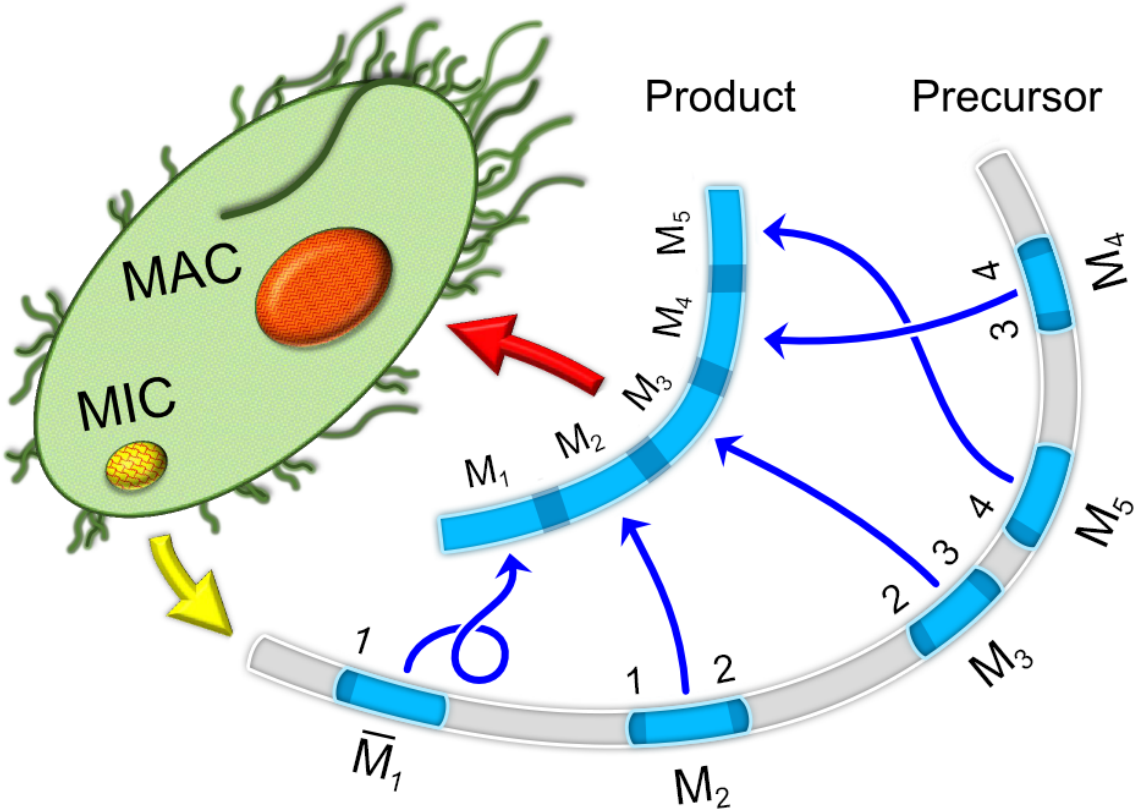
Transformations on Double Occurrence Words Motivated by DNA Rearrangement

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	) $w = 1123434$

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 $= \{0, 1, \ldots, 9\}$,

$w = 015164443$ is a word over

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 $[w] = \{0, 1, 3, 4, 5, 6\}$

The set of all words over is and includes the empty word.

Preliminaries

Given alphabet $= f0; 1; : : : ; 9g$,

$w = 015164443$ is a word over

The length of w is 9, written $jwj = 9$

$w^R = 344461510$ is the reverse of w

The set of symbols used in w is $[w] = f0; 1; 3; 4; 5; 6g$

The set of all words over is and includes the empty word.

De nition

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

$$11; \quad 1221; \quad 112234342 \quad _{DOW}$$

Single occurrence words (SOWs) and $_{SOW}$ are similarly de ned.

