# Recovering minimal L systems from words and lengths sequences

G. Hunault, Université d'Angers

L. Picouleau, Angers

**Covered topics.** - *Formal languages*
- *Algorithms for biological computing*
- *Lindenmayer grammars, L systems*

**Motivations.** *Working for many years with biologists of the national research institute INRA, we have met on several occasions the need for modeling tools that would build parallel grammars from observations data.*

**Main results and their significance.** *First, it is decidable to build a minimal parallel grammar from a sequence of words (with a constructive proof). So biologists have a tool that automates the making of such a grammar. Second, it is possible to build a minimal parallel grammar from a sequence of numbers (interpreted as words lengths); the proof is not constructive but for some polynomial sequences (which are very common for biologists) we have found partial formulas, here again easily programmable, to obtain one (if not the only one) minimal parallel grammar corresponding to the data.*

**Abstract.** *Formal languages are sets of words. Classical generative devices for such languages are grammars, either sequential or parallel. We deal here with special parallel propagating deterministic grammars called* Lindenmayer grammars *or "*L systems*" for short and we will show how to build such grammars with a minimal alphabet given the words or their length at each rewriting step. Section 1 is devoted to notations and definitions for grammars whereas section 2 does the same for "unilinear" recurent sequences of positive integers. Section 3 gives our main theorems and section 4 shows an application of the theory therefore giving to biologists a way to model and extend biological data.*

# 1. Languages and Grammars

Let $E$ be an alphabet, that is, a set whose elements are called symbols. A word $m$ on $E$ is a finite sequence of symbols of $E$. The empty sequence, also known as the empty word will be denoted by $\omega$. The set of all words on $E$ is denoted by $W(E)$. The concatenation of two sequences (one sequence following the other) endows $W(E)$ with the structure of a non abelian semigroup, whose neutral element is $\omega$. A language on $E$ is a subset of $W(E)$. It is usual to note $a^n$ the (con)catenation of $n$ consecutive symbols $a$.

On way to build such languages, possibly infinite, is to use rewriting systems. A rewriting system $S$ on $E$ is a triple $(E, A, \mathcal{R})$ where $A$ is a word on an alphabet $E$ called the axiom and $\mathcal{R}$ a finite set of couples $(a_i, b_i)$ of words on $E$ called rewriting rules and usually written $a_i \ \rightarrow b_i$. Applying the same rule for all occurences of a word gives parallel rewriting. Applying possibly different rules for the occurences of the same word gives sequential rewriting. For instance, parallel rewriting of $a\,a$ with the two rules $a \rightarrow b$ and $a \rightarrow c$ leads only to $b\,b$ or $c\,c$ whereas sequential rewriting with the same rules leads also to $b\,c$ and $c\,b$. If $a \rightarrow x_1$, $x_1 \rightarrow x_2$,... $x_n \rightarrow b$ is a sequence of rewritings then $b$ is said to be derived from $a$, denoted by $a \Rightarrow b$. The language generated by the rewriting system $S = (A, \mathcal{R})$ is the set of all words that can be derived from the axiom :

$$\mathcal{L}(S) = \{\ m \in W(E)\ ;\ A \Rightarrow m\ \}$$

Classical formal languages distinguish between "good" words build upon terminal symbols and "bad" or temporary words build upon terminal and non terminal symbols. Thus a generative phrase structure grammar is a quadruple $G = (\mathcal{T}, \mathcal{N}, B, \mathcal{R})$ where $\mathcal{T}$ is the set of terminal symbols, $\mathcal{N}$ the set of non terminal symbols, $B$ is a (non terminal beginning) symbol, $\mathcal{R}$ a set of sequential rules. The language generated by $G$ is defined by the set of words built upon terminal symbols, derived from the beginning symbol : $\mathcal{L}(G) = \{m \in W(\mathcal{T})\ ;\ B \Rightarrow m\}$.

A propagating deterministic Lindenmayer system with no interaction (**PD0L**) or **L system** for short in this article is a triple $\mathcal{G} = (\mathcal{S}, A, \mathcal{R})$ where $\mathcal{S}$ is a set of symbols called alphabet, $A$ is a word called the axiom, $\mathcal{R}$ a set of parallel rules such that

- no rule has $\omega$ as right hand side,
- each left hand side of the rule consists of exactly one symbol,
- there is one and only one rule for each symbol.

In other words, an L system is the parallel equivalent of a deterministic context free non erasing traditional grammar. It has to be noted that a PD0L with an alphabet of $n$ symbols $s_i$ is defined by exactly $n$ rules $(r_i)$ : $s_i \ \rightarrow t_i$ where each word $t_i$ has length at least 1. From now on, we will suppose that the alphabet is ordered.

