Watson-Crick D0L Systems

Valeria Mihalache
Arto Salomaa

Turku Centre for Computer Science
TUCS Technical Report No 108
April 1997
ISBN 951-650-996-7
ISSN 1239-1891

Abstract

D0L systems constitute the simplest and most widely studied type of Lindenmayer systems. They have the remarkable property of generating their language as a (word) sequence and, consequently, are very suitable for modeling growth properties. In this paper a new type of D0L systems is introduced, where the parallelism presented in L systems is combined with the paradigm of (Watson-Crick) complementarity characteristic for DNA computing. The new Watson-Crick D0L systems are still deterministic and generate their language as a sequence. However, the capability of modeling growth is greatly enhanced: the new growth functions are not necessarily even $\mathbb{Z}$-rational.

TUCS Research Group
Mathematical Structures of Computer Science

1 Introduction

Adleman demonstrated, [1], how methods of molecular biology can be applied to solve a computationally difficult problem. Since then the interest in "DNA computing" has been growing rapidly, for instance, see [5], [7] and their references. That DNA computing is relevant also for the theory of formal languages can be concluded from many chapters of the recent Handbook, [10]. On the other hand, there are still considerable obstructions to creating a practical molecular computer, and also very pessimistic views have been expressed, [4].

DNA (deoxyribonucleic acid) is found in all living organisms as the storage medium for genetic information. It consists of polymer chains, customarily referred to as *DNA strands*. A chain is composed of *nucleotides*, also referred to as *bases*. The chains are also referred to as *oligonucleotides*, briefly *oligos*. The four DNA nucleotides or bases are customarily denoted by A (adenine), C (cytosine), G (guanine), and T (thymine). The *DNA alphabet* $\Sigma_{DNA} = \{A, C, G, T\}$ will be important in our subsequent considerations. We often use lower-case letters instead of capital ones.

Thus, DNA strands may be viewed as words over the DNA alphabet. According to a chemical convention, each strand has a " 5' end" and a " 3' end", for instance,

$$5'\ ATTAGCAT\ 3' \quad \text{or} \quad 3'\ TAATCGTA\ 5',$$

making the words oriented. However, this orientation is irrelevant for our purposes and will be ignored in the sequel.

The familiar double helix of DNA arises by the bondage of two separate strands. In the formation of such double strands a phenomenon known as *Watson-Crick complementarity* comes into the picture. Bonding happens by the pairwise attraction of bases: A bonds with T, and C bonds with G. This is the reason why the unordered pairs (A, T) and (C, G) are referred to as *complementary* pairs of bases. Bonding occurs only if the bases in the corresponding positions in the two strands are complementary. (Moreover, they have to have opposite orientation but, as already pointed out, we will ignore here the orientation.) Thus, the two strands mentioned above will form the double strand

$$\begin{array}{c} ATTAGCAT \\ TAATCGTA \end{array}$$

Such double strands form a data structure of a new type, a data structure characteristic for the theory of DNA computing. The very nature of this data structure is essential for DNA computing. It is the source of the strength of

DNA computing because, in some sense, it makes the powerful *twin-shuffle language* "freely" available. (The universality of the twin-shuffle language is due to [3], see also [11] and [2]. The interconnection between DNA double strands and the twin-shuffle languages was pointed out in [9].) To illustrate the power of complementarity present in double strands, we now briefly recall Adleman's famous experiment, [1], of solving an instance of the Hamiltonian path problem, HPP.

Given a directed graph, we want to solve its HPP. Each vertex is encoded by an oligo of even length. (Thus, in our considerations oligos are simply words over the DNA alphabet. Adleman used words of length 20.) Each edge is encoded by an oligo of the same length as was used for the vertices. The encoding of the edges is determined by the encoding of the vertices in the following fashion. Consider the endomorphism h of Σ_{DNA}^* mapping each letter into its Watson-Crick complement:

$$h(A) = T, \quad h(T) = A, \quad h(C) = G, \quad h(G) = C.$$

(We will refer to h as the *Watson-Crick morphism.*)

Let (x, y) be an edge, where x is the oligo encoding the vertex from which the edge emanates and, similarly, y encodes the target vertex. Write x in the form $x = x_1x_2$, $|x_1| = |x_2|$ and, similarly, $y = y_1y_2$, $|y_1| = |y_2|$. The edge (x, y) is now encoded by the oligo $h(x_2y_1)$.

