International Journal of Foundations of Computer Science
Vol. 31, No. 7 (2020) 915–928

DOI: 10.1142/S0129054120500343

World Scientific
www.worldscientific.com

Symbol Separation in Double Occurrence Words

Nataša Jonoska* and Masahico Saito†

Department of Mathematics and Statistics
University of South Florida, 4202 E. Fowler Ave.
Tampa, FL 33620, USA
**jonoska@mail.usf.edu*
†*saito@usf.edu*

Hwee Kim‡

Department of Computer Science and Engineering
Incheon National University, 119 Academy-ro, Yeonsu-gu
Incheon 22012, Republic of Korea
hweekim@inu.ac.kr

Brad Mostowski

Department of Mathematics
Florida State University, 600 W. College Ave.
Tallahassee, FL 32306, USA
bmostows@math.fsu.edu

Received 5 December 2019
Accepted 7 April 2020
Published 10 November 2020
Communicated by Mario de Jesus Perez-Jimenez

A double occurrence word (DOW) is a word in which every symbol appears exactly twice. We define the symbol separation of a DOW w to be the number of letters between the two copies of a symbol, and the separation of w to be the sum of separations over all symbols in w. We then analyze relationship among size, reducibility and separation of DOWs. Specifically, we provide tight bounds of separations of DOWs with a given size and characterize the words that attain those bounds. We show that all separation numbers within the bounds can be realized. We present recursive formulas for counting the numbers of DOWs with a given separation under various restrictions, such as the number of irreducible factors. These formulas can be obtained by inductive construction of all DOWs with the given separation.

Keywords: Double occurrence words; symbol separation; recursive word generation.

‡Corresponding author.

1. Introduction

A word w over an alphabet Σ is a double occurrence word (DOW) if each element (symbol) of Σ appears either zero or two times. DOWs have been studied in relation to knot theory [8, 14], mathematical logic [5], and algebraic combinatorics [12]. DOWs are also known as Gauss words and are closely related to linear diagrams, chord diagrams, and circle graphs. In the context of genomics, DOWs and operations on DOWs have been used in studies of DNA rearrangement [1, 3, 7]. By modeling the DNA rearrangement process using DOWs, it was observed that over 95% of the scrambled genome of the ciliate *Oxytricha trifallax* could be described by iterative insertions of certain patterns in DOWs [3, 9].

In [1], a DOW w corresponds to an arrangement of recombination sites (modeled by the symbols of w) in a genetic sequence. The simple recombination, just a deletion of a segment, is shown experimentally to be the earliest process during the rearrangement. This occurs when the two recombination sites are 'next' to each other, or when the DOW w has a factor of a form xx. The number of other symbols between two occurrences of x in w, which we call the *separation* of x in w, can be considered as a complexity measure of the recombination at the cite corresponding to x. The sum of the separations over all symbols that appear in w can be seen as a complexity measure of the rearrangements of the gene corresponding to w. In knot theory, symbols in a DOW represent crossings of a knot diagram. The separation is even for every symbol in a classical (versus virtual) knot diagram, but it can be odd in virtual knot diagrams. Odd separations of symbols were used to define an invariant for virtual knots in [10]. The separation, thus, seems to contain important information of DOWs and corresponding chord diagrams.

With this paper we provide general combinatorial studies for separation properties of DOWs. We provide tight bounds of separations of DOWs with a given size, and show that all separation values within those bounds can be realized. Moreover, we characterize the words that achieve the separation bounds. We present formulas for counting the numbers of DOWs with a given separation under various restrictions, such as irreducibility (when DOWs cannot be factored in smaller DOWs). The proofs of these generating formulas are constructive and show how one can obtain all words with a given separation having those properties.

More specific results and the organization of the paper follow. After an overview of definitions, conventions and background material in Sec. 2, we define the separation for DOWs in Sec. 3. This section also includes lemmas on basic properties of the separation that are used in later sections. In particular, it is proved that the separations are even numbers for all DOWs. We provide a range of separations for DOWs with a given size, and prove the realization of every even number in this range. In Sec. 4, the relation between the number of irreducible factors of DOWs and the possible separation ranges of DOWs together with their realizations is studied. Furthermore, in the case of minimum and maximum separations, corresponding DOWs are characterized. The formulas for the numbers of DOWs with

various constraints are established in Sec. 5. We observe that the position of the last appearing symbol (called the *last symbol index*) in DOW plays a key role, and use it to provide counting formulas. In particular, formulas are presented for the number of irreducible DOWs of a given size with a given separation. Realization results under given constraints are also proved.

2. Preliminaries

An *alphabet* Σ is a set of *symbols*. We denote the *empty symbol* by ϵ. Let Σ be an alphabet. A *word* u over Σ is a finite sequence $u_1u_2\cdots u_n$ of symbols, where $u_i \in \Sigma$ for each i. We call each i the index of u_i. Elements of u are called *letters* and denoted by $\Sigma[u]$. The number of letters in u is called the *length* of u and denoted by $|u|$. A subsequence $v = u_iu_{i+1}\cdots u_{j-1}u_j$ of a word $u = u_1u_2\cdots u_n$ is called a *factor* of u. The set of all words over Σ is denoted by Σ^*. In this paper, we assume that Σ is an ordered alphabet and take $\Sigma = \mathbb{N}$ with its order.

The word $u = u_1u_2\cdots u_n$ over Σ is called a *double occurrence word*, or a *DOW*, if for any $b \in \Sigma$, b occurs in u zero or two times. The set of all double occurrence words over Σ is denoted by Σ^*_{DOW}. We define the *size* of a DOW u by $\mathrm{size}(u) = |\Sigma[u]|$, which is the half of $|u|$.