De nition: Repeat and Return Words

De nition

Given $w \in$ and SO $W u \in^{+} =$ n f g,

the word uu is a repeat word in w if $w = z_1uz_2uz_3$ for some $z_1; z_2; z_3 \in$

the word uu^R is a return word in w if $w = z_1uz_2u^Rz_3$ for some $z_1; z_2; z_3 \in$

w	Repeat words
11234552367887 6	234234, 2323, 88, etc.

w	Return words
1123455346788 76	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

$$w_0 = 56677889ab1239a12344 5bcc$$

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

$w_0 = 5\underline{6}6\underline{7}7\underline{8}89ab1239a123\underline{4}\underline{4}5b\underline{c}c$

$w_1 = 59ab1239a1235b$

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

$w_0 = 5667788 9ab1239a12344 5bcc$

$w_1 = 59ab1239a1235b$

$w_2 = 5b5b$

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

$w_0 = 5667788 9ab1239a12344 5bcc$

$w_1 = 59ab1239a1235b$

$w_2 = 5b5b$

$w_3 =$

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.

De nition: Repeat and Return Insertions

De nition

Given $w = a_1 \quad a_n \in \quad^{DOW}$,

let $1 \quad k \quad ` \quad n+1$,

let $I \in \{\ ;\ \}$ be a symbol not in $\ $, and

let $u \in \quad^{+}$ be a SOW such that $[u] \setminus \quad [\ w] = \ ;$.

Then $I(u;k;`)$ is an insertion on w which acts as follows:

$w \ ? \ I(u;k;`) = a_1 \quad a_{k\ 1}(u a_k \quad a_{`\ 1} u^0 a_` \quad a_n$ where

$$u^0 = \begin{cases} u & \text{if} \quad I = \\ u^R & \text{if} \quad I = \ : \end{cases}$$

$$1232314554 \ !^{(abc;4;6)} \ 123abc23abc14554$$

$$1232314554 \ !^{(abc;7;10)} \ 123231abc4554cba$$

De nition: Equivalence

De nition

Words $v; w \in$ are equivalentif there exists a bijection $f : \to$ such that $f(v) = w$; in this case, we write v w.

Examples of equivalent pairs of words:

123123	1	2	3	1221?	$(ab; 3; 3) =$	12abba21	1	2	a	b
	#	#	#				#	#	#	#
321321	3	2	1	1221?	$(ab; 1; 5) =$	ab1221ba	a	b	1	2

The following words are not equivalent:

$$1232314554? \quad (abc; 4; 6) = 123abc23abc14554$$
$$\# \quad \#$$
$$1232314554? \quad (abc; 7; 10) = 123231abc4554cba$$

When Do Insertions Yield Equivalent Words?

Let $w = a_1 \quad a_n \in \quad_{DOW}$ be given.

1. If $w_1 = w\,?\,(u_1; k_1; `_1)$ and $w_2 = w\,?\,(u_2; k_2; `_2)$, then is it possible for w_1 and w_2 to be equivalent?

When Do Insertions Yield Equivalent Words?

Let $w = a_1 \quad a_n \in \quad_{DOW}$ be given.

1. If $w_1 = w ? \ (u_1; k_1; `_1)$ and $w_2 = w ? \ (u_2; k_2; `_2)$, then is it possible for w_1 and w_2 to be equivalent? Yes.

$$w_1 = 1212 ? \ (a; 3; 5) = 12a12a$$
$$w_2 = 1212 ? \ (a; 1; 3) = a12a12$$

1	2	a
#	#	#
a	1	2

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

2. In general if $w_1 = w ? I_1(u_1; k_1; `_1) \quad w ? I_2(u_2; k_2; `_2) = w_2$, what can we say about w if the insertions are "distinct"?

$$1221 ? \ (ab; 3; 3) = 12abba21$$
$$1221 ? \ (ab; 1; 5) = ab1221ba$$

1	2	a	b
#	#	#	#
a	b	1	2

De nition: Distinct Insertions

De nition

Two insertions $I_1(u_1; k_1; \ell_1)$ and $I_2(u_2; k_2; \ell_2)$ on $w \in DOW$ are distinct if at least one of the following holds:

$$(k_1; \ell_1) \neq (k_2; \ell_2); \quad I_1 \neq I_2; \quad \text{or} \quad |u_1| \neq |u_2|$$

If $w_1 \sim w_2$, then $|u_1| = |u_2|$. What if $w_1 \sim w_2$ but only $I_1 \neq I_2$?

