

Figure 9: A segment $\Gamma_{(2)}^4$ for $d_i = 9$ with height $(9, 1, 1, 1)$.

$\sum_{i=1}^{\ell} t_i = k-1$. Observe that the numbers of copies of 1s are equal:

$$1+\sum_{i=1}^{\ell}\frac{d_i-3}{2}=\frac{1}{2}\left[2-3\ell+\sum_{i=1}^{\ell}d_i\right]=\frac{1}{2}\left[(2-3\ell)+(2k+\ell-2)\right]=k-\ell.$$

Note that ${\Gamma_{(2)}}^{t_i}$ has a Hamiltonian set of polygonal paths of height $(2t_i+1,1,\dots,1)$ with t_i-1 copies of 1s (see Figure 9 for the case $t_i=4$). Then $\Gamma_{(k)}={\Gamma_{(2)}}^{t_1}\circ\cdots\circ{\Gamma_{(2)}}^{t_\ell}\circ\Gamma_{(1)}$ has a Hamiltoniam set of polygonal paths of height $D=(d_1,\dots,d_\ell,1,\dots,1)$. □

Theorem 5.6 *For any assembly graph Γ with $|\Gamma|=n$, the loop-saturated graph $\tilde{\Gamma}$ obtained from Γ has assembly number $n+1$ and the interior loop-saturated graph $\tilde{\Gamma}^{\circ}$ has assembly number $n-1$.*

Proof. Let $\widetilde{\Gamma}$ be loop-saturated graph obtained from Γ with $|\Gamma|=n$. Because Γ has $2n+1$ edges, $|\widetilde{\Gamma}|=3n+1$. We show that $\mathrm{An}(\widetilde{\Gamma})=n+1$. Let $\gamma=\{\gamma_1,\gamma_2,\gamma_3,\dots,\gamma_s\}$ be a minimal Hamiltonian set of polygonal paths of $\widetilde{\Gamma}$ and set $|\gamma_i|=d_i$ for all $i=1,\dots,s$. Suppose m is the number of singletons in γ. The remaining $(s-m)$ polygonal paths visit all vertices of Γ because every path in γ starts and ends at a loop as shown in the proof of Lemma 5.4.

We assume that each polygonal path γ_i visits k_i vertices in $V(\Gamma)$ for $i=1,2,3,\dots,(s-m)$. Then $\sum_{i=1}^{s-m}k_i=n$. Hence $3n+1=\sum_{i=1}^{s}d_i=\sum_{i=1}^{s-m}d_i+\sum_{i=s-m+1}^{s}d_i=\sum_{i=1}^{s-m}(2k_i+1)+m$. This implies that $\sum_{i=1}^{s-m}(2k_i+1)=2\sum_{i=1}^{s-m}k_i+s-m=3n+1-m$. Consequently $2n+s=3n+1$ and $s=n+1=\mathrm{An}(\widetilde{\Gamma})$.

The claim for $\tilde{\Gamma}^{\circ}$ follows similarly as $|\tilde{\Gamma}^{\circ}|=3n-1$. □

Graphs that necessarily increase the assembly number upon composition represent "forbidden" structures in the graphs that encode a single scrambled gene. For the rest of this section, we investigate such assembly graphs.

Definition 5.7 An assembly graph Γ is *left-* and *right-additive*, respectively, if for any assembly graph Γ',

$$\mathrm{An}(\Gamma\circ\Gamma')=\mathrm{An}(\Gamma)+\mathrm{An}(\Gamma') \quad\text{and}\quad \mathrm{An}(\Gamma'\circ\Gamma)=\mathrm{An}(\Gamma')+\mathrm{An}(\Gamma),$$

respectively. If an assembly graph is both left- and right-additive, it is called *additive.*

If an assembly graph Γ satisfies $\mathrm{An}(\Gamma_1\circ\Gamma\circ\Gamma_2)=\mathrm{An}(\Gamma_1)+\mathrm{An}(\Gamma)+\mathrm{An}(\Gamma_2)$ for any assembly graphs Γ_1 and Γ_2, then it is called *middle additive.*

Remark 5.8 If an assembly graph is middle additive then it is also left- and right-additive. Given a middle additive assembly graph Γ, consider Γ_1 and Γ_2 such that either Γ_1 or Γ_2 is the trivial assembly graph with no 4-valent vertices. If Γ_1 is trivial, then $\Gamma_1\circ\Gamma\circ\Gamma_2=\Gamma\circ\Gamma_2$ and $\mathrm{An}(\Gamma_1\circ\Gamma\circ\Gamma_2)=\mathrm{An}(\Gamma)+\mathrm{An}(\Gamma_2)=\mathrm{An}(\Gamma\circ\Gamma_2)$. Hence Γ is left-additive. A symmetric argument shows that Γ is right-additive.