Adleman's experiment begins by forming a "DNA soup" containing, in large quantities oligos of the vertices, as well as of the edges of the graph. The ligation reaction resulting from the Watson-Crick complementarity now will link together compatible edges. This domino game continues, identifying longer and longer paths. In this way DNA molecules encoding random paths through the graph are formed. By a filtering procedure not of interest for us here, one can check whether or not paths satisfying HPP are present.

Adleman's experiment shows the computational strength of the Watson-Crick complementarity. It is a consequence of the universality of the twin-shuffle language that any model of DNA computing, where the filtering procedures available can simulate gsm mappings, is capable of Turing machine computations.

Although the real practical feasibility of molecular computers remains still in doubt, computer scientists and mathematicians are already looking for new models of computation along these lines. Such models could be called *Watson-Crick machines.* We are not going to specify here explicitly any such model. It seems obvious that in any such model use must be made of the following two advantages stemming from DNA molecules: (i) Watson-Crick complementarity which always renders the powerful twin-shuffle languages available, and (ii) the multitude of DNA molecules which brings the massive parallelism to the computing scene. It seems also that theoretical studies have so far been focused on the point (i). We will now discuss complementarity as a general language-theoretic principle.

2 Complementarity in language theory

The notion of *complementarity* based on the idea of Watson-Crick has so far been very little observed or investigated in language theory. Yet the notion is very simple and natural. We would like to formulate some of the issues in the following way.

Paradigms of complementarity.

(i) A string induces the complementary string, either randomly or guided by a control device.

(ii) The complementarity of two strings leads to some phenomenon such as bondage. Thus, the occurrence of the phenomenon guarantees that the strings involved are complementary.

Adleman's experiment makes use of Paradigm (ii). Another type of use of Paradigm (ii), dealing with Prolog, is given in [6]. In the sequel we will be concerned with Paradigm (i). We have so far considered complementarity only in connection with the DNA alphabet Σ_{DNA}. The following generalization is straightforward.

A *DNA-like alphabet* Σ is an alphabet with even cardinality $2n,\ n \geq 1$, where the letters are enumerated as follows:

$$\Sigma = \{a_1, \ldots, a_n, \overline{a_1}, \ldots, \overline{a_n}\}.$$

Thus, each of the non-barred letters $a_i,\ 1 \leq i \leq n$, has its barred version $\overline{a_i}$. We say that a_i and $\overline{a_i}$ are *complementary.* The letter-to-letter endomorphism of Σ^* mapping each letter to the complementary letter is referred to as the *Watson-Crick morphism.*

When the DNA alphabet Σ_{DNA} is viewed as DNA-like, we consider the *purines* A and G as non-barred letters: $a_1 = A$ and $a_2 = G$. Hence the *pyrimidines* T and C are their barred versions: $\overline{a_1} = T$ and $\overline{a_2} = C$.

It is both natural and sufficient for our purposes to consider only the "mild" generalization of the alphabet Σ_{DNA} introduced above. Of course, a more general notion of a DNA-like alphabet would be the pair (Σ, ρ), where ρ is a binary relation on Σ, satisfying some suitably chosen restrictive conditions.

The paradigms of complementarity can now be understood with respect to complementarity in DNA-like alphabets. As already mentioned, we will be concerned with Paradigm (i). The paradigm can be applied to any language-theoretic device: the transition to the complementary string happens either freely or under some control. However, the paradigm seems very natural in connection with *L systems*. According to an L system, the whole string is rewritten. This is also exactly what the Watson-Crick morphism does.

We are fully aware that complementarity opens a vast area for language-theoretic research; many kinds of questions can be asked along these lines. Our purpose here is to make only some observations. It seems to us that a very important generalization of the DNA-like alphabet, as well as complementarity based on it, consists of considering a disjoint union of a DNA-like alphabet Σ_1 and another alphabet Σ_2. Only the letters of Σ_1 are affected in the transition to the complementary string, while the letters of Σ_2 remain unchanged. However, as already pointed out, in this paper the mild generalization of the alphabet Σ_{DNA}, as well as the resulting notion of complementarity, will be sufficient for our purposes.

