UNIVERSITAT POLITÈCNICA DE VALÈNCIA

DEPARTAMENTO DE SISTEMAS INFORMÁTICOS Y COMPUTACIÓN

Duplication languages: characterizations and open problems

José M. Sempere
(jsempere@dsic.upv.es)

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

DSIC DEPARTAMENTO DE SISTEMAS INFORMÁTICOS Y COMPUTACIÓN

The genomic duplication is a DNA recombination problem due to unequal crossing over

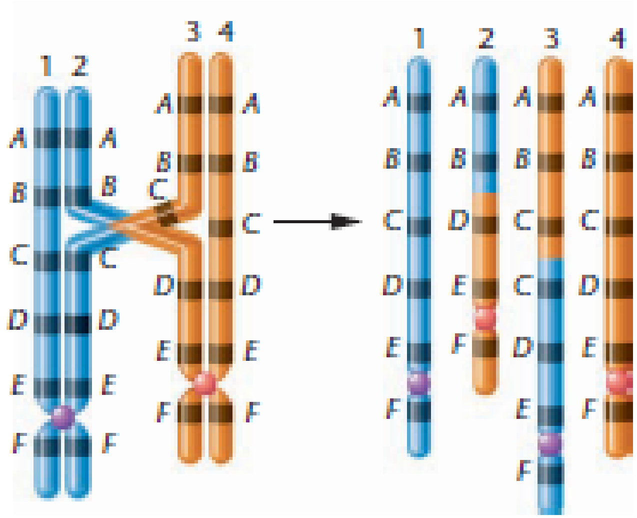


FIGURE 8–17 The origin of duplicated and deficient regions of chromosomes as a result of unequal crossing over. The tetrad at the left is mispaired during synapsis. A single crossover between chromatids 2 and 3 results in deficient (chromosome 2) and duplicated (chromosome 3) chromosomal regions. The two chromosomes uninvolved in the crossover event remain normal in their gene sequence and content.

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

$$\text{abcde} \downarrow \text{abcdbcde}$$

A first view at duplication

abcde

↓

abcdbcde

A first view at duplication

$$\text{abcde} \downarrow \text{abcdbcde}$$

A first view at duplication

abcde

↓

abcdbcde

↓

abcdbccdbcde

A first view at duplication

abcde

↓

abcdbcde

↓

abcdbccdbcde

A first view at duplication

abcde

abcdbcde

abcdbccdbcde

A first view at duplication

abcde

↓

abcdbcde

↓

abcd**bcc**dbcde

↓

abcd**bccbcc**dbcde

A first view at duplication

abcde

↓

abcdbcde

↓

abcdbccdbcde

↓

abcdbccbccdbcde

A first view at duplication

abcde

↓

abcdbcde

↓

abcdbccdbcde

↓

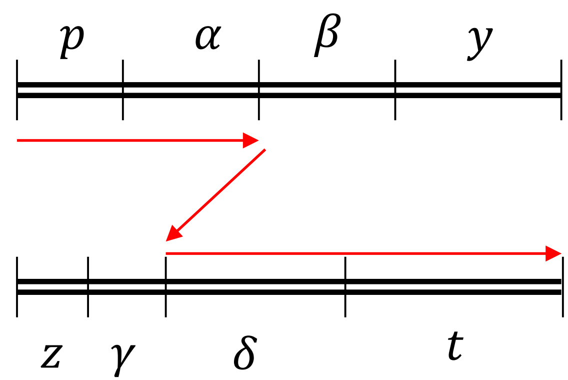
abcdbccbccdbcde

⋮

Duplication Languages D*(w)

Let V be an alphabet

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



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

$$\boxed{D^*(w) = \bigcup_{i \geq 0} D^i(w)}$$

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

$$w = a_1 a_2 \dots a_n$$

$$\begin{aligned} D^*(w) = \{ & w_1 a_1 w_2 a_2 \dots w_{n-1} a_{n-1} w_n a_n \mid \\ & w_1 \in a_1^*, \\ & w_{n+1} \in a_n^*, \\ & w_i \in a_i^* \text{ for } a_{i-1} = a_i, \\ & w_i \in V^* \text{ for } a_{i-1} \neq a_i, 2 \leq i \leq n \} \end{aligned}$$

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.