Being deterministic, not erasing and without context, an L system produces only one word at each rewriting step.

For such grammars, the rewriting rules define a *morphism* $f$ on $W(\mathcal{S})$: let $f(t_1t_2...t_n) = f(t_1)f(t_2)...f(t_n)$ where the $t_i$ are elements of the alphabet; $f^n(A)$ will denote the $n$-th rewriting of the axiom with the usual convention $f^0(A) = A$. The **words sequence** of the L system is defined by the function $n \to f^n(A)$ and its **lengths sequence** is then defined by the function $n \to |f^n(A)|$ where $|m|$ is the length of the word $m$, that is, its number of symbols. In a similar way, for a sequence $(w_n)$ of words, the associated lengths sequences is $(|w_n|)$.

Let $s_1$, $s_2...s_p$ be the ordered elements of the alphabet. The *canonical form* of a word $m$ is $s_1^{n_1(m)}s_2^{n_2(m)}...s_p^{n_p(m)}$ where $n_i(m)$ is the number of occurences of symbol $s_i$ in $m$. One may see this canonical form as a kind of factorization. The **canonical words sequence** is the sequence of the words in canonical form. The vector $(n_1(m), n_2(m), ...n_p(m))$ is called the *counting vector* of the word $m$, usually denoted by $C_v(m)$.

For instance the L system over the alphabet $\{a, b\}$ with axiom $a$ and the two rules $a \to b$, $b \to ba$ produces successively the words $b$, $ba$, $bab$, $babba$, $babbabab$... So the canonical words sequence is $b$, $ab$, $ab^2$, $a^2b^3$, $a^3b^5$ and the lengths sequence (including the axiom) gives the *Fibonacci* numbers 1,1,2,3,5,8... which verify the linear relation $F_n = F_{n-1} + F_{n-2}$ with $F(0) = F(1) = 1$. The induced morphism is here defined by $f(a) = b$, $f(b) = ba$.

For further details on formal languages and L systems, see [SaLo73,RoSa80].

## 2. Sequences and Grammars

For the rest of this article, every sequence of numbers will be a sequence of strictly positive numbers which is not decreasing except explicity mentioned.

**Defininition 1.** A sequence of numbers $(u_n)_{n\in N}$ is called a *Lindenmayer growth sequence* (or LG) if there exist an integer $T$, a square matrix $M$ of size $T$ with integer positive coefficients and a vector $L$ of $T$ non negative integers such that

$$\forall n,\ u_n \ = \ L\ .\ M^n\ .\ R$$

where $R$ is the column-vector $(1, 1...1)$ of length $T$. The couple (L,M) is called a *Parikh* representation of $(u_n)_{n\in N}$. The size $T$ is called the *Parikh* dimension of the sequence and will be noted $dim_P(\ (u_n)\ )$.

It is a restriction of the classical of the definition of a $\mathbb{N}$-rational function used in Formal Power Series Theory. See for instance [SaSo78,RoSa80].

For instance, the sequence defined by $L = (1, 0, 0)$ and $M = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 3 \end{pmatrix}$ is a LG sequence of *Parikh* dimension 3 whose values correspond to $n \mapsto 3^n$.

**Defininition 2.** A sequence $(u_n)_{n\in N}$ is called an *unilinear recurent sequence* (or UR) of order $R$ if there exists a minimal positive integer $R$ such that

$$\exists a_1, a_2 ... a_R \in \mathbb{Z}\ ;\ a_1 \neq 0 \ \text{ and } \forall n,\ n\ >\ R\ \Rightarrow\ u_n\ =\ \sum_{k=1}^{R}\ a_k\ .\ u_{n-k}$$

Here again, it is a restriction of the definition of a classical linear recurent sequence with constant coefficients where we impose

- minimality (for we will want minimal alphabets),
- positivity (in order to deal with word lengths),
- integer coefficients (for they lead to counts of letters),
- divisibility (since $u_n$ is the length number $n$).

The integer $R$ is called the recurence dimension of the sequence and will be noted $dim_R(\ (u_n)\ )$.

The condition $a_1 \neq 0$ is there to ensure that we use the same relation for all $u_i$ with a coherent beginning (look at $z_n$ underneath). So the sequence $u_n = 3u_{n-1} + 2u_{n-2} - u_{n-3}$ may be UR depending on $u_0$, $u_1$ and $u_2$ but the sequences $v_n$, $w_n$, $z_n$ defined by

| | | | |
|---|---|---|---|
| $v_0$=1, | $v_1$=2, | $v_2$=3 | $v_n = -3v_{n-1} + 2v_{n-2} + v_{n-3}$ |
| $w_0$=1, | $w_1$=2, | $w_2$=3 | $w_n = (w_{n-1} + w_{n-2} + w_{n-3})/3$ |
| $z_0$=1, | $z_1$=2 | | $z_n = z_{n-2}$ |

can not be UR sequences since $v_n$ leads to negative values, the $w_n$ values are not integers, and for $z_n$, $R$=2 but $a_1 = 0$. For similar reasons, the sequences

```
2, 3, 4, 8, 16, 32, 64, 128,...
1, 5, 7, 9, 11, 33, 99, 990, 9900, 99000, 990000...
```

are (ultimately) linear recurent sequences that are not UR sequences because of the required minimality of $R$.

