EVOLVING UNDER SMALL DISRUPTION

Jürgen Dassow[1], Gema M. Martín[2] and Francisco J. Vico[2]

[1]Fakultät für Informatik, Otto-von-Guericke-Universität Magdeburg
PSF 4120; D-39016 Magdeburg; Germany
Email: dassow@iws.cs.uni-magdeburg.de

[2]Departamento de Lenguajes y Ciencias de la Computación, Universidad de Málaga
Severo Ochoa, 4, Parque Tecnológico de Andalucía,
E-29590 Campanillas - Málaga; Spain
Email: gema,fjv@geb.uma.es

Abstract
We extend the edit operators of substitution, deletion, and insertion of a symbol over a word by introducing two new operators (partial copy and partial elimination) inspired by biological gene duplication. We define a disruption measure for an operator over a word and prove that whereas the traditional edit operators are disruptive, partial copy and partial elimination are non-disruptive. Moreover, we show that the application of only edit operators does not generate (with low disruption) all the words over a binary alphabet, but this can indeed be done by combining partial copy and partial elimination with the substitution operator.

1. Introduction

Edit operators of substitution, deletion, and insertion of a symbol over a word have been extensively studied in literature and have been applied to many different kinds of problems. These are biologically inspired operators that are also known as point mutation operators [2, 3].

They have been applied to the problem of transforming a word of finite length into another word. Moreover, this very case has been studied expanding the set of edit operators. For example, in [10], the set of edit operators is extended to include the squashing and expansion operators. Whereas in the squashing operator two (or more) contiguous symbols of the first word can be transformed into a single symbol of the second word, in the expansion operator a single symbol in the first word may be expanded into two or more contiguous symbols of the second word. In [11], the edit operators together with the straightforward transposition of adjacent symbols are used in pattern recognition. The theory of error-correcting codes of variable lengths treats errors that can be modelled as substitutions, insertions or deletions of symbols ([7, 5]).

Furthermore, there are many studies that endeavour to explain a number of bioinspired evolutionary processes using edit operators. In [3], the concept of an evolutionary system is introduced. This is a language generating device inspired by the evolution of cell populations, and it is based on edit operators and string divisions. The purpose of this system is to model

some properties of evolving cell communities at the syntactical level. In [2], a computational device called network of evolutionary processors is proposed. It is based on evolutionary rules and communication within a network. Such evolutionary rules are substitution, deletion, and insertion rules. The generative power of evolutionary networks where only two types of such rules are allowed is discussed in [1]. There have been several studies of molecular evolution models that incorporate base substitutions, insertions, and deletions ([13, 8]).

However, to our knowledge, there are not many studies that analyze the disruptive effects of the edit operators. Since non-random search methods benefits from a low disruption in the application of operators to refine solutions, an analysis of how disruptive these operators are, and the proposal of new low disruptive operators is necessary.

In this paper, such a study of disruption of the edit operators is done. In order to be able to use the edit operators, we need to use devices that can be represented as words. One of the simplest devices that can be represented in this way is the cyclic unary deterministic finite automata (CUDFAs, for short). With the purpose of studying the disruption of the edit operators, we define a disruptive measure by using the similarity measure for CUDFAs that has been introduced in [4]. A CUDFA will be given directly from its graphical expression (Figure 1), and represented as a binary word, $w \in \{0,1\}^+$, where the zeros represent the non-accepting states of the automaton, and the ones represent the accepting states of the automaton. We define the disruption of an operator over a CUDFA w as the portion of words that are accepted by the initial CUDFA and are not accepted by the resultant CUDFA after applying an operator and vice versa. We show that by iterative application of edit operators we cannot generate all words if we require that any operator is accompanied by small disruption. We define two new non-disruptive bioinspired operators. If we combine them with the substitution operator, then starting from any $w \in \{0,1\}^+$, we obtain all the words $v \in \{0,1\}^+$ that accept a non-empty language where each step has low disruption.

The proposed non-disruptive operators have been inspired by gene duplication, an important genetic mechanism that plays an important role in evolution [9, 14]. Considering the binary word as a genome, duplication simply adds redundant information (in our case, to $w \in \{0,1\}^+$), keeping the associated phenotype (the language accepted by w) unchanged. The genomic portion gained after gene duplication provides a substrate for coding new functions (proteins, in biology) by future alterations: mutations, additions, deletions, or even being totally or partially copied/eliminated again. In particular, partial copy/elimination may introduce significant differences in the genome, but keeping the fitting level of the phenotype.

2. Cyclic unary deterministic automata

The reader is assumed to be familiar with the basic concepts of formal language theory. For further information the reader is referred to [12]. Here, only some notations used in this paper will be recalled.

In the sequel, we will consider that $0 \in \mathbb{N}$. For the cases in which zero is not included, we will

write $\mathbb{N}^+$. Throughout the paper, $V = \{0, 1\}$ and h is the mapping $V \to V$ with $h(1) = 0$ and $h(0) = 1$. For $w \in V^*$ and $x \in V$, we denote the length of w and the number of occurrences of x in w by $|w|$ and $|w|_x$.

In this paper we work with languages over a unary alphabet. Let A be a deterministic finite automaton over a unary alphabet (for short, UDFA) that represents an infinite regular language. As the alphabet is unary, each UDFA will have the structure that is shown in Figure 1. Its states are divided into two groups, the first one, that we call initial phase, will contain the states from the first state to the $i-1$ state, the second one, that we call loop, will contain the rest of the states. The initial phase can be empty in those automata, whose last state transits to its initial state. A UDFA can be represented as a vector (v, w) where $v \in \{0, 1\}^*$ describes

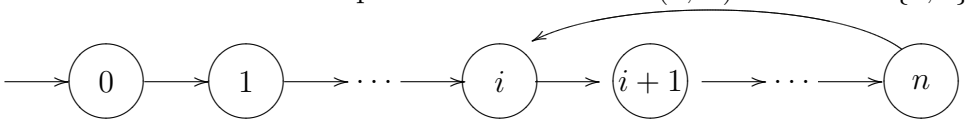

Figure 1: Structure of a UDFA

the initial phase and $w \in \{0, 1\}^+$ describes the loop. The zeros represent the non-accepting states of the automaton, and the ones represent the accepting states of the automaton.

