

Note

Note on the topological structure of random strings*

Cristian Calude and Cezar Câmpeanu

Department of Mathematics, University of Bucharest, Romania

Communicated by A. Salomaa

Received July 1991

Revised April 1992

Abstract

Calude, C. and C. Câmpeanu, Note on the topological structure of random strings, *Theoretical Computer Science* 112 (1993) 383–390.

A string x is *random* according to Kolmogorov [10] if, given its length, there is no string y , sensibly shorter than x , by means of which a universal partial recursive function could produce x . This remarkable definition has been validated in several ways (see [12, 14, 2, 11]), including a topological one [13].

Our present aim is to develop a constructive topological analysis of the “size” of the set of random strings in order to show to what extent they are incompressible. A substring of an incompressible string can be compressible [11] (conforming a well-known fact from probability theory: every sufficiently long binary random string must contain long runs of zeros). The converse operation makes sense and we may ask the question: can a compressible string be “padded” in order to be a substring of a random string? The answer depends upon the way we “pad” the initial string: for instance, if we add only arbitrary long prefixes (suffixes), then the answer is no, but if we pad from both directions, the answer is yes.

1. Preliminaries

The set of natural numbers will be denoted by $\mathbb{N} = \{0, 1, 2, \dots\}$. We work with a finite alphabet $X = \{a_1, a_2, \dots, a_p\}$, with $p \geq 2$ elements. The free monoid generated

Correspondence to: C. Calude, Computer Science Department, The University of Auckland, Private Bag 92019, Auckland, New Zealand. Email: c_calude@cs.auckland.ac.nz.

* Research partially done during the visit of the first author at Turku University, as a guest of the Academy of Finland.

by X under concatenation is X^* (λ is the empty string). The length of a string $x = x_1 x_2 \dots x_n$ is $l(x) = n$ ($l(\lambda) = 0$). The set X^* is quasi-lexicographically ordered by $\lambda < a_1 < a_2 < \dots < a_p < a_1 a_1 < \dots < a_1 a_p < \dots$; let $y(n)$ be the n th string in this order.

For recursion function theory and general notation see [2]. For every partial recursive (p.r.) function $\phi: X^* \times \mathbb{N} \xrightarrow{0} X^*$ we define the Kolmogorov complexity induced by ϕ to be the function $K_\phi: X^* \times \mathbb{N} \rightarrow \mathbb{N} \cup \{\infty\}$ defined by $K_\phi(x|m) = \min\{l(z) \mid z \in X^*, \phi(z, m) = x\}$ ($\min \emptyset = \infty$). Fix a *universal Kolmogorov algorithm* $\psi: X^* \times \mathbb{N} \xrightarrow{0} X^*$, i.e. a p.r. function such that for every p.r. function ϕ there exists a natural c (depending upon ψ and ϕ) such that $K_\psi(x|m) \leq K_\phi(x|m) + c$, for all $x \in X^*$, $m \in \mathbb{N}$ (Kolmogorov's Theorem); denote by K the complexity K_ψ . A string $x \in X^*$ is called t -random (with respect to ψ) if $K(x|l(x)) \geq l(x) - t$ (here $t \in \mathbb{N}$). The 0-random strings are called *random strings*. The set of t -random strings is denoted by RAND_t .

For all natural $n \geq t \geq 0$, one has

$$\text{card}\{x \in X^* \mid l(x) = n, K(x|n) \geq n - t\} \geq p^n(1 - p^{-t}) / (p - 1) > 0$$

(see [3, 2]). Consequently, there exist random strings of every length and, moreover, from a quantitative point of view, most strings of fixed length are t -random ($t \geq 0$); see [2] for other estimations.

Let $<$ be a partial order on X^* which is recursive, i.e. " $u < v$ " is a binary recursive predicate. Denote by $\tau(<)$ the topology generated by the family $(U_w)_{w \in X^*}$, $U_w = \{x \in X^* \mid w < x\}$. Note that $\tau(<)$ is a T_0 -space (which is not T_1 in case it is not trivial). The closure operator in this space acts as follows: $A \subset X^*$, $\bar{A} = \{x \in X^* \mid x < z, \text{ for some } z \in A\}$. For every $A \subset X^*$ and $w \in X^*$ the following three statements are equivalent: (i) $A \cap U_w = \emptyset$, (ii) $\bar{A} \cap U_w = \emptyset$, (iii) $w \notin \bar{A}$. A set $A \subset X^*$ is *dense* if $\bar{A} = X^*$. See [9] for more topological facts.

