Growth Functions and Length Sets of Replicating Systems

Valeria Mihalache
Arto Salomaa

Turku Centre for Computer Science
TUCS Technical Report No 23
June 1996
ISBN 951-650-774-3
ISSN 1239-1891

Abstract

Growth functions and length sets are studied for classes of replicating systems. The so-called *deterministic* classes of replicating systems, which are systems for which one can define growth functions, are fully characterized. Their growth is either exponential, or linear. For *nondeterministic* classes, where length sets rather than growth functions are considered, we obtain detailed characterizations in many cases, while some details remain open in other cases.

TUCS Research Group
Mathematical Structures of Computer Science

1 Introduction

Replication, introduced in [2], is an operation of generating strings by an insertion subjected to some additional constraints. The reader is referred to [2] for interconnections with several research areas: molecular biology (DNA recombination, a particular type of splicing), linguistics (insertion grammars), language theory, combinatorics on words.

The basic set-up is the following: there are a starting string (called replicating string), say w, over a finite alphabet, and a pair of strings, (u, v) (called insertion context), over the same alphabet. If the string uv appears as a substring of w, then one can insert in-between u and v any substring of w which starts with v and ends with u. A more intuitive representation for this is in the next figure.

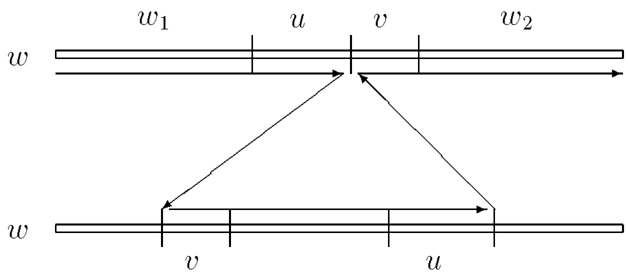

Figure 1

Several variants arise whith respect to the string to be inserted or to the place where insertion is performed. So far only restrictions on the string to be inserted have been taken into account (i. e. insertion is allowed to be performed in-between u and v in any position where uv occurrs as substring of the current string). Moreover, insertion contexts investigated so far have consisted of one pair of letters. In this paper, we restrict ourselves to these same variants.

The subject matter in [2] was mainly the generative power of replication systems, with comparisons to one another or to generative grammars in the regulated rewriting area. In [3], closure properties with respect to sets of replicating strings or sets of insertion contexts were investigated.

The aim of this paper is to find intrinsic properties of the strings obtained by replication. More precisely, following the approach in Lindenmayer systems theory, we study *growth functions* for the strings generated in a chain of replication steps. This can be done in the *deterministic* case, when the lengths of resulting strings are uniquely determined. For the *nondeterministic* case, we study *length sets*.

2 Basic Definitions

As general formal language notation, we use: V^* = the free monoid generated by the alphabet V, λ = the empty string, $V^+ = V^* - \{\lambda\}$, $|x|$ = the length of $x \in V^*$, $|x|_a$ = the number of occurrences of $a \in V$ in $x \in V^*$, $Pref(x)$ $(Sub(x), Suf(x))$ = the set of prefixes (subwords, suffixes) of $x \in V^*$, $alph(L) = \{a \mid |x|_a > 0$ for some $x \in L\}$. If $V = \{a_1, \ldots, a_n\}$ and $x \in V^*$, its Parikh vector is $\Psi_V(x) = (|x|_{a_1}, \ldots, |x|_{a_n})$. The mapping Ψ_V is extended in the natural way to languages. For a vector $v = (a_1, a_2, \ldots, a_n) \in \mathbb{N}^n$, we denote by $|v| = \sum_{i=1}^{n} a_i$, and $v(i) = a_i$, for any $i, 1 \leq i \leq n$. The family of regular and context-sensitive languages are denoted by REG, CS, respectively. Further elements of formal language theory can be found in [5]. For Lindenmayer systems we refer to [4].

Definition 1 *A replicating system is a triple*

$$\sigma = (V, w, (a, b)),$$

where V is an alphabet, $w \in V^+$ is a replication string, and $(a, b) \in V \times V$ is an insertion context.

Definition 2 *With respect to a replicating system as above, for $x, y \in V^*$ we define the direct replication relation as*

$$x \rightsquigarrow y \quad \text{iff} \quad \begin{array}{ll} (1) & x = x_1abx_2, \;\; x_1, x_2 \in V^*, \\ (2) & y = x_1azbx_2, \;\text{ for } z = bz' = z''a, \\ (3) & z \in Sub(x). \end{array}$$

No restriction is imposed about the position of the substring ab in x (condition (1)) or on the way z was selected from $Sub(x)$ (condition (3)). Ten possibilities can be pointed out when considering restrictions on condition (3).