A UDFA is cyclic (for short CUDFA) if its initial phase is empty. Then, instead of (λ, w), we represent the CUDFA as a word $w \in \{0, 1\}^+$. A language accepted by some CUDFA will be called a cyclic unary regular language (for short CURL).

For a word $w = x_1 x_2 \dots x_n \in \{0, 1\}^+$, we set $B(w) = \{i \mid x_i = 1\}$. Let w describe a CUDFA, and let $B(w) = \{b_1, b_2, \dots, b_m\}$. It is clear that $b_i < |w|$ for $1 \leq i \leq m$, and the regular set accepted by the CUDFA is

$$M = \{b_i + |w|k \mid 1 \leq i \leq m,\ k \in \mathbb{N}\}.$$

That is, M is union of a finite set of disjoint successions of natural numbers. In the sequel we use the notation

$$M = \{\{b_i + |w|k\}_{k \in \mathbb{N}}\}_{i=1,\dots,m}. \tag{1}$$

If M is the union of the successions $A_1, A_2, \dots, A_m$, then we also say that A_i is an element of M. In this paper we consider a CUDFA as a genotype, and its accepted language as the corresponding phenotype.

3. Definitions

We first define some operators over CUDFAs which are inspired by mutations, insertions, deletions and copying of molecules which occur in the evolution of biological systems.

For any natural numbers $m, p > 0$, we set

$$T(m, p) = \{w \mid w = (x_1 x_2 \dots x_m)^p,\ x_i \in V \text{ for } 1 \leq i \leq m\}.$$

Definition 1. *For any natural numbers $n, m, p > 0$, i with $1 \leq i \leq n$, $q > 1$, and $y \in V$ we define*

- *the addition operator $A_{i,y} : V^n \longrightarrow V^{n+1}$ as*

$$A_{i,y}(x_1x_2 \dots x_n) = x_1x_2 \dots x_iyx_{i+1} \dots x_n,$$

- *the partial copy operator $PC_p : T(m,p) \longrightarrow T(m,p+1)$ as*

$$PC_p((x_1x_2 \dots x_m)^p) = (x_1x_2 \dots x_m)^{p+1},$$

- *the elimination operator $E_i : V^n \longrightarrow V^{n-1}$ as*

$$E_i(x_1x_2 \dots x_n) = x_1x_2 \dots x_{i-1}x_{i+1} \dots x_n,$$

- *the partial elimination operator $PE_q : T(m,q) \longrightarrow T(m,q-1)$ as*

$$PE_q((x_1x_2 \dots x_m)^q) = (x_1x_2 \dots x_m)^{q-1},$$

- *the mutation operator $M_i : V^n \longrightarrow V^n$ as*

$$M_i(x_1x_2 \dots x_n) = x_1x_2 \dots h(x_i) \dots x_n.$$

Let $\mathcal{A}$, $\mathcal{E}$, $\mathcal{M}$, $\mathcal{PC}$, and $\mathcal{PE}$, be the sets of all addition, elimination, mutation, partial copy, and partial elimination operators, respectively. The operators in $\mathcal{A}$, $\mathcal{E}$ and $\mathcal{M}$ are called the edit operators.

We mention that all these operators are defined on the genotype. In nature, the rate of fixation of those low-disruption mutations is higher than the rate of fixation of those mutations that change the original phenotype too much. Therefore, we need a measure for the similarity of CURLs which represent the phenotypes.

In order to define the disruption of an operator over an automaton, we use the measure of similarity for CURLs defined in [4]. According to this measure, the disruptiveness of applying an operation to an automaton A to obtain an automaton B will be described by two rational numbers. The first one represents the portion of the words accepted by A but not by B, and the second one represents the portion of words accepted by B but not by A. This is analogous to the concepts of Recall and Precision in Information Retrieval. The precision is the fraction of the documents retrieved that are relevant to the user's information needs, while the recall is the fraction of the documents that are relevant to the query and are successfully retrieved.

Definition 2. *For two successions $A = \{a + bn\}_{n\in\mathbb{N}}$ and $B = \{c + dk\}_{k\in\mathbb{N}}$, the overlap $ISO_{A,B}$ of A and B (for Infinite Successions Overlap) is defined as:*

$$ISO_{A,B} = \begin{cases} \dfrac{gcd(b,d)}{d} & \text{if } A \cap B \neq \emptyset \\ 0 & \text{in other case} \end{cases}.$$

Given two CURLs M and N, we have that $M \cap N \neq \emptyset$ if and only if there exist at least $A \in M$ and $B \in N$ such that $A \cap B \neq \emptyset$.

Definition 3. *Let M and N be two CURLs, and let n be the number of successions of M. We define the overlap $URLO_{M,N}$ of M with N (for Unary Regular Languages Overlap) as*

$$URLO_{M,N} = \begin{cases} \frac{1}{n} \sum\limits_{\substack{A \in M \\ B \in N}} ISO_{A,B} & \text{if } M \cap N \neq \emptyset \\ 0 & \text{in other case} \end{cases}.$$

Then, we can say that the measure $ISO_{M,N}$ between CURLs gives the portion of strings in M which also belong to N. In [4], it is proven that the previous definition is independent on the choice of the successions used to represents languages M and N and also [4] includes the following statement.

Lemma 1. *Let M and N be CURLs. $URLO_{M,N} = 1$ if and only if $M \subseteq N$.*

Now we are in the position to define a notion which measures the change of the phenotypes obtained though the application of an operator.