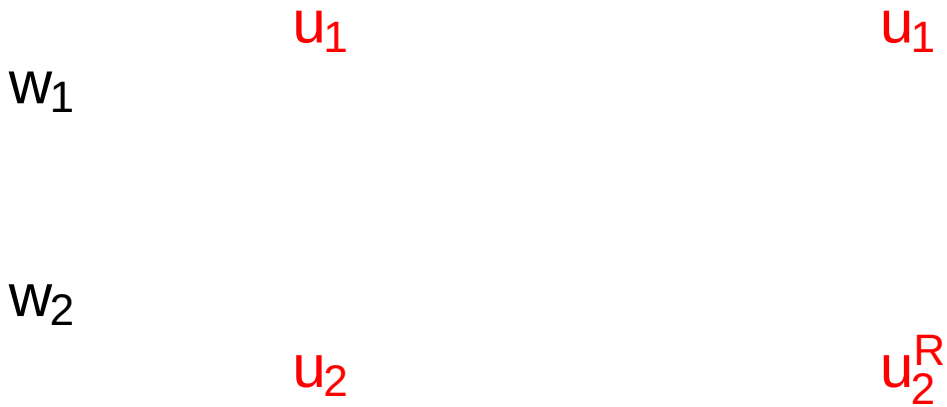
De nition: Distinct Insertions

De nition

Two insertions $I_1(u_1; k_1; \ell_1)$ and $I_2(u_2; k_2; \ell_2)$ on $w \in DOW$ are distinct if at least one of the following holds:

$$(k_1; \ell_1) \neq (k_2; \ell_2); \quad I_1 \neq I_2; \quad \text{or} \quad |u_1| \neq |u_2|$$

If $w_1 \quad w_2$, then $|u_1| = |u_2|$. What if $w_1 \quad w_2$ but only $I_1 \neq I_2$?



$u_1 \quad u_2$ and $u_1 \quad u_2^R$) $|u_2| = 1$ since $u_2 \in SOW$

Distinct Insertions and Equivalent DOWs

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

u_1 u_1'

w_1

w_2

u_2 u_2'

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

u_1 u_1' u_1 u_1' u_1 u_1'

u_2 u_2' u_2 u_2' u_2 u_2'