Condition (3) can be replaced by more restrictive ones as follows (in all cases, $z = bz' = z''a$):

1. $z = x$ (*total*; *t*),

2. $z \in Pref(x)$ (*arbitrary prefix; ap*),

3. $z \in Pref(x)$ and z is maximal (if $x = z_1u_1, z_1 = bz_1' = z_1''a$, then $|z| \geq |z_1|$) (*maximal prefix; Mp*),

4. $z \in Pref(x)$ and z is minimal (if $x = z_1u_1, z_1 = bz_1' = z_1''a$, then $|z| \leq |z_1|$) (*minimal prefix; mp*),

5. $z \in Sub(x)$ and z is leftmost (if $x = u_1 z u_2, x = u_1' z_1 u_2', z_1 = bz_1' = z_1'' a$, then $|u_1| \leq |u_1'|$) (*arbitrary leftmost; al*),

6. $z \in Sub(x)$, z is leftmost and maximal ($x = u_1 z u_2$ and if $x = u_1' z_1 u_2', z_1 = bz_1' = z_1'' a$, then $|u_1| \leq |u_1'|$; moreover, if $x = u_1 z u_2 = u_1 z_1 u_2', z_1 = bz_1' = z_1'' a$, then $|z| \geq |z_1|$) (*maximal leftmost; Ml*),

7. $z \in Sub(x)$, z is leftmost and minimal ($x = u_1 z u_2$ and if $x = u_1' z_1 u_2', z_1 = bz_1' = z_1'' a$, then $|u_1| \leq |u_1'|$; moreover, if $x = u_1 z u_2 = u_1 z_1 u_2', z_1 = bz_1' = z_1'' a$, then $|z| \leq |z_1|$) (*minimal leftmost; ml*),

8. $z \in Sub(x)$ and z is maximal (if $x = u_1 z u_2$ and $x = u_1' z_1 u_2', z_1 = bz_1' = z_1'' a$, $|u_1'| \leq |u_1|, |u_2'| \leq |u_2|$, then $z = z_1$) (*arbitrary maximal; aM*),

9. $z \in Sub(x)$ and z is minimal (if $x = u_1 z u_2$ and $x = u_1' z_1 u_2', z_1 = bz_1' = z_1'' a$, $|u_1'| \geq |u_1|, |u_2'| \geq |u_2|$, then $z = z_1$) (*arbitrary minimal; am*).

The case in Definition 2 corresponds to

10. $z \in Sub(x)$ (*any* subword, *free; af*).

For $g \in \{t, ap, Mp, mp, al, Ml, ml, af, aM, am\} = D$, we write $x \rightsquigarrow_g y$ if $x \rightsquigarrow y$ and the restrictions required by the g mode of replication are satisfied. $\rightsquigarrow_g^*$ denotes the reflexive and transitive closure of the relation $\rightsquigarrow_g$.

The *language generated* by the replicating system $\sigma = (V, w, (a, b))$ in the mode $g \in D$ is defined by

$$L_g(\sigma) = \{z \in V^* \mid w \rightsquigarrow_g^* z\}.$$

We denote by $\mathbf{SF}(g)$ the family of languages of the form $L_g(\sigma)$, $g \in D$ (SF stands for "snake family". By a "snake language" we mean in the sequel any language in any SF-family.)

Consider in the following a replicating system $\sigma = (V, w, (a, b))$. A "snake sequence" of σ with respect to the $g \in D$ mode of replication,

$$E(\sigma, g) = \ w_0, w_1, \ldots, w_n, \ldots,$$

determines in a natural way a so-called *growth function*, defined, just as in the case of Lindenmayer systems, by the length of the strings in the sequence. More precisely, the growth function associated to a snake sequence is the function f defined on $\mathbb{N}$ and valued in $\mathbb{N}$, such that $f(n) = |w_n|$, for any $n \geq 0$. If the replication mode considered uniquely determines the string to be inserted to obtain a next string in a snake sequence, at each step, that is, the snake sequence lengths are uniquely determined (we say that such a replication mode is *deterministic*), then we can associate the growth function as considered above to the replicating system and that mode of replication. In the theory of L systems, this corresponds to DOL systems, respectively to DOL sequences. Yet the snake sequence is not uniquely determined, because the insertion of a string can be performed in several positions. However, observe that the special effects on determinism are due to the presence of *only one insertion context*. For instance, in the aM mode, z must always begin with the leftmost occurrence of b and end with the rightmost occurrence of a. No matter between which pair (a, b) such a z is inserted, the leftmost b and rightmost a are uniquely positioned also for the next step. Thus, although the snake sequence may vary according to the positioning of z's, the length sequence remains unique. The situation is quite different in the presence of two insertion contexts. Then, apart from the t mode, z is not in general unique.

When the replication mode is not a deterministic one, then we do not have a growth function as above associated to a replicating system with respect to that replication mode. However, following the approach in the theory of L systems for the nondeterministic case (see, for example, [1]), for a more involved study on the nondeterministic replication modes as well, we associate to any replicating system its *length set* with respect to a given replicating mode.

Definition 3 *For any replicating system $\sigma = (V, w, (a, b))$ and for any replication mode $g \in D$, we define the* length set associated to σ with respect to the g mode of replication *as*

$$LS_g(\sigma) = \{|z| \mid z \in L_g(\sigma)\}.$$

A length set $N \subseteq \mathbb{N}$ is a $\mathbf{SF}(g)$ length set *if there exists a language $L \in \mathbf{SF}(g)$ such that $N = \{|z| \mid z \in L\}$. The* family of $\mathbf{SF}(g)$ length sets *is denoted by $\mathcal{LS}(SF(g))$.*

It was remarked in [2] that a snake language is either singleton or infinite, the case of singleton languages being the trivial one, when the replication operator cannot be applied to the initial replicating string. Therefore, we consider in the following only infinite snake languages.