Definition 4. *Let $w \in V^+$ be a CUDFA and $O \in \mathcal{M} \cup \mathcal{A} \cup \mathcal{E} \cup \mathcal{PC} \cup \mathcal{PE}$ be an operator such that $O(w)$ is defined. Let L and L' be the CURLs represented by w and $O(w)$, respectively. We define the disruption $D(O, w)$ of the operator O over w as*

$$D(O, w) = (1 - URLO_{L,L'}, 1 - URLO_{L',L}).$$

That is, the disruption of an operator O over w is a pair (a, b) with $a, b \in \mathbb{R}$, where a is the portion of words that are accepted by w and are not accepted by $O(w)$ and b is the portion of words that are accepted by $O(w)$ and are not accepted by w.

When $D(O, w) = (0, 0)$ for a given operator O and all w, we will say that the operator O is not disruptive or not destructive.

4. Determination of the disruption of the operators

In this section, we study the disruption of the operators that have been defined in the previous section. First of all, let us see a result that we will use in the sequel.

Lemma 2. *Let $w \in V^+$ be a CUDFA. The CURLs represented by w and by w^n, with $n \in \mathbb{N}$ and $n > 1$, are the same.*

Proof. Let $C = \{\{a_i + |w|k\}_{k \in \mathbb{N}}\}_{i=1,\ldots,m}$ be the CURL represented by w, where $a_i\{0, \ldots, |w|-1\}$ for any $1 \leq i \leq m$. Therefore, the CURL represented by w^n is

$$C' = \bigcup_{j=0}^{n-1} \{\{(a_i + j|w|) + n|w|k\}_{k \in \mathbb{N}}\}_{i=1,\ldots,m}.$$

Since $URLO_{C,C'} = 1$ and $URLO_{C',C} = 1$, by Lemma 1, $C = C'$. □

Note 1. *A string with length s is accepted by a CUDFA $w = x_0 \dots x_{k-1}$ with $x_i \in \{0,1\}$ for any $i = 0, \dots, k-1$ if and only if $x_{s \pmod k}$ is an accepting state. If $w' = w^n$ for some natural number $n > 1$, then the acceptance is given by $x_{s \pmod{kn}} = x_{(s \pmod{kn}) \pmod k} = x_{s \pmod k}$.*

The next corollaries follow immediately.

Corollary 1. *Let $w \in V^+$ be a CUDFA. The regular languages represented by w^n and by w^m, $n, m \in \mathbb{N}$ and $n, m > 1$, coincide.*

Corollary 2. *For any $p \geq 1$ and $q > 1$, PC_p and PE_q are not disruptive operators.*

Let us study the disruption of the remaining operators.

Lemma 3. *Let $w \in V^+$ be a CUDFA and i a natural number with $1 \leq i \leq |w|$. If $|w|_1 = m$, then*

- $D(M_i, w) = (0, \frac{1}{m+1})$ *if we mutate a zero into a one,*
- $D(M_i, w) = (\frac{1}{m+1}, 0)$ *if we mutate a one into a zero.*

Proof. Let $C = \{\{a_i + |w|k\}_{k\in\mathbb{N}}\}_{i=1,\dots,m}$ be the CURL represented by w, where $a_i \in \{0, \dots, |w|-1\}$ for any $1 \leq i \leq m$.

If we mutate a zero in the position i, the CURL represented by $M_i(w)$ is

$$C' = \{\{a_i + |w|k\}_{k\in\mathbb{N}}\}_{i=1,\dots,m} \cup \{b + |w|k\}_{k\in\mathbb{N}}$$

with $0 \leq b \leq |w| - 1$. In this case, since a non-accepting state has been changed into an accepting state in w, a portion of new words has been added to C. Since, $URLO_{C',C} = \frac{m}{m+1}$ (because $gcd(|w|, |w|) = |w|$ and C has only m subsuccessions of the $m+1$ that C' has) and $URLO_{C,C'} = 1$ (because $gcd(|w|, |w|) = |w|$ and C' has m subsuccessions of the m that C has),

$$D(M_i, w) = (0, \frac{1}{m+1}).$$

If we mutate a one in the position i, the CURL represented by $M_i(w)$ is

$$\begin{aligned} C' &= \{a_1 + |w|k\}_{k\in\mathbb{N}} \cup \dots \cup \{a_{i-1} + |w|k\}_{k\in\mathbb{N}} \\ &\quad \cup \{a_{i+1} + |w|k\}_{k\in\mathbb{N}} \cup \dots \cup \{a_m + |w|k\}_{k\in\mathbb{N}}. \end{aligned}$$

In this case, since an accepting state has been changed into a non-accepting state in w, a portion of words has been removed from C. Since, $URLO_{C,C'} = \frac{m}{m+1}$ and $URLO_{C',C} = 1$,

$$D(M_i, w)) = (\frac{1}{m+1}, 0).$$

□

Lemma 4. *For any CUDFA $w \in V^+$ with $|w|_1 = m$, any natural number i with $1 \leq i \leq |w|$, and any $y \in V$, $D(A_{i,y}, w) = (1 - \frac{m+y}{|w|+1}, 1 - \frac{m}{|w|})$.*

Proof. Let $C = \{\{a_i + |w|k\}_{k\in\mathbb{N}}\}_{i=1,\ldots,s} \cup \{\{b_j + |w|k\}_{k\in\mathbb{N}}\}_{j=1,\ldots,r}$ be the CURL represented by w, where $a_i \in \{0, \ldots, i-1\}$ for any $1 \leq i \leq s$ and $b_j \in \{i, \ldots, |w|-1\}$ for any $1 \leq j \leq r$. Then, $m = s + r$.

Case $y = 1$. The CURL represented by $A_{i,1}(w)$ is

$$\begin{aligned} C' &= \{\{a_i + (|w|+1)k\}_{k\in\mathbb{N}}\}_{i=1,\ldots,s} \cup \{i + (|w|+1)k\}_{k\in\mathbb{N}} \\ &\quad \cup \{\{(b_j+1) + (|w|+1)k\}_{k\in\mathbb{N}}\}_{j=1,2,\ldots,r}. \end{aligned}$$

Let us compute that portion of words accepted by w that are still accepted by $A_{i,1}(w)$. We get $URLO_{C,C'} = \frac{m+1}{|w|+1}$ (because $gcd(|w|, |w|+1) = 1$, $A \cap B \neq \emptyset$ for any $A \in C$ and any $B \in C'$ and C' has $m+1$ subsuccessions).