Example 5.9 Let Γ_i, $i=1,2,3$, be assembly graphs corresponding to the words 122133, 112332 and 12213443, respectively. Then Γ_1 is right-additive but not left-additive, Γ_2 is left-additive but not right-additive, and Γ_3 is additive. Yet, Γ_3 is not middle additive, as $\mathrm{An}(\Gamma_{(1)}\circ\Gamma_3\circ\Gamma_{(1)})=2$.

Lemma 5.10 *Let Γ be an assembly graph. If for every minimal Hamiltonian set of polygonal paths $\{\gamma_1,\dots,\gamma_k\}$ none of γ_is starts or ends at a vertex adjacent to the initial endpoint* (*or terminal endpoint, respectively*)*, then Γ is right-additive* (*left-additive, respectively*)*.*

Proof. Let Γ be an assembly graph such that any minimal Hamiltonian set of paths $\{\gamma_1, \dots, \gamma_k\}$, where k is a positive integer, has the property that none of γ_is ends at a vertex adjacent to the initial endpoint of Γ. We will show that Γ is right-additive.

It is known [1] that $\mathrm{An}(\Gamma_1 \circ \Gamma_2) = \mathrm{An}(\Gamma_1) + \mathrm{An}(\Gamma_2)$ or $\mathrm{An}(\Gamma_1) + \mathrm{An}(\Gamma_2) - 1$ for any assembly graphs Γ_1 and Γ_2. Hence we derive a contradiction after assuming that there is an assembly graph Γ' with $\mathrm{An}(\Gamma') = m$ such that $\mathrm{An}(\Gamma' \circ \Gamma) = k + m - 1$.

Let $\gamma = \{\gamma_1, \dots, \gamma_{k+m-1}\}$ be a Hamiltonian set of polygonal paths of $\Gamma' \circ \Gamma$. Then from the condition that $\mathrm{An}(\Gamma) = k$ and $\mathrm{An}(\Gamma') = m$, there is a path, say γ_1, that goes through the point connecting Γ' to Γ. Assume without loss of generality that the set $\{\gamma_2, \dots, \gamma_{k'}\}$ consists of the other paths contained in Γ. Cut the path γ_1 into two paths: γ_1' contained in Γ' and γ_1'' contained in Γ. Then $\{\gamma_1'', \gamma_2, \dots, \gamma_{k'}\}$ is a Hamiltonian set of polygonal paths of Γ. But γ'' passes through the initial point of Γ and therefore this set of polygonal paths is not minimal. Since $\mathrm{An}(\Gamma) = k$, we have $k' + 1 \geq k$, so that $k' - k > 0$. Then Γ' has a Hamiltonian set of polygonal paths $\{\gamma_1', \gamma_{k'+1}, \dots, \gamma_{k+m-1}\}$ whose cardinality is $(k + m - 1) - k' + 1 = m - (k' - k) < m$ which contradicts $\mathrm{An}(\Gamma') = m$. □

Lemma 5.11 *Let Γ be an assembly graph and let $\Gamma_{(1)}$ be the loop* 11*. Then*

- Γ *is right-additive if and only if* $\mathrm{An}(\Gamma_{(1)} \circ \Gamma) = \mathrm{An}(\Gamma) + 1$*,*
- Γ *is left-additive if and only if* $\mathrm{An}(\Gamma \circ \Gamma_{(1)}) = \mathrm{An}(\Gamma) + 1$*,*
- Γ *is middle additive if and only if* $\mathrm{An}(\Gamma_{(1)} \circ \Gamma \circ \Gamma_{(1)}) = \mathrm{An}(\Gamma) + 2$*.*

Proof. For each of the three statements, one of the implications is immediate from the definition of left-, right- and middle additive. We will prove the converse of the last statement, since the other two statements can be shown similarly. Suppose that $\mathrm{An}(\Gamma_{(1)} \circ \Gamma \circ \Gamma_{(1)}) = \mathrm{An}(\Gamma) + 2$, and let Γ', Γ'' be two assembly graphs with assembly numbers k_1 and k_2, respectively. Let $\mathrm{An}(\Gamma) = k$. We have that $k_1 + k + k_2 - 2 \leq \mathrm{An}(\Gamma' \circ \Gamma \circ \Gamma'') \leq k_1 + k + k_2$. We show that the equality on the right hand side must hold.