For the deterministic replication modes, that is, $g \in \{t, Mp, Ml, aM, mp, ml\}$, we have the following characterizations:

Theorem 1 *A sequence $u(n)$ of nonnegative integers is the growth function for a replicating system σ with respect to the t mode of replication if and only if $u(n)$ is a geometric progression with ratio 2 and with the initial element not equal to 1.*

Proof: Let
$$u(n) = \ l, 2l, 2^2l, \ldots, 2^nl, \ldots$$
Consider $w = a^l$, $\sigma = (\{a\}, w, (a,a))$. Because the replication mode is total, at any step in a replication sequence the entire current string is inserted in-between some consecutive occurrences of a, therefore the length of the resulted string being the double of the length of the current one, i. e. $|w_{n+1}| = 2|w_n|$, for any $n \geq 0$. Since $|w_0\ (= w)| = l$, it then follows that the growth function associated to a snake sequence of σ with respect to the total mode of replication is exactly $f(n) = u(n)$.

Conversely, let $\sigma = (V, w, (a,b))$ be a replicating system and denote $l = |w|$. Observe $l \geq 2$ (l can be 2 or 3 only in case $a = b$). With the same arguments as above, the growth function associated to any snake sequence of σ with respect to t is
$$f(n) = l, 2l, 2^2l, \ldots, 2^nl, \ldots,$$
and hence $f(n) = u(n)$. □

Theorem 2 *A sequence $u(n)$ of nonnegative integers is the growth function for a replication system σ with respect to the g mode of replication, where $g \in \{Mp,\ Ml,\ aM\}$, if and only if $u(n)$ is of the form $u(n) = l + 2^nk$, where $k \geq 2, l \geq 0$.*

Proof: First of all, note that the case $l = 0, k \geq 2$, is the one outlined in the preceding theorem. So we have to consider only the situation $l \geq 1, k \geq 2$.

Let $u(n)$ be a sequence of such nonnegative integers, let a, b, c be distinct symbols, $w = bc^{k-2}abc^{l-1}$, and $\sigma = (\{a,b,c\}, w, (a,b))$. In any of the maximal modes of replication (Mp, Ml, aM), the string to be inserted at the first step is $z = bc^{k-2}a, |z| = k$, w resulting in a string $w' = b\alpha abc^{l-1}$. The substring of w' to be used in replication in a maximal mode is now $z' = b\alpha a, |z'| = 2|z|$. Moreover, this property of doubling the length of the string to be inserted in a maximal replication mode is preserved at any replicating step, therefore we get a replicating sequence $w_0 = w, w_1, \ldots, w_n, w_{n+1}, \ldots$ having the property $|w_{n+1}| = 2(|w_n| - l) + l, |w_0| = l + k$. This implies that the growth function associated to it is just $f(n) = u(n)$.

Conversely, let $\sigma = (V, w, (a,b))$ be a replicating system. We select the string to be inserted in w in a maximal mode of replication, that is, $w = zy, z = bz' = z''a, y = by'$. Denote $|z| = k, |y| = l$. With the same observations as above, it follows that the growth function associated to σ with respect to any maximal mode of replication is $f(n) = u(n)$. □

As for the families of length sets, we have the following immediate corollary.

Corollary 1 *i)* $\mathcal{LS}(SF(Mp)) = \mathcal{LS}(SF(Ml)) = \mathcal{LS}(SF(aM))$;

ii) $\mathcal{LS}(SF(t)) \subset \mathcal{LS}(SF(Mp))$, *strict inclusion.*

Theorem 3 *A sequence $u(n)$ of nonnegative integers is the growth function for a replicating system σ with respect to the g mode of replication, where $g \in \{mp, ml\}$, if and only if $u(n)$ is an arithmetical progression $u(n) = l+nk$, with $1 \leq k < l$.*

Proof: Let $u(n) = l + nk$ be an aritmetical progression with the initial term l and ratio $k, 1 \leq k < l$. Consider first the case $k \geq 2$. Let a, b, c, be distinct symbols, let $w = bc^{k-2}abc^{l-k-1}$ and let $\sigma = (\{a, b, c\}, w, (a, b))$ be a replicating system. In both the minimal prefix and the leftmost minimal modes of replication, the string to be inserted at any replicating step is $z = bc^{k-2}a$, with $|z| = k$. Hence in any replicating sequence $w_0 = w, w_1, \ldots, w_n, \ldots$, for any $n \geq 0$, the relation $|w_{n+1}| = |w_n| + k$. Since $|w_0| = l$, it follows that the growth function satisfies $f(n) = u(n)$.

In case $k = 1$, consider a, b distinct symbols, $w = aab^{l-2}$, and the replicating system $\sigma = (\{a, b\}, w, (a, a))$. The conclusion follows with the same arguments as in the preceding situation.