Let us compute that portion of words accepted by $A_{i,1}(w)$ that are also accepted by w. We get $URLO_{C',C} = \frac{m}{|w|}$ (because $gcd(|w|, |w|+1) = 1$, $A \cap B \neq \emptyset$ for any $A \in C$ and any $B \in C'$ and C has m subsuccessions). Therefore, $D(A_{i,1}, w) = (1 - \frac{m+1}{|w|+1}, 1 - \frac{m}{|w|})$.

Case $y = 0$. The CURL represented by $A_{i,0}(w)$ is

$$C' = \{\{a_i + (|w|+1)k\}_{k\in\mathbb{N}}\}_{i=1,\ldots,s} \cup \{\{(b_j+1) + (|w|+1)k\}_{k\in\mathbb{N}}\}_{j=1,\ldots,r}.$$

Let us compute that portion of words accepted by w that are still continue being accepted by $A_{i,0}(w)$. We get $URLO_{C,C'} = \frac{m}{|w|+1}$.

Let us compute that portion of words accepted by $A_{i,0}(w)$ that are also accepted by w. We get $URLO_{C',C} = \frac{m}{|w|}$. Therefore, $D(A_{i,0}, w) = (1 - \frac{m}{|w|+1}, 1 - \frac{m}{|w|})$. □

Lemma 5. *Let $w \in V^+$ be a CUDFA, $|w|_1 = m \geq 1$, i a natural number with $1 \leq n \leq |w|$, and y the i-th letter of w. Then $D(E_i, w) = (1 - \frac{m-y}{|w|-1}, 1 - \frac{m}{|w|})$.*

Proof. Let $C = \{\{a_i + |w|k\}_{k\in\mathbb{N}}\}_{i=1,\dots,s} \cup \{\{b_j + |w|k\}_{k\in\mathbb{N}}\}_{j=1,\dots,r}$ be the CURL represented by w, where $a_i \in \{0,\dots,i-1\}$ for $1 \leq i \leq s$ and $b_j \in \{i,\dots,|w|-1\}$ for $1 \leq j \leq r$. Then, $m = s + r$.

If we eliminate a one at position i, that is, $y = 1$, the CURL represented by $E_i(w)$ is

$$C' = \{\{a_i + (|w|-1)k\}_{k\in\mathbb{N}}\}_{i=1,\dots,s-1} \cup \{\{(b_j - 1) + (|w|-1)k\}_{k\in\mathbb{N}}\}_{j=1,\dots,r}.$$

Let us compute that portion of words accepted by w that are still accepted by $E_i(w)$. We get $URLO_{C,C'} = \dfrac{m-1}{|w|-1}$ because $gcd(|w|, |w|-1) = 1$, $A \cap B \neq \emptyset$ for any $A \in C$ and any $B \in C'$ and C' has $m-1$ subsuccessions).

Furthermore, $URLO_{C',C} = \dfrac{m}{|w|}$ because $gcd(|w|, |w|-1) = 1$, $A \cap B \neq \emptyset$ for any $A \in C$ and any $B \in C'$ and C has m subsuccessions). Therefore, $D(E_i, w) = (1 - \dfrac{m-1}{|w|-1}, 1 - \dfrac{m}{|w|})$.

If we eliminate a zero in position i, that is, $y = 0$, the CURL represented by $E_i(w)$ is

$$C' = \{\{a_i + (|w|-1)k\}_{k\in\mathbb{N}}\}_{i=1,\dots,s} \cup \{\{(b_j - 1) + (|w|-1)k\}_{k\in\mathbb{N}}\}_{j=1,\dots,r}.$$

Then we have $URLO_{C,C'} = \dfrac{m}{|w|-1}$ and $URLO_{C',C} = \dfrac{m}{|w|}$. Therefore, we get $D(E_i, w) = (1 - \dfrac{m}{|w|+1}, 1 - \dfrac{m}{|w|})$. □

Therefore, the edit operators are disruptive operators. Moreover, for an edit operator, the disruption is decreasing as the number of ones in the word is increasing.

5. Small disruptions and iterated application of operators

We now define the central notion of the paper.

Definition 5. *Let the CUDFA $w \in V^+$, $\mathcal{O} \subseteq \mathcal{M} \cup \mathcal{A} \cup \mathcal{E} \cup \mathcal{PC} \cup \mathcal{PE}$, and a real number λ, $0 < \lambda < 1$ be given.*

i) We say that a word v can be obtained with a disruption strictly less than λ from w using $\mathcal{O}$ if there exist operators $O_1, O_2, \dots, O_p \in \mathcal{O}$, $p \geq 0$, such that

- $v = O_p(O_{p-1} \dots (O_2(O_1(w))) \dots)$ *and*
- $D(O_i, O_{i-1}(\dots(O_2(O_1(w))\dots)) < (\lambda, \lambda)$ *for any $1 \leq i \leq p$.*

ii) By $LD(w, \mathcal{O}, \lambda)$ we denote the set of all words v which can be obtained with a disruption strictly less than λ from w using $\mathcal{O}$.

An important branch of the biological community supports the idea that during evolution gradual accumulations of small genetic changes occur resulting in producing small alterations in the phenotype; this permits the individual to stay adapted to the environment. From this point of view, those words which can be obtained in such a way that in each step a low disruption occurs are the most interesting of the set of all words which can be obtained from w by iterated applications of operations from $\mathcal{O}$ (e.g. [3], [1] and other papers).

In Definition 5, we have made the natural supposition $0 < \lambda < 1$. If $\lambda = 1$, then any sequence of operators is an evolution with disruption at most 1, i.e., we allow all sequences which coincides with the situation studied in previous papers. If $\lambda = 0$, no change of the phenotype is possible, which is not of interest from the biological point of view. By the biological motivation, we are only interested in the case of small λ, for instance $\lambda = \frac{1}{100}$. In the sequel we require $0 < \lambda \leq \frac{1}{2}$, which is sufficient from the mathematical point of view to guarantee a low disruption.