Let $\gamma = \{\gamma_1, \dots, \gamma_m\}$ be a Hamiltonian set for the composition $\hat{\Gamma} = \Gamma' \circ \Gamma \circ \Gamma''$. The paths in γ contain vertices only from Γ, or only from Γ' or only from Γ'', except in the following three mutually exclusive cases: (1) there is one path in γ that contains vertices from all three graphs $\Gamma', \Gamma, \Gamma''$, (2) there are two paths in γ, one that contains vertices from Γ' and Γ and the other that contains vertices from Γ and Γ'', (3) there is exactly one path in γ that contains vertices only from two of the graphs Γ, Γ', and Γ''. If none of the cases (1)–(3) appear, then γ is a disjoint union of Hamiltonian sets for each of Γ', Γ and Γ'' and hence $m \geq k_1 + k + k_2$.

We provide the argument for case (1), since the cases (2), (3) can be proved similarly. Without loss of generality assume that γ_1 is of the form $\gamma_1 = \beta' e_1 \beta e_2 \beta''$ such that β' is a polygonal path in Γ', β is a path in Γ, and β'' is a path in Γ''. The edges e_1 and e_2 are obtained by joining the graphs' endpoints through the composition $\Gamma' \circ \Gamma$ and $\Gamma \circ \Gamma''$, respectively. Let $\gamma_2, \dots, \gamma_s$ be the paths in γ that contain vertices only from Γ. The rest of the paths $\{\gamma_{s+1}, \dots, \gamma_m\}$ can be partitioned to form two Hamiltonian sets $\{\beta', \gamma_{s+1}, \dots, \gamma_t\}$ for Γ' and $\{\beta'', \gamma_{t+1}, \dots, \gamma_m\}$ for Γ''. Then $m - s + 2 \geq k_1 + k_2$. Consider $\Gamma_{(1)} \circ \Gamma \circ \Gamma_{(1)}$ where v' and v'' are the vertices incident to the two loops obtained by composing $\Gamma_{(1)}$ at the two endpoints of Γ, respectively. Let $\gamma' = \{\gamma_1', \gamma_2, \dots, \gamma_s\}$ be a Hamiltonian set of polygonal paths where $\gamma_1' = v' e_1 \beta e_2 v''$. Since $\mathrm{An}(\Gamma_{(1)} \circ \Gamma \circ \Gamma_{(1)}) = k + 2$, we have that $s \geq k + 2$. Then we have $m + 2 \geq k_1 + k_2 + s \geq k_1 + k_2 + k + 2$ or $m \geq k_1 + k_2 + k$. □

Remark 5.12 We observe that the converse of Lemma 5.10 does not hold. The graph Γ that corresponds to 122344513665 is realizable (i.e., $\mathrm{An}(\Gamma) = 1$) with a unique minimal Hamiltonian set which consists of the path $1 \to 2 \to 3 \to 6 \to 5 \to 4$. This path starts at 1 which is adjacent to the initial endpoint, but the edge between 1 and 2 used by this path is not a neighbor to the edge incident to the initial endpoint. If we compose a loop to the initial endpoint of Γ, we obtain a graph with assembly number 2. Hence by Lemma 5.11, Γ is right-additive. In fact Γ is additive by Lemma 5.10. It is left-additive because the Hamiltonian path does not end at the vertex adjacent to the terminal endpoint.

Corollary 5.13 *The interior loop-saturated graphs are middle additive.*

Proof. Let $\tilde{\Gamma}$ and $\tilde{\Gamma}^{\circ}$ be the loop-saturated and the interior loop-saturated graphs obtained from Γ respectively. Then $\tilde{\Gamma} = \Gamma_{(1)} \circ \tilde{\Gamma}^{\circ} \circ \Gamma_{(1)}$ and the corollary follows from Theorem 5.6 and Lemma 5.11. □

Definition 5.14 Let $\mathcal{A}_{\min}(k)$ be the set of middle additive assembly graphs Γ with $\mathrm{An}(\Gamma) = k$ and $|\Gamma|$ is minimum.