Conversely, let $\sigma = (V, w, (a, b))$ be a replicating system. Let $w = \alpha z \beta$, where z is the leftmost minimal substring of w to be inserted in the g mode of replication, $g \in \{mp, ml\}$ (for $g = mp$, note that α should be the empty word). Let $|w| = l, |z| = k$. Observing that the string to be inserted is z, at each replicating step, for both modes considered here, the growth function for the system, with respect to the g mode of replication, is $f(n) = l + kn$. □

Corollary 2 *i)* $\mathcal{LS}(SF(mp)) = \mathcal{LS}(SF(ml))$;

ii) $\mathcal{LS}(SF(mp))$ *is incomparable with any of the families* $\mathcal{LS}(SF(g))$ *with* $g \in \{t, Mp, aM, Ml\}$.

Note that the Parikh languages associated to the replication modes considered above resemble the shapes of the growth functions, respectively. That is, we have :

Proposition 1 *A set of vectors $H \subseteq \mathbb{N}^p$ is the Parikh set for a language $L \in \mathbf{SF}(g), g \in \{t, Mp, Ml, aM, mp, ml\}, V = alph(L), card(V) = p$, if and only if:*

i) $H = \{2^n v_1 \mid n \geq 0\}$, *where* $v_1 \in \mathbb{N}^p$, $|v_1| \geq 2$, *in case* $g = t$;

ii) $H = \{v_2 + 2^n v_3 \mid n \geq 0\}$, *where* $v_2, v_3 \in \mathbb{N}^p$, $|v_3| \geq 2$, *in case* $g \in \{Mp, Ml, aM\}$;

iii) $H = \{v_1 + nv_3 \mid n \geq 0\}$, *where* $v_1, v_3 \in \mathbb{N}^p$, $1 \leq |v_3| < |v_1|$, *in case* $g \in \{mp, ml\}$.

We mention that for the total or a maximal mode of replication, the shape of the Parikh set associated to a snake language was already pointed out in [2].

As it was remarked in the beginning of this section, for nondeterministic replication modes one cannot speak about growth functions. However, length sets can be studied. Also we can present properties of the associated Parikh sets.

The *arbitrary minimal* mode is fully characterized, with respect to its length and Parikh sets, by the two properties that follow.

Theorem 4 *A set of nonnegative integers* $N \subseteq \mathbb{N}$ *is the length set for a replication system* σ *with respect to the am mode of replication, if and only if*

either there exist nonnegative integers l, r, *and* $k_1, k_2, \ldots k_r \geq 2$, *with* $l \geq \sum_{i=1}^{r} k_i$, *such that* $N = \{l + c_1k_1 + c_2k_2 + \ldots + c_rk_r \mid c_1, \ldots, c_r \in \mathbb{N}\}$,

or there exists $l \in \mathbb{N}, l \geq 2$, *such that* $N = \{n \mid n \geq l\}$.

Proof: If N is a set of nonnegative integers defined as $N = \{n \mid n \geq l\}$, for some $l \geq 2$, then consider the replicating system $\sigma = (\{a\}, a^l, (a, a))$. The string to be inserted in the arbitrary minimal mode is $z = a$, at any replication step, therefore $|w_{n+1}| = |w_n| + 1$, for any $n \geq 0$. Together with $|w_0| = l$, this implies that the growth function associated to σ with respect to the *am* mode of replication (which is then well defined, in a similar manner as for deterministic replication modes) is $f(n) = l + n$. Therefore, $LS_{am}(\sigma) = N$.

Consider now the case $N = \{l + c_1k_1 + c_2k_2 + \ldots + c_rk_r \mid c_1, \ldots, c_r \in \mathbb{N}\}$, where $l, r, k_1, \ldots k_r \in \mathbb{N}$, and $k_1, k_2, \ldots k_r \geq 2$, with $l \geq \sum_{i=1}^{r} k_i$.

Denote $d = l - \sum_{i=1}^{r} k_i$, and consider $z_i = bc^{k_i - 2}a$, for any $i, 1 \leq i \leq r$, and $w = z_1 z_2 \ldots z_r c^d$. Let $\sigma = (\{a, b, c\}, w, (a, b))$. One can observe that at any replication step in the arbitrary minimal mode, the string to be inserted is a $z_i, 1 \leq i \leq r$, therefore $|w_{n+1}| = |w_n| + k_i$, for an $i, 1 \leq i \leq r$. This results in $LS_{am}(\sigma) = \{l + \sum_{j=1, i_j \in \{1, \ldots, r\}}^{n} k_{i_j} \mid n \geq 0\} = \{l + c_1k_1 + c_2k_2 + \ldots + c_rk_r \mid c_1, \ldots, c_r \in \mathbb{N}\} = \mathrm{N}$.

Conversely, one can observe in a similar manner that the length set of a replicating system is of either one of the forms in the statement of the theorem. When the insertion context is (a, a), the arbitrary minimal replication mode works like a deterministic one. $\square$

Note that for any replication sequence with respect to the am mode of replication and with the insertion context (a, b), we have $|w_n|_a = |w_0|_a + n, |w_n|_b = |w_0|_b + n$, for any $n \geq 0$.

We still want to point out that in case of only one string to be inserted in the am mode of replication, the growth function of such a system, with respect to the am mode, can be characterized by an arithmetical progression of nonnegative integers, just as in the case of an mp or ml replication. This immediately implies the following corollary.

Corollary 3 *i)* $\mathcal{LS}(SF(mp)) \subset \mathcal{LS}(SF(am))$, *strict inclusion;*

ii) $\mathcal{LS}(SF(am))$ *is incomparable with any of the families* $\mathcal{LS}(SF(g)), g \in \{t,\ Mp,\ aM,\ Ml\}$.