2. Results

In a topological space, a set A is rare if its closure contains no nonempty open set. So, a set A in $\tau(<)$ is rare if $U_w \not\subset \bar{A}$, for every $w \in X^*$. A set A is *recursively rare* if for every $w \in X^*$ we can obtain, in a recursive way, a witness which certifies that $U_w \not\subset \bar{A}$, i.e. a string $w < v$, $v \notin \bar{A}$. Thus, we obtain the following definition inspired by [1] (and used in [13] in case of the prefix order).

Definition 2.1. A set $A \subset X^*$ is *recursively rare* if there exists a recursive function $r: \mathbb{N} \rightarrow \mathbb{N}$ such that the following two conditions hold for all $n \geq 0$:

- (1) $y(n) < y(r(n))$,
- (2) $A \cap U_{y(r(n))} = \emptyset$.

Remark. The family of recursively rare sets is closed under subset. Every recursively rare set is rare.

Example 2.2. Each basic neighborhood U_w is not (recursively) rare.

Remark. Let $A \subset X^*$. The following assertions are equivalent: (i) A is recursively rare, (ii) \bar{A} is recursively rare, and (iii) there exists a recursive function $r: \mathbb{N} \rightarrow \mathbb{N}$ such that for all natural $n \geq 0$, $y(n) < y(r(n))$ and $y(r(n)) \notin A$.

Definition 2.3. A partial order $<$ on X^* is *unbounded* if for every $x \in X^*$ and every natural $n > l(x)$, there exists a string y of length $l(y) \geq n$ such that $x < y$.

Example 2.4. The following partial orders on X^* are unbounded and recursive (here $w = w_1 w_2 \dots w_n$, $l(w) = n$ and $v = v_1 v_2 \dots v_m$, $l(v) = m$; $a_1 < a_2 < \dots < a_p$ is the order on X):

- (1) $w <_p v$ iff $v = wu$, for some $u \in X^*$ (prefix order),
- (2) $w <_s v$ iff $v = uw$, for some $u \in X^*$ (suffix order),
- (3) $w <_i v$ iff $v = xwu$, for some $x, u \in X^*$ (infix order),
- (4) $w <_h v$ iff $v = u_1 w_1 u_2 \dots u_n w_n u_{n+1}$, for some $u_1, u_2, \dots, u_{n+1} \in X^*$ (embedding order),
- (5) $w <_m v$ iff $w_{n-i} < v_{m-i}$, for all $0 \leq i \leq \min(m, n) - 1$ and if $n > m$, then $w_j = a_1$, for all $1 \leq j \leq n - m$ (masking order),
- (6) $w <_{pm} v$ iff $l(w) \leq l(v)$ and $w_i < v_i$, for all i , $1 \leq i \leq l(w)$ (prefix-masking order),
- (7) $w <_d v$ iff $w <_p v$ and $w <_s v$ (2-ps-codes order),
- (8) $w <_1 v$ iff $w <_p v$ or $w = xa_i y$, $v = xa_j z$ with $i < j$, for some $x, y, z \in X^*$ (lexicographical order). \square

Remark. See [8, 7] for relevance of the above partial orders.

Example 2.5. If $<$ is a partial recursive (unbounded) order on X^* and $f: X^* \rightarrow X^*$ is a recursive bijection, then the partial order: $x <_f y$ iff $f(x) < f(y)$, is recursive (unbounded). For instance, $<_s$ is obtained from $<_p$ using the mirror function $\text{mir}: X^* \rightarrow X^*$, $\text{mir}(\lambda) = \lambda$, $\text{mir}(x) = x$, $x \in X$, $\text{mir}(xy) = \text{mir}(y) \text{mir}(x)$, $x \in X^*$, $y \in X$.

Proposition 2.6. Assume that $<$ is recursive and unbounded. A set $A \subset X^*$ is recursively rare iff there exist a natural i and a recursive function $f: \mathbb{N} \rightarrow \mathbb{N}$ such that $y(n) < y(f(n))$, for every $n \in \mathbb{N}$ and $U_{y(f(n))} \cap A = \emptyset$, for all strings with $l(y(n)) > i$.

Proof. Define the recursive function $q: \mathbb{N} \rightarrow \mathbb{N}$ by $q(n) = \min\{m \geq 0 \mid y(n) < y(m) \text{ and } l(y(m)) > i\}$. Take $r = f \circ q$. Clearly, $y(n) < y(q(n)) < y(f(q(n))) = y(r(n))$. Finally, $l(y(q(n))) > i$ implies $U_{y(r(n))} \cap A = U_{y(f(q(n)))} \cap A = \emptyset$. \square

Theorem 2.7. Assume that $<$ is recursive and unbounded and suppose that there exists a recursive function $s: \mathbb{N} \rightarrow X^*$ such that

- (*) for all natural i, j , if $s(i) < x$, $s(j) < x$, for some string x , then $i = j$;

then we can find a rare set which is not recursively rare.