Lemma 5.15 *If Γ is a simple assembly graph and $|\Gamma| \leq 4$, then Γ is not middle additive.*

Proof. Let f_1 and f_2 be edges incident to the initial and terminal endpoints of Γ, respectively. Let $\Gamma_{(1)}$ be the loop that corresponds to $w = 11$. Consider the assembly graph $\Gamma^{'} = \Gamma_{(1)} \circ \Gamma \circ \Gamma_{(1)}$. Consequently $|\Gamma^{'}| \leq 6$ and because $R_{min}(3) = 7$ (see Table 3), $An(\Gamma^{'}) < 3$. □

The following proposition says that the set of graphs with minimal number of vertices is precisely the interior loop-saturated graphs obtained from those with two 4-valent vertices. The preceding lemma shows that $\mathcal{A}_{\min}(1)$ contains only graphs with 5 vertices.

Proposition 5.16 $\mathcal{A}_{\min}(1) = \mathcal{G}_2$.

Proof. There are only three assembly graphs with exactly two 4-valent vertices. The set of assembly graphs $\mathcal{G}_2$ obtained from these graphs by interior loop-saturation consists of the graph corresponding to 1223443551, 1221334554, and 1223441553 (see Figure 7 (2A), (2B) and (2C)). By Corollary 5.13 these three graphs are middle additive.

Let $\Gamma \in \mathcal{A}_{\min}(1)$. Then Γ is a simple assembly graph that satisfies middle additive property and since $\mathcal{G}_2 \subseteq \mathcal{A}_{\min}(1), |\Gamma| = 5$. We show that $\Gamma \in \mathcal{G}_2$. Let f_1 and f_2 be initial and terminal edges of Γ, respectively. Because Γ is middle additive, if we attach two loops to both edges f_1 and f_2 we obtain an assembly graph $\Gamma^{'}$ of assembly number 3. Every simple assembly graph $\Gamma^{'}$ with $|\Gamma^{'}| = 7$ and $\mathrm{An}(\Gamma^{'}) = 3$ corresponds to the loop-saturated graphs obtained from graphs with two 4-valent vertices, by Theorem 5.6 and Table 3 . Hence Γ must be interior loop-saturated and $\Gamma \in \mathcal{G}_2$, i.e., $\mathcal{A}_{\min}(1) \subseteq \mathcal{G}_1$. □

Conjecture 5.17 $\mathcal{A}_{\min}(k) = \mathcal{G}_{k+1}$ for any positive integer k.

For a positive integer k, recall from Section 2 that the *minimal realization number for the assembly number k* is defined by $R_{\min}(k) = \min\{|\Gamma| \, : \, \mathrm{An}(\Gamma) = k\}$, where $|\Gamma|$ is the number of 4-valent vertices in Γ. A graph Γ such that $R_{\min}(k) = |\Gamma|$ is called *a realization of the assembly number k*, or a *minimal realization graph.* Let $\mathcal{R}_{\min}(k)$ denote the set of minimal realization graphs for k.

Lemma 5.18 [1] *$R_{\min}(k) \leq 3k - 2$.*

This lemma was proved by constructing a specific assembly graph $\Gamma_{(k)}$ for any positive integer k such that $\mathrm{An}(\Gamma_{(k)}) = k$ and $|\Gamma_{(k)}| = 3k-2$. Here we observe that if $|\Gamma| = n$ and $\tilde{\Gamma}$ is loop-saturated graph obtained from Γ then, by Theorem 5.6, $\mathrm{An}(\tilde{\Gamma}) = n + 1$. Since $\tilde{\Gamma}$ has $3n+1$ vertices, the lemma follows.

Conjecture 5.19 The set $\mathcal{R}_{\min}(k)$ consists of graphs obtained by loop saturation. More specifically: (1) Any graph obtained by loop saturation is a minimal realization graph. (2) Any minimal realization graph is obtained by loop saturation.

The conjecture is true for $k = 1, 2, 3$ from Table 3.

6 Assembly polynomials

The number of components after smoothing of all of the vertices in an assembly graph corresponds to the number of molecules after the recombination [2]. Experimentally, circular molecules excised from the micronuclear DNA sequence have been observed after the rearrangements [4]. Such molecules correspond to the cyclic components of the assembly graphs after smoothing of the vertices [1]. Motivated by polynomial invariants in knot theory (see, for example, [8]), we define the following polynomial that encodes the number of components in an organized manner.