The aim of the remaining part of this paper is the study of the sets $LD(w, \mathcal{O}, \lambda)$. We start with two easy examples.

Let $w = 10^n$ for some $n \geq 2$, $0 < \lambda < \frac{1}{2}$ and $\mathcal{O} = \mathcal{M} \cup \mathcal{A} \cup \mathcal{E}$. Then by Lemmas 3, 4, and 5, for any operator from $\mathcal{O}$, we have $D(w, O(w)) = (a, b)$ with $a \geq \frac{1}{2}$ or $b \geq \frac{1}{2}$. Thus no word can be obtained with a disruption at most λ using $\mathcal{O}$ from w. Since we allow that no operator has to be used, $LD(w, \mathcal{O}, \lambda) = \{w\}$.

Let $w = 0^n$ for some $n \geq 1$, $0 < \lambda < \frac{1}{2}$ and $\mathcal{O} = \mathcal{M} \cup \mathcal{A} \cup \mathcal{E} \cup \mathcal{PC} \cup \mathcal{PE}$. It is easy to see that operators from $\mathcal{M}$ and of the form $A_{i,1}$ applied to w have a disruption at least $\frac{1}{2}$. Moreover, by operators from $\mathcal{PC}$ and $\mathcal{PE}$ we can get all words only consisting of zeros with no disruption (see Corollary 2). Hence $LD(w, \mathcal{O}, \lambda) = \{0^m \mid m \geq 1\}$.

Obviously, the reason that in the first example no operator has small disruption comes from the very small number of ones. If we change this situation, $LD(w, \mathcal{O}, \lambda)$ can be non-empty and contain infinitely many words, as can be seen from the following theorem.

Theorem 1. *Let $w \in V^+$ be a CUDFA and $0 < \lambda \leq \frac{1}{2}$ such that $\frac{1}{|w|_1+1} < \lambda$, and let $\mathcal{O} = \mathcal{M} \cup \mathcal{A} \cup \mathcal{E}$. Then*

$$LD(w, \mathcal{O}, \lambda) = \{v \mid |v|_0 > 0,\ \frac{1}{|v|_1 + 2} < \lambda\} \cup \{1^m \mid m \geq 1\} \cup \{w\}.$$

Proof. Let us suppose $|w|_0 = t$ and $|v|_0 = q$ for some $t, q \geq 0$ and let us consider the following finite sequence of operators:

- By $O_1, O_2, \ldots, O_t \in \mathcal{M}$ we mutate all the zeros of w. Therefore, $O_t(O_{t-1} \ldots (O_1(w)) \ldots) = 1^{|w|}$.
- Let $b = ||v| - |w||$.

  - If $|w| \leq |v|$, we choose $O_{t+1}, O_{t+2}, \ldots, O_{t+b} \in \mathcal{A}$ and get $O_{t+b}(\ldots(O_{t+1}(1^{|w|}))\ldots) = 1^{|v|}$.
  - If $|w| > |v|$, we choose $O_{t+1}, O_{t+2}, \ldots, O_{t+b} \in \mathcal{E}$ and obtain $O_{t+b}(\ldots(O_{t+1}(1^{|w|}))\ldots) = 1^{|v|}$.

- By $O_{t+b+1}, O_{t+b+2}, \ldots, O_{t+b+q} \in \mathcal{M}$ we mutate all the positions in which $1^{|v|}$ has a one and v has a zero and get $O_{t+b+q}(\ldots(O_{t+b+1}(1^{|v|}))\ldots) = v$.

Therefore, we have $O_{t+b+q}(O_{t+b+q-1}(\ldots(O_2(O_1(w)))\ldots)) = v$.

Let us calculate the disruption each time that we apply one of the operators given above.

If $1 \leq j \leq t$, then O_j increases the numbers of ones by 1. Thus, for $1 \leq i \leq t$, we have $|O_{i-1}(\ldots(O_2(O_1(w)))\ldots)|_1 > |w|_1$ and hence

$$\begin{aligned} D(O_i, O_{i-1}(\ldots(O_2(O_1(w)))\ldots)) &= (0, \frac{1}{|O_{i-1}(\ldots(O_2(O_1(w)))\ldots)|_1 + 1}) \\ &\leq (0, \frac{1}{|w|_1 + 1}) < (\lambda, \lambda). \end{aligned}$$

Let $|w| \leq |v| = b$. Then, for $t+1 \leq i \leq t+b$, $O_i \in \mathcal{A}$ adds a one to a word 1^k for some k. Thus O_i can be interpreted as a partial copy. By Corollary 2,

$$D(O_i, O_{i-1}(\ldots(O_2(O_1(w)))\ldots)) = (0,0) < (\lambda, \lambda).$$

Let $|w| > |v| = b$. Then, for $t+1 \leq i \leq t+b$, $O_i \in \mathcal{E}$ can be interpreted as a partial elimination. By Corollary 2,

$$D(O_i, O_{i-1}(\ldots(O_2(O_1(w)))\ldots)) = (0,0) < (\lambda, \lambda).$$

For $t+b+1 \leq j \leq t+b+q$, the operator O_i does not change the $|v|_1$ ones of v. Thus $|O_{i-1}(\ldots(O_2(O_1(w)))\ldots)|_1 \geq |v|_1 + 1$ and hence

$$\begin{aligned} D(O_i, O_{i-1}(\ldots(O_2(O_1(w)))\ldots)) &= (\frac{1}{|O_{i-1}(\ldots(O_2(O_1(w)))\ldots)|_1 + 1}, 0) \\ &\leq (\frac{1}{|v|_1 + 2}, 0) < (\lambda, \lambda). \end{aligned}$$

Therefore, for $1 \leq i \leq t+b+q$, we have $D(O_i, O_{i-1}(\ldots(O_2(O_1(w)))\ldots)) < (\lambda, \lambda)$.

