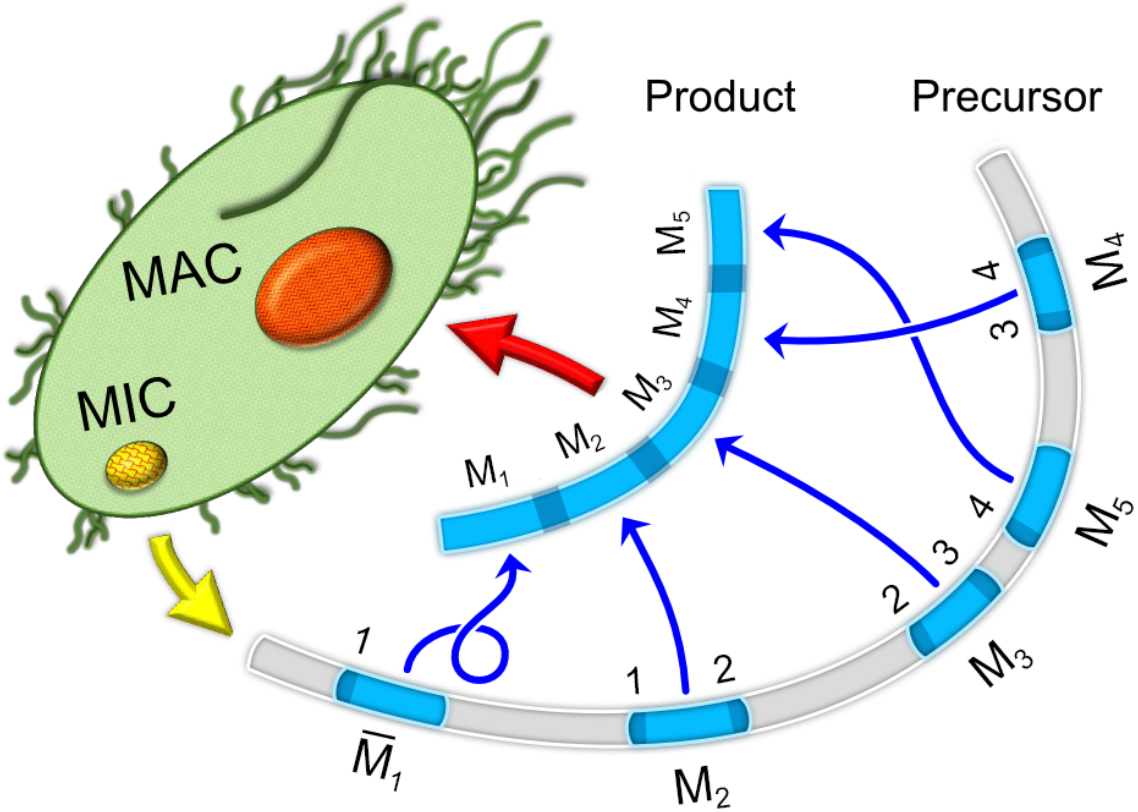
Transformations on Double Occurrence Words Motivated by DNA Rearrangement

Daniel A. Cruz, Margherita Maria Ferrari, Nataša Jonoska, Lukas Nabergall, and Masahico Saito
University of South Florida

dcruz@mail.usf.edu
December 4, 2018

Motivation: Analysis of DNA Scrambling in Ciliates



$\overline{M}_1$	M_2	M_3	M_5	M_4	
1	12	23	4	34	$\Rightarrow w = 1123434$

Jonoska, N. et al. Patterns and Distances in Words Related to DNA Rearrangement. Fundamenta Informaticae **154**:1-4 (2017) pp 225-238.

Preliminaries

Given *alphabet* $\Sigma = \{0, 1, \ldots, 9\}$,

- $w = 015164443$ is a *word* over Σ
- The *length* of w is 9, written $|w| = 9$
- $w^R = 344461510$ is the *reverse* of w
- The set of symbols used in w is $\Sigma[w] = \{0, 1, 3, 4, 5, 6\}$

The set of all words over Σ is Σ^* and includes the empty word ϵ.

Preliminaries

Given *alphabet* $\Sigma = \{0, 1, \ldots, 9\}$,

- $w = 015164443$ is a *word* over Σ
- The *length* of w is 9, written $|w| = 9$
- $w^R = 344461510$ is the *reverse* of w
- The set of symbols used in w is $\Sigma[w] = \{0, 1, 3, 4, 5, 6\}$

The set of all words over Σ is Σ^* and includes the empty word ϵ.

Definition

The word w is a *double occurrence word (DOW)* if each symbol in Σ appears 0 or 2 times in w. The set of all DOWs is Σ_{DOW}.

$$11, \quad 1221, \quad 11223434 \in \Sigma_{DOW}$$

Single occurrence words (SOWs) and Σ_{SOW} are similarly defined.

Definition: Repeat and Return Words

Definition

Given $w \in \Sigma^*$ and SOW $u \in \Sigma^+ = \Sigma^* \setminus \{\epsilon\}$,

- the word uu is a *repeat word in* w if $w = z_1 u z_2 u z_3$ for some $z_1, z_2, z_3 \in \Sigma^*$
- the word uu^R is a *return word in* w if $w = z_1 u z_2 u^R z_3$ for some $z_1, z_2, z_3 \in \Sigma^*$

w	Repeat words
11234552346788 76	234234, 2323, 88, etc.

w	Return words
1123455234678876	678876, 6776, 22, etc.

A repeat word uu or return word uu^R is *trivial* if $|u| = 1$.

Repeat and Return Words in Ciliate DNA

M_6	M_7	M_8	M_9	M_{11}	M_1	M_3	M_{10}	M_2	M_4	M_5	M_{12}	M_{13}
56	67	78	89	ab	1	23	$9a$	12	34	45	bc	c

$$w_0 = 56677889ab1239a123445bcc$$

Burns, J. et al. Recurring patterns among scrambled genes in the encrypted genome of the ciliate Oxytricha trifallax. Journal of Theoretical Biology **410** (2016) pp 171-180.

Repeat and Return Words in Ciliate DNA

M_6	M_7	M_8	M_9	M_{11}	M_1	M_3	M_{10}	M_2	M_4	M_5	M_{12}	M_{13}
56	67	78	89	ab	1	23	$9a$	12	34	45	bc	c

$$w_0 = 56677889ab1239a12344 5bcc$$

$$w_1 = 59ab1239a1235b$$

Burns, J. et al. Recurring patterns among scrambled genes in the encrypted genome of the ciliate Oxytricha trifallax. Journal of Theoretical Biology **410** (2016) pp 171-180.

Repeat and Return Words in Ciliate DNA

M_6	M_7	M_8	M_9	M_{11}	M_1	M_3	M_{10}	M_2	M_4	M_5	M_{12}	M_{13}
56	67	78	89	ab	1	23	$9a$	12	34	45	bc	c

$$w_0 = 56677889ab1239a123445bcc$$
$$w_1 = 59ab1239a12355b$$
$$w_2 = 5b5b$$

Burns, J. et al. Recurring patterns among scrambled genes in the encrypted genome of the ciliate Oxytricha trifallax. Journal of Theoretical Biology **410** (2016) pp 171-180.

Repeat and Return Words in Ciliate DNA

M_6	M_7	M_8	M_9	M_{11}	M_1	M_3	M_{10}	M_2	M_4	M_5	M_{12}	M_{13}
56	67	78	89	ab	1	23	$9a$	12	34	45	bc	c

$$w_0 = 5667788\,9ab1239a12344\,5bcc$$

$$w_1 = 59ab1239a1235b$$

$$w_2 = 5b5b$$

$$w_3 = \epsilon$$

Burns, J. et al. Recurring patterns among scrambled genes in the encrypted genome of the ciliate Oxytricha trifallax. Journal of Theoretical Biology **410** (2016) pp 171-180.

Jonoska, N. et al. Patterns and Distances in Words Related to DNA Rearrangement. Fundamenta Informaticae **154**:1-4 (2017) pp 225-238.