We know that the language generated by a replicating system in the arbitrary minimal mode is regular ([2]), therefore, we expect its Parikh set to be at least semilinear. But actually we can obtain more than that: we can characterize it by a linear set.

Proposition 2 *A linear set* $H = \{v_0 + c_1v_1 + \ldots + c_rv_r \mid c_i \in \mathbb{N}\}$, *where* $v_i \in \mathbb{N}^p$, *for any* $i, 0 \leq i \leq r$ *and* $p \geq 1$, *is the Parikh set of a replicating system with respect to the arbitrary minimal mode of replication if and only if* $v_0 \geq \sum_{i=1}^{r} v_i$, *and there exist an* $s \in \{1, 2\}$ *and* $j_1, \ldots, j_s, 1 \leq j_1 < \ldots < j_s \leq p$, *such that for any* $i, 1 \leq i \leq r, v_i(j_1) = \ldots = v_i(j_s) = 1$ *and, in addition, if* $s = 1$, *then* $v_0(j_1) \geq 2$.

Proof: The fact that the Parikh set for a replicating system with respect to the arbitrary minimal mode of replication is H alike follows with similar arguments as in the proof of the preceding theorem, by considering $v_i = \psi_V(z_i)$, for $1 \leq i \leq r$, and $v_0 = \psi_V(w)$.

Conversely, let H be as in the statement of the proposition, and consider first $s = 2$. Let $V = \{a_1, \ldots, a_p\}$. Without loss of generality, one can assume that $j_1 = 1, j_2 = 2$. Denote $v_{r+1} = v_0 - \sum_{i=1}^{r} v_i$, and $a_j^{(i,j)} = a_i^{v_i(j)}$, for any $i, j, 1 \leq i \leq r+1, 1 \leq j \leq p$. For any $i, 1 \leq i \leq r$, consider $z_i = a_2^{(i,2)} a_3^{(i,3)} \ldots a_p^{(i,p)} a_1^{(i,1)}$, $z_{r+1} = a_1^{(r+1,1)} a_2^{(r+1,2)} \ldots a_p^{(r+1,p)}$, $w = z_1z_2 \ldots z_rz_{r+1}$, and $\sigma = (V, w, (a_1, a_2))$. Note that when replicating in the arbitrary minimal mode, the strings to be inserted are $z_1, \ldots, z_r$ (each of them contains exactly one occurrence of a_1 and one occurrence of a_2, in the right positions) and only they. By similar observations as in the proof of the above theorem it follows that $\Psi_V(L) = H$.

The case $s = 1, v_0(j_1) \geq 2$ can be treated similarly, considering an insertion context (a_{j_1}, a_{j_1}). $\square$

For a closer study of *arbitrary prefix* and *arbitrary leftmost* modes of replication, we first consider the following lemma:

Lemma 1 *Let $\sigma = (V, w, (a, b))$ be an arbitrary replicating system. Then there exist $l, t, k_1, k_2, \ldots, k_t \in \mathbb{N}$ with the property that for any $w' \in L_g(\sigma)$ ($g \in \{al, ap\}$), there exist $c_1, \ldots, c_t \in \mathbb{N}$, $c_1 \geq c_2 \geq \ldots \geq c_t$, such that $|w'| = l + \sum_{j=1}^{t} c_j k_j$. Moreover, any string z allowable to be inserted in w' according to the replicating mode g has the length $|z| = \sum_{j=1}^{i} c'_j k_j$ for an $i, 1 \leq i \leq t$ and for some $c'_1, \ldots, c'_i \in \mathbb{N}$, $c'_1 \geq c'_2 \geq \ldots \geq c'_i$.*

Proof: One can write $w = \alpha b \alpha_1 a \alpha_2 a \ldots \alpha_t a \alpha'$, where $\alpha \in (V - \{b\})^*, \alpha' \in (V - \{a\})^*$, and for any $i, 1 \leq i \leq t$, $\alpha_i \in (V - \{a\})^*$. Denote $l = |w|$, $k_1 = |\alpha_1| + 2$, and $k_i = |\alpha_i| + 1$, for any $i, 2 \leq i \leq t$.

We prove the statement for these $l, t, k_1, \ldots, k_t$, by induction on the length of the replicating chain $w \rightsquigarrow_g^n w'$.

If $n = 0$, then $w' = w$ and therefore the statement is trivially true, with $c_j = 0$ for any $j, 1 \leq j \leq t$ and with i having any value $1 \leq i \leq t$, and $c'_j = 1$ for any $j, 1 \leq j \leq i$.

Suppose the statement holds true for n and consider the replicating chain $w \rightsquigarrow_g^n w_1 \rightsquigarrow_g w_2$. One can easily observe that $w_1 = \alpha b \beta_1 a \beta_2 a \ldots a \beta_m a \beta_{m+1}$, for an $m \geq t$, where $\beta_{m+1} = \alpha'$. By the induction hypothesis, $|w_1| = l + \sum_{j=1}^{t} c_j k_j$ for some $c_1, \ldots, c_t \in \mathbb{N}$, $c_1 \geq c_2 \geq \ldots \geq c_t$.