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

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

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

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

u_1 z_1 z_2 u_1 z_3

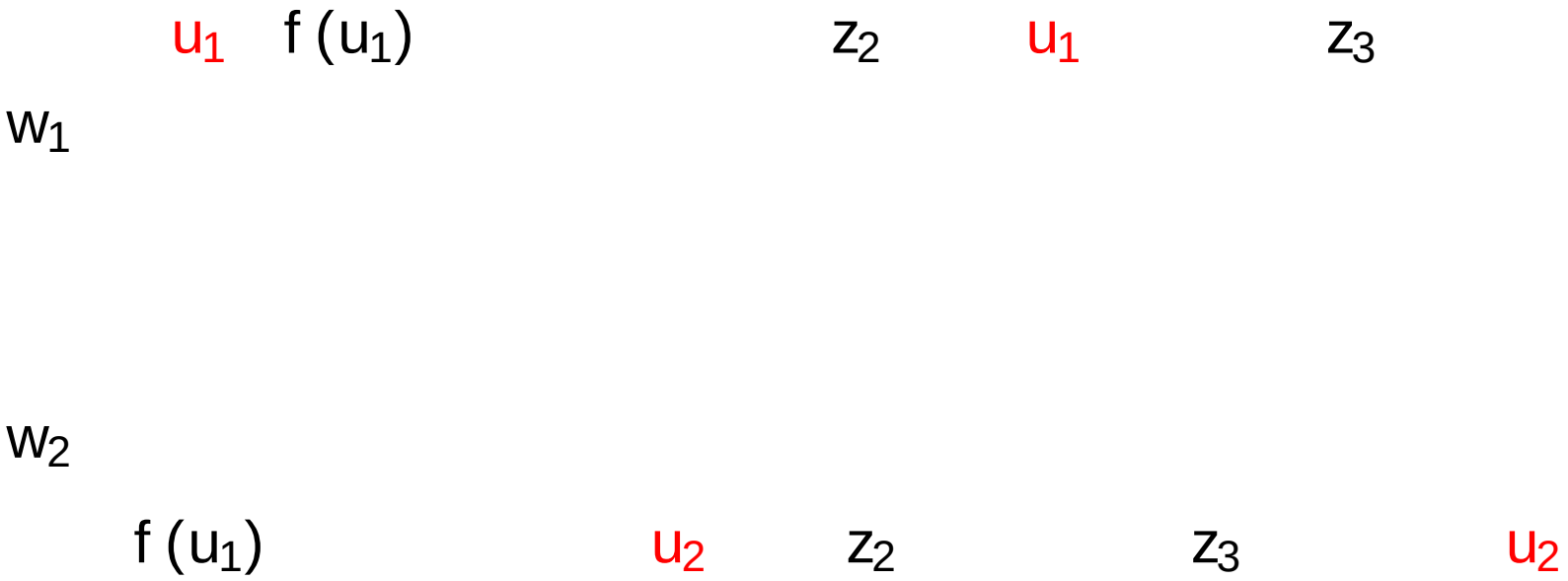
w_1

w_2

z_1 u_2 z_2 z_3 u_2

Note that $u_1 z_1$ $z_1 u_2$. We consider $l_1 =$ $l_2 =$ to start.

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



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

u_1 $f(u_1)$ $f^h(u_1)$ z_2 u_1 $f(u_1)$ $f^h(u_1)$

w_1

w_2

$f(u_1)$ $f^2(u_1)$ u_2 z_2 $f(u_1)$ u_2

We adapt a result by Lyndon and Schützenberger:

Lemma

If $xz = zy$ and $x \neq \varepsilon$, then $x = st$, $z = (st)^h s$, and $y = ts$ for some $s, t \geq 2$ 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 ($k_1 < k_2$, $\ell_1 < \ell_2$)

u_1 $f(u_1)$ $f^h(u_1)$ z_2 u_1^0 $f(u_1^0)$ $f^h(u_1^0)$

w_1

w_2

$f(u_1)$ $f^2(u_1)$ u_2 z_2 $f(u_1^0)$ u_2^0

Proposition (Interleaving)

If $l_1 = l_2 =$, then $z_1 z_3$ is a repeat word.

If $l_1 = l_2 =$, then $z_1 z_3$ $R(k_2 \quad k_1; |u_1|)$.

$R(h; q) = x_1 x_2 \quad x_h x_1^R x_2^R \quad x_h^R$ where each $x_i x_i^R$ is a return word and $|x_i| = q$ for $1 \quad i \quad h$.

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

u_1 $f(u_1)$ $f^h(u_1)$ z_2 u_1^0

w_1

w_2

$f(u_1)$ $f^2(u_1)$ u_2 z_2 u_2^0

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

u_1 $f(u_1)$ $f^h(u_1)$ z_2 $f^h(u_1^0)$ $f(u_1^0)$ u_1^0

w_1

w_2

$f(u_1)$ $f^2(u_1)$ u_2 z_2 u_2^0 $f^2(u_1^0)$ $f(u_1^0)$

Proposition (Nested)

If $l_1 = l_2 =$, then $z_1 z_3$ $T(k_2 \quad k_1; |u_1|)$.

If $l_1 = l_2 =$, then $z_1 z_3$ is a return word.

$T(h; q) = x_1 x_2 \quad x_{h-1} x_h x_h x_{h-1} \quad x_2 x_1$ where each $x_i x_i$ is a repeat word and $|x_i| = q$ for $1 \quad i \quad h$.