Definition: Repeat and Return Insertions

Definition

Given $w = a_1 \cdots a_n \in \Sigma_{DOW}$,

- let $1 \leq k \leq \ell \leq n+1$,
- let $\mathcal{I} \in \{\rho, \tau\}$ be a symbol not in Σ, and
- let $u \in \Sigma^+$ be a SOW such that $\Sigma[u] \cap \Sigma[w] = \emptyset$.

Then $\mathcal{I}(u,k,\ell)$ is an *insertion on* w which acts as follows: $w \star \mathcal{I}(u,k,\ell) = a_1 \cdots a_{k-1} u a_k \cdots a_{\ell-1} u' a_\ell \cdots a_n$ where

$$u' = \begin{cases} u & \text{if} \quad \mathcal{I} = \rho \\ u^R & \text{if} \quad \mathcal{I} = \tau \ . \end{cases}$$

$$1232314554 \xrightarrow{\rho(abc,4,6)} 123abc23abc14554$$

$$1232314554 \xrightarrow{\tau(abc,7,10)} 123231abc4554cba$$

Definition: Equivalence

Definition

Words $v, w \in \Sigma^*$ are *equivalent* if there exists a bijection $f : \Sigma \to \Sigma$ such that $f(v) = w$; in this case, we write $v \sim w$.

Examples of equivalent pairs of words:

$$\begin{array}{c|c|c|c|c|c|c|c|c} 123123 & 1 & 2 & 3 & 1221 \star \tau(ab,3,3) = 12abba21 & 1 & 2 & a & b \\ & \downarrow & \downarrow & \downarrow & & \downarrow & \downarrow & \downarrow & \downarrow \\ 321321 & 3 & 2 & 1 & 1221 \star \tau(ab,1,5) = ab1221ba & a & b & 1 & 2 \end{array}$$

The following words are not equivalent:

$$\begin{array}{rl} 1232314554 \star \rho(abc,4,6) = & \underline{1}23ab\underline{c}23abc14554 \\ & \downarrow \quad \downarrow \\ 1232314554 \star \tau(abc,7,10) = & \underline{1}23231\underline{1}abc4554cba \end{array}$$

When Do Insertions Yield Equivalent Words?

Let $w = a_1 \cdots a_n \in \Sigma_{DOW}$ be given.

1. If $w_1 = w \star \rho(u_1, k_1, \ell_1)$ and $w_2 = w \star \tau(u_2, k_2, \ell_2)$, then is it possible for w_1 and w_2 to be equivalent?

When Do Insertions Yield Equivalent Words?

Let $w = a_1 \cdots a_n \in \Sigma_{DOW}$ be given.

1. If $w_1 = w \star \rho(u_1, k_1, \ell_1)$ and $w_2 = w \star \tau(u_2, k_2, \ell_2)$, then is it possible for w_1 and w_2 to be equivalent? Yes.

$$\begin{array}{l|ccc} w_1 = 1212 \star \tau(a, 3, 5) = 12a12a & 1 & 2 & a \\ & \downarrow & \downarrow & \downarrow \\ w_2 = 1212 \star \rho(a, 1, 3) = a12a12 & a & 1 & 2 \end{array}$$

   But what if $|u_1| = |u_2| \neq 1$?

2. In general if $w_1 = w \star \mathcal{I}_1(u_1, k_1, \ell_1) \sim w \star \mathcal{I}_2(u_2, k_2, \ell_2) = w_2$, what can we say about w if the insertions are "distinct"?

$$\begin{array}{l|cccc} 1221 \star \tau(ab, 3, 3) = 12abba21 & 1 & 2 & a & b \\ & \downarrow & \downarrow & \downarrow & \downarrow \\ 1221 \star \tau(ab, 1, 5) = ab1221ba & a & b & 1 & 2 \end{array}$$

Definition: Distinct Insertions

Definition

Two insertions $\mathcal{I}_1(u_1, k_1, \ell_1)$ and $\mathcal{I}_2(u_2, k_2, \ell_2)$ on $w \in \Sigma_{DOW}$ are *distinct* if at least one of the following holds:

$$\bullet\ (k_1, \ell_1) \neq (k_2, \ell_2), \quad \bullet\ \mathcal{I}_1 \neq \mathcal{I}_2, \quad \text{or} \quad \bullet\ |u_1| \neq |u_2|$$

If $w_1 \sim w_2$, then $|u_1| = |u_2|$. What if $w_1 \sim w_2$ but only $\mathcal{I}_1 \neq \mathcal{I}_2$?

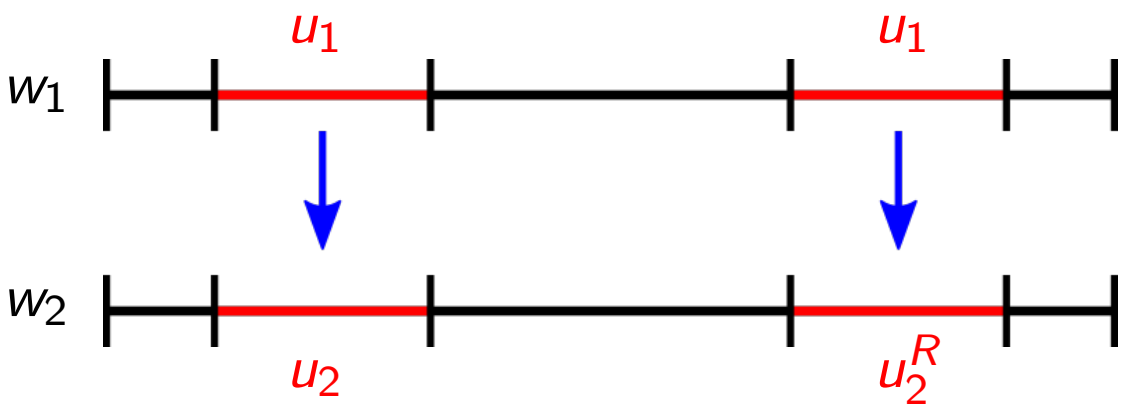
Definition: Distinct Insertions

Definition

Two insertions $\mathcal{I}_1(u_1, k_1, \ell_1)$ and $\mathcal{I}_2(u_2, k_2, \ell_2)$ on $w \in \Sigma_{DOW}$ are *distinct* if at least one of the following holds:

$$\bullet\ (k_1, \ell_1) \neq (k_2, \ell_2), \quad \bullet\ \mathcal{I}_1 \neq \mathcal{I}_2, \quad \text{or} \quad \bullet\ |u_1| \neq |u_2|$$

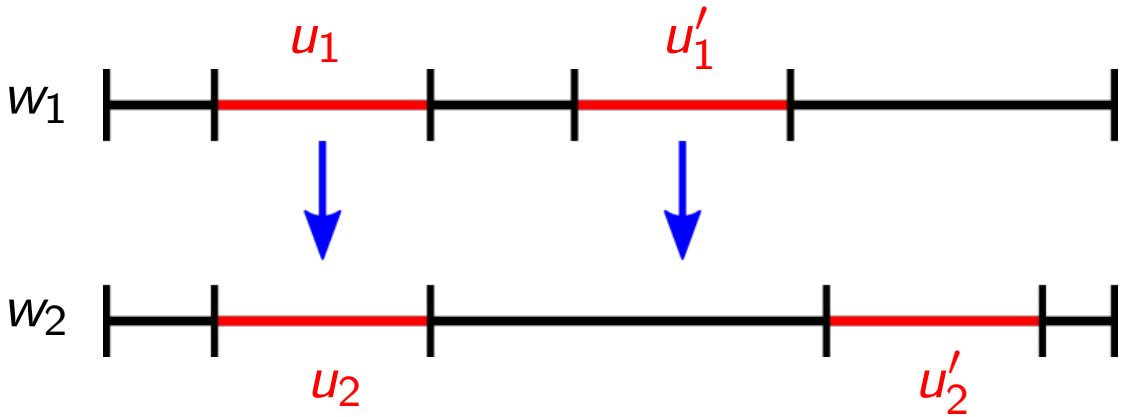
If $w_1 \sim w_2$, then $|u_1| = |u_2|$. What if $w_1 \sim w_2$ but only $\mathcal{I}_1 \neq \mathcal{I}_2$?