Two words u and w are *equivalent* if there is a symbol-to-symbol bijection mapping u to w, in which case we write $u \sim w$. A word $u = u_1u_2\cdots u_n \in \Sigma^+$ is in *ascending order* if $u_1 = 1$ and the first appearance of a symbol is one greater than the largest of all preceding symbols in the word [2]. Every assembly word class $[u]$ contains a unique ascending order DOW v. Thus, we use v to identify $[u]$, and call v an *assembly word*. For instance, two DOWs 566577 and 133122 belong to the same assembly word class identified by the assembly word 122133. For convenience, we consider assembly words for further analysis of DOWs, and the following theorems and lemmas for assembly words hold the same for DOWs. We define the *last symbol index* $\ell(w)$ of an assembly word w of size n to be the index of the first appearance of the symbol n. Then, $n \le \ell(w) \le 2n-1$. Note that any assembly word w of size $n+1$ can be constructed by inserting two new symbols $b = n+1$ after the last symbol index of an assembly word w' of size n, and such insertion always results in an ascending order word. We say that w is constructed from w', and always use such insertion to incrementally construct an assembly word from the assembly word 11.

Let $u \in \Sigma^*_{DOW}$. Then, u is called a *repeat word* (respectively *return word*) of size n if $u \sim 12\cdots n12\cdots n$ (respectively $u \sim 12\cdots nn\cdots 21$). Also, u is called *flat* if $u \sim 11$. We define a *tangled cord* recursively as follows:

(1) A flat word is a tangled cord.
(2) If $u = u_1\cdots u_{2n}$ is a tangled cord, then $u' = u_1\cdots bu_{2n}b$ for a symbol $b \notin \Sigma[u]$ is a tangled cord.

These notions of repeat, return and tangled cord words can be generalized as in [6].

3. Separations of DOWs

In this section we define the separation value of a DOW, which gives a measure of how "scrambled" a DOW appears.

Definition 1. Given a DOW u and a symbol $b \in \Sigma[u]$, suppose $u = xbybz$ for some $x, y, z \in \Sigma[u]^*$. The *separation of* b *in* u is $\mathrm{sep}_b(u) = |y|$. The value $\mathrm{sep}(u) = \sum_{b\in\Sigma[u]} \mathrm{sep}_b(u)$ is called the *separation* of u.

Lemma 2. *Given two assembly words* $w' = xyz$ *of size* $n-1$ *and* $w = xbybz$ *of size* n *for some* $x, y, z \in \Sigma^*$ *and* $b = n$, $\mathrm{sep}(w) = \mathrm{sep}(w') + 2(|y| + |z|)$.

Proof. Since w is constructed from w' by insertion of b, $|x| \geq \ell(w')$, $\Sigma[x] = \Sigma[w']$ and all symbols in y and z are distinct. Then, we can trace changes of the separation of each symbol c in w' (see Fig. 1).

(1) If the second appearance of c is in x, then there is no change of the separation.
(2) If the second appearance of c is in y, the separation of c increases by 1.
(3) If the second appearance of c is in z, the separation of c increases by 2.

In addition, $\mathrm{sep}_b(w) = |y|$. Thus, $\mathrm{sep}(w) = \mathrm{sep}(w') + |y| + 2|z| + |y| = \mathrm{sep}(w) + 2(|y| + |z|)$. □

Corollary 3. *Given an assembly word* w, $\mathrm{sep}(w)$ *is even.*

Next, we define a permutation word.

Definition 4. Let u be a DOW. The word u is called a *permutation word* if

$$u \sim 1 \cdots n\sigma(1) \cdots \sigma(n)$$

for a bijection $\sigma : \Sigma \to \Sigma$.

Note that there are $n!$ distinct permutation assembly words of size n, and $\ell(w) = n$ if and only if w is a permutation assembly word. Repeat and return words are special cases of permutation words.

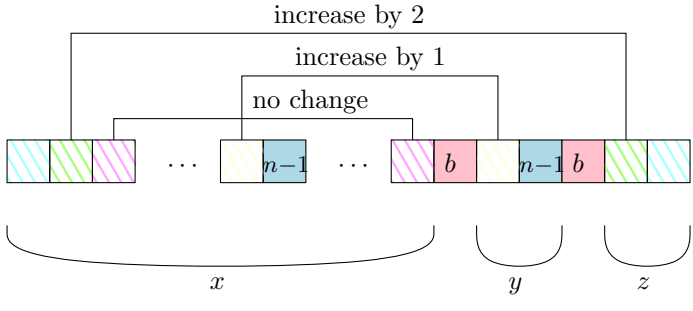

Fig. 1. Insertion of two b's and three different cases of changes of the separation.

Lemma 5. *Let w be an assembly word of size n. Then,* $\mathrm{sep}(w) \leq n(n-1)$. *In particular,* $\mathrm{sep}(w) = n(n-1)$ *if and only if w is a permutation word, hence there are $n!$ assembly words of size n and separation $n(n-1)$.*

Proof. We prove the statement by induction on n. When $n = 1$, the unique assembly word 11 satisfies the statement. Now, assume that the statement holds for all sizes less than n. Let w' be an permutation (assembly) word of size $n-1$ and separation $(n-1)(n-2)$. For an assembly word w of size n constructed from w', $\mathrm{sep}(w) = (n-1)(n-2)+2(|y|+|z|)$ from Lemma 2. The maximum value of $|y|+|z|$ is $n-1$ when the new symbol is inserted right after the first $n-1$. Then, w is a permutation word and $\mathrm{sep}(w) = (n-1)(n-2) + 2(n-1) = n(n-1)$. □