We now present an example of a "*Watson-Crick 0L system*", without giving the full formal definition of the notion. Consider first an ordinary $0L$ system, say, the system G with the alphabet $\Sigma = \{a, g, t, c\}$, the axiom gac and the productions

$$a \to \lambda, a \to ca, g \to cat, c \to ta, t \to \lambda.$$

(Observe that we use the DNA alphabet but prefer the lower-case letters. For all unexplained notions about L systems we refer to [8].)

The $0L$ system G is now transformed into a *Watson-Crick* $0L$ *system* G_w by supplementing G with a *trigger for complementarity transition.* In our case the trigger will be the appearance of the subword $catcat$. The subword $catcat$ is not permitted to appear; whenever it would appear in the normal course of the $0L$ derivation, the complementary word must be taken. Thus,

$$gac \Rightarrow catcata$$

is a derivation step according to G but must be replaced by the derivation step

$$gac \Rightarrow_w \ gtagtat$$

according to G_w.

Formally, the yield relation $\Rightarrow_w$ is defined from the yield relation $\Rightarrow$ of the $0L$ system G as follows. Let h be the Watson-Crick morphism. Assume that $\alpha \in \Sigma^*$ is a word not containing $catcat$ as a subword, that is, $\alpha \notin \Sigma^* catcat \Sigma^*$. Then $\alpha \Rightarrow_w \beta$ if $\alpha \Rightarrow \beta$ and $\beta \notin \Sigma^* catcat \Sigma^*$. Also $\alpha \Rightarrow_w h(\beta)$ if $\alpha \Rightarrow \beta$, $\beta \in \Sigma^* catcat \Sigma^*$ and $h(\beta) \notin \Sigma^* catcat \Sigma^*$.

Observe that the relation $\Rightarrow_w$ is not defined between α and β if either α contains $catcat$ as a subword, or else $\alpha \Rightarrow \beta$ and both β and $h(\beta)$ contain $catcat$ as a subword. It can be verified that in our particular example, starting from the axiom gac, such blocking situations do not occur. Indeed, if $catcat$ is a subword of $h(\beta)$, then $gtagta$ is a subword of β. This is not possible for the simple reason that g does not appear on the right sides of the productions. The same observation tells us also that, with our productions and trigger for complementarity transition, blocking can occur only if it occurs in the axiom, that is, the axiom contains both $catcat$ and $gtagta$ as subwords.

Observe further that

$$gac \Rightarrow_w catta \Rightarrow_w tacaca \Rightarrow_w catacataca,$$

$$gac \Rightarrow_w gtagtat \Rightarrow_w catcacat,$$

$$gac \Rightarrow_w gtagtat \Rightarrow_w gtagtagt \Rightarrow_w gtagtagtgta$$

are all the derivations according to G_w. Continuing the last one and applying λ-rules for t and a between two g's, we obtain derivations with arbitrarily many complementarity transitions.

The complementarity transition dramatically changes the behaviour of a $0L$ system, also with respect to fundamental properties. Thus, it is well-known that if $G = (\Sigma, h, w)$ is a $0L$ system, then for any nonnegative integer n and for any words $x_1, x_2, y_1, y_2 \in \Sigma^*$, if $x_1 \Rightarrow^n y_1$ and $x_2 \Rightarrow^n y_2$, then $x_1 x_2 \Rightarrow^n y_1 y_2$, and conversely (by "$\Rightarrow^n$" we have denoted the relation "derives in n steps"). This property does not hold anymore for Watson-Crick $0L$ systems. In the above example, we have $ga \Rightarrow_w catca$ and $c \Rightarrow_w ta$. However, due to the trigger for complementarity transition, $gac \not\Rightarrow_w catcata$, but $gac \Rightarrow_w gtagtat$ instead. Conversely, $gtagtat$ cannot be decomposed as to derive from two subwords composing gac.

In general, the appearance of a particular subword is one possible way of triggering the complementarity transition. Natural decision problems in this case are: (i) Does a complementarity transition occur in a given Watson-Crick $0L$ system? (ii) Does a transition occur unboundedly many times in the same derivation according to a given system?