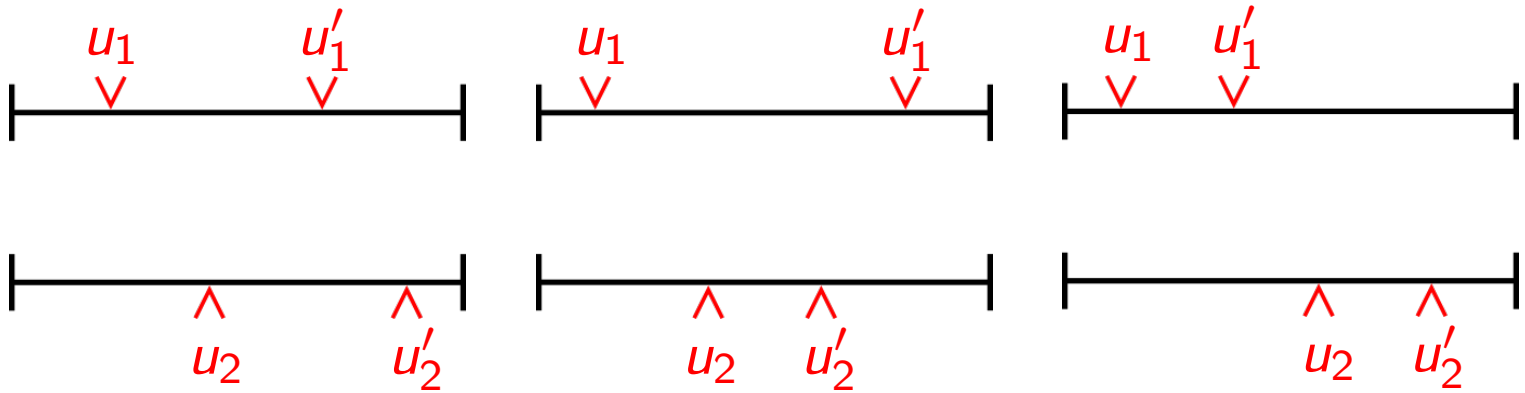
$u_1 \sim u_2$ and $u_1 \sim u_2^R \Rightarrow |u_2| = 1$ since $u_2 \in \Sigma_{SOW}$

Distinct Insertions and Equivalent DOWs

Without loss of generality, we take $k_1 \leq k_2$. Suppose that $k_1 = k_2$:



Thus, $k_1 \neq k_2$; similarly $\ell_1 \neq \ell_2$. We have three cases:

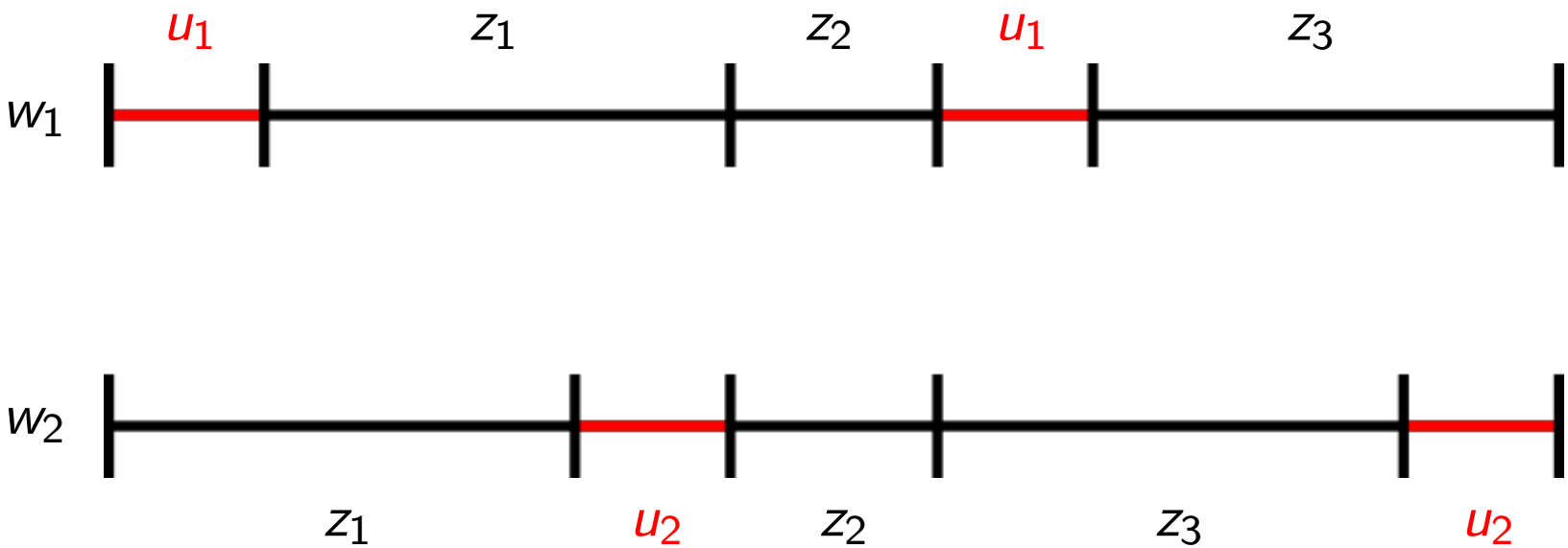


Interleaving
$(k_1 < k_2 \leq \ell_1 < \ell_2)$

Nested
$(k_1 < k_2 \leq \ell_2 < \ell_1)$

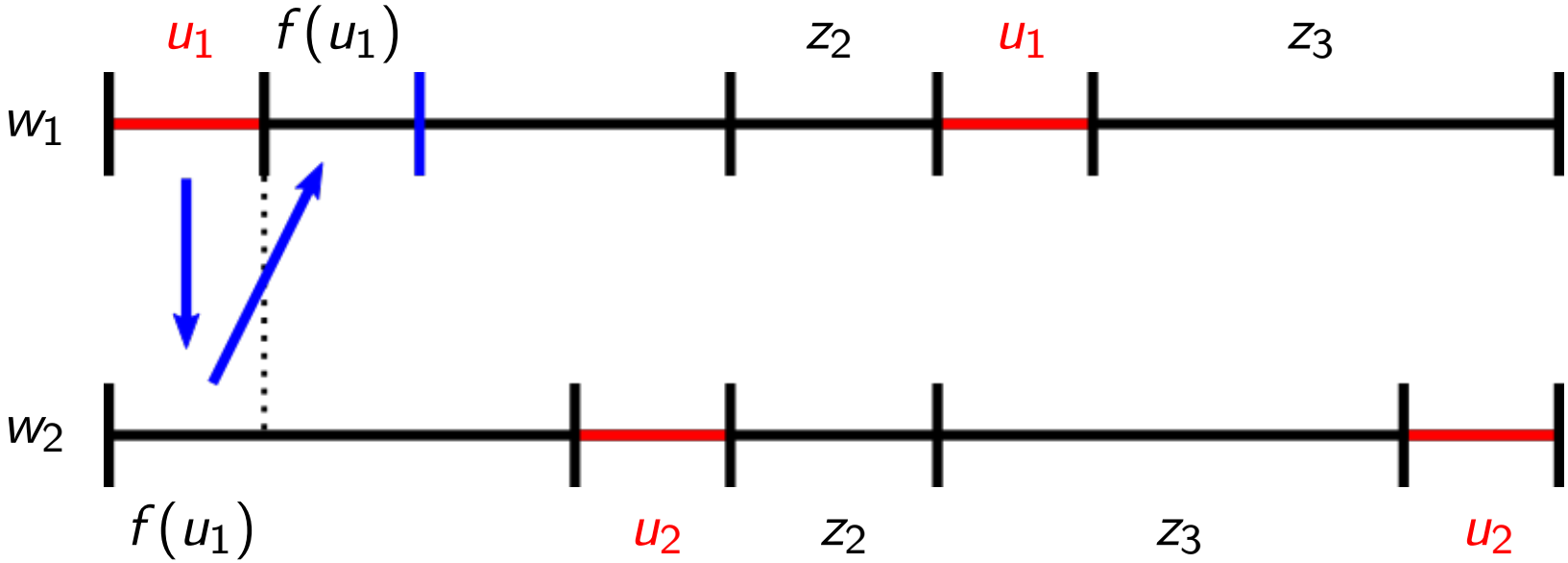
Sequential
$(k_1 \leq \ell_1 < k_2 \leq \ell_2)$