Sequential Insertions $k_1(\quad \grave{}_1 < k_2 \quad \grave{}_2)$

z_1

w_1 u_1 $f^h(u_1)$ u_1

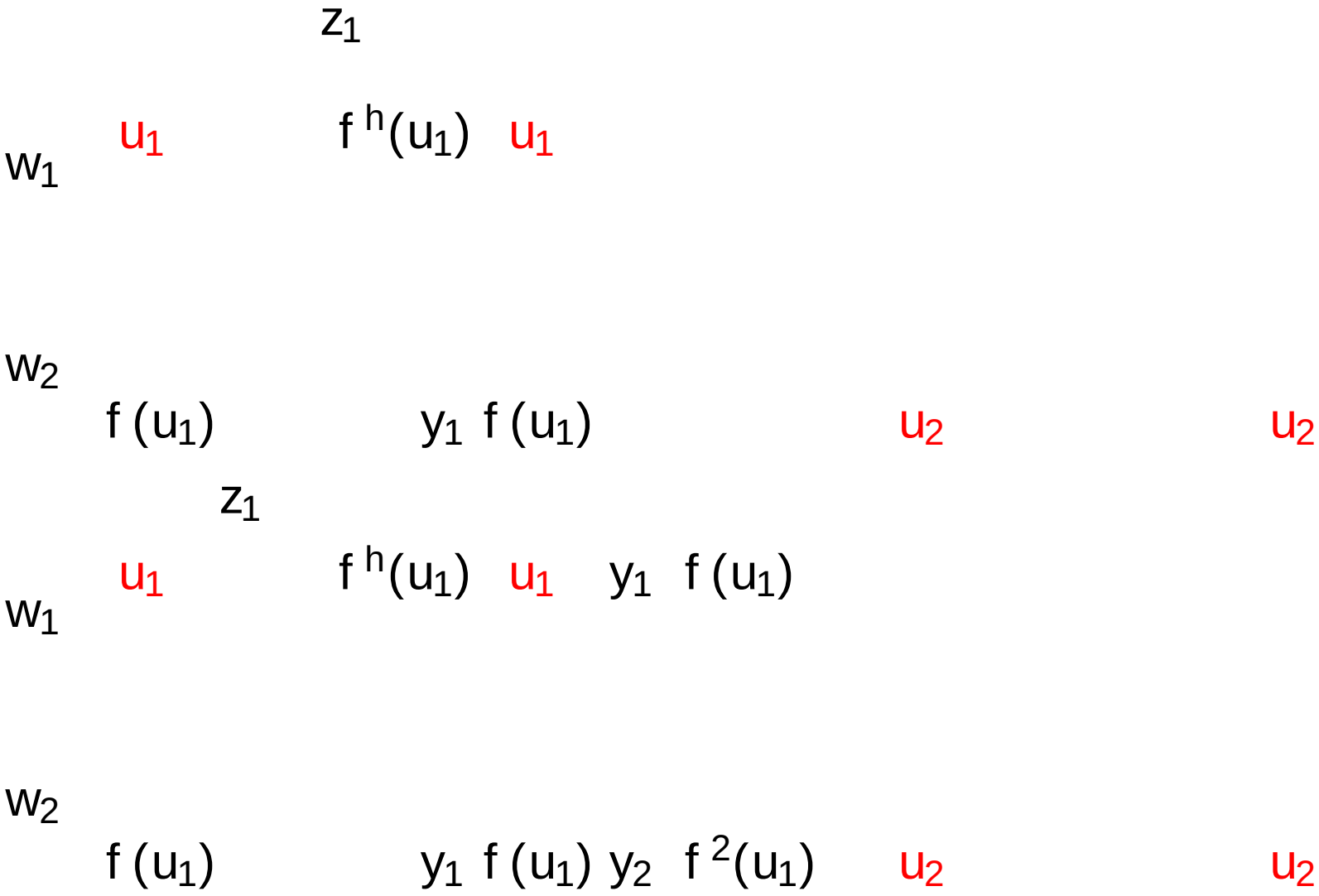
w_2

$f(u_1)$ y_1 $f(u_1)$ u_2 u_2

z_1

Note that $jy_1j = ju_1j$. We consider $l_1 = l_2 =$ to start.

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



Sequential Insertions ($k_1 < k_2$)

u_1 $f^h(u_1)$ u_1 $y_1 f(u_1)$ $y_p f^p(u_1)$ y_p

w_2

$f(u_1)$ $y_1 f(u_1) y_2$ u_2 u_2

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

For example:

$12345612781234567878(ab; 1; 5)$ $12345612781234567878(ab; 13; 17)$

$w = 12\ 34\ 56\ 12\ 78\ 12\ 34\ 56\ 78$

$x_1\ x_2\ y_1\ x_1\ y_2\ x_2\ x_2\ y_1\ y_2$

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

Consider the following words:

$$v_0 = 123123 \qquad |v_0| = 2\ \ 3 = 6$$
$$v_1 = 123a123a = v_0 ?\ (a; |v_0|\ \ 3; |v_0| + 1)$$
$$v_2 = 123a1b23ab = v_1 ?\ (b; |v_1|\ \ 3; |v_1| + 1)$$
$$v_3 = 123a1b2c3abc = v_2 ?\ (c; |v_2|\ \ 3; |v_2| + 1)$$
$$v_4 = 123a1b2c3dabcd = v_3 ?\ (d; |v_3|\ \ 3; |v_3| + 1)$$

