Duplication languages: characterizations and open problems

José M. Sempere
(jsempere@dsic.upv.es)
The genomic duplication is a DNA recombination problem due to unequal crossing over.
Previous approaches to duplication languages:

• J. Dassow and V. Mitrana, 1997 (together with A. Salomaa, 2002): genomic operations over strings and languages (transpositions, duplications, inversions, cross over, etc)

• C. Martín-Vide and Gh. Paun, 1999: Duplication Grammars

• V. Mitrana and G. Rozenberg, 1999: Properties of Duplication Grammars

• J. Dassow, V. Mitrana, G. Paun, 1999: Regularity of duplication closure

• M. Wang, 2000: Irregularity of duplication closure

• P. Leupold, V. Mitrana, J.M. Sempere, 2004: Properties of (restricted) duplication languages

• P. Leupold, C. Martín-Vide, V. Mitrana, 2005: Uniformly Bounded duplication languages

• P. Leupold, V. Mitrana, 2007: Uniformly Bounded duplication codes
A first view at duplication

abcde
A first view at duplication

abcde
A first view at duplication

\[ \text{abcde} \]
\[ \rightarrow \]
\[ \text{abcdbcde} \]
A first view at duplication

abcde

abcdbcde
A first view at duplication

\[
\text{abcde}
\]

\[
\text{abcd} \text{bcde}
\]
A first view at duplication

```
abcde
```
```
abcdbcde
```
```
abcdbccdbcde
```
A first view at duplication

abcde

→

abcdbcde

→

abcdbcdbcde
A first view at duplication

```
abcde

abcdbcde

abcdbccdbede
```


A first view at duplication

ABCDE

\[ \begin{align*}
\text{abcdbbcdbcde} \\
\text{abcdbccdbcde} \\
\text{abcdbccbccdbcde}
\end{align*}\]
A first view at duplication

```
ABCDE

ACDCE

ACDBDACE

ACDBDCDBACE

ACDDBCCDBACE
```

abcde
A first view at duplication

abcde

→

abcdbcde

→

abcdbccdbbcde

→

abcdbccbccdbcde
Duplication Languages $D^*(w)$

Let $V$ be an alphabet

$$D(w) = \{uxxv \mid w = uxv, u, x, v \in V^* \}$$

- $D^0(w) = w$
- $D^i(w) = \bigcup_{x \in D^{i-1}(w)} D(x) \quad i \geq 1$

$$D^*(w) = \bigcup_{i \geq 0} D^i(w)$$
On the regularity of duplication language $D^*(w)$
(Dassow, Mitrana & Paun, 1999)

**Th.** If $w$ is a string over a two-letter alphabet then $D^*(w)$ is a regular language

$$w = a_1a_2 \ldots a_n$$

$$D^*(w) = \{w_1a_1w_2a_2 \ldots w_{n-1}a_{n-1}w_n a_n \mid w_1 \in a_1^*, \ w_{n+1} \in a_n^*, \ w_i \in a_i^* \text{ for } a_{i-1} = a_i, \ w_i \in V^* \text{ for } a_{i-1} \neq a_i, 2 \leq i \leq n\}$$
On the irregularity of duplication language $D^*(w)$  
(M. Wang, 2000)

**Th.** Suppose $w$ is a word containing at least three distinct letters. Then $D^*(w)$ is not regular.

If $u = abc$ and $V$ is an alphabet.
If $u = abcv$ then there exists $v' \in V^*$ such that $uv' \in D^*(w)$

Suppose $u = abcu'$ is square free. Let $v$ be a shortest word such that $uv \in D^*(w)$. Then

$$|v| \geq \log_2(|u|/3)$$

Through Myhill-Nerode’s characterization we can construct an infinite sequence of pairwise inequivalent strings
Bounded and unbounded duplication languages

(Leupold, Mitrana & Sempere, 2004)

An equivalent definition of (un)bounded duplication languages

\[ X \in \{ \mathbb{N} \} \cup \{ [k] | k \geq 1 \} \]

\[ D_X(w) = \{ uxxv | w = u xv, u, x, v \in V^*, |x| \in X \} \]