Interleaving Insertions $(k_1 < k_2 \leq \ell_1 < \ell_2)$

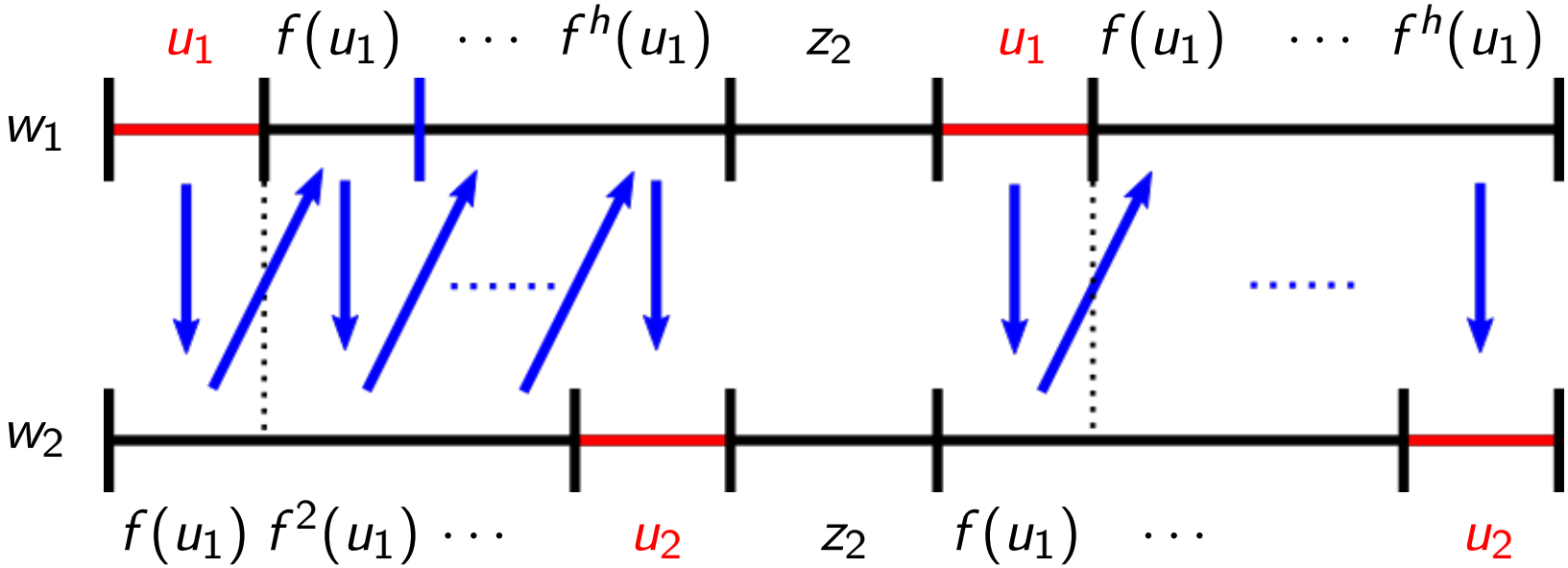


Note that $u_1 z_1 \sim z_1 u_2$. We consider $\mathcal{I}_1 = \mathcal{I}_2 = \rho$ to start.

Interleaving Insertions $(k_1 < k_2 \leq \ell_1 < \ell_2)$



Interleaving Insertions ($k_1 < k_2 \leq \ell_1 < \ell_2$)



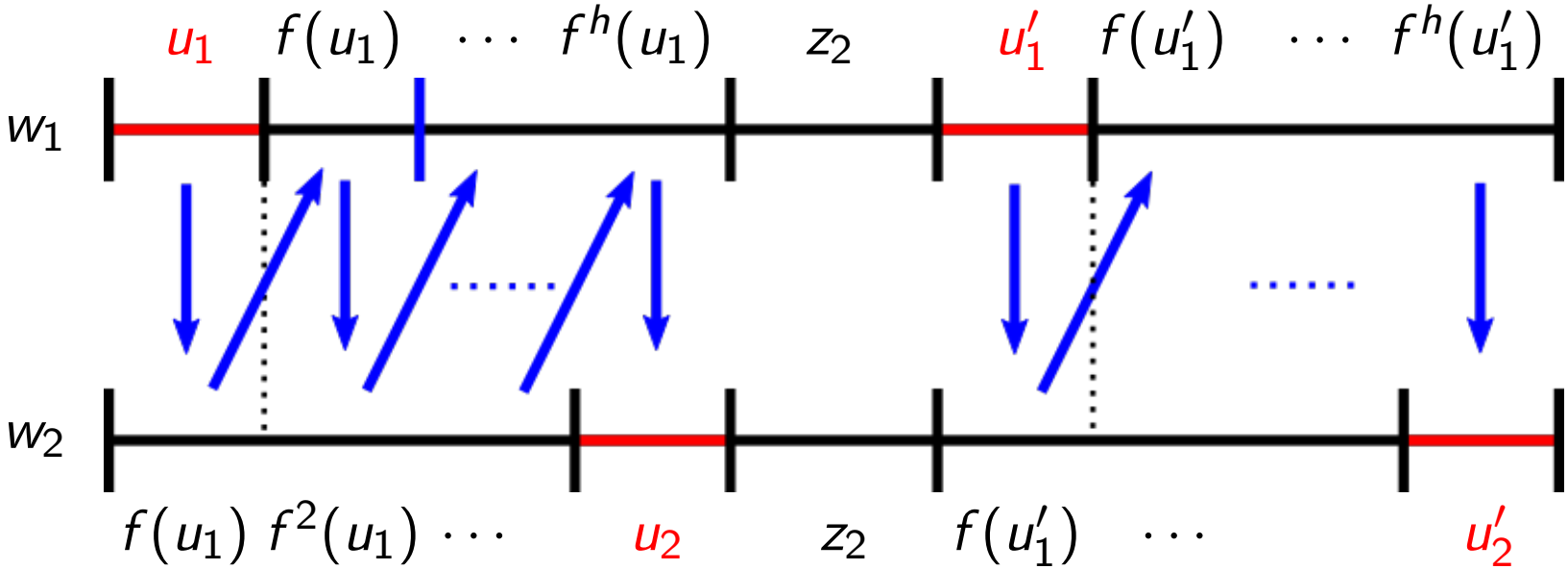
We adapt a result by Lyndon and Schützenberger:

Lemma

If $xz = zy$ and $x \neq \epsilon$, then $x = st$, $z = (st)^h s$, and $y = ts$ for some $s, t \in \Sigma^*$ and $h \geq 0$.

Lyndon, R.C., and Schützenberger, M.-P. "The equation $a^M = b^N c^P$ in a free group." The Michigan Mathematical Journal **9**:4 (1962) pp 289-298.