A UR sequence $(u_n)_{n\in N}$ of order $R$ will be noted $[\ V\ ;\ A\ ]$ where $V$ is the vector $(u_1, ..., u_R)$ of the first $R$ terms of $(u_n)_{n\in N}$ and $A$ is the vector $(a_1, .., a_R)$ of the $R$ coefficients in the recurence relation. For instance the UR sequence defined by $u_0$=1, $u_1$=2, $u_n = 3u_{n-1} + 4u_{n-2}$ is noted $[\ (1,2)\ ;\ (3,4)\ ]$.

**Defininition 3.** The *Hankel matrix* with size $P$ of a sequence $(u_n)_{n\in N}$ is the square matrix of size $P$ whose element at row $i$, column $j$ is $u_{i+j-2}$.

**Defininition 4.** The *Hankel dimension* of a UR sequence $(u_n)$ is the smallest integer $D$ such that for $k \geq D$ the *Hankel matrix* with size $k$ of $(u_n)_{n \in N}$ have constant rank equal to $D$. This dimension will be noted $dim_H(\ (u_n)\ )$.

As an example, let $u_n$ be the sequence defined by $u_n = 5n^2 + 3n + 2$ ; then the first values of $u_n$ are $2, 10, 28, 56, 94, 142, 200....$ So the Hankel matrices of $u_n$ with size 1, 2, 3, 4 are

$$(\ 2\ ) \quad \begin{pmatrix} 2 & 10 \\ 10 & 28 \end{pmatrix} \quad \begin{pmatrix} 2 & 10 & 28 \\ 10 & 28 & 56 \\ 28 & 56 & 94 \end{pmatrix} \quad \begin{pmatrix} 2 & 10 & 28 & 56 \\ 10 & 28 & 56 & 94 \\ 28 & 56 & 94 & 142 \\ 56 & 94 & 142 & 200 \end{pmatrix}$$

whose determinant are respectively 2, -44, -1000, 0 and whose rank are respectively 1,2,3,3. Since $u_n$ is a polynom in $n$ of degree 2, it satisfies the unilinear recurence relation $u_n = 3u_{n-1} - 3u_{n-2} + u_{n-3}$. Then for $k > 3$ the *Hankel* matrix with size $k$ of $u_n$ have zero determinant and rank 3. Hence $dim_H(\ (u_n)\ ) = 3$.

**Remark** : Please note that the *Hankel* dimension is not defined as the size of the first *Hankel* matrix whose determinant is 0 for we want a "stable" result.

To understand our choice, please consider the following example : let $(u_n)$ be the sequence defined by the function

$$n \ \rightarrow \ \frac{1}{120}n^5 + \frac{1}{8}n^3 + \frac{13}{15}n + 1$$

The first values of $u_n$ are $1, 2, 4, 9, 21, 47, 98, 190, 345$ which corresponds to the *Parikh* representation

$$\text{L} = (1,0,0,0,0,0) \qquad \text{M} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

The first 8 *Hankel* determinants are `1,0,-1,4,-3,-1,0,0` misleading to a possible dimension of 2 whereas the first *Hankel* ranks are `1,1,3,4,5,6,6,6` giving the correct value $dim_H(\ (u_n)\ ) = 6$.

Using two consecutive determinants whose value is zero is also not a correct definition. Consider the polynomial whose value at $n$ is

$$\frac{1}{40320}n^8 - \frac{1}{3360}n^7 + \frac{7}{2880}n^6 - \frac{1}{240}n^5 + \frac{167}{5760}n^4 + \frac{53}{480}n^3 + \frac{4723}{10080}n^2 + \frac{331}{840}n + 1$$

Its first values are 1, 2, 5, 12, 27, 58, 121, 248, 502, 1003... The twelve first *Hankel* determinants are 1, 1, -2, 0, 0, -2, -3, -1, 1, 0, 0, 0.. and the first *Hankel* ranks are 1, 2, 3, 3, 4, 6, 7, 8, 9, 9, 9... corresponding to the correct value $dim_H(\ (u_n)\ ) = 9$ since we used a matrix of size 9 to build this example.

## 3. Main theorems

**Theorem 1.** The lengths sequence of an L system is an LG sequence whose *Parikh* dimension is the size of the alphabet.