Let Γ be a simple oriented assembly graph and v be a rigid vertex of Γ. We consider each edge of Γ as a directed edge, oriented according to the orientation of Γ's transversal. Consequently, at each 4-valent vertex there are two incoming and two outgoing edges. Let e_1, e_2, e_3, e_4 be the edges incident to v in the order they are encountered by the transversal (we call such an order *natural*). Then edges e_1 and e_3 are incoming to v and e_2 and e_4 are outgoing from v. Consider two graphs Γ_1 and Γ_2 defined as follows. The graph Γ_1 is obtained from Γ by replacing the rigid vertex v with two vertices v' and v'' such that v' is incident to e_1 and e_4, and v'' is incident to e_2 and e_3 (i.e., the indegree and outdegree of both new vertices v' and v'' is 1). Similarly, the graph Γ_2 is obtained from Γ such that v' is incident to e_1 and e_3, and v'' is incident to e_2 and e_4 (see Figure 10). We say that Γ_1 is obtained from Γ by a p-smoothing of the vertex v and Γ_2 is obtained from Γ by an n-smoothing of the vertex v. In this case, we write $\Gamma_1 = p_v(\Gamma)$ and $\Gamma_2 = n_v(\Gamma)$ (see Figure 10). Observe that by preserving the directions of the edges after smoothing some vertices, the directions of the edges at each remaining 4-valent vertex are unchanged. Hence, the p and n-smoothing are well defined. We call a directed graph obtained from an assembly graph by smoothing of rigid vertices an intermediate. More formally, a graph G is called an *intermediate* of Γ if $G = \Gamma$ or there exists a sequence of distinct rigid vertices $v_1, v_2, \ldots, v_k$ of Γ such that $G = s_{v_k}(\cdots s_{v_2}(s_{v_1}(\Gamma)))$ where s_{v_i} is either n_{v_i} or p_{v_i}. Note that the previous expression is independent of the order of the rigid vertices and every permutation of $s_{v_1}, \ldots, s_{v_k}$ gives G. Note that an intermediate is not necessarily an assembly graph because it is directed and may contain 2-valent vertices.

Let G be an intermediate Γ with 4-valent rigid vertices $w_1, w_2, \ldots, w_m$. We represent possible smoothings of these rigid vertices by m-tuples of symbols from $S = \{p, n\}$. Given such m-tuple $t \in S^m$, if $t_i = p$, then w_i is smoothed with a p-smoothing, and if $t_i = n$, then w_i is smoothed with

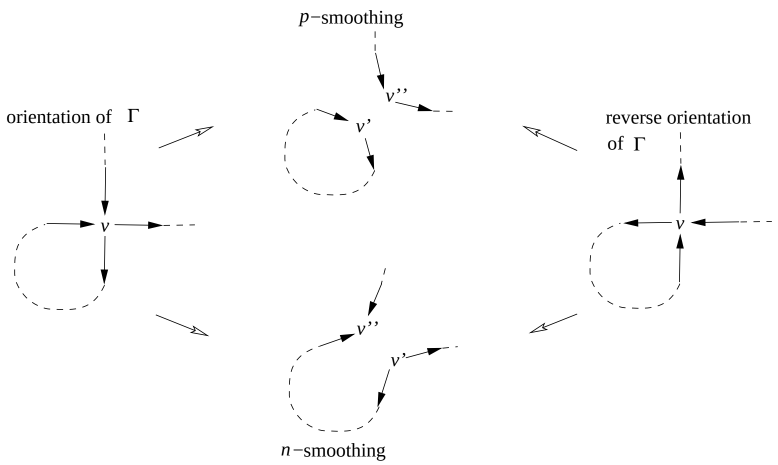


Figure 10: p-smoothing and n-smoothing of a vertex.

an n-smoothing. We abuse the notation and call such an m-tuple a *smoothing* of G. For a given smoothing $s = (s_1, \ldots, s_m)$ of an intermediate G, define

$$\begin{aligned} \pi_G(s) = & \text{ number of } p\text{-smoothings in } s \\ \mu_G(s) = & \text{ number of connected components in a graph resulting} \\ & \text{ from smoothing every rigid vertex } v_i \text{ with } s_i\text{-smoothing} \end{aligned}$$

We write $\pi(s)$ and $\mu(s)$ when the graph is clear from the context.