Let z be the string which is inserted in w_1 when resulting into w_2. Then z is of the form $z = b\beta_1 a \beta_2 a \ldots \beta_s a$, for an $s, 1 \leq s \leq m$. By the inductive assumption, $|z| = \sum_{j=1}^{i} c'_j k_j$ for an $i, 1 \leq i \leq t$ and for some $c'_1, \ldots, c'_i \in \mathbb{N}$, $c'_1 \geq c'_2 \geq \ldots \geq c'_i$. Therefore, w_2 satisfies $|w_2| = |w_1| + |z| = l + \sum_{j=1}^{t} c_j k_j + \sum_{j=1}^{i} c'_j k_j = l + \sum_{j=1}^{t} c''_j k_j$, where $c''_j = c_j + c'_j$, for any $j, 1 \leq j \leq i$, and $c''_j = c_j$ for any $j, i+1 \leq j \leq t$. Still note that $c''_1 \geq c''_2 \geq \ldots \geq c''_t$.

In order to determine the length of the strings to be inserted in w_2, we need to point out the places where the insertion was performed when replicating w_1 into w_2. One can notice two possible situations:

case a): the prefix α of w_1 is of the form $\alpha = \gamma a$, and the insertion is performed in-between this occurrence of a and the occurrence of b which follows it (note that this case possibly occurrs only when $g = al$). This implies $w_2 = \gamma a b \beta_1 a \beta_2 a \ldots \beta_s a b \beta_1 a \beta_2 a \ldots \beta_m a \beta_{m+1}$.

The strings to be inserted in w_2 are either of the form $z' = b\beta_1 a\beta_2 a \dots \beta_p a$, for a $p, 1 \leq p \leq s$, and then such a z' is a prefix of z and also a string to be inserted in w_1, hence it satisfies the restriction in the assertion (by inductive assumption), or $z' = zz''$, with $z'' = b\beta_1 a\beta_2 a \dots \beta_p a$, for a $p, 1 \leq p \leq m$. In this case, one can observe that $z''$is a string allowed to be inserted in w_1, and therefore $|z''| = \sum_{j=1}^{r} c''_j k_j$ for some $r, 1 \leq r \leq t, c''_1 \geq c''_2 \dots c''_r$. We have $|z'| = |z| + |z''| = \sum_{j=1}^{i} c'_j k_j + \sum_{j=1}^{r} c''_j k_j = \sum_{j=1}^{v} \overline{c}_j k_j$, where $v = max\{i, r\}$, and $\overline{c}_j$ is defined as

$$\overline{c}_j = \begin{cases} c'_j + c''_j, & \text{for } 1 \leq j \leq min\{i, r\}, \\ c'j & \text{for } i+1 \leq j \leq v, \\ c''j & \text{for } r+1 \leq j \leq v \end{cases}$$

Moreover, one can note that $\overline{c}_1 \geq \overline{c}_2 \geq \dots \geq \overline{c}_v$, and hence the assertion follows for this case.

case b): $\beta_{i+1} = b\beta'_{i+1}$, and the insertion is performed in-between the occurrence of a immediately preceding this occurrence of b and this b.

Denote $z'' = b\beta_1 a\beta_2 a \dots \beta_i a$. Then z'' is a string possible to be inserted in w_1, and hence, by the inductive assumption, $|z''| = \sum_{j=1}^{r} c''_j k_j$ for an $r, 1 \leq r \leq t$, and $c''_1 \geq \dots \geq c''_r$.

As for the strings z' to be inserted in w_2, one can note that they are of one of the following forms:

b.1): $z' = b\beta_1 a\beta_2 a \dots \beta_p a$, for a $p, 1 \leq p \leq i$. Then such a z' is also a string allowed to be inserted in w_1, and therefore the assertion follows from the inductive assumption.

b.2): $z' = z''\overline{z}$, with $\overline{z} = b\beta_1 a\beta_2 a \dots \beta_p a$, for a $p, 1 \leq p \leq s$. One can note that $\overline{z}$ is a string allowed to be inserted in w_1, and then the assertion follows similarly as in case a).

b.3): $z' = z''zz'''$, with $z''' = \beta_{i+1} a\beta_{i+2} a \dots \beta_p a$, for a $p, i+1 \leq p \leq m$. One can note that the string $\overline{z} = z''z'''$ is allowed to be inserted in w_1, and still $|w_2| = |z| + |z''| + |z'''| = |z| + |\overline{z}|$, and then the assertion follows similarly as in case *a)*.

Therefore, by the inductive principle, the assertion stated follows. $\square$

Now we can predict a superset of the length sets for the ap and al case. Moreover, we can precisely characterize these sets for replication systems whose starting strings are subjected to some restrictions.

Theorem 5 *For any replicating system $\sigma = (V, w, (a, b))$ and for a replicating mode $g \in \{ap, al\}$, there exist nonnegative integers $l, t, k_1, k_2, \dots, k_t$, such that for the set $N \subseteq \mathbb{N}$ defined as $N = \{l + c_1 k_1 + c_2 k_2 + \dots + c_t k_t \mid$ for any $i, 1 \leq i \leq t, c_i \in \mathbb{N}$, and $c_i \geq c_{i+1}, 1 \leq i < t\}$ we have* *i)* $LS_g(\sigma) \subseteq N$

ii) moreover, if $w = \gamma ab\gamma'$, *with* $|\gamma'|_a = 0$, *then the equality holds, i. e.* $LS_g(\sigma) = N$.