*Proof.* Let $\mathcal{L}$ be an L system, $T$ the size of its alphabet $s_1$,...$s_T$, $A$ its axiom and $R$ the vector $(1, 1...1)$ of length $T$. Let $M_f$ be the matrix of the induced morphism, that is, the matrix such that its $i$-th line is the counting vector for the rewriting of the $i$-th symbol : $M_f(i,j) = \alpha_j$ if $f(s_i) = \prod s_j^{\alpha_j}$. It is easy to check that $C_v(f(s_i)) = C_v(s_i).M_f$. Then, since $f$ is a morphism, $C_v(f(w)) = C_v(w).M_f$ and by induction, with $C_v(A) = C_v(A).M_f^0$ one can conclude that $C_v(f^n(A)) = C_v(A).M_f^n$. Now, $|w| = |C_v(w)| = C_v(w).R$ so $u_n = |f^n(A)| = C_v(A).M_f^n.R$. $\diamondsuit$

**Corollary.** Every LG sequence $(u_n)$ induces an L system $\mathcal{L}$ such that the lengths sequence of $\mathcal{L}$ is $(u_n)$.

*Proof.* If $u_n = L\,.\,M^n\,.\,R$ then define the axiom $A$ as the canonical word whose counting vector is $L$ and take for the rewriting of $i$-th symbol $s_i$ the canonical word whose counting vector is the $i$-th line of $M$. $\diamondsuit$

So from now on, we will call *Parikh representation* of an L system the couple $(L, M)$ where $L$ and $M$ are defined as in the corollary. It is immediate that two L systems with the same alphabet and whose rules differ only by the order of the symbols share the same *Parikh* representation.

**Theorem 2.** Let $(w_n)$ be a finite sequence of $t+1$ words whose alphabet has $t$ symbols and whose lengths sequence is not decreasing. It is decidable whether there exists at least one L system whose words sequence is $(w_n)$. Moreover, if there exists only one such L system, it can be effectively and easily constructed.

*Proof.* Let $V_i$ be the counting vector of $w_i$, $M_1$ the square matrix of size $t$ whose lines are $V_1, V_3...V_t$, and let $M_2$ be the square matrix of size $t$ whose lines are $V_2, V_3...V_{t+1}$. Since $\beta = f(\alpha)$ implies $C_v(\beta) = C_v(\alpha)M_f$, if there is an L system whose words sequence is $(w_n)$ then the matrix $M$ of its morphism satisfies the $t^2$ equations $V_{i+1} = V_i M$ so $M$ can be computed by $M_1^{-1}M_2$. So our problem is equivalent

a) to decide if a linear integer numeric matrix system has at least one non negative integer solution $M$

b) to find, whenever such a solution exists, an L system whose *Parikh* representation is $(w_0, M)$ that gives exactly the words $t_i$.

Condition a) is a simple linear algebra problem and condition b) can be done by trying all the possibilities on the symbols of the alphabet compatible with the words. This process is finite since each rewriting rule has finite length.$\diamondsuit$

*Remark :* Every matrix solution may not be an acceptable solution if the symbols for the words generated by the L system are not at the same position as in the words $w_i$. For instance, consider the following sequence of 5 words :

$$dad, dbad, dcdbad, ddbdcdbad, ddcdddbdcdbad, dddbdddcdddbdcdbad$$

whose alphabet has $t$=4 symbols. The first word $w_0$ has the counting vector $C_v(w_0)$=(1,0,0,2). The second counting vector is $(1, 1, 0, 2)$ and the five first words lead to the four equations $C_v(w_{i+1}) = C_v(w_i)M$ whose unique solution is the matrix

$$M = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Thus $a$ rewrites to a word whose canonical form is $a\ b$ (this may be either $a\ b$ or $b\ a$), $b$ may be rewritten either as $c\ d$ or as $d\ c$, $c$ may be rewritten either as $b\ d$ or as $d\ b$ and $d$ rewrites to $d$. Now, the first occurence of symbol $b$ has position 2 in word $w_0$, just after $d$. Since rule 4 has length 1 and rule 2 has length 2, $b$ rewrites as symbols 2 and 3 of word $w_1$, that is, $b \rightarrow c\,d$. Similarly, $a$ rewrites as symbols 4 and 5 of word $w_1$ which are $b\ a$ and finally, $c$ rewrites to $d\ b$. It is easy to check that with $w_0$ as axiom, one gets the same other words $t_i$.

Now, if $w_0$ had been *add* instead of *dad*, the equations would have been the same but $M$ would not be an acceptable solution.

**Theorem 3.** If $(u_n)_{n\in N}$ is a LG sequence then $(u_n)_{n\in N}$ is a UR sequence and

$$dim_R(\ (u_n)\ ) \leq dim_P(\ (u_n)\ )$$

*Proof.* See Appendix.