Interleaving Insertions $(k_1 < k_2 \leq \ell_1 < \ell_2)$

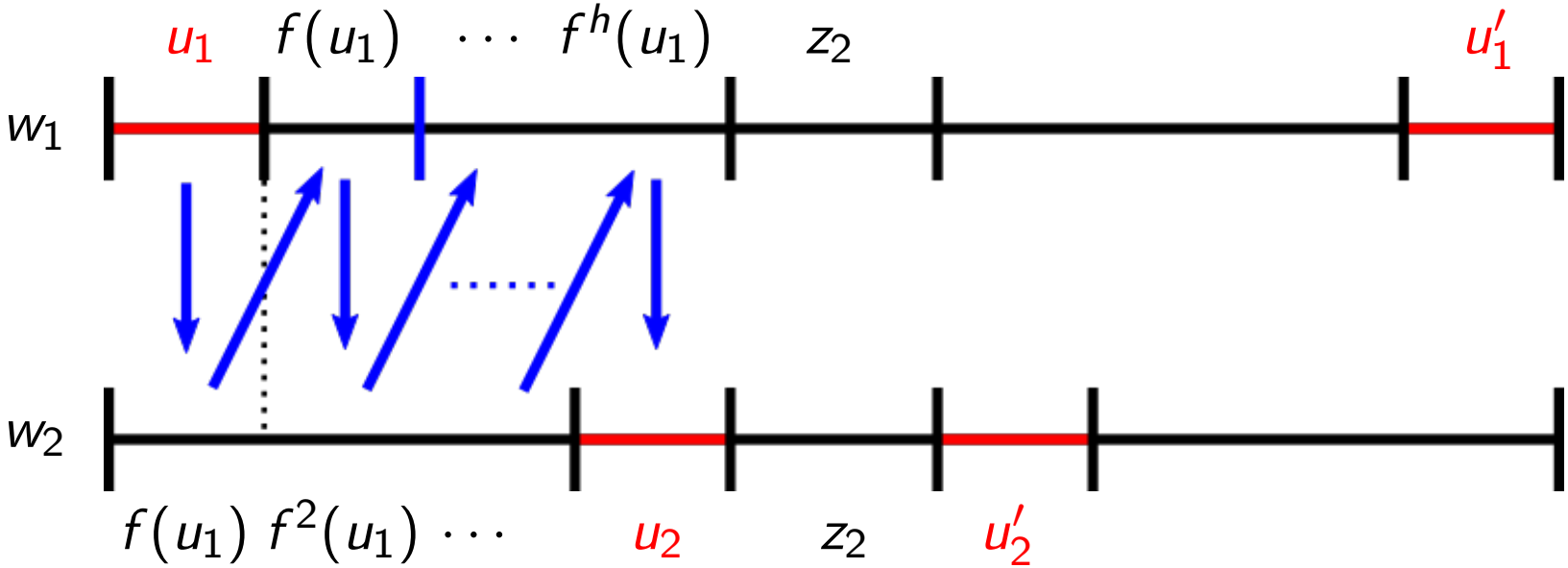


Proposition (Interleaving)

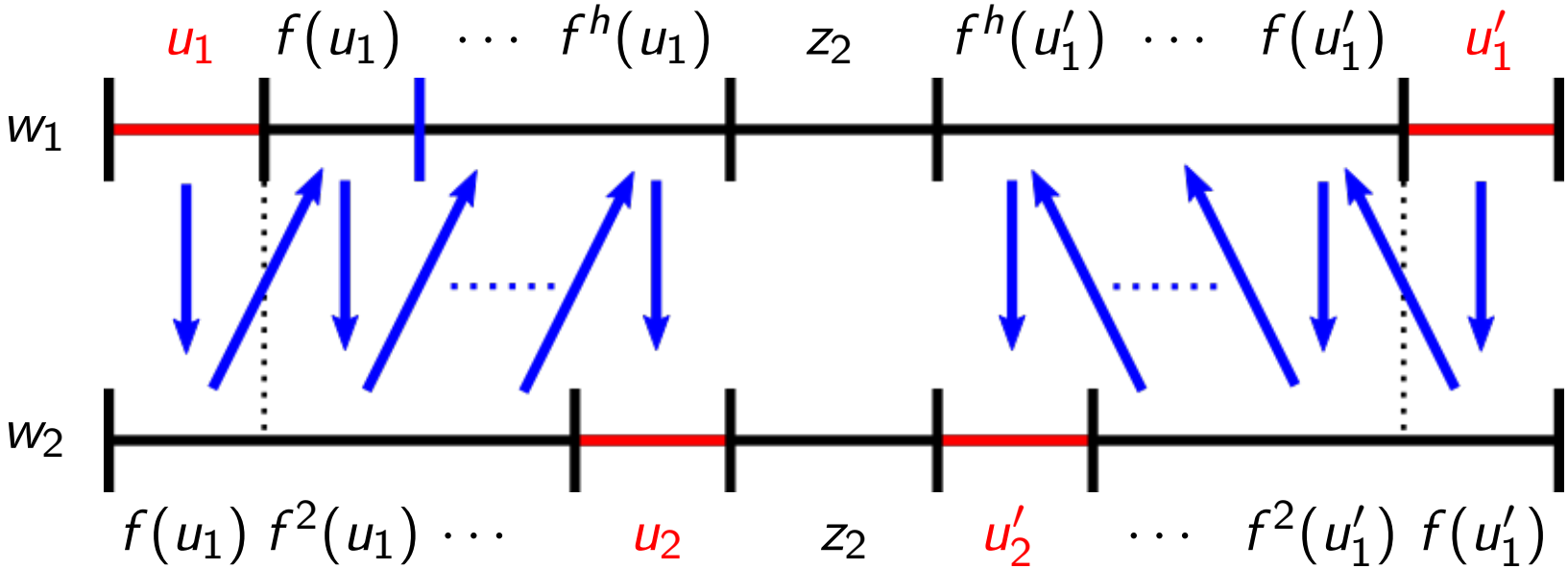
- If $\mathcal{I}_1 = \mathcal{I}_2 = \rho$, then $z_1 z_3$ is a repeat word.
- If $\mathcal{I}_1 = \mathcal{I}_2 = \tau$, then $z_1 z_3 \sim R_\tau(k_2 - k_1, |u_1|)$.

$R_\tau(h, q) = x_1 x_2 \cdots x_h x_1^R x_2^R \cdots x_h^R$ where each $x_i x_i^R$ is a return word and $|x_i| = q$ for $1 \leq i \leq h$.

Nested Insertions ($k_1 < k_2 \leq \ell_2 < \ell_1$)



Nested Insertions ($k_1 < k_2 \leq \ell_2 < \ell_1$)