Proof. Let $(\phi_n)_{n \geq 0}$, $\phi_n: \mathbb{N} \overset{0}{\rightarrow} \mathbb{N}$ be an acceptable Gödel numbering of the unary p.r. functions. Define the set $A = \{y(t_n) \mid n \geq 0\}$, where t_n is defined only in case $\phi_n(n) \neq \infty$ and $t_n = \min \{j \in \mathbb{N} \mid s(n) < y(j), l(y(j)) \geq l(s(n)) + \phi_n(n)\}$.

The set A is rare. Assume, by *reductio ad absurdum*, that $U_x \subset \bar{A}$, for some $x \in X^*$. So, there exists $n \geq 0$ such that $x < y(t_n)$, $s(n) < y(t_n)$, $l(y(t_n)) \geq l(s(n)) + \phi_n(n)$. Pick a string z with $y(t_n) < z$ and $l(z) > l(y(t_n))$. Clearly, $z \in U_x$. We shall prove that $z \notin \bar{A}$, a contradiction. For, if $z \in \bar{A}$, there exists $m \geq 0$ and w such that $s(m) < w$, $z < w$, $l(w) \geq l(s(m)) + \phi_m(m)$ and w is the least string (according to the quasi-lexicographical order) having the above properties. So, $s(n) < y(t_n) < z < w$, $s(m) < w$; by (*), $n = m$, i.e. $l(y(t_n)) = l(z)$.

Next we prove that A is not recursively rare. Again we proceed by *reductio ad absurdum*. Suppose that, for all $n \geq 0$, $y(n) < y(r(n))$ and $A \cap U_{y(r(n))} = \emptyset$, for some fixed recursive function $r: \mathbb{N} \rightarrow \mathbb{N}$. Let $f, g: \mathbb{N} \rightarrow \mathbb{N}$ be the recursive functions given by $y(f(n)) = s(n)$ and $g(n) = l(y(r(f(n)))) - l(y(f(n)))$.

First note that $y(r(f(n))) \notin A$, for all $n \geq 0$, since $y(r(f(n))) \in U_{y(r(f(n)))}$ and $A \cap U_{y(r(f(n)))} = \emptyset$.

Secondly, $g(n) \neq \phi_n(n)$, for all $n \geq 0$. If $g(n) = \phi_n(n)$, for some $n \geq 0$, then choose the least $j \geq 0$, with $s(n) < y(j)$ and $l(y(j)) \geq l(s(n)) + \phi_n(n) = l(y(r(f(n))))$; one has $y(j) = y(r(f(n)))$; so, $y(r(f(n))) \in A$.

Finally, $g = \phi_i$, for some $i \geq 0$; since g is total, one has $g(i) = \phi_i(i) \neq \infty$, a contradiction. \square

Example 2.8. (a) The prefix and suffix orders satisfy the hypothesis of Theorem 2.7. For instance, in case of suffix order take $s(i) = a_1 a_2^i$. (b) If $<$ is a partial recursive, unbounded order having the property (*) with respect to s , then the partial recursive order $<_f$ has the same property for $f^{-1} \circ s$.

Remark. Theorem 2.7 was proved in [13] for the prefix order.

Proposition 2.9. Assume that $<$ is recursive, unbounded and for all strings x, y there exists a string z with $x < z$ and $y < z$. Then (i) each rare set is recursively rare and (ii) every nonrare set is dense.

Proof. (i): Let $z \in X^*$ with $U_z \cap A = \emptyset$ and define the recursive function $f: \mathbb{N} \rightarrow \mathbb{N}$ by $f(n) = \min \{i \in \mathbb{N} \mid z < y(i) \text{ and } y(n) < y(i)\}$. Clearly, $y(n) < y(f(n))$ and $U_{y(f(n))} \cap A \subset U_z \cap A = \emptyset$.

(ii): Let $A \subset X^*$ be a nonrare set, i.e. $U_w \subset \bar{A}$ for some $w \in X^*$. Take $x \in X^*$ and pick a string y such that $w < y$ and $x < y$. One has $x \in U_y \subset \bar{U}_w \subset \bar{A} = \bar{A}$. \square

Example 2.10. The infix, embedding, masking and prefix-masking orders satisfy the hypothesis of Proposition 2.9.