We can not say better than $dim_R(\ (u_n)\ ) \leq dim_P(\ (u_n)\ )$ since for our example following definition 1, $dim_R(\ (u_n)\ ) = 1$, $dim_P(\ (u_n)\ )$=3 and for the *Fibonacci* sequence $dim_R(\ (u_n)\ ) = 2$, $dim_P(\ (u_n)\ )$=2.

**Theorem 4.** If $(u_n)$ is a UR sequence then $dim_H(\ (u_n)\ ) = dim_R(\ (u_n)\ ) + 1$.

*Proof.* See Appendix.

**Theorem 5.** If $(u_n)_{n\in N}$ is a LG sequence with *Parikh* representation (L,M) then

$$dim_H(\ (u_n)\ ) \leq dim_P(\ (u_n)\ )$$

*Proof.* See Appendix.

Remark : It would be tempting (since it is the case for a lot of examples) to think that $dim_P(\ (u_n)\ ) = dim_H(\ (u_n)\ ) + 1$ but unfortunately, here is a counter-example: $u_n = n^3/6 - n^2/2 + 4n/3 + 1$ whose first values are 1,2,3,5,9,16,27,43. $dim_H(\ (u_n)\ ) = 4$ since $u_n$ is a polynomial of degree 3 but there is no L system with an alphabet of 4 symbols whose length sequence is $(u_n)$.

The proof is easy but lengthy: since the first length is 1, the axiom is reduced to one symbol, say $a$. Since the second length is 2, $a$ rewrites to $a\,a$, $a\,b$, $b\,b$ or $b\,c$. $a \to a\,a$ is not possible for it would give a third length of 4 and the correct length is 3. Let's try the second solution : $a \to a\,b$ ; the third legnth is 4 and since we have already $ab$, $b$ rewrites to only symbol. It can't be $a$ so try $b$, $c$ etc. Using a computer program to be sure that no case is forgotten, it is possible to conclude that for this example $dim_P(\ (u_n)\ ) > dim_H(\ (u_n)\ ) + 1$.

**Theorem 6.** For every LG sequence there exists at least one *Parikh* representation with minimal size.

*Proof.* Let $u_n$ be a LG sequence and consider the set $S$ of all L systems whose lengths sequence is $u_n$. There is at least one element in $S$, namely the canonical representation $[L; M]$ induced by $(L, M)$. $S$ is a finite set since the the equations $u_n = |f^n(A)|$ are integer relations on positive unknonws of fixed sums. So the set $\{dim_P(L); L \in S\}$ has a smaller element. $\diamondsuit$

*Remark:* The theorem does not reveal how to build the canonical representation $[L; M]$. Neither does it give the exact dimension of this representation. The reason of it is simple: to deal with only lengths sequences of words is much harder than to work with words sequences and can lead only to rules in canonical form. However, we have found some partial solutions mainly in the polynomial case and when the inequality of theorem 5 reduces to an equality. For these cases, the solution with smallest dimension can be exhibited from a sometimes tedious computation, especially for polynomials but to our knowledge, a general algorithm to get it is still to be found. To get a glimpse of the difficulty of the problem, we leave it to the reader to prove the following assumptions as exercises.

EXERCISE 1.

The LG lengths sequence $n \to (n+1)^t$ is given, in the smallest dimension $t+1$, by the left vector $L$ such that $L_1 = 1$, $L_i = 0$ for $i > 1$ and the matrix $M$ such that $M(1,1) = 1$, if $i > j$ then $M(i,j) = 0$, else if $i = 1$ then $M(i,j) = C_t^{j-2}$ and finally else $M(i,j) = C_{t+1-i}^{j-i}$ where the value $C_n^p$ is the classical binomial coefficient $n!/p!(n-p)!$.

*Lemma*

If $s_k(n) = \sum_{j=1}^{n} j^k$ then $(n+1)^p = 1 + \sum_{k=0}^{p-1} C_p^k \, s_k(n)$ for $k \geq 0$ and $n \geq 1$.

*Proof*

Develop $(n+1)^p$ as $(a+b)^p$ and take the first term $n^p$ from the sum. So

$$(n+1)^p = n^p + \sum_{k=1}^{p} C_p^k \, n^{p-k} = n^p + \sum_{k=0}^{p-1} C_p^k \, n^k$$

Do the same for $n^p$, $(n-1)^p$,...$2^p$ and add term to term. One gets

$$(n+1)^p = 1 + \sum_{j=1}^{n} C_p^k \left( \sum_{k=0}^{p-1} j^k \right)$$

Now, permute the two sigmas : $(n+1)^p = 1 + \sum_{k=0}^{p-1} C_p^k \left( \sum_{j=1}^{n} j^k \right)$. which is what we wanted, using the definition of $s_k(n)$. $\diamondsuit$

*Solution to Exercise 1*

From the definition of the matrix, $M$ is triangular, with only 0 under the diagonal. $\diamondsuit$

EXERCISE 2.