Clearly, (i) is decidable because the emptiness of the intersection between a $0L$ language and a regular language is decidable. Problem (ii) is trickier, and we have no direct answer.

3 Watson-Crick $D0L$ systems: general definition

From now on we will focus the attention on $D0L$ systems. We will also define the notions explicitly. The case of *Watson-Crick $D0L$ systems* will be particularly interesting because, similarly as in case of ordinary $D0L$ systems, the language is generated as a *sequence*. In what follows, Σ *will be a DNA-like alphabet and h the Watson-Crick morphism.*

By definition, a *Watson-Crick $D0L$ system* is a pair $W = (G, \phi)$, where $G = (\Sigma, f, w_0)$ is a $D0L$ system (f is the morphism and $w_0 \in \Sigma^*$ is the axiom) and $\phi : \Sigma^* \to \{0, 1\}$ is a mapping such that $\phi(w_0) = \phi(\lambda) = 0$ and, for all words $w \in \Sigma^+$,

$$\phi(w) = 0 \ \text{ iff } \ \phi(h(w)) = 1.$$

According to the terminology of the preceding section, the condition $\phi(w) = 1$ is the *trigger for complementarity transition.* If in the course of a derivation a word w is encountered such that $\phi(w) = 1$, then w is replaced by $h(w)$. The condition imposed on ϕ guarantees that in such a case necessarily $\phi(h(w)) = 0$, that is, the blocking situation discussed in the preceding section never occurs. If f is defined in terms of productions, the definition here is in full accordance with that in the preceding section.

Formally, the *sequence* $S(W)$ of the system W consists of the words $w_0, w_1, w_2, \ldots$, where for each $i \geq 0$,

$$w_{i+1} = \begin{cases} f(w_i) & \text{if } \phi(f(w_i)) = 0, \\ h(f(w_i)) & \text{if } \phi(f(w_i)) = 1. \end{cases}$$

The *language* $L(W)$ of the system W consists of all words in the sequence $S(W)$. Two Watson-Crick $D0L$ systems W_1 and W_2 are termed *sequence* (resp. *language*) *equivalent* if $S(W_1) = S(W_2)$ (resp. $L(W_1) = L(W_2)$).

The definition guarantees that, for all $i \geq 0$, $\phi(w_i) = 0$. For instance, consider the DNA alphabet $\Sigma = \{a, g, t, c\}$ and define $\phi(w) = 0$ iff w begins with a purine. In other words, for all w,

$$\phi(aw) = \phi(gw) = 0, \quad \phi(cw) = \phi(tw) = 1.$$

If $G = (\Sigma, f, w_0)$ is any $D0L$ system such that w_0 begins with a or g, then (G, ϕ) is a Watson-Crick $D0L$ system. This definition of ϕ is conspicuously *local*: the value $\phi(w)$ is determined by the first letter of w.

The decidability of sequence or language equivalence has been a celebrated problem for $D0L$ systems. The same problems can be posed for Watson-Crick $D0L$ systems; of course, one has to define ϕ more explicitely. We now give a very natural explicit definition.

Let Σ be a DNA-like alphabet. A subset Σ_ϕ of Σ is termed a *selector* if it contains exactly one letter from each pair $(a, \overline{a})$, $a \in \Sigma$. Observe that the cardinality of a selector equals the cardinality of Σ divided by 2. For instance, the set of purines $\{a, g\}$ constitutes a selector for the DNA alphabet $\{a, g, t, c\}$.

We use the customary notation $|w|_{\Sigma_\phi}$ to denote the length of the scattered subword of w, consisting of letters of Σ_ϕ. We now define the function ϕ by

$$\phi(w) = \begin{cases} 0 & \text{if } |w|_{\Sigma_\phi} \geq \frac{|w|}{2} \ , \\ \\ 1 & \text{if } |w|_{\Sigma_\phi} < \frac{|w|}{2} \ . \end{cases}$$

It is immediately verified that the required condition is satisfied: $\phi(w) = 0$ iff $\phi(h(w)) = 1$. Thus, each selector can be used to define a Watson-Crick $D0L$ system, provided the axiom is chosen in the proper way. We refer to such Watson-Crick $D0L$ systems as *selector-based*.