Words v_i are generalized -tangled cords, denoted $C(m; q; i)$ with $m = 3$ and $q = 1$. Tangled cords, $C(1; 1; i)$, 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 ($k_1 < k_2$)

u_1 $f^h(u_1)$ u_1^0 y_1 $f(u_1^0)$ y_p $f^p(u_1^0)$ y_p^0

w_2

$f(u_1)$ y_1 $f(u_1^0)$ y_2 u_2 u_2^0

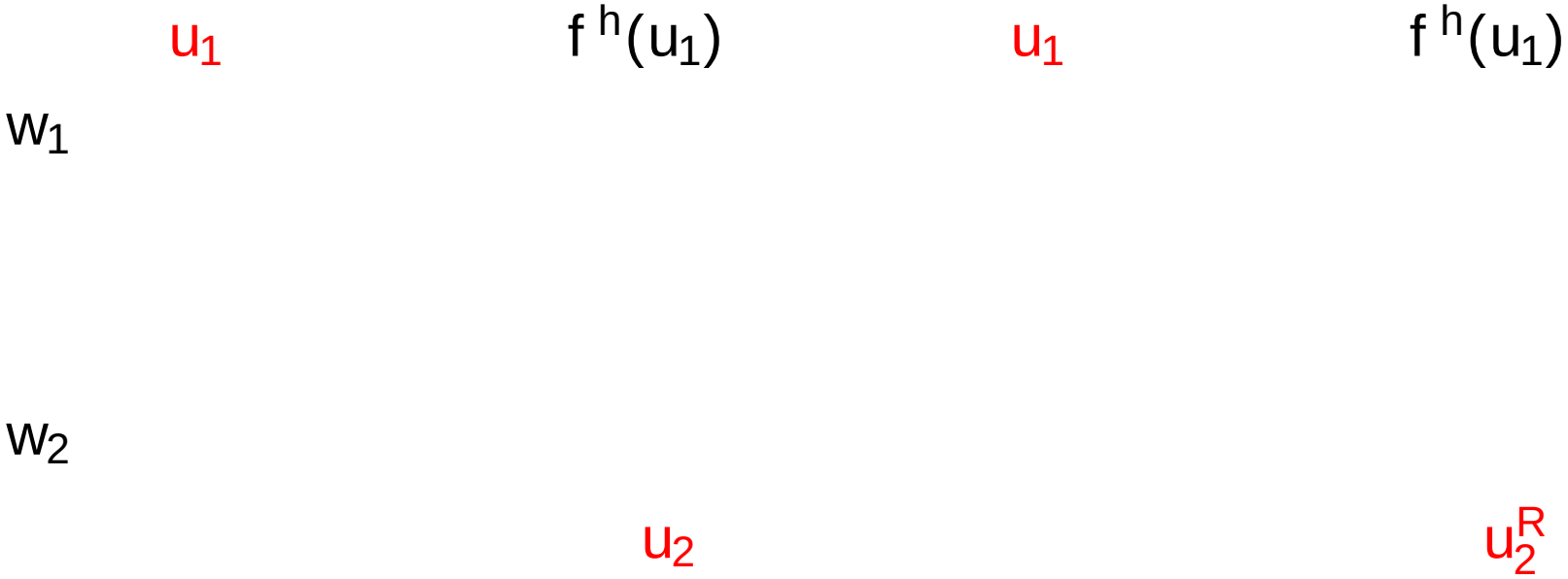
Proposition (Sequential)

If $l_1 = l_2 =$, then $z_1 z_2 z_3$ C $k_1; ju_1j; \frac{k_2}{2ju_1j}$.

If $l_1 = l_2 =$, then $z_1 z_2 z_3$ C $k_1; ju_1j; \frac{k_2}{2ju_1j}$.

Generalized -tangled cord $C(m; q; j)$ is de ned similarly.

Repeat and Return Insertions



Lemma

If uu and vv^R are repeat and return words in w 2 such that $[u] \setminus [v]$ 6 ; , then $juj = 1$ or $jvj = 1$.

Proposition (Repeat and Return)

Suppose that l_1 6 l_2. If w_1 w_2, then $ju_1j = ju_2j = 1$.

: 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