The set of LG sequences $n \to p(n)$ such that the smallest representation of $p(n)$ is found by the following method includes all polynomials $p(n) = \sum_{i=0}^{d} a_i n^i$ of degree $d$ whose coefficients are either $i$) all positive or $ii$) all positive but $a_{d-1}$.

The method is: take $L$ as the vector $(1, 0, 0...a_0 - 1)$ of length $d+1$ or by $a_0$ if $d$=0. Define $M$ as a square matrix of size $d+1$ by $M(i,i)$=1, $M(i,i+1)$=1, $M(i,d+1)$=-1+ $\sum_{j=i}^{d} L(i,j)a_j$ and by 0 anywhere else, where the $L$ function is defined by $L(i,j) = \sum_{k=1}^{i} (-1)^{i+k} C_i^k k^j$ for $j > 0$ and $L(i,0) = 0$.

The following theorem and its corollaries are easy to prove by a direct calculus.

**Theorem 7.** Let $\mathcal{S}$(T) be the set of the LG sequences whose Parikh dimension is less than T, $\mathcal{U}$(T) be the set of the UR sequences whose Hankel dimension is less than T, let $\mathcal{S}$ be the union of all $\mathcal{S}$(T) and $\mathcal{U}$ the union of all $\mathcal{U}$(T). Morever, if $(u_n)_{n\in N}$=$[G_1, M_1]$ and $(v_n)_{n\in N}$=$[G_2, M_2]$ are in $\mathcal{S}$, then

$$\begin{array}{lllll} a) & (u_n + v_n)_{n\in N} & = & [\ conc(G_1, G_2)\ ,\ blad(M_1, M_2)\ ], \\ b) & (a.u_n)_{n\in N} & = & [\ a.G_1\ ,\ M_1\ ] \text{ for } a \in N, \\ c) & (u_n.v_n)_{n\in N} & = & [\ (G_1 \otimes G_2)\ ,\ M_1 \otimes M_2\ ] \end{array}$$

where *conc* stands for vector concatenation, *blad* is the block addition of matrices and $\otimes$ is the usual tensor product.

**Corollary 1.** $\mathcal{S}$ and $\mathcal{U}$ are stable for the addition, the multiplication by a positive constant and for Hadamard's multiplication (term to term multiplication).

**Corollary 2.** If $dimH(\ (u_n)\ ) = d_u$ and $dimH(\ (v_n)\ ) = d_v$ then

$$\begin{array}{lllll} a) & dimH(\ (u_n) + (v_n)\ ) & \geq & d_u + d_v, \\ b) & dimH(\ k.(u_n)\ ) & = & d_u, \\ c) & dimH(\ (u_n).(v_n)\ ) & \geq & d_u.d_v. \end{array}$$

# 4. A biological detailed example

Consider the following words sequence, which comes, slightly modified from LindenMayer's Mathematical models (see [Lind68]) for the red alga *Callithamnion roseum*.

```
word 1  :  1 1 3 1 4
word 2  :  1 1 1 4 1 5 4
word 3  :  1 1 1 5 4 1 6 5 4
word 4  :  1 1 1 6 5 4 1 7 6 5 4
word 5  :  1 1 1 7 6 5 4 1 8 x 2 y 7 6 5 4
word 6  :  1 1 1 8 x 2 y 7 6 5 4 1 8 x 1 3 y 8 x 2 y 7 6 5 4
```

```
word 7  :  1 1 1 8 x 1 3 y 8 x 2 y 7 6 5 4 1 8 x 1 1 4 y 8 x 1 3 y 8 x 2 y 7 6 5 4
word 8  :  1 1 1 8 x 1 1 4 y 8 x 1 3 y 8 x 2 y 7 6 5 4 1 8 x 1 1 5 4 y 8 x 1 1 4 y
           8 x 1 3 y 8 x 2 y 7 6 5 4
word 9  :  1 1 1 8 x 1 1 5 4 y 8 x 1 1 4 y 8 x 1 3 y 8 x 2 y 7 6 5 4 1 8 x 1 1 6 5
           4 y 8 x 1 1 5 4 y 8 x 1 1 4 y 8 x 1 3 y 8 x 2 y 7 6 5 4
word 10 :  1 1 1 8 x 1 1 6 5 4 y 8 x 1 1 5 4 y 8 x 1 1 4 y 8 x 1 3 y 8 x 2 y 7 6 5
           4 1 8 x 1 1 7 6 5 4 y 8 x 1 1 6 5 4 y 8 x 1 1 5 4 y 8 x 1 1 4 y 8 x 1 3
           y 8 x 2 y 7 6 5 4
word 11 :  1 1 1 8 x 1 1 7 6 5 4 y 8 x 1 1 6 5 4 y 8 x 1 1 5 4 y 8 x 1 1 4 y 8 x 1
           3 y 8 x 2 y 7 6 5 4 1 8 x 1 1 8 x 2 y 7 6 5 4 y 8 x 1 1 7 6 5 4 y 8 x 1
           1 6 5 4 y 8 x 1 1 5 4 y 8 x 1 1 4 y 8 x 1 3 y 8 x 2 y 7 6 5 4
word 12 :  1 1 1 8 x 1 1 8 x 2 y 7 6 5 4 y 8 x 1 1 7 6 5 4 y 8 x 1 1 6 5 4 y 8 x 1
           1 5 4 y 8 x 1 1 4 y 8 x 1 3 y 8 x 2 y 7 6 5 4 1 8 x 1 1 8 x 1 3 y 8 x 2
           y 7 6 5 4 y 8 x 1 1 8 x 2 y 7 6 5 4 y 8 x 1 1 7 6 5 4 y 8 x 1 1 6 5 4 y
           8 x 1 1 5 4 y 8 x 1 1 4 y 8 x 1 3 y 8 x 2 y 7 6 5 4
```