Lemma 6. *Let $n \in \mathbb{N}$. Then, for every even k where $0 \leq k \leq n(n-1)$, there exists an assembly word w of size n and separation k.*

Proof. We prove the statement by induction on n. The result obviously holds for $n = 1$. Suppose the statement holds for all size less than n. Then, for every even k where $0 \leq k \leq (n-1)(n-2)$, there exists an assembly word w of size $n-1$ and separation k. From Lemma 2, separation can increase by an even number from 0 to $2(n-1)$ in construction of an assembly word of size n by choosing appropriate y and z. Thus, for every even k where $0 \leq k \leq n(n-1)$, there exists an assembly word w of size n and separation k. □

Note that for an assembly word w with separation k and size $m < n$, there exists an assembly word w' with separation k and size n, constructed by attaching $n-m$ flat words before or after w. We call such construction *padding*, and call w' a *padding word* of w. Since $\mathrm{sep}(11) = 0$, the only word of size n and separation 0 is $1122\cdots nn$.

4. Irreducibility and Separation

In this section, we define irreducible DOWs and show how they play an important role in determining the separation of a DOW.

Definition 7. Let u be a DOW. If $u = vw$ for some DOW factors v and w, then u is called *reducible*. Otherwise, it is *irreducible*. A *decomposition* of u is a t-tuple $(u_1, \ldots, u_t)$ of irreducible DOWs, or simply $u_1 \cdots u_t$, if $u = u_1 \cdots u_t$.

Observation 8. *If u is a DOW, there exists a unique decomposition of u. If an irreducible assembly word w is constructed from w', then w' is irreducible.*

Since the decomposition of a DOW is unique, we say that a DOW u is t-reducible for $1 \leq t \leq \mathrm{size}(u)$ if it can be decomposed to t factors. Note that permutation words are special cases of irreducible words, and there exists a strongly irreducible word which is not a permutation word, such as a tangled cord, i.e. 121323. Also, if u is t-reducible and can be decomposed to $u_1 \cdots u_t$, then $\mathrm{sep}(u) = \sum_{i=1}^{t} \mathrm{sep}(u_i)$.

Now, we give the generalized relationship among size, separation and the number of irreducible factors.

Lemma 9. *For given $n \geq 1$, every irreducible assembly word w with size n has separation* $\mathrm{sep}(w) \geq 2(n-1)$.

Proof. We prove the lemma by induction on n. When $n = 1$, we have one assembly word 11, that satisfies the given condition. Now, suppose that the lemma holds for all sizes less than n by choosing y and z appropriately. We construct a given irreducible assembly word w of size n from an assembly word w' of size $n-1$. Let $w' = xyz$ and $w = xbybz$. From Lemma 2, $\mathrm{sep}(w) \geq 2(n-2) + 2(|y| + |z|)$. The minimum value of $|y|$ is 0.

(1) When $|y| = 0$, $|z| \geq 1$ to make w irreducible. Thus, $\mathrm{sep}(w) \geq 2(n-2) + 2 = 2(n-1)$.
(2) When $|y| \geq 1$, $\mathrm{sep}(w) \geq 2(n-1)$. □

Corollary 10. *For given $n \geq 1$ and $1 \leq t \leq n$, a t-reducible assembly word w with size n has separation* $2(n-t) \leq \mathrm{sep}(w) \leq (n-t)(n-t+1)$.

Proof. Let the ith irreducible factor of w be v_i with size m_i for $1 \leq i \leq t$. Then, $\mathrm{sep}(w) = \sum_{i=1}^{t} \mathrm{sep}(v_i)$. From Lemmas 5 and 9, $2(m_i - 1) \leq \mathrm{sep}(v_i) \leq m_i(m_i - 1)$. Thus, $\mathrm{sep}(w) \geq \sum_{i=1}^{t} 2(m_i - 1) = 2(n-t)$. On the other hand, since $p(p-1) + q(q-1) \leq (p+q-1)(p+q-2)$ for $p, q \geq 1$, $\mathrm{sep}(w) \leq \sum_{i=1}^{t} m_i(m_i - 1) \leq (n-t+1)(n-t)$. □

Lemma 11. *For given $n \geq 1$ and an even number k where $2(n-1) \leq k \leq n(n-1)$, there exists an irreducible assembly word w with size n. In particular,* $\mathrm{sep}(w) = 2(n-1)$ *if and only if deleting all flat factors of w results in a tangled cord.*

Proof. We prove the statement by induction on n. When $n = 1$, $w = 11$ satisfies the given statement. Now, suppose the statement holds for all sizes less than n. For even number k where $2(n-2) \leq k \leq (n-1)(n-2)$, there exists an irreducible assembly word w' with size $n-1$. We construct an irreducible assembly word w of size n from w'. From Lemma 2, for appropriate x and y, $\mathrm{sep}(w) = \mathrm{sep}(w') + 2(|y| + |z|)$. Note that if $|y| = |z| = 0$, w becomes reducible. Thus, for even number k where $\mathrm{sep}(w') + 2 \leq k \leq \mathrm{sep}(w') + 2(n-1)$, there exists an irreducible assembly word w of size n by choosing appropriate y and z. Combined with the condition for w', for even number k where $2(n-1) \leq k \leq n(n-1)$, there exists an irreducible assembly word w of size n.

Let $\mathrm{sep}(w') = 2(n-2)$. The condition $\mathrm{sep}(w) = 2(n-1)$ is satisfied by increasing the separation by 2, which implies $|y| + |z| = 1$. There are two cases:

(1) If $|y| = 0$ and $|z| = 1$, two new symbols are inserted right before the last symbol of w'. Then, deleting all flat factors of w results in a tangled cord.