Theorem 2.11. *Let $<$ be recursive and unbounded. Then, there exists a natural $c > 0$ such that for all naturals m and d , with $d \geq c$, the set*

$$A(m, d) = \{x \in X^* \mid l(x) \geq m, K(x \mid l(x)) \leq d\}$$

is dense.

Proof. Define the recursive function $f: \mathbb{N} \rightarrow \mathbb{N}$ by $f(n) = \min\{i \geq 0 \mid l(y(i)) \geq n, y(n) < y(i)\}$. Put $B(m) = \{y(f(n)) \mid n \geq m\}$ and construct the p.r. function $\phi: X^* \times \mathbb{N} \xrightarrow{0} X^*$, $\phi(x, l(f(n))) = y(f(n))$, for all $x \in X^*$ and $n \geq m$.

Clearly, $K_\phi(y(f(n)) \mid l(y(f(n)))) = 0$, for all $n \geq m$; so, according to Kolmogorov's Theorem, there exists a constant $c > 0$ such that $K(y(f(n)) \mid l(y(f(n)))) \leq c$, for all $n \geq m$.

Next we show that for every $d \geq c$, $B(m) \subset A(m, d)$. Indeed, if $n \geq m$, then $l(y(f(n))) \geq n \geq m$ and $K(y(f(n)) \mid l(y(f(n)))) \leq c \leq d$.

Finally, to prove that $\overline{B(m)} = X^*$ we show that for every $x \in X^*$ there exists $n \geq m$ such that $x < y(f(n))$. If $x = y(k)$, $k \geq m$, take $n = k$ (since $x < y(f(k))$, $k \geq m$). If $x = y(k)$ with $k < m$, then take $y(i)$ with $x < y(i)$ and $l(y(i)) \geq m$: $x < y(i) < y(f(i))$ and $i \geq m$. \square

Corollary 2.12. *For every natural $t \geq 0$, $\text{non-RAND}_t = \{x \in X^* \mid K(x \mid l(x)) < l(x) - t\}$ is dense in case $<$ is recursive and unbounded.*

Proof. For every $d \geq 0$, $A(1 + d + t, d) \subset \text{non-RAND}_t$ (here $A(1 + d + t, d)$ comes from Theorem 2.11). Pick $d \geq c$, where c also comes from Theorem 2.11. \square

Remarks. (a) A stronger form of the above statement can be easily obtained: for every increasing, unbounded (not necessarily recursive) function $f: \mathbb{N} \rightarrow \mathbb{N}$, the set $T(f) = \{x \in X^* \mid K(x \mid l(x)) \leq f(l(x))\}$ is dense. Indeed, pick a natural D such that $f(m) > d$ whenever $m \geq D$ (here d comes from Theorem 2.11). If $x \in X^*$, $l(x) \geq D$, then $f(l(x)) > d$; so, $A(D, d) \subset T(f)$ a.s.o.

(b) We can interpret Corollary 2.12 as follows: each section of the universal Martin-Löf test $V(\psi) = \{(x, m) \in X^* \times \mathbb{N} \mid K(x, l(x)) < l(x) - m\}$ is dense: see, for details, [5, 2].

So, every set non-RAND_t is "large" with respect to all topologies considered in Examples 2.4 and 2.5 (for unbounded $<$). Now we pass to the study of RAND_t .

A routine verification shows the validity of Lemma 2.13.

Lemma 2.13. *A set $A \subset X^*$ is rare (recursively rare, dense) in $\tau(<)$ iff $f(A) = \{f(x) \mid x \in A\}$ is rare (recursively rare, dense) in $\tau(<_f)$, where $f: X^* \rightarrow X^*$ is recursive and bijective.*

Corollary 2.14. *For every $t \geq 0$, RAND_t is recursively rare in $\tau(<_p), \tau(<_s), \tau(<_d)$.*

Proof. The first part comes from [13, Theorem 4]; the second follows from Lemma 2.13 and Example 2.5. For the third let rs and rp be the recursive functions satisfying Definition 2.1(1) and 2.1(2) for RAND_t in $\tau(<_p), \tau(<_s)$, respectively; the recursive function $r(n) = \min\{k \geq 0 \mid y(rp(n)) <_p y(k) \text{ and } y(rs(n)) <_s y(k)\}$ will work for RAND_t in $\tau(<_d)$. \square

Proposition 2.15. *For every $t \geq 0$, RAND_t is recursively rare in $\tau(<_m)$.*

Proof. Define the recursive function $f: \mathbb{N} \rightarrow \mathbb{N}$ by $y(f(n)) = a_p^{l(y(n))}$ and the p.r. function $\phi: X^* \times \mathbb{N} \xrightarrow{0} X^*$, $\phi(x, n) = xa_p^{n-l(x)}$, in case $n \geq l(x)$.