Yokomori's algorithm ([Yoko92]) to identify the L system that produces these words since "PDOL languages are identifiable in the limit from positive data" does not simply apply here, but our method of theorem 2 does not give either directly a solution : $M_1$ has determinant 0. But since the linear equations corresponding to $V_{i+1} = V_i M$ have many real solutions, one could use a few nearly blind trials or a systematic computer programs to detect the integer solution.

However, it is possible to do better and quicker with the help of a little extra information to get a straightforward solution. We are using a branching structure, modeled by a bracketed grammar (see [Lind71]): `8` and `y` or `x` and `y` are the only candidates as bracketing symbols that can read as [ and ] respectively. The fact that the structure is apical (see [PrKa96]) is not usefull, though. So we know already the last two lines of the solution matrix for `x` and `y` are the last two symbols of the aphabet. Using this partial information and simplifying the equations (such as $u{+}v{=}0$ leads only to $u{=}v{=}0$, $2u{+}3v{+}w{=}1$ leads only to $u{=}v{=}0$, $w{=}1$...), our resolution leads to only one simple parametric rule, namely $1 \to 1^i$ with $i > 0$ and two equations of the form $u + v{=}1$ which have to be solved with non negative integers. So it is quick and easy to find the correct ten rules

| | | | | | |
|---|---|---|---|---|---|
| 1 | → | 1 | 2 | → | 1 3 |
| 3 | → | 1 4 | 4 | → | 5 4 |
| 5 | → | 6 | 6 | → | 7 |
| 7 | → | 8 x 2 y | 8 | → | 8 |
| x | → | x | y | → | y |

and to use the first word `1 1 3 1 4` as axiom.

The resolution of this problem takes a couple of seconds with *Maple*, even on a P.C. This is a great improvement, compared to brute combinatorial exploration for the 8 rules and 8 symbols.

Now, let's try to see if we are able to derive the same solution knowing only the

numbers of symbols for each word, that is, using only the sequence

$$5, 7, 9, 11, 16, 25, 36, 49, 64, 81, 103, 134$$

without even knowing that we have 10 symbols (for now we don't even know that the structure is bracketed). Even with so few values the unilinear relation is detected with an $dim_R$ dimension of 7 :

$$u_n = 2u_{n-1} - u_{n-2} + u_{n-6} - u_{n-7} \text{ for } n > 7$$

But the closed form is not a polynomial and so, without a mathematics resolution, only a brute "try and check" algorithm that tries all matrices of size 7, 8, 9, 10, 11 is able to find the *Parikh* representation whose left vector is

$$V = \begin{pmatrix} 3 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

and whose matrix is

$$M = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

With this solution, unique up to a permutation of symbols the biologist can be happy: there are only three couples of possible bracketing symbols, the matrix has a small size (and for those who know the origin of the problem, it is the correct minimal size). There is some more interpretation to be done on this canonical solution, but this is the best that computers program can do without extra (biological) knowledge. Please note that with a "simpler" sequence with the same $dim_R$ dimension, namely $3, 7, 15, 31, 63, 127, 254, 501, 967, 1815, 3301$ the method described as exercise 2 is able to find immediately a correct *Parikh* representation which is

$$V = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 2 \end{pmatrix}$$

$$M = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 3 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 3 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 3 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

# 5. Conclusion

We dealt with two problems, that is, to find a parallel grammar with a minimal alphabet given either a finite sequence of words or a sequence of numbers. We have shown that for the first problem, it is decidable to know if there is a solution and we have also given a method to compute it (which has been implemented with *Maple* on our computers). For the second problem, we had to restrict ourselves to a special class of sequences, called LG sequences and our existence theorem is not constructive. However, for some cases we have partial formulas that are programmable which we showed and used in the examples. Combined with geometric programs to visualize biological data, these methods of inference and heuristics are new usefull modelizing tools, especially for biologists.