Thus it is shown that all words v with $|v|_0 > 0$ and $\frac{1}{|v|_1+2} < \lambda$ or 1^m, $m \geq 1$ (in this case the operators $O_1, O_2, \ldots, O_{t+b}$ are sufficient), can be obtained.

It remains to show that further words cannot be generated by iterated applications of operators from $\mathcal{O}$, i.e., that words v with $v \neq w$, $|v|_0 > 0$ and $\frac{1}{|v|_1+2} \geq \lambda$ cannot be obtained.

Assume that $LD(w, \mathcal{O}, \lambda)$ contains a word v with $v \neq w$, $|v|_0 > 0$ and $\frac{1}{|v|_1+2} \geq \lambda$. Let Z be the set of all such v. We introduce a partial order on Z by $v_1 \prec v_2$ if and only if
– $|v_1| < |v_2|$ or
– $|v_1| = |v_2|$ and $|v_1|_1 < |v_2|_1$. Let v be a minimal word with respect to $\prec$ in Z. Let $O_1, O_2, \ldots, O_p$ be the operators from $\mathcal{O}$ such that $O_p(\ldots O_2(O_1(w))\ldots) = v$ and $D(O_j, O_{j-1}(\ldots O_2(O_1(w))\ldots)) \leq (\lambda, \lambda)$ for $1 \leq j \leq p$. We consider the step $v = O_p(x)$ where $x = O_{p-1}(\ldots O_2(O_1(w))\ldots)$. Let $m = |x|_1$.

We discuss some cases for O_p.

Case 1. $O_p = A_{i,0}$ for some i. If $x \neq 1^m$ for all $m \geq 1$, then $|x| < |v|$ and $|x|_0 > 0$ in contrast to our choice of v. Therefore $x = 1^m$ for some $m \geq 1$. Then $|x| = |x|_1 = m$ and

$$D(O_p, x) = (1 - \frac{m}{m+1}, 1 - \frac{m}{m}) = (\frac{1}{m+1}, 0) \geq (\frac{1}{m+2}, 0) \geq (\lambda, 0),$$

i.e., the last step does not satisfy the requirement for a disruption at most λ.

Case 2. $O_p = A_{i,1}$. Then x satisfies $|x| < |v|$ and $|x|_0 > 0$ which contradicts our choice of v.

Case 3. $O_p = E_i$ for some i.

If we cancel a letter 1, then $m = |v|_1 + 1$ and $|x|_0 \geq 1$ and $|x| \geq |x|_1 + 1 = m + 1$.

Because $m(|x| - m - 1) > -1$ or equivalently $1 - \frac{m-1}{|x|-1} > \frac{1}{m+1}$, the first component of $D(O_p, x)$ satisfies

$$1 - \frac{m-1}{|x|-1} > \frac{1}{m+1} = \frac{1}{|v|_1+2} > \lambda$$

in contrast to the choice of the operators.

If we cancel a zero, then $|x|_0 \geq 2$ and hence $|x| \geq m + 2$.Moreover, $|v|_1 = m$. Because $m(|x| - m - 2) \geq -|x|$ or equivalently $1 - \frac{m}{|x|} > \frac{1}{m+2}$, the second component of $D(O_p, x)$ satisfies

$$1 - \frac{m}{|x|} > \frac{1}{m+2} = \frac{1}{|v|_1+2} > \lambda,$$

which contradicts our assumption again.

Case 4. $O_p = M_i$. By the choice of v, we have to change a one into a zero. Hence $m = |v|_1 + 1$. Moreover, the first component of $D(O_p, x)$ satisfies $\frac{1}{m+1} = \frac{1}{|v|_1+2} > \lambda$. We have a contradiction, again.

Since we got a contradiction in each case, $Z = \emptyset$. □

From a biological point of view, the tendency of the complexity through the evolution has been a increasing tendency. For that reason, we could think that in order to find a parallelism with

biology, it is logical that we have to increase the length of the words. Therefore we give the following corollaries.

Corollary 3. *i) Let $w \in V^+$ be a CUDFA and $0 < \lambda \leq \frac{1}{2}$ such that $\frac{1}{|w|_1+1} < \lambda$, and let $\mathcal{O} = \mathcal{M} \cup \mathcal{A}$. Then*

$$LD(w, \mathcal{O}, \lambda) = \{v \mid |w| < |v|, |v|_0 > 0, \ \frac{1}{|v|_1 + 2} < \lambda\} \cup \{1^m \mid m \geq 1\} \cup \{w\}.$$

ii) Let $w \in V^+$ be a CUDFA and $0 < \lambda \leq \frac{1}{2}$ such that $\frac{1}{|w|_1+1} < \lambda$, and let $\mathcal{O} = \mathcal{M} \cup \mathcal{PC}$. Then

$$LD(w, \mathcal{O}, \lambda) = \{v \mid |w| < |v|, |v|_0 > 0, \ \frac{1}{|v|_1 + 2} < \lambda\} \cup \{1^m \mid m \geq 1\} \cup \{w\}.$$

Proof. i) For$|w| < |v|$, we have used only operators from $\mathcal{M} \cup \mathcal{A}$ in the proof of Theorem 1.

ii) The addition operators used add a 1 to a word only consisting of ones. Hence, there is an operator from $\mathcal{PC}$ which has the same effect. □

If we allow operators of $\mathcal{PE}$ in addition to those from $\mathcal{M} \cup \mathcal{PC}$, we get a case where all words of interest (i.e., all words describing a CUDFA which accepts a non-empty language) can be obtained with low disruptions from a given word w.

Theorem 2. *Let $w \in V^+$ be a CUDFA and $0 < \lambda \leq \frac{1}{2}$, and let $\mathcal{O} = \mathcal{M} \cup \mathcal{PC} \cup \mathcal{PE}$. Then*

$$LD(w, \mathcal{O}, \lambda) = V^+ \setminus \{0^m \mid m \geq 1\}.$$

Proof. Let $w \in V^+$ be a word with $|w| = m$ and $|w|_1 = r > 0$ and let $v \in V^+$ be a word with $|v| = n$ and $|v|_1 = s > 0$.