Definition 6.1 Given an intermediate G of size $m \geq 1$, an *assembly polynomial* of G is

$$S_G(p,t) = \sum_{s \in S^m} p^{\pi(s)} t^{\mu(s)-1},$$

where the sum is taken over all possible smoothings of G. If G does not have rigid vertices, then $S_G(p,t) = t^{c-1}$ where c is the number of connected components of G.

Example 6.2 An assembly graph corresponding to the word 122313 is depicted in Figure 11(A). In Figure 11(B), markers indicating p-smoothing at all vertices are shown. The corresponding smoothed curve is the left-most diagram in (C) which gives the monomial p^3t. All other possible smoothings of the graph are depicted in (C) in a projection of a hypercube (as in [3]). Two diagrams are connected by an edge if they differ by a single smoothing type. Each smoothing makes a monomial contribution to the assembly polynomial. Thus the assembly polynomial of this graph is $p^3t + 2p^2t + pt^2 + p^2 + 2p + t$.

Lemma 6.3 *The assembly polynomial of a simple assembly graph does not depend on the orientation of the transversal.*

Proof. As can be seen in Figure 10, the number of components resulting from a smoothing of a rigid vertex is independent of the transversal orientation corresponding to the given assembly

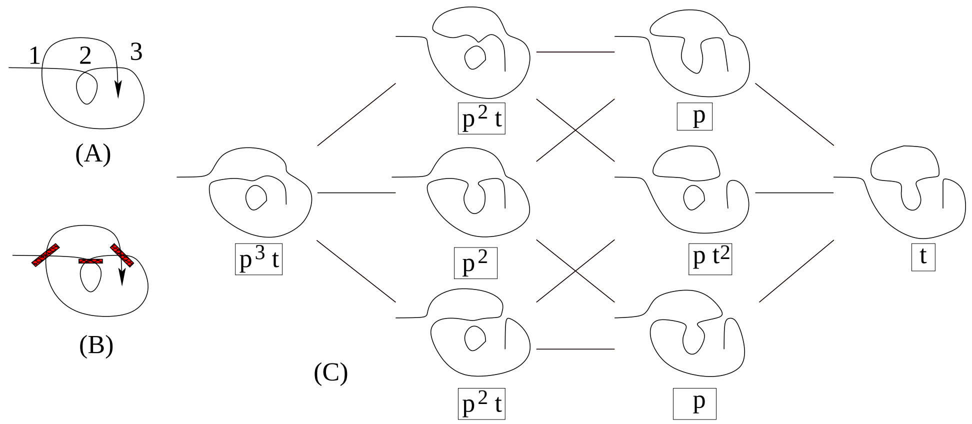


Figure 11: A cube of smoothings.

graph. Hence the number of resulting components in both cases (regardless of the orientation of Γ) depends only on the type of smoothing. □

Table 4 contains the assembly polynomials of graphs corresponding to the assembly words of length 3.

Assembly word	Assembly Polynomial
112233	$1 + 3pt + 3p^2t^2 + p^3t^3$
121233	$t + 2p + p^2 + pt^2 + 2tp^2 + tp^3$
122133	$1 + 3pt + 3p^2t^2 + p^3t^3$
122313	$t + 2p + p^2 + pt^2 + 2tp^2 + tp^3$
122331	$1 + 3pt + 3p^2t^2 + p^3t^3$
112323	$t + 2p + p^2 + pt^2 + 2tp^2 + tp^3$
121323	$1 + 2p + pt + 2p^2 + tp^2 + tp^3$
123123	$3pt + t^2 + 3p^2 + tp^3$
123213	$1 + 2p + pt + 2p^2 + tp^2 + tp^3$
123231	$t + 2p + p^2 + pt^2 + 2tp^2 + tp^3$
112332	$1 + 3pt + 3p^2t^2 + p^3t^3$
121332	$t + 2p + p^2 + pt^2 + 2tp^2 + tp^3$
123132	$1 + 2p + pt + 2p^2 + tp^2 + tp^3$
123312	$t + 2p + p^2 + pt^2 + 2tp^2 + tp^3$
123321	$1 + 3pt + 3p^2t^2 + p^3t^3$

Table 4: Assembly words with 3 distinct symbols and corresponding assembly polynomials.