$w = abc$ 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

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

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

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

$$\boxed{D_X^*(w) = \bigcup_{i \geq 0} D_X^i(w)}$$

$D_{\mathbb{N}}^*(w)$ the unbounded duplication language

$D_{[k]}^*(w)$ 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

$$\begin{bmatrix} \mu \\ v \\ w \end{bmatrix} \begin{array}{l} \rightarrow \text{guess} \\ \rightarrow \text{memory} \\ \rightarrow \text{pattern} \end{array}$$

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

DEPARTAMENTO DE SISTEMAS INFORMÁTICOS Y COMPUTACIÓN

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

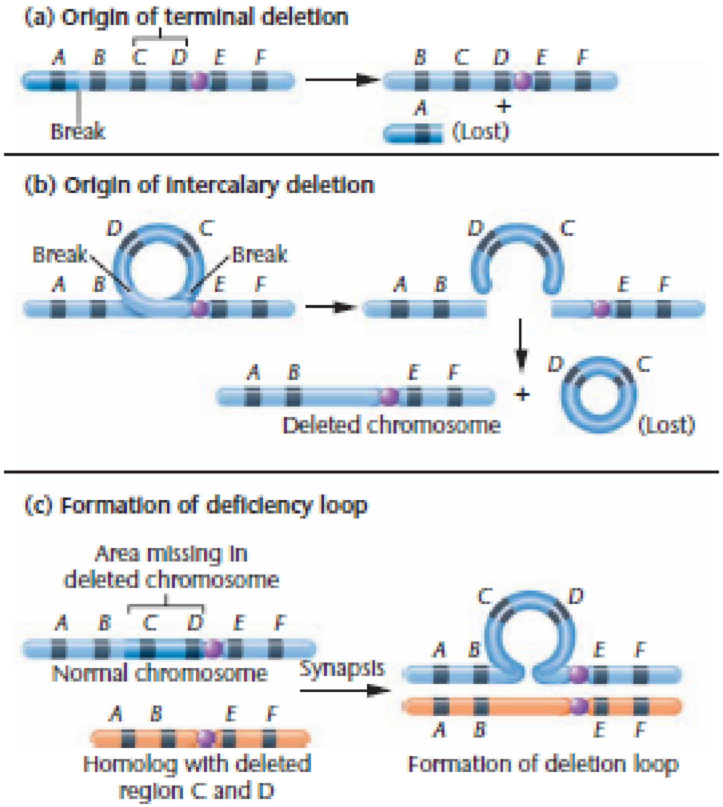
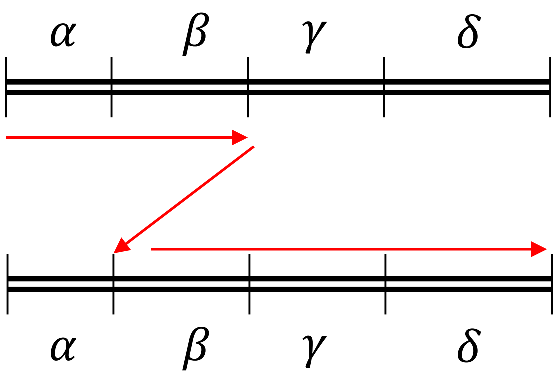


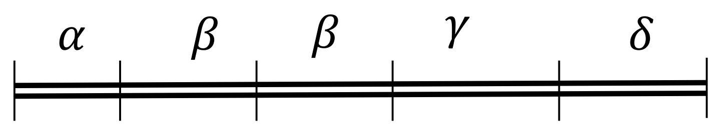
FIGURE 8–14 Origins of (a) a terminal and (b) an intercalary deletion. Part (c) shows that pairing can occur between a normal chromosome and one with an intercalary deletion if the undeleted portion buckles out to form a deletion (or a compensation) loop.

An scheme for compensation (deletion) loops formation

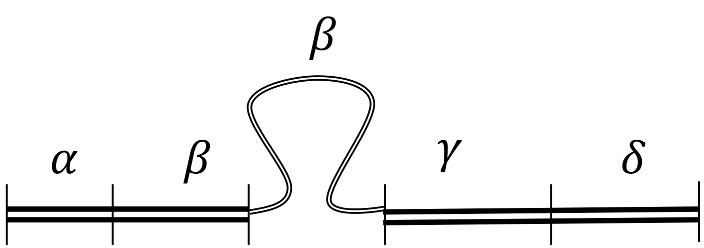
(1)



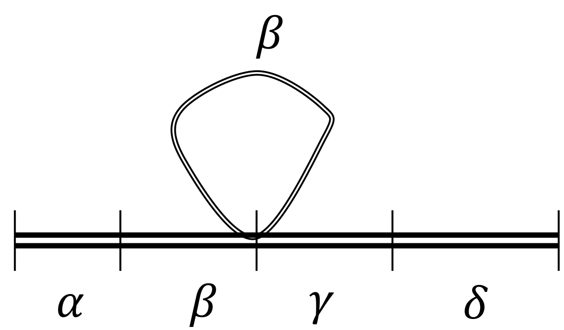
(2)