(2) If $|y| = 1$ and $|z| = 0$, two new symbols are inserted right before and right after the last symbol of w'. Then, deleting all flat factors of w results in a tangled cord. □

Corollary 12. *For given $n \geq 1$, $1 \leq t \leq n$ and an even number k where $2(n-t) \leq k \leq (n-t)(n-t+1)$, there exists a t-reducible assembly word w with size n and separation k. In particular,*

(1) $\mathrm{sep}(w) = 2(n-t)$ *if and only if deleting all flat factors of w results in tangled cord factors, and*
(2) $\mathrm{sep}(w) = (n-t)(n-t+1)$ *if and only if w is a padding word of a permutation word of size $n-t+1$.*

Proof. Let a t-reducible assembly word w of size n be factored into $v_1, v_2, \ldots, v_t$ with sizes $m_1, m_2, \ldots, m_t$ respectively. From Lemma 11, for any even number k_i such that $2(m_i-1) \leq k_i \leq m_i(m_i-1)$, there exists an irreducible assembly word u_i such that $\mathrm{sep}(v_i) = k_i$. Thus, for any even number $k = \sum_{i=1}^{t} k_i$, $\mathrm{sep}(w) = k$ holds when $v_i = u_i$ for all $1 \leq i \leq t$. Now, the minimum separation of w is achieved by $\sum_{i=1}^{t} 2(m_i - 1) = 2(n-t)$, and only can be achieved by an assembly word where deleting all flat factors of w results in tangled cord factors. The maximum separation of w is achieved by $\max(\sum_{i=1}^{t} m_i(m_i-1)) = (n-t)(n-t+1)$. Note that $p(p-1) + q(q-1) = (p+q-1)(p+q-2)$ only when p or q is 1. Thus, the maximum separation can only be achieved by an assembly word which is a padding word of a permutation word of size $n-t+1$ by Lemma 5. □

Corollary 13. *Every size n assembly word of separation $(n-1)(n-2) < k \leq n(n-1)$ is irreducible.*

Corollary 14. *Let w be an assembly word of size n. Then,* $\mathrm{sep}(w) = 2$ *if and only if w is a padding word of a size two repeat or return word. There are $2(n-1)$ assembly words of separation 2.*

Observation 15. *The set of all irreducible assembly words of size 3 and separation 4 is given by* $\{122331, 121332, 122313, 121323\}$.

Corollary 16. *Let u be an assembly word with* $\mathrm{sep}(u) = 4$ *and* $\mathrm{size}(u) = n$. *Then, either of the following two cases hold*:

(1) *u consists of two padding words of a size two repeat or return word. The number of assembly words of this form is $4\binom{n-2}{2} = 2(n-2)(n-3)$.*
(2) *u is a padding word of one irreducible factor v of separation 4 from Observation 15. The number of assembly words of size n and separation 4 that are of this form is $4(n-2)$.*

Consequently, the number of assembly words of size n and separation 4 is $2(n-1)(n-2)$.

5. Relationship among Size, Reducibility and Separation of DOWs

We propose formulas to count the number of assembly words according to the size, reducibility, separation and last symbol index, and analyze relationship among these properties of DOWs.

Proposition 17 ([11, 13]). *The number W_n of assembly words of size n is $(2n-1)!!$.*

We refer to the number of assembly words of size n, t-reducibility, separation k and last symbol index ℓ as $W_{(n,t,k,\ell)}$. For these four arguments, we use $*$ to state arguments that are irreverent in counting. For example, $W_{(n,t,*,*)}$ denotes the number of t-reducible assembly words with size n.

Lemma 18. *The number $W_{(n,t,*,*)}$ of t-reducible assembly words of size n is given by the following recursion*:

- *For $t \geq 2$, $W_{(n,t,*,*)} = \sum_{i=1}^{n-t+1}(W_{(n-i,t-1,*,*)} \cdot W_{(i,1,*,*)})$.*
- *For $t = 1$, $W_{(n,1,*,*)} = (2n-1)!! - \sum_{i=1}^{n-1}((2(n-i)-1)!! \cdot W_{(i,1,*,*)})$ [2, 4].*

Proof. For $t \geq 2$, a t-reducible assembly word of size n can be partitioned into two factors: a $t-1$-reducible prefix of size $n-i$ and an irreducible suffix of size i for $1 \leq i \leq n-t+1$. Thus, the given formula for $t \geq 2$ holds. □

Theorem 19. *The number $W_{(n,1,k,\ell)}$ of irreducible assembly words of size n, separation k and last symbol index ℓ can be recursively calculated by*

$$W_{(n,1,k,\ell)} = (2n-\ell) \cdot \sum_{\ell'=n-1}^{\ell-1} (W_{(n-1,1,k-4n+2\ell+2,\ell')})$$

for $n \geq 3$. For $n = 1$ or 2, $W_{(n,1,k,\ell)} = 0$ except $W_{(1,1,0,1)} = 1$ and $W_{(2,1,2,2)} = 2$.

Proof. For each assembly word w' of size $n-1$, we insert two copies of a new symbol $b = n$ into w' to construct an irreducible assembly word w of size n. Since w is irreducible, w' is also irreducible from Observation 8. Suppose that the first b is inserted after the index $2(n-1)-j$. From Lemma 2, $\mathrm{sep}(w) = \mathrm{sep}(w') + 2j$. Since there are $j+1$ different possible indices for the second b, we may say that $W_{(n-1,1,k',\ell)} \cdot (j+1)$ is added to $W_{(n,1,k'+2j,2(n-1)-j+1)}$ (see Fig. 2).