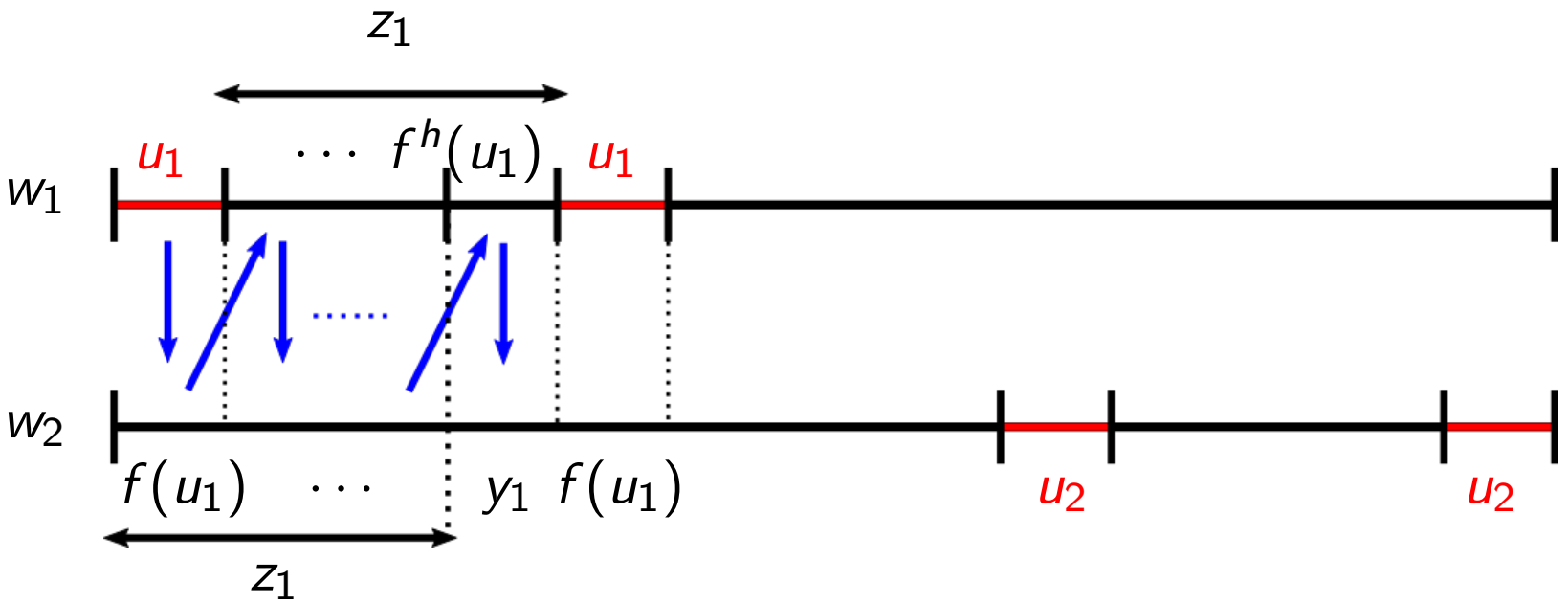


Proposition (Nested)

- If $\mathcal{I}_1 = \mathcal{I}_2 = \rho$, then $z_1 z_3 \sim T_\rho(k_2 - k_1, |u_1|)$.
- If $\mathcal{I}_1 = \mathcal{I}_2 = \tau$, then $z_1 z_3$ is a return word.

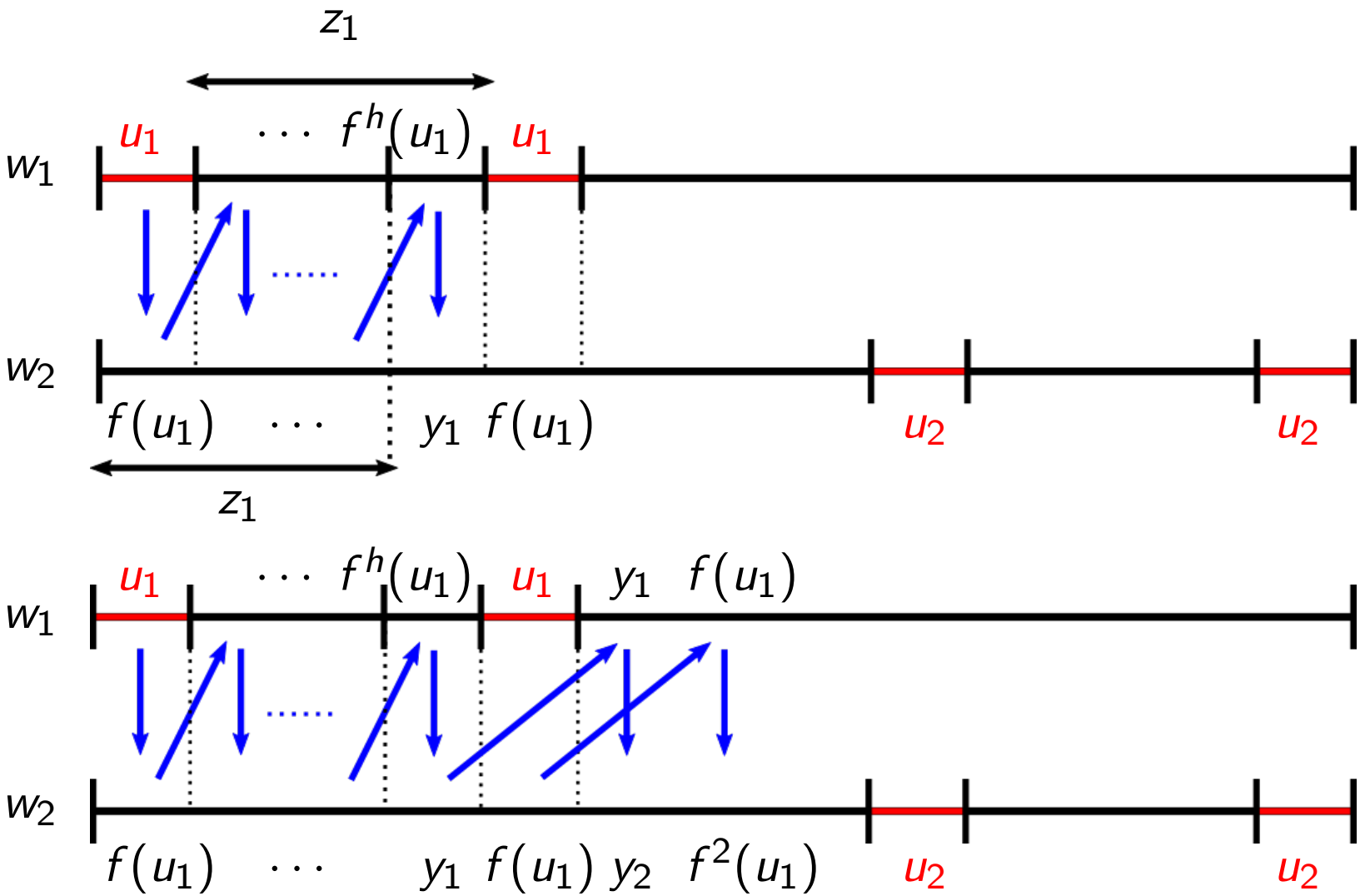
$T_\rho(h, q) = x_1 x_2 \cdots x_{h-1} x_h x_h x_{h-1} \cdots x_2 x_1$ where each $x_i x_i$ is a repeat word and $|x_i| = q$ for $1 \leq i \leq h$.

Sequential Insertions ($k_1 \leq \ell_1 < k_2 \leq \ell_2$)

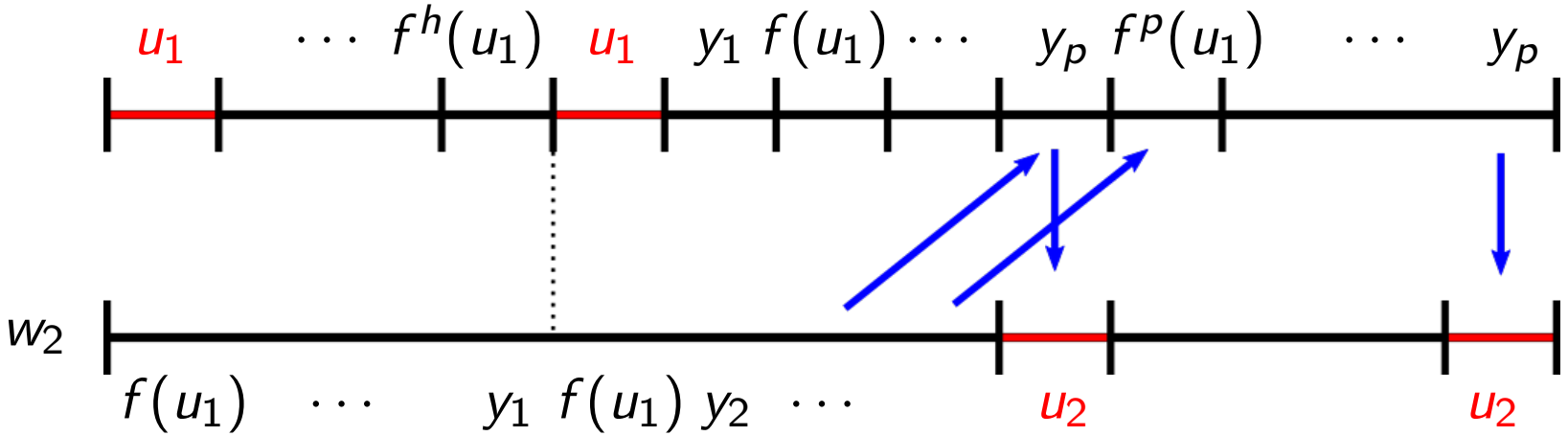


Note that $|y_1| = |u_1|$. We consider $\mathcal{I}_1 = \mathcal{I}_2 = \rho$ to start.