(3)



(4)



Duplication Languages with compensation loops $D^*_{cl}(w)$

Let V be an alphabet without bracket symbols [and]

$$w = x_1[w_1]x_2[w_2]\dots x_n[w_n]$$
$$z = x_1[z_1]x_2[z_2]\dots x_n[z_n]$$

shuffle with common segments

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

- $D^0_{cl}(w) = \{ux[x]v \mid w = uxv, \;\; u,x,v \in V^*\}$
- $D^i_{cl}(w) = \{scs(x,y) \mid x,y \in D^{i-1}_{cl}(w)\}, \quad i \geq 1\}$

$$\boxed{D^*_{cl}(w) = \bigcup_{i \geq 0} D^i_{cl}(w)}$$

Duplication Languages with compensation loops $D^*_{cl}(w)$

An example

$$w = abc$$

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

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

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

$$h(a[a]b[ab]c) = aababc$$

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 $w = w_1 w_2 \ldots w_n$ with $w_i \in V$

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

$$w_1(w_1)^* w_2(w_2 + w_1 w_2)^* w_3(w_3 + w_2 w_3 + w_1 w_2 w_3)^* \ldots w_n(w_n + \cdots + w_1 w_2 \ldots w_n)^*$$

$$w = abc$$

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

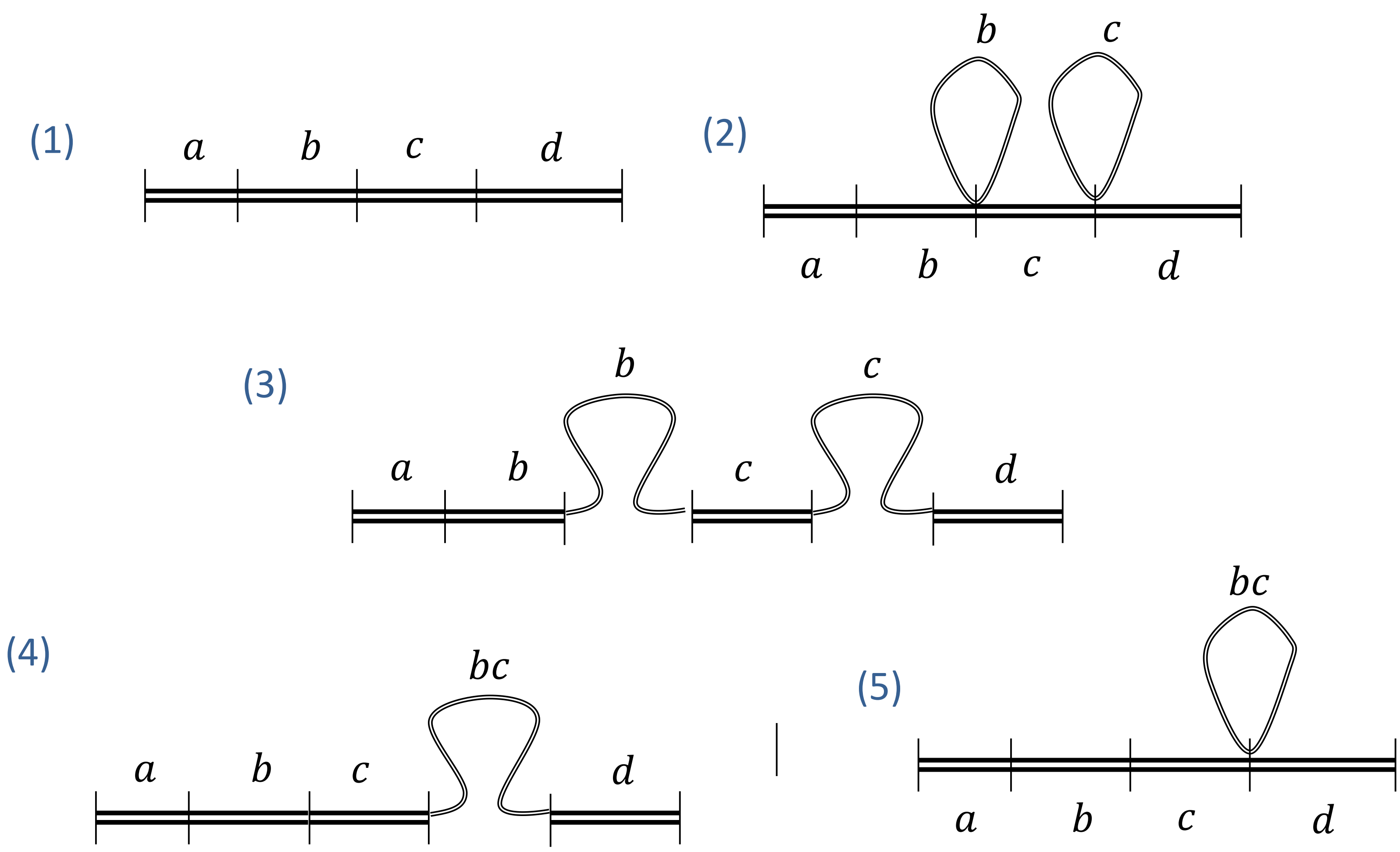
1) is trivial given that if $w = a^n$ then $h\left(D^*_{cl}(a^n)\right) = D^*(a^n) = a^n a^*$