Based on this analysis, we can construct a partial recursive formula for $W_{(n,1,k,\ell)}$ from $W_{(n',1,k',\ell')}$: $n' = n-1$ and $k' = k-2j$. Thus, $W_{(n-1,1,k-2j,\ell_0)} \cdot (j+1)$ is added to $W_{(n,1,k,\ell)}$. Since the first b is inserted before the index $2n-j$, we know that $j = 2n-\ell-1$. Since we insert the first b after ℓ_0, we know that $n-1 \leq \ell' \leq \ell-1$. Moreover, since w' is irreducible, $\ell' \leq 2n-4$ for $n \geq 3$. Combining the conditions,

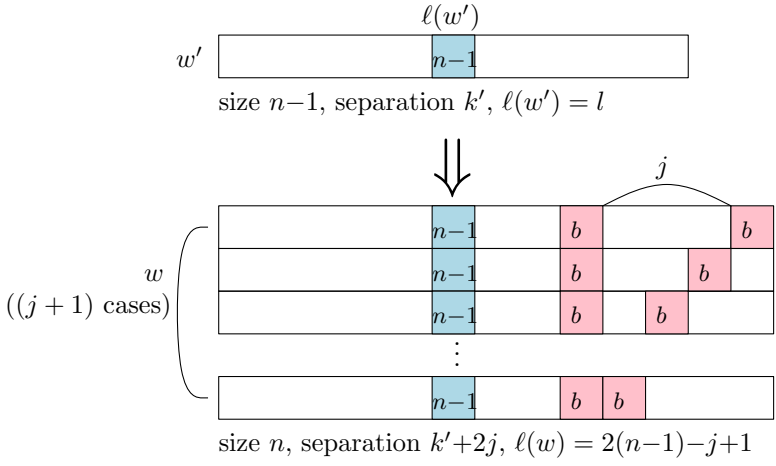

Fig. 2. The general case where $\ell(w) = 2(n-1) - j + 1$.

we have the following recursive formula:

$$W_{(n,1,k,\ell)} = (2n-\ell) \cdot \sum_{\ell'=n-1}^{\min(2n-4,\ell-1)} (W_{(n-1,1,k-4n+2\ell+2,\ell')}).$$

We need boundary conditions for arguments while calculating $W_{(n,1,k,\ell)}$. For n, one simple condition $n \geq 1$ is necessary. For k, we have the condition $0 \leq k \leq n(n-1)$. For ℓ, we have $n \leq \ell \leq 2n-2$ for $n \geq 2$. For all $W_{(n,1,k,\ell)}$ that fall out of these boundary conditions when $n \geq 2$, we may regard $W_{(n,1,k,\ell)} = 0$. Now, according to these conditions, we claim that we can rewrite the upper bound $\min(2n-4, \ell-1)$ to $\ell-1$. The only situation where these two different conditions make difference is when $\ell = 2n-2$, and $W_{(n-1,1,k-4n+2\ell+2,\ell'=2n-3)}$ is added to the latter case. However, given the size $n-1$ and the last symbol $2n-3$, we know that such W equals zero from boundary conditions. Thus, we may successfully use the upper bound $\ell-1$ for ℓ'. For initial conditions when $n \leq 2$, we have $W_{(1,1,0,1)} = 1$ and $W_{(2,1,2,2)} = 2$. □

For example, we can calculate $W_{(4,1,6,6)}$ as follows:

$$\begin{aligned}
W_{(4,1,6,6)} &= 2 \cdot \sum_{\ell'=3}^{5} W_{(3,1,4,\ell')} \\
&= 2 \cdot [W_{(3,1,4,3)} + W_{(3,1,4,4)} + W_{(3,1,4,5)}] \\
&= 2 \cdot \left[3 \cdot \sum_{\ell'=2}^{2} W_{(2,1,0,\ell')} + 2 \cdot \sum_{\ell'=2}^{2} W_{(2,1,2,\ell')} + 2 \cdot \sum_{\ell'=2}^{2} W_{(2,1,4,\ell')} \right] \\
&= 2 \cdot [3 \cdot W_{(2,1,0,2)} + 2 \cdot W_{(2,1,2,2)} + 2 \cdot W_{(2,1,4,2)}]
\end{aligned}$$

$$= 2 \cdot [3 \cdot 0 + 2 \cdot 2 + 2 \cdot 0]$$
$$= 8.$$

We propose an alternate way of counting $W_{(n,1,k,\ell)}$ by simulating recursive construction of words by insertion of symbols, starting from the word 11. The goal is to construct an irreducible assembly word of size n, separation k and last symbol index ℓ after recursively inserting pairs of i's for $2 \leq i \leq n$. Let w_i be an irreducible assembly word after the ith iteration with separation k_i and the last symbol index ℓ_i. Also, let $s_i = 2i - \ell_i$, which denotes the reversal index of the first symbol i in w_i. Then, the following observations hold:

(1) For each w_{i-1}, there are $(2i - \ell_i)$ w_i's that share the same ℓ_i.
(2) $2 \leq s_i \leq s_{i-1} + 1$, since $\ell_{i-1} \leq \ell_i$.
(3) $k_i = k_{i-1} + 2(s_i - 1)$ and $k = 2 \cdot \sum_{i=2}^{n}(s_i - 1)$.