# References


**[Lind68] A. Lindenmayer**
*Mathematical models for cellular interaction in development.*
Journal of Theoretical Biology; **18** :280-315, 1968.

**[Lind71] A. Lindenmayer**
*Developmental systems without cellular interaction, their languages and grammar.*
Journal of TheoreticalBiology; **30** :455-484, 1971.

**[PrKa96] P. Prusinkiewicz & Lila Kari**
*Subapical bracketed L-systems, Grammars and their application to computer science.*
Lecture Notes in Computer Science, volume **1073** :550-564.
Springer-Verlag 1996.

**[RoSa80] G. Rozenberg et A. Salomaa**
*The mathematical theory of L systems.*
Academic Press, 1980.

**[Salo73] A. Salomaa**
*Formal Languages.*
Academic Press, 1973.

**[SaSo78] A. Salomaa, M. Soittola**
*Automata-Theoretic Aspects of Formal Power Series.*
Springer-Verlag, 1978.

**[Yoko92] T. Yokomori**
*Inductive inference of 0L Languages.*
Lindenmayer Systems, Impacts on Theoretical Computer Science,

Computer Graphics and Developmental Biology.
Springer-Verlag, 1992.

[**SaSo78**] **A. Salomaa, M. Soittola**
*Automata-Theoretic Aspects of Formal Power Series.*
Springer-Verlag, 1978.

# Appendix

## Proof of theorem 3.

If $(u_n)_{n\in N}$ is a LG sequence with *Parikh* representation $(L, M)$ then let $P$ be the characteristic polynomial of $M$, call $a_i$ its coefficients and let $D = dim_P(\ (u_n)\ )$. Since $P(X) = det(X.Id - M)$, $P$ has degree $D$, the $a_i$ are integers and $P$ has leading coefficient 1. *Cayley-Hamilton*'s theorem states that $M$ is a root of $P$ so

$$M^D \ + \ \sum_{i=\ 0}^{D-1} \ a_i \ . \ M^i \ = \ 0$$

Multiplying both sides on the left by $L$ and on the right by $R$, one gets

$$L \ . \ M^D \ . \ R \ + \ \sum_{i=\ 0}^{D-1} \ a_i \ . \ L \ . \ M^i \ . \ R \ = \ 0$$

and since $L \ . \ M^i \ . \ R \ = \ u_i$, we have

$$u_D \ = \sum_{i=1}^{D-1} \ -a_i \ . \ u_{n-i}$$

so $(u_n)$ is UR with order at most $D = dim_P(\ (u_n)\ )$.$\diamondsuit$

## Proof of theorem 4.

Let $R = dim_R(\ (u_n)\ )$ and $d = dim_H(\ (u_n)\ )$. The relation $u_n \ = \sum_{k=1}^{R} \ a_k \ . \ u_{n-k}$ for $n > R$ shows that the lines $R+1$, $R+2$... of $H_d$ with $d > R$ and $i > 0$ are a linear combination (CL) of the previous lines.
So the determinant of $H_d$ with $d > R$ is 0 and its rank is constant equal to $R+1$. Whence $dim_H(\ (u_n)\ ) = dim_R(\ (u_n)\ ) + 1$ results from minimality of $dim_H(\ (u_n)\ )$ and $dim_R(\ (u_n)\ )$. $\diamondsuit$

## Proof of theorem 5.

Let $d = dim_H(\ (u_n)\ )$ and call $H_d$ its *Hankel* matrix of size $d$. Since $det(H_d + 1) = 0$, the last line of $H_d$ is a linear combination (LC) of the previous lines. So the first term of it, $u_{d-1}$ is a LC of $u_0, u_1 ... u_{d-2}$. But $u_n \ = \ L \ . \ M^n \ . \ R$ for all $n$, so, multiplying by $M^i$, we have that $M^{d+i-1}$ is a LC of $M^i, M^{i+2} ... M^{i+d-2}$ for all $i$ which means that $u_{d+i-1}$ is a LC of $u_i, u_{i+1} ... u_{i+d-2}$. So $(u_n)$ is UR with $dim_R(\ u_n\ ) \leq \ d-1$. $\diamondsuit$