The condition determining the value of ϕ is *global* in selector-based systems. Yet the condition is very simple and, thus, the definition of the resulting word sequence is a natural generalization of the ordinary $D0L$ word sequence. The customary notions and problems dealing with $D0L$ systems are readily extended to concern (selector-based) Watson-Crick $D0L$ systems. In particular, one can consider word length and *growth functions*. The growth functions of a $D0L$ system is defined by the condition $f(i) = |w_i|$, where $w_0, w_1, w_2, \ldots$ is the sequence of the system. In case of $D0L$ systems this definition leads to a simple *matrix representation* of f. Consequently, because the classical Perron-Frobenius theory is applicable, the theory of $D0L$ growth functions is quite well understood, [8], [11]. The same does not hold for Watson-Crick $D0L$ systems. Even in the simple case of selector-based systems the complementarity transition invalidates the matrix representation. The following examples are intended to illustrate this phenomenon. They show also that (selector-based) Watson-Crick $D0L$ systems constitute a really substantial generalization of $D0L$ systems. That they constitute a generalization is obvious: ordinary $D0L$ systems are obtained as a special case by using only non-barred letters, both in the system and in the selector.

In the following example we consider the DNA-like alphabet $\Sigma = \{a_1, a_2, a_3, \overline{a_1}, \overline{a_2}, \overline{a_3}\}$ and the selector set $\Sigma_\phi = \{a_1, a_2, a_3\}$. Our selector-based Watson-Crick $D0L$ system has the axiom $a_1a_2\overline{a_3}$ and its morphism is defined by the rules

$$a_1 \to a_1,\ a_2 \to a_2,\ a_3 \to a_3,\ \overline{a_1} \to \overline{a_1}\,\overline{a_2},\ \overline{a_2} \to \overline{a_2},\ \overline{a_3} \to \overline{a_3}^3.$$

The beginning of the sequence $S(W)$ looks as follows, where bold characters indicate that complementarity transition has taken place:

$$a_1a_2\overline{a_3},\ \mathbf{\overline{a_1}\,\overline{a_2}a_3^3},\ \overline{a_1}\,\overline{a_2}^2a_3^3,\ \mathbf{a_1a_2^3\overline{a_3}^3},\ \mathbf{\overline{a_1}\,\overline{a_2}^3a_3^9},\ \overline{a_1}\,\overline{a_2}^4a_3^9,\ \overline{a_1}\,\overline{a_2}^5a_3^9,$$

$$\overline{a_1}\,\overline{a_2}^6a_3^9,\ \overline{a_1}\,\overline{a_2}^7a_3^9,\ \overline{a_1}\,\overline{a_2}^8a_3^9,\ \mathbf{a_1a_2^9\overline{a_3}^9},\ \mathbf{\overline{a_1}\,\overline{a_2}^9a_3^{27}},\ \overline{a_1}\,\overline{a_2}^{10}a_3^{27},$$

after which there are 16 words before the next complementarity transition.

Observe that only the barred letters induce growth. The length sequence is strictly growing:

$$3, \mathbf{5}, 6, \mathbf{7}, \mathbf{13}, 14, 15, 16, 17, 18, \mathbf{19}, \mathbf{37}, \mathbf{38}, \ldots$$

The growth fluctuates between linear growth (from n to $n+1$) and exponential jumps (from $1 + 2 \times 3^n$ to $1 + 3^n + 3^{n+1}$). Indeed, when $f(i)$ is of the form $1 + 2 \times 3^n$, $n \geq 0$, then $f(i+1) = 1 + 3^n + 3^{n+1}$. When $f(i)$ (> 3) is not of the form $1 + 2 \times 3^n$, then $f(i+1) = f(i) + 1$. Thus, the strictly increasing growth function f assumes all values ≥ 3, except the values strictly between the numbers $1 + 2 \times 3^n$ and $1 + 3^n + 3^{n+1}$, for all $n \geq 0$. Consequently, the growth function f

(i) is linearly bounded, indeed $f(i) < 4(i+1)$ for all $i \geq 0$, and

(ii) the differences $f(i+1) - f(i)$ have no finite upper bound, that is, there is no integer k such that $f(i+1) - f(i) \leq k$ holds for all k.