Motivated from skein relations in knot theory [8], we show the following.

Lemma 6.4 *Let G be an intermediate of an assembly graph Γ. If v is a rigid vertex of G then*

$$S_G(p,t) = pS_{p_v(G)}(p,t) + S_{n_v(G)}(p,t).$$

Proof. By a direct computation, if G has more than one rigid vertex, then

$$\begin{aligned}S_G(s,t) &= \sum_{\substack{s \text{ is a}\\ \text{smoothing}\\ \text{of } G}} p^{\pi(s)}t^{\mu(s)-1} = \sum_{\substack{s:s(v)=p\\ \text{smoothing}\\ \text{of } G}} p^{\pi(s)}t^{\mu(s)-1} + \sum_{\substack{s:s(v)=n\\ \text{smoothing}\\ \text{of } G}} p^{\pi(s)}t^{\mu(s)-1}\\ &= p\sum_{\substack{s \text{ is a}\\ \text{smoothing}\\ \text{of } p_v(G)}} p^{\pi(s)}t^{\mu(s)-1} + \sum_{\substack{s \text{ is a}\\ \text{smoothing}\\ \text{of } n_v(G)}} p^{\pi(s)}t^{\mu(s)-1}\\ &= pS_{p_v(G)}(p,t) + S_{n_v(G)}(p,t).\end{aligned}$$

If G has one rigid vertex, then the statement follows directly from the definition of an assembly polynomial. □

Let Γ be a simple assembly graph with endpoints, and Γ' be an assembly graph without endpoints obtained from Γ by identifying the two endpoints. The graph obtained from Γ' by smoothing all of its vertices is a collection of cycle components because each one of its vertices has degree 2. On the other hand, the result of smoothing all vertices of Γ is a collection of cycle components and a "line" component. This line component is obtained by cutting the cyclic component from smoothing Γ' containing the two identified endpoints. The number of connected components after smoothing is the same for both Γ and Γ'. Thus if w is an assembly word and Γ' is a directed assembly graph without endpoints defined by w and Γ_w is a directed assembly graph with two endpoints, then the assembly polynomials of both graphs are the same.

Lemma 6.5 *If w is an assembly word that is not strongly-irreducible, then $S_{\Gamma_w}(p,t) = S_{\Gamma_u}(p,t) \cdot S_{\Gamma_v}(p,t)$ for some assembly words u and v.*

Proof. Suppose w is an assembly word that is not strongly-irreducible. By the paragraph preceding this lemma, the assembly polynomial of Γ_w is the same regardless whether Γ_w has, or doesn't have endpoints. Therefore, we can consider Γ_w to be an assembly graph without endpoints. By Remark 3.9, after cyclic permutation (if necessary) we may assume that $w = uv$ for some assembly words u and v. Because $w = uv$ and both u and v are double occurrence words, Γ_w contains exactly two edges e_1 and e_2, each being incident to vertices labeled by a symbol from u and by a symbol from v. For $i = 1, 2$, suppose e_i is incident to a_i and b_i such that a_i is a symbol in u and b_i is a symbol in v. Further, if we remove edges e_1 and e_2 and add new edges f_1 and f_2 such that f_1 is incident to a_1, a_2 and f_2 is incident to b_1, b_2 (preserving the cyclic order of the edges with respect to the vertices a_1, a_2, b_1 and b_2) then we obtain two simple assembly graphs Γ_u and Γ_v (see Figure 12).

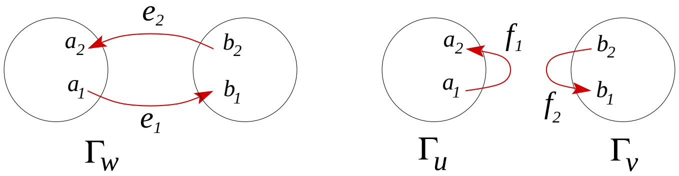


Figure 12: An assembly graph corresponding to a non strongly-irreducible and its decomposition in two assembly graphs.