Then, the number of all w_n's can be represented by the following alternate formula:

$$W_{(n,1,k,\ell)} = \sum_{(s_2,\ldots,s_n)} \left(\prod_{j=2}^{n} s_j \right),$$

where $s_2 = 2$, $2 \leq s_i \leq s_{i-1} + 1$ for $3 \leq i \leq n-1$, $s_n = (2n - \ell)$ and $k = 2 \cdot \sum_{i=2}^{n}(s_i - 1)$. This observation provides an alternative way to compute $W_{(n,1,k,\ell)}$.

Theorem 20. *The number $W_{(n,1,k,\ell)}$ of assembly words of size n, separation k and last symbol index ℓ can be calculated by*

$$W_{(n,1,k,\ell)} = \sum_{(s_2,\ldots,s_n)} \left(\prod_{j=2}^{n} s_j \right),$$

where $s_2 = 2$, $2 \leq s_i \leq s_{i-1} + 1$ for $3 \leq i \leq n-1$, $s_n = (2n - \ell)$ and $k = 2 \cdot \sum_{i=2}^{n}(s_i - 1)$.

Figure 3 shows a tree describing the counting $W_{(5,1,10,8)}$ by recursive insertion of symbols.

Based on Theorems 19 and 20, we have the following corollary that provides the count of irreducible assembly words of n symbols with separation k.

Corollary 21. *The number $W_{(n,1,k,*)}$ of irreducible assembly words of size n and separation k can be recursively calculated by*

$$W_{(n,1,k,*)} = \sum_{\ell=n}^{2n-2} \left((2n - \ell) \cdot \sum_{\ell'=n-1}^{\ell-1} (W_{(n-1,1,k-4n+2\ell+2,\ell')}) \right)$$

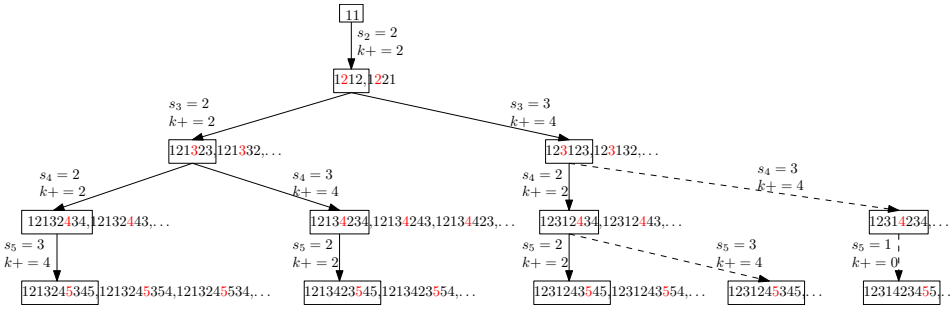

Fig. 3. Counting $W_{(5,1,10,8)} = \sum_{(s_2,\ldots,s_5)} \left(\prod_{j=2}^{5} s_j\right) = 2\times2\times2\times3+2\times2\times3\times2+2\times3\times2\times2$. Symbols at ℓ_i's are colored in red. The symbol $k+$ denotes increase of separation at each step, which is equal to $2(s_i - 1)$. All combinations of s_i's that do not match the conditions of Theorem 20 are trimmed. The dashed lines in the figure are examples of branches that are trimmed. For example, we cannot take the second leaf from the right in the formula, because the resulting separation is 12 which is > 10. Also, we cannot take the rightmost branch since its leaf has a reducible word.

for $n \geq 3$. *For* $n = 1$ *or* 2, $W_{(n,1,k,*)} = 0$ *except* $W_{(1,1,0,*)} = 1$ *and* $W_{(2,1,2,*)} = 2$. *The formula can be represented by an alternate form*

$$W_{(n,1,k,*)} = \sum_{(s_2,\ldots,s_n)} \left(\prod_{j=2}^{n} s_j\right),$$

where $s_2 = 2$, $2 \leq s_i \leq s_{i-1}+1$ *for* $3 \leq i \leq n-1$, $2 \leq s_n \leq n$ *and* $k = 2\cdot\sum_{i=2}^{n}(s_i-1)$.

Corollary 21 and the fact that the separation of a t-reducible word is a sum of separations of its irreducible factors leads to the following observation.

Corollary 22. *The number* $W_{(n,t,k,*)}$ *of* t*-reducible assembly words of size* n *and separation* k *is given by the following recursion formula*:

$$W_{(n,t,k,*)} = \sum_{(n_1,\cdots,n_t,k_1,\cdots,k_t)} \left(\prod_{i=1}^{t} W_{(n_i,1,k_i,*)}\right)$$

where $\sum_{i=1}^{t} n_i = n$, $\sum_{i=1}^{t} k_i = k$ *and* $2(n_i - 1) \leq k_i \leq n_i(n_i - 1)$ *for* $1 \leq i \leq t$.

Now, we prove the necessary and sufficient condition of k when n and ℓ are given. Note that we already have relationship between n, t and k from Corollary 12, and t and ℓ are not independent because $\ell \geq 2t + 1$ holds.

Proposition 23. *For given* $n \geq 1$, $n \leq \ell \leq 2n - 1$ *and an even number* k *where* $\mathrm{F}(n,\ell) = (2n-\ell)(2n-\ell-1) \leq k \leq \mathrm{P}(n,\ell) = n(n+1)-2\ell$, *there exists an assembly word* w *with size* n, *separation* k *and last symbol index* ℓ. *In particular*,

(1) $\mathrm{sep}(w) = \mathrm{F}(n,\ell)$ *if and only if* w *consists of* $\ell - n$ *flat words followed by a permutation word of size* $2n - \ell$.
(2) $\mathrm{sep}(w) = \mathrm{P}(n,\ell)$ *if and only if deleting the last symbols results in a permutation word of size* $n - 1$.