No function satisfying (i) and (ii) can be a $D0L$ growth function. Indeed, if a linearly bounded function is a $D0L$ growth function, then it can be *decomposed* (see [8] or [10]) into linear functions of the form $ai + b$. At least one component of the decomposition would have to satisfy (ii), which is not possible.

In fact, even a stronger result concerning the growth function f of our example system can be obtained: f is not a $\mathbb{Z}$*-rational* function. ($D0L$ growth functions are a special case of $\mathbb{N}$*-rational* functions. $\mathbb{N}$-rational functions coincide with growth functions of $HD0L$ systems, that is, morphic images of $D0L$ systems. $\mathbb{Z}$-rational functions coincide with differences of two $\mathbb{N}$-rational functions. See [8] or [10] for further details.) This stronger result can be established by the following simple argument. If f were $\mathbb{Z}$-rational, then also the function $f_1(i) = f(i+1) - f(i)$ would be $\mathbb{Z}$-rational. However, f_1 assumes the same value ($= 1$) in arbitrarily long intervals without being ultimately constant. By the "Lemma of Long Constant Intervals", f_1 cannot be $\mathbb{Z}$-rational. We want to present these results as a formal theorem.

Theorem. *The class of growth functions of selector-based Watson-Crick D0L systems includes strictly the class of D0L growth functions. Moreover, the former class contains functions that are not $\mathbb{Z}$-rational.*

Our intention in the sequel is to give an explicit formula for the above growth function. To this aim, denote by i_n the $(n+1)$-st "exponential jumping" position that occurs in the values of f, for arbitrary $n \geq 0$. More precisely, i_n is such that

(1) $f(i_n - 1) = 1 + 2 \times 3^n$,

(2) $f(i_n) = 1 + 3^n + 3^{n+1}$.

Particular values are $i_0 = 1,\ i_1 = 4,\ i_2 = 11$.

The value of the growth function f with respect to the argument i_n is

$$\begin{array}{lcl} f(i_n + k) & = & 1 + 3^n + 3^{n+1} + k, \quad \text{for } 0 \leq k \leq 2 \times 3^n - 1 \\ f(i_n + 2 \times 3^n) & = & 1 + 2 \times 3^{n+1} \end{array}$$

(the upper limit for k was obtained as $(1 + 2 \times 3^{n+1} - 1) - (1 + 3^n + 3^{n+1}) = 2 \times 3^n - 1$.)

From here we deduce the recurrent relation for the sequence of numbers $i_n,\ n \geq 0$,

$$\begin{array}{lcl} i_0 & = & 1, \\ i_{n+1} & = & i_n + 2 \times 3^n + 1, \quad \text{for any } n \geq 0. \end{array}$$

The solution of this recurrence is $i_n = i_0 + 2 \sum_{k=0}^{n-1} 3^k + n$, i.e. $i_n = 3^n + n$ for any $n \geq 0$.

The growth function f is therefore given by

$$f(i) = \begin{cases} 3, & \text{if } i = 0, \\ 1 + 3^n + 3^{n+1} + k, & \text{if } i = 3^n + n + k, \\ & \quad \text{for any } n \geq 0,\ 0 \leq k \leq 2 \times 3^n - 1, \\ 1 + 2 \times 3^{n+1}, & \text{if } i = 3^n + n + 2 \times 3^n, \\ & \quad \text{for any } n \geq 0. \end{cases}$$

Just as in case of $D0L$ systems, rewriting being deterministic gives rise to certain *periodicities* in the Watson-Crick $D0L$ sequence.

Suppose that a word occurs twice in the sequence of a Watson-Crick $D0L$ system, that is, $w_i = w_{i+j}$ for some $i, j \geq 0, j \neq 0$. Let $w' = f(w_i)$. Due to the determinism, then also $f(w_{i+j}) = w'$. Therefore, if $\phi(w') = 0$, then $w_{i+1} = w'$, $w_{i+j+1} = w'$, whereas if $\phi(w') = 1$, then $w_{i+1} = \overline{w'}, w_{i+j+1} = \overline{w'}$. Thus $w_i = w_{i+j}$ implies $w_{i+1} = w_{i+j+1}$, whence the words start repeating periodically in the sequence. It then follows that the language generated by such a Watson-Crick $D0L$ system is finite. Moreover, the converse of this property is true as well, i.e., the language generated by a Watson-Crick $D0L$ system is finite if and only if a repetition occurs in the sequence of the system.