Sequential Insertions ($k_1 \leq \ell_1 < k_2 \leq \ell_2$)



Sequential Insertions $(k_1 \leq \ell_1 < k_2 \leq \ell_2)$



$$w = f(u_1)\cdots f^h(u_1)y_1 f(u_1)y_2 f(u_2)\cdots y_p f^p(u_1)\cdots y_p$$

For example:

$$12345612781234567 8 \star \rho(ab, 1, 5) \sim 12345612781234567 8 \star \rho(ab, 13, 17)$$

$w =$	12	34	56	12	78	12	34	56	78
	x_1	x_2	y_1	x_1	y_2	x_2	x_2	y_1	y_2

Sequential Insertions ($k_1 \leq \ell_1 < k_2 \leq \ell_2$)

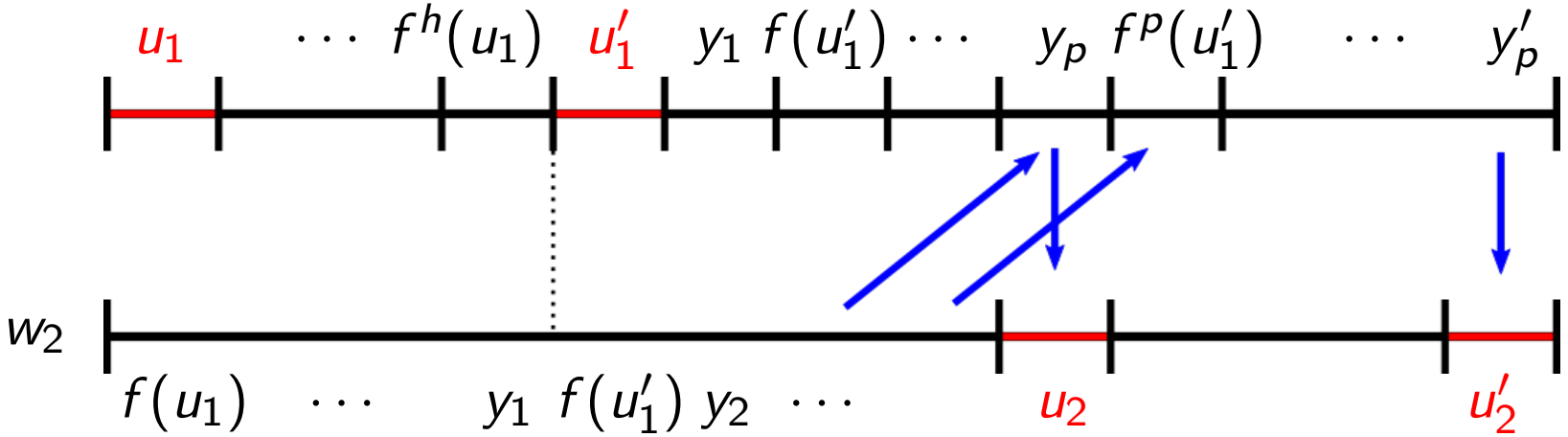
Consider the following words:

$$
\begin{aligned}
v_0 &= 123123 && |v_0| = 2 \cdot 3 = 6 \\
v_1 &= 123a123a && = v_0 \star \rho(a, |v_0| - 3, |v_0| + 1) \\
v_2 &= 123a1b23ab && = v_1 \star \rho(b, |v_1| - 3, |v_1| + 1) \\
v_3 &= 123a1b2c3abc && = v_2 \star \rho(c, |v_2| - 3, |v_2| + 1) \\
v_4 &= 123a1b2c3dabcd && = v_3 \star \rho(d, |v_3| - 3, |v_3| + 1)
\end{aligned}
$$

Words v_i are *generalized ρ-tangled cords*, denoted $C_\rho(m, q, i)$ with $m = 3$ and $q = 1$. *Tangled cords*, $C_\rho(1, 1, i)$, were introduced in:

Burns, J. et al. Four-regular graphs with rigid vertices associated to DNA recombination. Discrete Applied Mathematics, **161**:10-11 (2013) pp 1378-1394.

Sequential Insertions ($k_1 \le \ell_1 < k_2 \le \ell_2$)

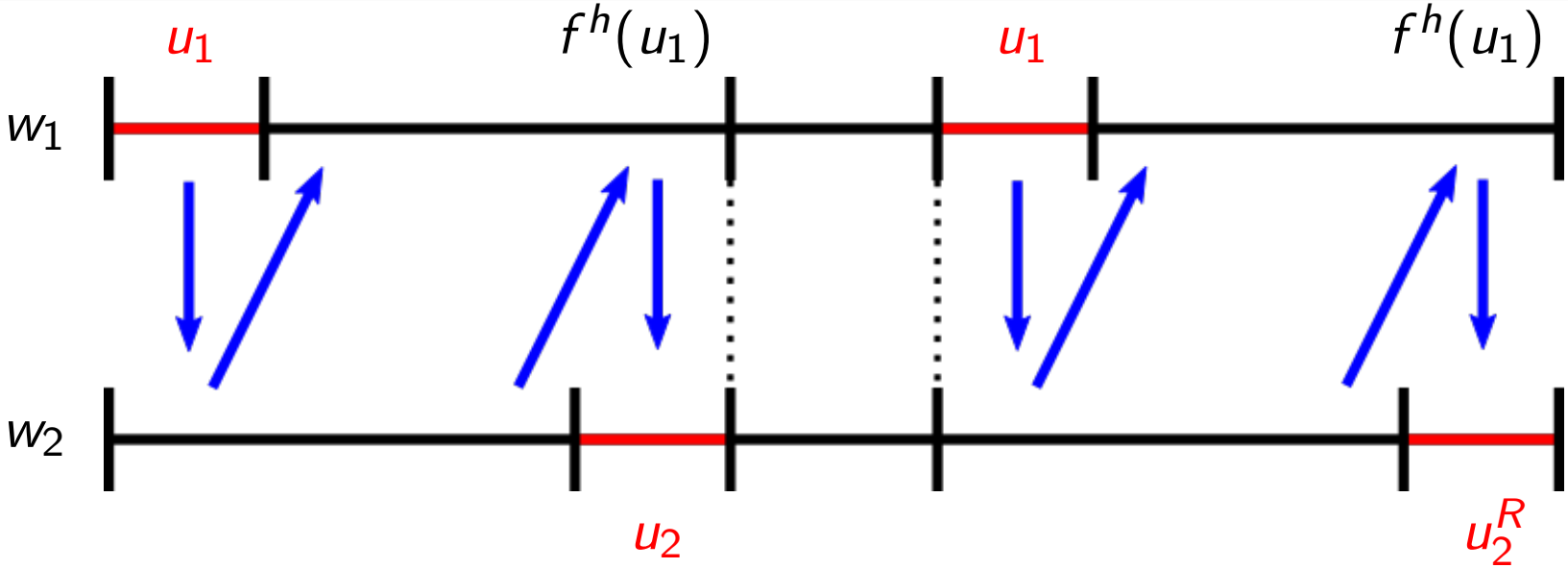


Proposition (Sequential)

- If $\mathcal{I}_1 = \mathcal{I}_2 = \rho$, then $z_1 z_2 z_3 \sim C_\rho\left(\ell_1 - k_1, |u_1|, \dfrac{k_2 - \ell_1}{2|u_1|}\right)$.
- If $\mathcal{I}_1 = \mathcal{I}_2 = \tau$, then $z_1 z_2 z_3 \sim C_\tau\left(\ell_1 - k_1, |u_1|, \dfrac{k_2 - \ell_1}{2|u_1|}\right)$.