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

$$aba \vDash abaaba \vDash abaabbaaba$$

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



Duplication Languages with dynamic compensation loops $D^*_{dcl}(w)$

Let V be an alphabet without bracket symbols [and]

$$w = [w_0]x_1[w_1]x_2[w_2]\ldots x_n[w_n]$$
$$z = [z_0]x_1[z_1]x_2[z_2]\ldots x_n[z_n]$$

generalized shuffle with common segments

$$gscs(w,z) = \{\,[\eta_0]x_1[\eta_1]x_2[\eta_2]\ldots x_n[\eta_n] \mid \eta_0 x_1 \eta_1 x_2 \eta_2 \ldots x_n \eta_n = w_0 z_0 x_1 w_1 z_1 \ldots x_n w_n z_n\}$$

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

$$\boxed{D^*_{dcl}(w) = \bigcup_{i\,\geq\,0} D^i_{dcl}(w)}$$

Duplication Languages with dynamic compensation loops $D^*_{dcl}(w)$

An example

$w = ab[ab]b[bb]c[bc]$

$\mathrm{z} = a[a]bbc[cc]$

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

$x_1 = a[a]b[ab]b[bb]c[bccc]$

$x_2 = [a]a[bab]bb[bcbcc]c$

$x_3 = a[aba]b[bbbc]b[cc]c$

$h(x_1) = h(x_2) = h(x_3) = aababbbbcbccc$

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 $aa^*b(b+ab)^*$

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

$$\left.\begin{array}{r} a[a]b[ab] \in gscs(a[a]b, ab[ab]) \\ a[aba]b \in gscs(a[a]b, ab[ab]) \end{array}\right\} D^1_{dcl}(ab)$$

$$a[abaaba]b \in gscs(a[aba]b, a[aba]b) \qquad D^2_{dcl}(ab)$$

$$aabaabab \in h\big(D^*_{dcl}(w)\big) - h(D^*_{cl}(w))$$

Duplication Languages with dynamic compensation loops $D^*_{dcl}(w)$

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

$$abc \vDash abcabc \vDash abcacabc$$

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

a symbol a cannot be inserted between the symbols c

$$abcacabc$$

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 $h\left(D^*_{dcl}(w)\right)$?
- For any string w with two different symbols is $h\left(D^*_{dcl}(w)\right) = D^*(w)$?

References

- **J. Dassow, V. Mitrana, On some operations suggested by the genome evolution. Pacific Symposium on Biocomputing'97, eds. R. Altman, K. Dunker, L. Hunter and T. Klein, (Hawaii, 1997), pp. 97-108.**
- **J. Dassow, V. Mitrana, Gh. Paun, On the regularity of duplication closure, Bull. EATCS, 69 (October 1999), pp. 133--136.**
- **J. Dassow, V. Mitrana, A. Salomaa, Operations and language generating devices suggested by the genome evolution, Theoretical Computer Science 270, Issue 1-2 (2002), pp 701-738.**
- **P. Leupold, C. Martín-Vide, V. Mitrana, Uniformly Bounded Duplication Languages. Discrete Appl. Math. 146 (2005), pp 301-310.**
- **P. Leupold, V. Mitrana, Uniformly Bounded Duplication Codes. RAIRO-Theor. Inf. Appl. 41 (2007), pp 411-424.**
- **P. Leupold, V. Mitrana, J.M. Sempere. Formal languages arising from DNA duplication. Aspects of Molecular Computing , eds. N. Jonoska, G. Paun, G. Rozenberg, (LNCS 2950, Springer, 2004), pp. 297-308.**
- **C. Martín-Vide, Gh. Paun, Duplication grammars. Acta Cybernetica, 14, 1(1999), pp. 101-113.**
- **V. Mitrana, G. Rozenberg, Some properties of duplication grammars. Acta Cybernetica, 14, 1(1999), pp. 165-177.**
- **M. Wang, On the irregularity of the duplication closure, Bull. EATCS, 70 (February 2000), pp. 162-163.**