However, due to the replacing of a string by its complementary in the generating process, the periodicity features here are not as strong as for usual deterministic Lindenmayer systems, where we have that the partial alphabets $\Sigma_i = alph(w_i)$ (the smallest set of letters Σ such that $w_i \in \Sigma^*$), $i \geq 0$, form an ultimately periodic sequence, and also prefixes and suffixes of any chosen length form an ultimately periodic sequence. As far as Watson-Crick $D0L$ systems are concerned, this is not the case anymore. As a proof observe that in the above example, we have that the alphabets $\Sigma_{i_n - 1} = \{a_1, a_2, \overline{a_3}\}$, whereas for any positive integer $k \neq i_n - 1$, $\Sigma_k = \{\overline{a_1}, \overline{a_2}, a_3\}$ (the notation i_n is with the same meaning as above, i.e., $i_n = 3^n + n$, for any $n \geq 0$.) More precisely, $\Sigma_i = \Sigma_j$ for any i, j of the form $i = 3^n + n - 1$, $j = 3^m + m - 1$, where $n, m \geq 0$ arbitrary, and $\Sigma_i = \Sigma_j$ for any $i, j > 0$ such that $i \neq 3^n + n - 1$, $j \neq 3^m + m - 1$, no matter which are $n, m \geq 0$. Then clearly Σ_i, $i \geq 0$, do not form an ultimately periodic sequence. Since either $\Sigma_i = \Sigma_j$ or $\Sigma_i \cap \Sigma_j = \emptyset$, for any $i, j \geq 0$, the non-periodicity of the alphabets Σ_i implies the non-periodicity of prefixes and suffixes of any chosen length.

4 Plain Watson-Crick $D0L$ Systems

We already came to the conclusion that selector-based systems constitute a very natural special case of Watson-Crick $D0L$ systems. Finally, we consider a really basic variant of selector-based systems: the alphabet is the DNA alphabet itself.

A selector-based Watson-Crick $D0L$ system is referred to as *plain* if its alphabet is the DNA alphabet $\{a, g, t, c\}$, and the selector is the set of purines $\{a, g\}$.

That we use purines as selectors is of course no loss of generality with respect to any important formal properties. On the other hand, plain Watson-Crick $D0L$ systems can no longer simulate arbitrary $D0L$ systems, for instance, with respect to growth properties. Yet they can be surprisingly complex, as seen in the following example.

Consider the plain Watson-Crick $D0L$ system W with the axiom atg and productions

$$a \to at,\ t \to t,\ g \to g,\ c \to c^{10}.$$

According to our choice of the selector set (purines), the complementarity transition has to take place whenever t and c would have the majority in a word. If we were dealing with an ordinary $D0L$ system, the generated sequence would be at^ig, $i \geq 1$, and the letter c would not be reachable. However, for our Watson-Crick $D0L$ system W, the sequence $S(W)$ is much more complicated. The first few words in the sequence are listed below. As before, bold characters indicate that complementarity transition has taken place. Recall that (a,t) and (c,g) are the complementary pairs:

$$atg,\ at^2g,\ \mathbf{ta^3c},\ \mathbf{a(ta)^3g^{10}},\ at(tat)^3g^{10},\ at^2(tat^2)^3g^{10},$$

$$\mathbf{ta^3(ata^3)^3c^{10}},\ \mathbf{a(ta)^3(ta^2(ta)^3)^3g^{100}}.$$

In the example in the preceding section, where the alphabet was bigger, we could affect the alternation between linear and exponential growth in a more conspicuous fashion. Here the construction is more complicated. Whenever a "bad" letter starts to be overwhelming, the complementarity transition changes it to a "good" letter. This may cause a loss in the other complementary pair which, however, is smaller than the gain.