Proof. We prove the claim by induction on n. When $n = 1$, the only assembly word is 11 where $\ell = 1$ and $k = 0$, and the given statement is satisfied. Now, assume that the statement holds for all sizes less than n. We construct an assembly word w of size n, separation k and last symbol index ℓ from an assembly word w' of size $n-1$, separation k' and last symbol index ℓ'. We can follow insertion of two new symbols $b = n$ as Fig. 2 in the proof of Theorem 19. The changes in the separation is determined by $j = 2n - \ell - 1$, and $k = k' + 2j$. Thus, if $W_{(n-1,*,k',\ell')} > 0$, then $W_{(n,*,k'+2j,2(n-1)-j+1)} > 0$ for all $0 \le j \le 2n - \ell' - 2$.

From the induction hypothesis, for each even k' where $\mathrm{F}(n-1,\ell') \le k' \le \mathrm{P}(n-1,\ell')$, $W_{(n-1,*,k',\ell')} > 0$. Thus, for given ℓ and ℓ' where $n-1 \le \ell' \le \ell - 1$, for each even k in $\mathrm{F}(n-1,\ell') + 2j \le k \le \mathrm{P}(n-1,\ell') + 2j$, $W_{(n,*,k,\ell)} > 0$ holds. Let $\mathrm{LB}(\ell')$ ($\mathrm{UB}(\ell')$) denote the lower (upper) bound of k for given ℓ'. Note that $\mathrm{LB}(\ell') = \ell'^2 - (4n-5)\ell' + (4n^2 - 6n + 4 - 2\ell)$ and $\mathrm{UB}(\ell') = -2\ell' + (n^2 + 3n - 2 - 2\ell)$. It is straightforward that $\mathrm{UB}(\ell')$ decreases as ℓ' increases, and $\mathrm{LB}(\ell')$ also decreases since the domain of ℓ' is in the decreasing part of the parabola. Thus, $\min(\mathrm{LB}(\ell'))$ is $(2n-\ell-1)(2n-\ell-2) + 2(2n-\ell-1) = (2n-\ell-1)(2n-\ell)$ when $\ell' = \ell - 1$. For an original word

$$w' = \underbrace{1122\cdots(\ell'-n+1)(\ell'-n+1)}_{\ell'-n+1 \text{ flat words}} \times \underbrace{(\ell'-n+2)\cdots(n-1)\sigma(\ell'-n+2)\cdots\sigma(n-1)}_{\text{permutation word of size } 2n-\ell'-2},$$

we observe that the resulting word

$$w = \underbrace{1122\cdots(\ell'-n+1)(\ell'-n+1)}_{\ell-n \text{ flat words}} \times \underbrace{(\ell'-n+2)\cdots(n-1)n\sigma(\ell'-n+2)\cdots n\cdots\sigma(n-1)}_{\text{permutation word of size } 2n-\ell}$$

follows the form in the given statement. On the other hand, $\max(\mathrm{UB}(\ell'))$ is $n(n-1) - 2(n-1) + 2(2n-\ell-1) = n(n+1) - 2\ell$ when $\ell' = n-1$. For an original word

$$w' = \underbrace{12\cdots(n-1)\sigma(1)\sigma(2)\cdots\sigma(n-1)}_{\text{permutation word of size } 2n-1},$$

we observe that the resulting word

$$w' = 12\cdots(n-1)\sigma(1)\sigma(2)\cdots n\cdots n\cdots\sigma(n-1)$$

follows the form in the given statement. Therefore, (1) and (2) in Proposition 23 are inductively proven.

Now, we prove that for every even number k in the interval $[\mathrm{F}(n,\ell), \mathrm{P}(n,\ell)]$, $W_{(n,*,k,\ell)} > 0$. By the induction hypothesis, for every even number k in intervals $[\mathrm{LB}(\ell'), \mathrm{UB}(\ell')]$, $W_{(n,*,k,\ell)} > 0$ as discussed above. Now, we claim that for all ℓ'

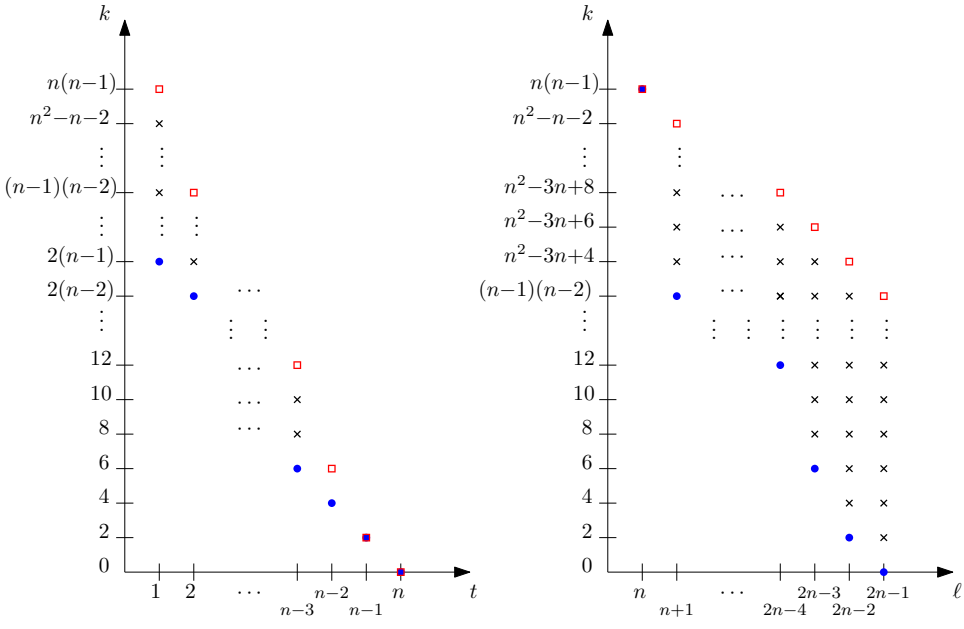

Fig. 4. (Left) Area that $W_{(n,t,k,*)} > 0$ holds. (Right) Area that $W_{(n,*,k,\ell)} > 0$ holds. Blue circles (red squares) represent the lower (upper) bound of k, and black crosses represent possible values of k strictly within the bounds.