For a multiple $y = lcm(m, n)z$, $z \in \mathbb{N}^+$, of the lowest common multiple of m and n, we set $z' = \frac{y}{m}$ and $z'' = \frac{y}{n}$. We choose y sufficiently large, i.e., z sufficient large, such that $\frac{1}{rz'} < \lambda$ and $\frac{1}{sz''} < \lambda$. We construct the following finite sequence of operators.

- We choose $O_1, O_2, \ldots, O_{\frac{y}{m}-1} \in \mathcal{PC}$ such that any O_i adds a copy of w. Therefore $O_{\frac{y}{m}-1}(O_{\frac{y}{m}-2} \ldots (O_1(w)) \ldots) = w^{z'}$.
- Let t be the number of positions in which $w^{z'}$ has a zero and $v^{z''}$ has a one. We choose $O_{\frac{y}{m}}, O_{\frac{y}{m}+1}, \ldots, O_{\frac{y}{m}+t-1} \in \mathcal{M}$, such that such zeros are changed into ones. Thus we obtain $\overline{w} = O_{\frac{y}{m}+t-1}(\ldots (O_{\frac{y}{m}}(w^{z'})) \ldots)$.
- Let q be the number of positions in which $\overline{w}$ has a one and $v^{z''}$ has a zero. We choose $O_{\frac{y}{m}+t}, O_{\frac{y}{m}+t+1}, \ldots, O_{\frac{y}{m}+t+q-1} \in \mathcal{M}$ such that such ones are mutated into zeros and obtain $v^{z''} = O_{\frac{y}{m}+t+q-1}(\ldots (O_{\frac{y}{m}+t}(\overline{w})) \ldots)$.

- We choose $O_{\frac{y}{m}+t+q}, O_{\frac{y}{m}+t+q+1}, \ldots, O_{\frac{y}{m}+t+q+\frac{y}{n}-2} \in \mathcal{PE}$ such that any of the operators $O_{\frac{y}{m}+t+q+j}$ cancels one copy of v. Obviously, then $v = O_{\frac{y}{m}+t+q+\frac{y}{n}-2}(\ldots(O_{\frac{y}{m}+t+q}(v^{z''}))\ldots)$.

Therefore, $O_{\frac{y}{m}+t+q+\frac{y}{n}-2}(O_{\frac{y}{m}+t+q+\frac{y}{n}-3}(\ldots(O_2(O_1(w)))\ldots)) = v$.

Let us calculate the disruption each time that we apply one of the previous operators. For $1 \leq i \leq \frac{y}{m} - 1$, $D(O_i, O_{i-1}(\ldots(O_2(O_1(w)))\ldots)) = (0,0) < (\lambda, \lambda)$ by Corollary 2.

Since any operator O_j, $\frac{y}{m} \leq j \leq \frac{y}{m} + t - 1$, changes a zero into a one, i.e., we add only ones, and $|w^{z'}|_1 = rz'$, we get $|O_{i-1}(\ldots(O_2(O_1(w)))\ldots)|_1 \geq rz'$ and

$$\begin{aligned} D(O_i, O_{i-1}(\ldots(O_2(O_1(w)))\ldots)) &= (0, \frac{1}{|O_{i-1}(\ldots(O_2(O_1(w)))\ldots)|_1 + 1}) \\ &< (0, \frac{1}{rz'}) < (\lambda, \lambda) \end{aligned}$$

for $\frac{y}{m} \leq i \leq \frac{y}{m} + t - 1$.

Since O_j, $\frac{y}{m} + t \leq j \leq \frac{y}{m} + t + q - 1$, changes a one into a zero, but sz'' ones of $\overline{w}$ are not changed, we have $|O_{i-1}(\ldots(O_2(O_1(w)))\ldots)|_1 \geq sz'$ and

$$\begin{aligned} D(O_i, O_{i-1}(\ldots(O_2(O_1(w)))\ldots)) &= (\frac{1}{|O_{i-1}(\ldots(O_2(O_1(w)))\ldots)|_1 + 1}, 0) \\ &< (\frac{1}{sz''}, 0) < (\lambda, \lambda) \end{aligned}$$

for $\frac{y}{m} + t \leq j \leq \frac{y}{m} + t + q - 1$.

For $\frac{y}{m} + t + q \leq i \leq \frac{y}{m} + t + q + \frac{y}{n} - 2$, $D(O_i, O_{i-1}(\ldots(O_2(O_1(w)))\ldots)) = (0,0) < (\lambda, \lambda)$ by Corollary 2.

Therefore, we have $D(O_i, O_{i-1}(\ldots(O_2(O_1(w)))\ldots)) < (\lambda, \lambda)$ for $1 \leq i \leq \frac{y}{m} + t + q + \frac{y}{n} - 2$.

It remains to show that we cannot obtain words 0^m for some m. If we assume the contrary, then there is a number k such that there are operators $O_1, O_2, \ldots, O_p \in \mathcal{PC} \cup \mathcal{PE} \cup \mathcal{M}$ with $0^k = O_p(O_{p-1}\ldots(O_2(O_1(w)))\ldots)$ and

$$D(O_i, O_{i-1}(\ldots(O_2(O_1(w))\ldots)) < (\lambda, \lambda) \text{ for } 1 \leq i \leq p. \tag{2}$$

Without loss of generality we can assume that $O_j(O_{j-1}\ldots(O_2(O_1(w)))\ldots) \notin \{0^m \mid m \geq 1\}$ for $1 \leq j < p$ (otherwise the word $O_j(O_{j-1}\ldots(O_2(O_1(w)))\ldots) \in \{0^m \mid m \geq 1\}$ is considered instead of 0^k). Therefore O_p is a mutation operator which replaces a 1 by a zero and $O_{p-1}(O_{p-2}\ldots(O_2(O_1(w)))\ldots)$ contains exactly once the letter 1. Therefore, we have $D(O_p, O_{p-1}(O_{p-2}\ldots(O_2(O_1(w)))\ldots) = (\frac{1}{2}, 0)$ by Lemma 3 which is a contradiction to (2). □

We note that the operators of $\mathcal{PE}$ are not so common in biology as the edit operators and those from $\mathcal{PC}$. Thus we now look for a result where we only use the edit operators together with that of $\mathcal{PC}$.