- \[ D_X^0(w) = w \]
- \[ D_X^i(w) = \bigcup_{x \in D_X^{i-1}(w)} D_X(x) \quad i \geq 1 \]

\[ D_X^*(w) = \bigcup_{i \geq 0} D_X^i(w) \]

\[ D_N^*(w) \quad \text{the unbounded duplication language} \]

\[ D_{[k]}^*(w) \quad \text{the k-bounded duplication language} \]
On the context-freeness of k-bounded duplication language $D^{*}(w)$
(Leupold, Mitrana & Sempere, 2004)

**Th.** For any word $w$ and any positive integer $k$, $D^{*}[k](w)$ is context-free

A pushdown automaton based on a set of states in the form

\[
\begin{bmatrix}
\mu \\
\nu \\
w
\end{bmatrix}
\]

- guess
- memory
- pattern
First Open Problem

For any string $w$ with at least three distinct symbols ... how is the (non-regular) duplication language $D^*(w)$? (is it context-free, context-sensitive, ... ?)
Compensation (deletion) loops

Figure 8-14 Origins of (a) a terminal and (b) an intercalary deletion. Part (c) shows that pairing can occur between a normal chromosome and one with an intercalary deletion if the undeleted portion buckles out to form a deletion (or a compensation) loop.

An scheme for compensation (deletion) loops formation

(1) \( \alpha \beta \gamma \delta \)

(2) \( \alpha \beta \beta \gamma \delta \)

(3) \( \alpha \beta \gamma \delta \)

(4) \( \alpha \beta \gamma \delta \)
Duplication Languages with compensation loops $D_{cl}^*(w)$

Let $V$ be an alphabet without bracket symbols $[ $ and $]$ 

$w = x_1[w_1]x_2[w_2] \ldots x_n[w_n]$  

shuffle with common segments

$z = x_1[z_1]x_2[z_2] \ldots x_n[z_n]$  

$scs(w, z) = x_1[w_1z_1]x_2[w_2z_2] \ldots x_n[w_nz_n]$  

• $D_{cl}^0(w) = \{ux[x]v \mid w = u xv, \; u, x, v \in V^*\}$

• $D_{cl}^i(w) = \{scs(x, y) \mid x, y \in D_{cl}^{i-1}(w)\}, \; i \geq 1$
Duplication Languages with compensation loops $D_{cl}^{*}(w)$

An example

$w = abc$

$D_{cl}^{0}(abc) = \{abc, a[ab]bc, ab[bc], abc[bc], ab[ab]c, abc[bc], abc[abc]\}$

$scs(a[ab]bc, ab[ab]c) = a[ab]b[ab]c \in D_{cl}^{1}(abc)$

An erasing morphism $h$ such that $h(\ ) = h(\ ) = \lambda$

$h(a[ab]b[ab]c) = aababc$
Duplication Languages with compensation loops $D^*_{cl}(w)$

**Property 1.** For any arbitrary alphabet $V$ and any string $w \in V^+$, $h(D^*_{cl}(w))$ is regular

Let $w = w_1w_2 \ldots w_n$ with $w_i \in V$

$h(D^*_{cl}(w))$ can be denoted by the following regular expression

$$w_1(w_1)^*w_2(w_2 + w_1w_2)^*w_3(w_3 + w_2w_3 + w_1w_2w_3)^* \ldots w_n(w_n + \ldots + w_1w_2 \ldots w_n)^*$$

Let $w = abc$

$h(D^*_{cl}(abc))$ is denoted by $aa^*b(b + ab)^*c(c + bc + abc)^*$
Duplication Languages with compensation loops $D^*_{cl}(w)$

Property 2. The following two statements are true

1) For any alphabet $V$ with $\text{card}(V)=1$ and $w \in V^+$, $h\left(D^*_{cl}(w)\right) = D^*(w)$
2) For any alphabet $V$ with $\text{card}(V) > 1$ there exists $w \in V^+$ such that $h\left(D^*_{cl}(w)\right) \subsetneq D^*(w)$