Proof: We consider $w, l, t, k_1, k_2, \ldots, k_t$ as in the proof of the preceding lemma.

Then part *i)* follows directly from this lemma. For part *ii)*, we have to prove only the inclusion $N \subseteq LS_g(\sigma)$. Without loss of generality, we can take into consideration only the *ap* mode of replication (the only difference between the two modes is that in the *al* mode, if, following the notations in the preceding proof, $\alpha = \alpha'' a$, then one can insert in-between this occurrence of a and the b following it; but since we want only to prove $N \subseteq LS_g(\sigma)$, then this inclusion will follow from N being included in the length set generated when we do not insert in this position).

Let N be as in the statement of the theorem. Following the notations in the proof of the lemma, we have $w = b\alpha_1 a\alpha_2 a \ldots a\alpha_t ab\beta$ (where $\alpha' = b\beta$).

We show that $N \subseteq LS_{ap}(\sigma)$.

Take an arbitrary element $n \in N$. Then there exist $c_1, \ldots, c_t \in \mathbb{N}$, with $c_i \geq c_{i+1}$ for any $i, 1 \leq i \leq t-1$. For any $i, 1 \leq i \leq t$, denote $\beta_i = b\alpha_1 a\alpha_2 a \ldots \alpha_i a$ ($|\beta_i| = \sum_{j=1}^{i} k_j$), $m_i = c_i - c_{i+1}$, for $1 \leq i < t$, $m_t = c_t$. We prove in the sequel that there exists $w' \in L_{ap}(\sigma)$ such that $|w'| = n$. More exactly, we show how w' can be constructed from w, by replicating in the *ap* mode.

The string β_t is allowed to be inserted in the *ap* mode in w, as well as in any string obtained from w by inserting β_t after the t-th occurrence of a in w or in a string generated from w in this way. Inserting in this fashion m_t $(= c_t)$ steps, we obtain

$$w \leadsto_{ap}^{m_t} b\alpha_1 a\alpha_2 a \ldots \alpha_t a(b\alpha_1 a\alpha_2 a \ldots \alpha_t a)^{c_t} b.$$

Denote

$$w_1 = b\alpha_1 a\alpha_2 a \ldots \alpha_t a(b\alpha_1 a\alpha_2 a \ldots \alpha_t a)^{c_t} b = b\alpha_1 a\alpha_2 a \ldots \alpha_t ab(\alpha_1 a\alpha_2 a \ldots \alpha_t ab)^{c_t},$$

with $|w_1| = l + c_t \sum_{i=1}^{t} k_i$.

One can note that the string β_{t-1} is allowed to be inserted in w_1, as well as in any string obtained from w_1 by inserting β_{t-1} after the t-th occurrence of a in such a string. Therefore, we obtain

$$w_1 \leadsto_{ap}^{m_{t-1}} b\alpha_1 a\alpha_2 a \ldots \alpha_t a(b\alpha_1 a\alpha_2 a \ldots \alpha_{t-1} a)^{m_{t-1}} b(\alpha_1 a\alpha_2 a \ldots \alpha_t ab)^{c_t}.$$

Denote the resulting string by w_2, and note that it can be rewritten as

$$w_2 = b\alpha_1 a\alpha_2 a \dots \alpha_t ab(\alpha_1 a\alpha_2 a \dots \alpha_{t-1} ab)^{m_{t-1}}(\alpha_1 a\alpha_2 a \dots \alpha_t ab)^{c_t},$$

and $|w_2| = l + (m_{t-1} + c_t)\sum_{i=1}^{t-1} k_i + c_t k_t = l + c_{t-1}\sum_{i=1}^{t-1} k_i + c_t k_t$.

Next we insert the string β_{t-2} in the same way, $c_{t-2} - c_{t-1}$ steps, resulting in a string w_3 with $|w_3| = l + c_{t-2}\sum_{i=1}^{t-2} k_i + c_{t-1}k_{t-1} + c_t k_t$.

Repeating this algorithm, we finally obtain the string

$$w_t = b\alpha_1 a\alpha_2 a \dots \alpha_t ab(\alpha_1 ab)^{m_1}(\alpha_1 a\alpha_2 ab)^{m_2} \dots$$

$$(\alpha_1 a\alpha_2 a \dots \alpha_{t-1} ab)^{m_{t-1}}(\alpha_1 a\alpha_2 a \dots \alpha_t ab)^{m_t},$$

with $|w_t| = l + m_1 k_1 + m_2(k_1 + k_2) + \dots + m_{t-1}(k_1 + k_2 + \dots + k_{t-1}) + m_t(k_1 + k_2 + \dots + k_t) = l + k_1\sum_{i=1}^{t} m_i + k_2\sum_{i=2}^{t} m_i + \dots + k_{t-1}\sum_{i=t-1}^{t} m_i + k_t m_t = l + k_1\sum_{i=1}^{t} m_i + k_2\sum_{i=2}^{t} m_i + \dots + k_{t-1}\sum_{i=t-1}^{t} m_i + k_t m_t = l + k_1 c_1 + k_2 c_2 + \dots + c_t k_t = n$. Thus we have obtained $N \subseteq LS_{ap}(\sigma)$. □