Theorem 3. *For any word w with $|w|_1 > 0$ and any λ with $0 < \lambda < \frac{1}{2}$,*

$$LD(w, \mathcal{PC} \cup \mathcal{M} \cup \mathcal{A} \cup \mathcal{E}, \lambda) = \{v \mid |v|_0 > 0,\ \frac{1}{|v|_1 + 2} < \lambda\} \cup \{1^m \mid m \geq 1\} \cup \{w\}.$$

Proof. Let $|w|_1 = m \geq 1$, and let v be a word with $\frac{1}{|v|_1} < \lambda$. Then there is a number $r \in \mathbb{N}^+$ such that $\frac{1}{mr} \leq \lambda$. Using $r-1$ times operators from $\mathcal{PC}$ which copy w, we get w^r. Moreover, $|w^r|_1 = mr$ and thus $\frac{1}{|w^r|_1} < \lambda$. All the disruptions of these operators are $(0,0)$ by Corollary 2.

Starting from w^r, by Theorem 1, we can construct a sequence of operators $O_1, O_2, \dots O_p$ such that

- $v = O_p(O_{p-1} \dots (O_2(O_1(w^r))) \dots)$ and
- $D(O_i, O_{i-1}(\dots (O_2(O_1(w)) \dots)) < (\lambda, \lambda)$ for any $1 \leq i \leq p$.

This proves $v \in LD(w, \mathcal{PC} \cup \mathcal{M} \cup \mathcal{A} \cup \mathcal{E}, \lambda)$. As in the part of the proof of Theorem 1, we can show that no further words can be generated (in the notation from that proof, the operators from $\mathcal{PC}$ cannot be used for O_p by the choice of v). □

Therefore, in this section we have proven that the expressive capability of the set of operators $\{\mathcal{M}, \mathcal{PC}, \mathcal{PE}\}$ while keeping a low disruption, is higher than the expressive capability of $\{\mathcal{M}, \mathcal{A}, \mathcal{E}\}$, $\{\mathcal{M}, \mathcal{A}\}$, $\{\mathcal{M}, \mathcal{PC}\}$ and $\{\mathcal{M}, \mathcal{A}, \mathcal{E}, \mathcal{PC}\}$. This is because with the set of operators $\{\mathcal{PC}, \mathcal{PE}\}$ any length can be obtained without disruption, and then with the operators $\mathcal{M}$, that in the most of cases have a very small disruption, we get the symbols in the right position.

6. Discussion

In this paper we started the investigation of iterated applications of some bioinspired operators with the additional requirement that the disruption is (very) small in each step. In one case (Theorem 2) we were able to generate all words which correspond to non-empty regular languages. However, from a biological point of view, the other results are also satisfactory because the genotypes have to contain a lot of information, i.e., the words under consideration have to be long and to contain a sufficiently large number of ones. This means that the assumptions of Theorem 1 are satisfied and all words of biological interest can be obtained by Theorems 1 and 3.

In the literature one can find nice algorithms to determine the minimal number of edit operators which transform a given word w into another given word v (see e.g. [6]). It remains to search for

good algorithms where the additional requirement of small disruption in any step is satisfied. Finally, a future research line will be to study whether the results presented in this paper are also satisfied for more complex devices than CUDFA.

Acknowledgments

The authors gratefully acknowledge the useful suggestions and comments of the unknown referees.

References

[1] ALHAZOW, A., DASSOW, J., MARTÍN-VIDE, C., ROGOZHIN, Y., TRUTHE, B., On Networks of Evolutionary Processors with Nodes of Two Types, Fundamenta Informaticae 91 (2009), 1–15.

[2] CASTELLANOS, J., MARTÍN-VIDE, C., MITRANA, V., SEMPERE, J., Solving NP-complete Problems with Networks of Evolutionary Processors, Lecture Notes in Computer Science 2084 (2001), 621–628.

[3] CSUHAJ-VARJÚ, E., MITRANA, V., Evolutionary Systems: A Language Generating Device Inspired by Evolving Communities of Cells, Acta Informatica 36 (2000), 913–926.

[4] DASSOW, J., MARTÍN, G., VICO, F., A Similarity Measure For Cyclic Unary Regular Languages, (submitted).

[5] DAVEY, M., MACKAY, D., Reliable Communication over Channels with Insertions, Deletions, and Substitutions, IEEE Transactions on Information Theory 47 (2001), 687–698.

[6] GUSFIELD, D., Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology, Cambridge University Press, New York, 1997.

[7] JÜRGENSEN, H., KONSTANTINIDIS, S., Error Correction for Channels with Substitutions, Insertions, and Deletions, Lecture Notes in Computer Science 1133 (1996), 149–163.

[8] MESSER, P., ARNDT, P., LÄSSIG, M., Solvable Sequence Evolution Models and Genomic Correlations, Phys. Rev. Lett. 94 (2005).

[9] OHNO, S., Evolution by Gene Duplication, Springer, 1970.

[10] OOMMEN, B., String Alignment with Substitution, Insertion, Deletion, Squashing, and Expansion Operations, Information Sciences 83 (1995), 89–107.

[11] OOMMEN, B., LOKE, R., Pattern Recognition of Strings with Substitutions, Insertions, Deletions and Generalized Transpositions, Pattern Recognition 30 (1997), 789–800.

[12] ROZENBERG, G., SALOMAA, A., Handbook of Formal Languages, Springer, 1997.

[13] SAAKIAN, D., Evolution Models with Base Substitutions, Insertions, Deletions, and Selection, Phys. Rev. 78 (2008).

[14] ZHANG, J., Evolution by Gene Duplication: An Update, Trends in Ecology and Evolution 18 (2003), 292–298.