1) is trivial given that if $w = a^n$ then $h\left(D^*_{cl}(a^n)\right) = D^*(a^n) = a^n a^*$

To prove 2), take $w = aba$ and the following duplicated string sequence

$$aba \ni abaaba \ni abaabbaaba$$

that does not belong to the language $h\left(D^*_{cl}(aba)\right) = aa^*b(b + ab)^*a(a + ba + aba)^*$
An scheme for dynamic compensation loops formation

(1) \[ a \rightarrow b \rightarrow c \rightarrow d \]

(2) \[ a \rightarrow b \rightarrow c \rightarrow d \]

(3) \[ a \rightarrow b \rightarrow c \rightarrow d \]

(4) \[ a \rightarrow b \rightarrow c \rightarrow d \]

(5) \[ a \rightarrow b \rightarrow c \rightarrow d \]
Duplication Languages with dynamic compensation loops $D_{dcl}^*(w)$

Let $V$ be an alphabet without bracket symbols [ and ]

\[
\begin{align*}
w &= [w_0]x_1[w_1]x_2[w_2] \ldots x_n[w_n] \\
z &= [z_0]x_1[z_1]x_2[z_2] \ldots x_n[z_n]
\end{align*}
\]

generalized shuffle with common segments
\[
gscs(w, z) = \{ [\eta_0]x_1[\eta_1]x_2[\eta_2] \ldots x_n[\eta_n] \mid \eta_0x_1\eta_1x_2\eta_2 \ldots x_n\eta_n = w_0z_0x_1w_1z_1 \ldots x_nw_nz_n \}
\]

- $D_{dcl}^0(w) = \{ux[x]v \mid w = uxv, \ u, x, v \in V^*\}$
- $D_{dcl}^i(w) = \bigcup_{x, y \in D_{dcl}^{i-1}(w)} gscs(x, y), \ i \geq 1$

\[
D_{dcl}^*(w) = \bigcup_{i \geq 0} D_{dcl}^i(w)
\]
Duplication Languages with dynamic compensation loops $D^*_dcl(w)$

An example

$$w = ab[ab]b[bb]c[bc]$$
$$z = a[a]bbc[cc]$$

$gscs(w, z)$ contains, among others, the following strings

$$x_1 = a[a]b[ab]b[bb]c[bc]c$$
$$x_2 = [a]a[bab]bb[bc]c$$
$$x_3 = a[aba]b[bbbc]b[cc]c$$

$$h(x_1) = h(x_2) = h(x_3) = aababbbbcabc$$
Duplication Languages with dynamic compensation loops $D_{dcl}^*(w)$

Property 3. Let $V$ be an alphabet with at least two symbols. Then there exists $w \in V^+$ such that $h(D_{cl}^*(w)) \subsetneq h(D_{dcl}^*(w))$

Let $w = ab$

$h(D_{cl}^*(w))$ can be denoted by the regular expression $aa^*b(b + ab)^*$

$ab[ab], a[a]b \in D_{dcl}^0(ab)$

$a[a]b[ab] \in gscs(a[a]b, ab[ab])$

$a[aba]b \in gscs(a[a]b, ab[ab])$

$D_{dcl}^1(ab)$

$a[abaaba]b \in gscs(a[aba]b, a[aba]b)$

$D_{dcl}^2(ab)$

$aabaabab \in h(D_{dcl}^*(w)) - h(D_{cl}^*(w))$
Property 4. Let $V$ be an arbitrary alphabet with at least three symbols. Then there exists $w \in V^+$ such that $h(D_{dcl}^*(w)) \subsetneq D^*(w)$

Let $w = abc$ and we obtain the following duplicated string

$$abc \equiv abcabc \equiv abcacabc$$

How can $abcacabc$ be obtained from $abc$ by duplication with dynamic compensating loops?

a symbol $a$ cannot be inserted between the symbols $c$

a symbol $c$ cannot be inserted between the symbols $a$
More Open Problem

• For any string $w \in V^+$ what is the most restrictive language class for $h(D_{dcl}^*(w))$?

• For any string $w$ with two different symbols is $h(D_{dcl}^*(w)) = D^*(w)$?
References