For the case of an arbitrary free replicating mode, we have:

Proposition 3 *Let $\sigma = (V, w, (a, b))$ be a replicating system. Then the Parikh set of the language L generated by σ with respect to the af mode of replication is linear, that is, $\Psi_V(L) = \{v_0 + c_1 v_1 + \dots + c_r v_r \mid c_i \in \mathbb{N}, 1 \leq i \leq r\}$, for an $r \geq 1$, and $v_0, \dots, v_r \in \mathbb{N}^p$, where $p = card(V)$.*

Proof: Let $z_1, \dots, z_r$, be all the substrings of w of the form $z_i = bz_i' = z_i''a, 1 \leq i \leq r$, not containing the substring ab, but z_i' possibly containing occurrences of b, z_i'' possibly containing occurrences of a. Denote $v_i = \Psi_V(z_i)$, for any $i, 1 \leq i \leq r$. A string to be inserted at an arbitrary replication step is either such a z_l, or a concatenation of several z_l's. Therefore, the Parikh set associated to the generated language is of the form given in the statement of the proposition. □

As we have pointed out in the Introduction section, the generative capacity of the replicating systems has been mainly dealt with in the paper where they are first considered. However, using their length sets, we can improve a result there, which states that they are all less powerful than context-sensitive grammars. We can prove now that the are strictly less powerful.

Theorem 6 *$\mathcal{LS}(SF(g)) \subset \mathcal{LS}(CS)$, strict inclusion, for any replication mode $g \in D$.*

Proof: Because any snake language is context-sensitive ([2]), the inclusion holds. In order to show that this inclusion is proper, consider $N = \{2^{2^n} \mid n \in \mathbb{N}\}$. It is well-known that this set it is a context-sensitive length set. We prove in the following that it is not a snake length set.

Assume that $\sigma = (V, w, (a, b))$ is a replicating system such that $LS_g(\sigma) = N$ for a $g \in D$. Since replication is a length-increasing operation, the length of w should be the least element of N, that is, $|w| = 2$. Then the only possibility for σ to generate an infinite language is $b = a$ and $w = aa$. Depending on g, the next string generated is either $w_1 = aaa$, with $|w_1| = 3$, or $w_2 = aaaa$, with $|w_2| = 4$. For the modes g generating w_1 (namely $g \in \{am, ml, mp, ap, al, af\}$, we then obtain $3 \in LS_g(\sigma)$, and hence $LS_g(\sigma) \neq N$. For the modes g generating w_2 only from w (namely $g \in \{t, aM, Mp, Ml\}$), the string w_2 results in at the next replication step is $w_3 = a^8$, with $|w_3| = 8 \notin N$. Therefore, also for this case $LS_g(\sigma) \neq N$. □

Note that for the replication modes for which the shape of the length sets is characterized, the above theorems could be deduced directly from those characterizations.

Theorem 7 $\mathbf{SF}(g) \subset CS$, *strict inclusion, for any* $g \in D$.

Proof: We have from [2] that $\mathbf{SF}(g) \subseteq CS$. Since $\mathcal{LS}(SF(g)) \subset \mathcal{LS}(CS)$, the strictness of the language inclusion holds as well. □

3 Final Remarks

One can note that replicating systems are similar to Lindenmayer systems in the sense that strings obtained after each step of applying the operation are considered as belonging to the generated language. Also, just as Lindenmayer systems, they can be used to model biological phenomena. Therefore, a study of such properties of replicating systems that are well known for Lindenmayer systems is worthwhile, from both language theory and molecular biology points of view. The present paper is a step in this direction, namely it deals with growth functions and length sets of the languages generated under the replication operation.

It has been proved here that growth functions (respectively length sets, Parikh sets) for the replicating systems studied so far are either exponential or linear. Nothing lies in-between. Therefore, it would be of interest to point out models with polynomial nonlinear growth.

Yet notice that the general case for arbitrary leftmost and arbitrary prefix modes of replication, as well as the arbitrary free mode, are not yet sufficiently characterized, as far as the length sets are concerned. We believe that in the first two cases mentioned, the length sets are based on exponential functions, but with some additional constraints, while in the last case it is a linear set, for which the coefficients satisfy some further relations.

References

[1] J. Karhumäki, *On Length Sets of L Systems*, Licentiate thesis, University of Turku, 1974

[2] V. Mihalache, Gh. Păun, G. Rozenberg, A. Salomaa, *Generating Strings by Replication: A Simple Case*, submitted

[3] V. Mihalache, A. Salomaa, *Language-Theoretic Aspects of String Replication*, submitted

[4] G. Rozenberg, A. Salomaa, *The Mathematical Theory of L Systems*, Academic Press, New York, London, 1980

[5] A. Salomaa, *Formal Languages*, Academic Press, New York, London, 1973

Turku Centre for Computer Science
Lemminkäisenkatu 14
FIN-20520 Turku
Finland

http://www.tucs.abo.fi

University of Turku
- **Department of Mathematical Sciences**

Åbo Akademi University
- **Department of Computer Science**
- **Institute for Advanced Management Systems Research**

Turku School of Economics and Business Administration
- **Institute of Information Systems Science**