We are basically interested only in the growth function of the system W and, hence, the order of letters in a word is immaterial. We now continue the sequence $S(W)$, indicating only the number of occurrences of each letter. According to this convention, the last word obtained above is written in the form $a^{19}t^{15}g^{100}$, and the sequence continues:

$$a^{19}t^{15}g^{100},\ a^{19}t^{34}g^{100},\ a^{19}t^{53}g^{100},\ a^{19}t^{72}g^{100},$$

$$a^{19}t^{91}g^{100},\ a^{19}t^{110}g^{100},\ \mathbf{a^{129}t^{19}c^{100}},\ \mathbf{a^{148}t^{129}g^{1000}},\ldots,$$

after which the growth is again linear before the next complementarity transition.

This example is different from the one in the preceding section. Here also the linear growth becomes faster, along with the exponential growth in c. Although the word sequence is very complicated from the point of view of L systems, the growth function of W is still probably a $D0L$ growth function. However, we do not enter here any discussion about the growth functions and other properties of plain Watson-Crick $D0L$ systems.

Let us have a look at one more example. Let us consider the plain Watson-Crick $D0L$ system W with the axiom ac and productions

$$a \to at,\ g \to g,\ t \to t, c \to c^3.$$

Again, we use purines (a and g) as selectors. The word sequence of W is (we stick with the convention of using bold characters whenever a complementarity transition occurs in the generated sequence):

$$ac,\ \mathbf{tcg^3},\ \mathbf{ag^3c^3},\ \mathbf{tc^4g^9},\ \mathbf{ag^{12}c^9},\ \mathbf{tc^{13}g^{27}}, \ldots$$

Observe that at each rewriting step a complementarity transition takes place in the fragment of W's sequence pointed out above! We can actually prove that this happens for *any* rewriting step, in the (infinite) word sequence of W.

Let $w = tc^ig^j$ be an arbitrary word satisfying $i+1 < j \leq 3i$. Then $f(w) = tc^{3i}g^j$, and since $|f(w)|_{\{a,g\}} = j < 3i+1 = |f(w)|_{\{c,t\}}$, a complementarity transition has to occur when rewriting w according to the productions of W. That is, if w is a word in the sequence of W, then the word following w will be $w' = ag^{3i}c^j$. Now $f(w') = ag^{3i+1}c^{3j}$, with $|f(w')|_{\{a,g\}} = 3i+2 < 3j = |f(w')|_{\{c,t\}}$, whence the word following w' in a word sequence of W has to be the complementary of $f(w')$, i.e. $w'' = tc^{3i+1}g^{3j}$. Denoting $k = 3i+1,\ l = 3j$, then $w'' = tc^kg^l$, and, moreover, k and l satisfy the same conditions as i and j satisfy in w, that is, $k+1 < l \leq 3k$. This fact, together with the observation that the second word generated in the sequence of W is of the form tc^ig^j, for $i = 1,\ j = 3$, and with $i+1 < j \leq 3i$, proves our claim that a complementarity transition takes place at any rewriting step in generating the word sequence of W.

References

[1] L. M. Adleman, Molecular computation of solutions to combinatorial problems. *Science*, 266 (1994), 1021-1024

[2] J. Engelfriet, Reverse twin shuffles. *Bulletin of EATCS* 60 (1996), 144

[3] J. Engelfriet, G. Rozenberg, Fixed point languages, equality languages and representation of recursively enumerable languages. *J. Assoc. Comput. Mach.* 27 (1980), 499-518

[4] J. Hartmanis, On Weight of Computations. *EATCS Bulletin* 55 (1995), 136-138

[5] L. Kari, DNA computing: the arrival of biological mathematics. *Mathematical Intelligencer*, to appear

[6] V. Mihalache, Prolog Approach to DNA Computing. *Proceedings IEEE ICEC'97*, Indianapolis, 1997, 249-254

[7] G. Păun, A. Salomaa, DNA computing based on the splicing operation. *Mathematica Japonica* 43 (1996), 607-632

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

[9] G. Rozenberg, A. Salomaa, Watson-Crick complementarity, universal computations and genetic engineering. *Leiden University, Computer Science Technical Report* 96-28 (1996)

[10] G. Rozenberg, A. Salomaa (ed.), *Handbook of Formal Languages, Vol. 1-3.* Springer-Verlag, Berlin, Heidelberg, New York (1997)

[11] A. Salomaa, *Jewels of Formal Language Theory.* Computer Science Press, Rockville, Md. (1981)

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**