where $n \le \ell' \le \ell-2$, it holds $\mathrm{LB}(\ell') \le \mathrm{UB}(\ell'+1)$. We have that $\mathrm{LB}(\ell')-\mathrm{UB}(\ell'+1) = (2n-2-\ell')(2n-3-\ell') - n(n-1) + 2\ell' + 2 = \ell'^2 - (4n-5)\ell' + (3n^2-9n+8)$. Since $n \le \ell' \le \ell-2 \le 2n-3 < \frac{4n-5}{2}$, the difference $\mathrm{LB}(\ell') - \mathrm{UB}(\ell'+1)$ decreases as n increases. For $\ell' = n$, the difference $\mathrm{LB}(\ell') - \mathrm{UB}(\ell'+1)$ becomes $\ell'^2 - (4n-5)\ell' + (3n^2-9n+8) = 8-4n$ which is ≤ 0 when $n \ge 2$. Thus, the claim holds for $n \ge 2$. This implies that the intervals $[\mathrm{LB}(\ell'), \mathrm{UB}(\ell')]$ and $[\mathrm{LB}(\ell'+1), \mathrm{UB}(\ell'+1)]$ overlap for all ℓ' satisfying $n \le \ell' \le \ell-2$, hence, for every even number k in the interval $[\mathrm{LB}(\ell-1), \mathrm{UB}(n)]$, we have $W_{(n,*,k,\ell)} > 0$. Combined with the fact that $\mathrm{P}(n,\ell) = \mathrm{LB}(n-1) = \mathrm{UB}(n-1) = n^2+n-2\ell$, $\mathrm{UB}(n) = n^2+n-2\ell-2$ and $\mathrm{LB}(\ell-1) = \mathrm{F}(n,\ell)$, we conclude that for every even k in the interval $[\mathrm{F}(n,\ell), \mathrm{P}(n,\ell)]$, the number of words with separation k of size n and last symbol appearing in location ℓ is $W_{(n,*,k,\ell)} > 0$. □

Figure 4 visualizes the possible range of separation k depending on the reducibility t and the last symbol index ℓ, from Corollary 10 and Proposition 23.

Acknowledgement

We thank Margherita Maria Ferrari for her comments on an earlier version of this paper. This research was (partially) supported by the grants NSF DMS-1800443/1764366 and the Southeast Center for Mathematics and Biology, an NSF-Simons

Research Center for Mathematics of Complex Biological Systems, under National Science Foundation Grant No. DMS-1764406 and Simons Foundation Grant No. 594594.

References

[1] A. Angeleska, N. Jonoska and M. Saito, DNA recombination through assembly graphs, *Discr. Appl. Math.* **157**(14) (2009) 3020–3037.

[2] J. Burns, E. Dolzhenko, N. Jonoska, T. Muche and M. Saito, Four-regular graphs with rigid vertices associated to DNA recombination, *Discr. Appl. Math.* **161**(10–11) (2013) 1378–1394.

[3] J. Burns, D. Kukushkin, X. Chen, L. Landweber, M. Saito and N. Jonoska, Recurring patterns among scrambled genes in the encrypted genome of the ciliate *oxytricha trifallax*, *J. Theoret. Biol.* **410** (2016) 171–180.

[4] J. Burns and T. Muche, Counting irreducible double occurrence words, *CoRR* **abs/1105.2926** (2011).

[5] B. Courcelle, Circle graphs and monadic second-order logic, *J. Appl. Logic* **6**(3) (2008) 416–442.

[6] D. A. Cruz, M. M. Ferrari, N. Jonoska, L. Nabergall and M. Saito, Insertions yielding equivalent double occurrence words, *CoRR* **abs/1811.11739** (2018).

[7] A. Ehrenfeucht, T. Harju, I. Petre, D. M. Prescott and G. Rozenberg, *Computation in Living Cells*: *Gene Assembly in Ciliates* (*Natural Computing Series*) (Springer, 2004).

[8] A. Gibson, Homotopy invariants of Gauss words, *Mathematische Annalen* **349**(4) (2011) 871–887.

[9] N. Jonoska, L. Nabergall and M. Saito, Patterns and distances in words related to DNA rearrangement, *Fundamenta Informaticae* **154**(1–4) (2017) 225–238.

[10] L. H. Kauffman, A self-linking invariant of virtual knots, *Fundamenta Mathematicae* **184** (2004) 135–158.

[11] M. Klazar, Non-P-recursiveness of numbers of matchings or linear chord diagrams with many crossings, *Adv. Appl. Math.* **30**(1–2) (2003) 126–136.

[12] B. Shtylla, L. Traldi and L. Zulli, On the realization of double occurrence words, *Discr. Math.* **309**(6) (2009) 1769–1773.

[13] P. R. Stein, On a class of linked diagrams, I. Enumeration, *J. Combin. Theory, Series A* **24**(3) (1978) 357–366.

[14] V. Turaev, Virtual strings, *Université de Grenoble. Annales de l'Institut Fourier* **54**(7) (2004) 2455–2525.