Generalized τ-tangled cord $C_\tau(m, q, j)$ is defined similarly.

Repeat and Return Insertions



Lemma*

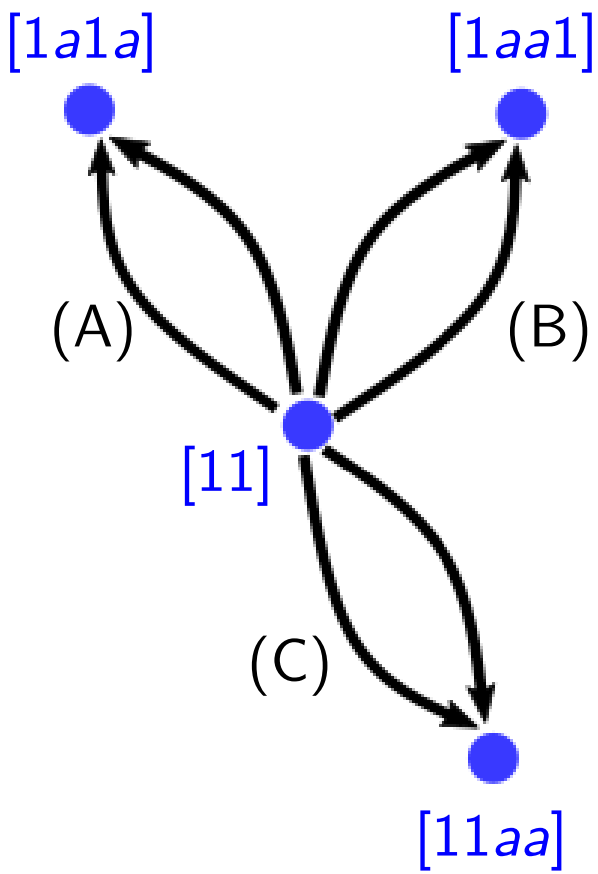
If uu and vv^R are repeat and return words in $w \in \Sigma^*$ such that $\Sigma[u] \cap \Sigma[v] \neq \emptyset$, then $|u| = 1$ or $|v| = 1$.

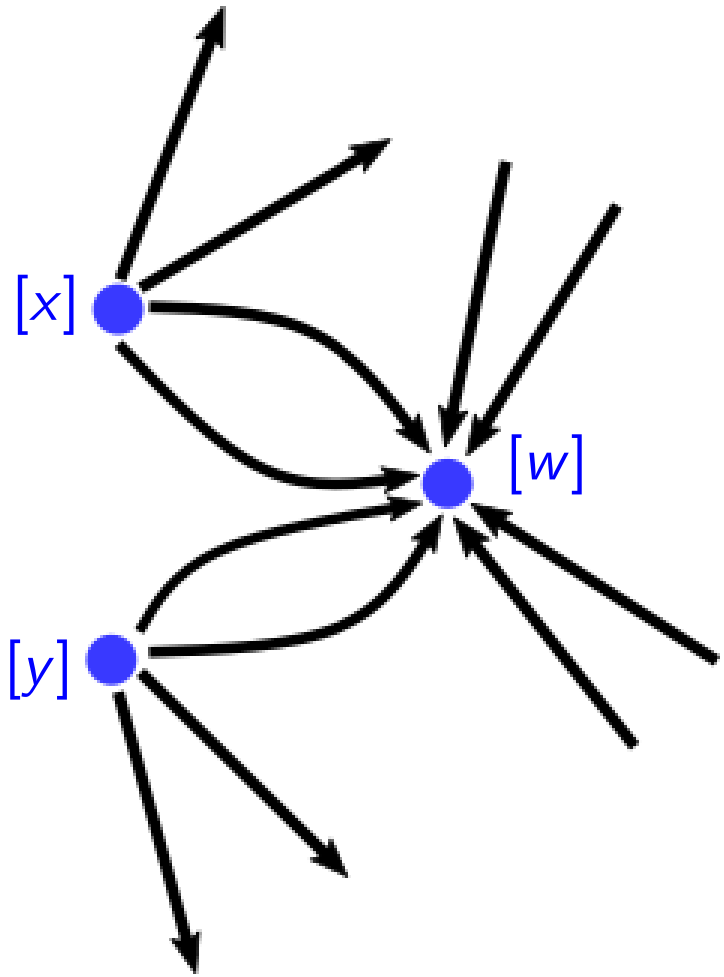
Proposition (Repeat and Return)

Suppose that $\mathcal{I}_1 \neq \mathcal{I}_2$. If $w_1 \sim w_2$, then $|u_1| = |u_2| = 1$.

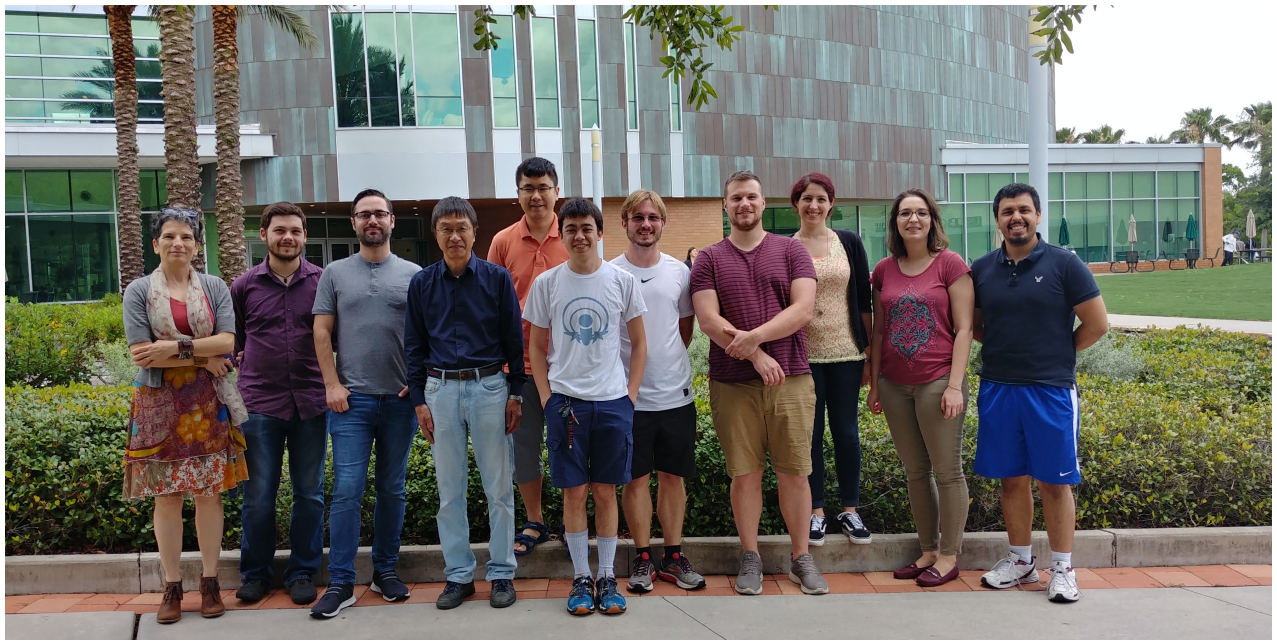
*: Jonoska, N. et al. Patterns and Distances in Words Related to DNA Rearrangement. Fundamenta Informaticae **154**:1-4 (2017) pp 225-238.

Future Work: Graph of Words





Thank You for Listening!



Work supported by NSF grants CCF-1526485, DMS-1800443, and DMS-1764366, and NIH grant R01 GM109459.

[1]: Burns, J. et al. Recurring patterns among scrambled genes in the encrypted genome of the ciliate Oxytricha trifallax. Journal of Theoretical Biology **410** (2016) pp 171-180.

[2]: Jonoska, N. et al. Patterns and Distances in Words Related to DNA Rearrangement. Fundamenta Informaticae **154**:1-4 (2017) pp 225-238.