For each smoothing s of vertices of Γ there are two corresponding smoothings s_1 and s_2 of Γ_u and Γ_v obtained by restricting s to the vertices of Γ_u and Γ_v, respectively. For each such s, s_1 and s_2 we have

$$\begin{aligned}\pi(s) &= \pi(s_1) + \pi(s_2),\\ \mu(s) &= \mu(s_1) + \mu(s_2) - 1.\end{aligned}$$

The first equality is straightforward and the second holds because of the following. Smoothing vertices in Γ_w with s produces a collection of cycles. All of these cycles correspond to cycles obtained by smoothing vertices of Γ_u according to s_1 or by smoothing vertices of Γ_v according to s_2, except for one cycle containing the edges e_1 and e_2 that contains vertices from both Γ_u and Γ_v. This "joint cycle" is split into two cycles by replacing e_1, e_2 with f_1, f_2, respectively; one cycle from smoothing vertices of Γ_u and the other cycle from smoothing vertices of Γ_v (see Figure 12). We have the following computation

$$\begin{aligned}S_{\Gamma_w}(p,t) &= \sum_s p^{\pi(s)} t^{\mu(s)-1} = \sum_{s_1}\sum_{s_2} p^{\pi(s_1)+\pi(s_2)} t^{\mu(s_1)+\mu(s_2)-2}\\ &= \sum_{s_1} p^{\pi(s_1)} t^{\mu(s_1)-1} \sum_{s_2} p^{\pi(s_2)} t^{\mu(s_2)-1} = S_{\Gamma_u}(p,t)\cdot S_{\Gamma_v}(p,t).\end{aligned}$$

The result follows by replacing v with the assembly word obtained by relabeling the symbols of v. □

Corollary 6.6 *If w is not strongly-irreducible, then S_{Γ_w} is a product of polynomials that correspond to strongly-irreducible assembly graphs.*

References

[1] A. Angeleska, N. Jonoska, M. Saito, DNA recombinations through assembly graphs, *Discrete Applied Mathematics* **157** (2009) 3020–3037.

[2] A. Angeleska, N. Jonoska, M. Saito, L.F. Landweber, RNA-guided DNA assembly, *Journal of Theoretical Biology* **248(4)** (2007) 706–720.

[3] D. Bar-Natan, Khovanov's homology for tangles and cobordisms, *Geom. Topol.* **9** (2005) 1443–1499.

[4] A.R.O. Cavalcanti, L.F. Landweber, Insights into a biological computer: detangling scrambled genes in ciliates, *Nanotechnology: Science and Computation* (J. Chen, N. Jonoska, G. Rozenberg eds.) Springer (2006) 349–360.

[5] A. Ehrenfeucht, T. Harju, I. Petre, D.M. Prescott, G. Rozenberg, *Computing in Living Cells* Springer 2005.

[6] A. Ehrenfeucht, T. Harju, G. Rozenberg, Gene assembly through cyclic graph decomposition, *Theoretical Computer Science* **281** (2002) 325–349.

[7] L. Kari, L.F. Landweber, Computational power of gene rearrangement, *DNA Based Computers* (E. winfree, D.K. Gifford eds.) AMS (1999) 207–216.

[8] L.H. Kauffman, *Knots and Physics* (Third Edition), World Scientific Publishing Co., 2001.

[9] M. Klazar, Non-P-recursiveness of numbers of matchings or linear chord diagrams with many crossings, *Advances in Applied Mathematics* , **30(1-2)** (2003) 126–136.

[10] M. Nowacki, V. Vijayan, Y. Zhou, K. Schotanus, T.G. Doak, L.F. Landweber, RNA-mediated epigenetic programming of a genome-rearrangement pathway, *Nature* **451** (10 Jan. 2008) 153–159.

[11] The On-Line Encyclopedia of Integer Sequences, http://oeis.org/.

[12] D.M. Prescott, A. Ehrenfeucht, G. Rozenberg, Template-guided recombination for IES elimination and unscrambling of genes in stichotrichous ciliates, *J. of Theoretical Biology* **222** (2003) 323–330.

[13] D.M. Prescott, A.F. Greslin, Scrambled Actin I gene in the micronucleus of *Oxytricha nova*, *Dev Genet* **13(1)** (1992) 66–74.

[14] R.R. Stein, C.J. Everett, On a class of linked diagrams I. Enumeration, J. Comb. Th. **A24** (1978) 357–366.

[15] J. Touchard, Sur une Problème de configurations et sur le fractions continues, *Canadian J. Math* **4** (1952) 2–25.

[16] http://jtburns.myweb.usf.edu/assembly/

[17] http://shell.cas.usf.edu/∼saito/DNAweb/SimpleAssemblyTable.txt