Let $i > t + c$, where c comes from Kolmogorov's Theorem applied to ψ and ϕ . Note that $y(n) <_m y(f(n))$; every $w \in U_{y(f(n))}$ with $l(y(n)) > i$, can be written as $w = xy(f(n))$, for some $x \in X^*$. One has $K(w \mid l(w)) \leq K_\phi(w \mid l(w)) + c \leq l(w) - l(y(f(n))) + c = l(w) - l(y(n)) + c < l(w) - l(y(n)) + i - t < l(w) - t$; so, $w \notin \text{RAND}_t$, i.e. $U_{y(f(n))} \cap \text{RAND}_t = \emptyset$. The result follows from Proposition 2.6. \square

Proposition 2.16. *For every $t \geq 0$, RAND_t is recursively rare in $\tau(<_{pm})$.*

Proof. Use the partial recursive function $\phi: X^* \times \mathbb{N} \xrightarrow{0} X^*$, $\phi(x, n) = a_p^{n-l(x)}x$, if $n \geq l(x)$, in a similar construction as that displayed in the proof of Proposition 2.15. \square

Lemma 2.17. *Assume that $<$ is a recursive partial order on X^* and let $A \subset X^*$. If for every $x \in X^*$ we can find a natural m and a string w , such that $x < w$ and $\text{card}\{y \in X^* \mid l(y) = m, w < y\} > \text{card}\{z \in X^* \mid l(z) = m, z \notin A\}$, then A is dense.*

Proof. Given a string x we can find m and w with the above properties. Accordingly, there exists $y \in A$, with $w < y$. Since $x < w$, it follows that $x < y$, i.e. $x \in \bar{A}$. \square

Corollary 2.18. *Let $t \geq 0$. If for every string x there exists a natural m and a string w , such that $x < w$ and $\text{card}\{y \in X^* \mid l(y) = m, w < y\} \cdot (p-1) \geq p^{m-t}$, then $\overline{\text{RAND}_t} = X^*$.*

Proof. It is known (see [2]) that $\text{card}\{y \in X^* \mid l(y) = m, K(y \mid m) < m - t\} \leq (p^{m-t} - 1)/(p - 1)$. \square

Theorem 2.19. *If $p > 2$ or $t > 0$, then RAND_t is dense with respect to the infix order.*

Proof. Recall that $p = \text{card } X$. We shall use Corollary 2.18; the proof will be divided into several steps.

A string $x \in X^*$ is called *unbordered* if for all strings y, z with $y \neq \lambda, x \neq yz$ [6] (unbordered strings are called variate in [4]).

Fact 2.20. Let x be an unbordered string of length $n \geq 3$. Let m be natural. Put $R(m, x) = p^m \cdot \text{card} \{y \in X^* \mid l(y) = m, x <_i y\}$. Then

$$R(m, x) = p^m, \quad 0 \leq m < n,$$

$$R(m+1, x) = p \cdot R(m, x) - R(m+1-n, x), \quad m \geq n.$$

Fact 2.21. For every unbordered string x of length $n \geq 3$ there is a natural M such that for every $m \geq M$, $R(m^2, x) < p^{m^2-m}/(p-1)$.

See [4] for the proofs of Facts 2.20 and 2.21.

Now, given a string x , we construct the unbordered string $v(x) = a_1^{l(x)} x a_2^{l(x)}$, $x <_i v(x)$. We shall prove the existence of a natural m such that $\text{card} \{y \in X^* \mid l(y) = m, y <_i v(x)\} \cdot (p-1) \geq p^{m-t}$, the condition required by Corollary 2.18 in order to assure that RAND_t is dense.

From Fact 2.21 it follows that, for every $i \geq M$,

$$R(i^2, v(x)) < p^{i^2-i}/(p-1).$$

Take $m \geq \max(M, t)$. The required inequality becomes

$$p^{m^2-m}/(p-1) \leq p^m(1 - 1/p^t(p-1)),$$

which is true in case $p > 2$ or $t > 0$. \square

Open problem. Is RAND dense with respect to the infix order in the binary case? In view of Proposition 2.9, RAND is rare or dense.

Corollary 2.22. For every $t \geq 0$. RAND_t is dense with respect to the uniform and embedding orders.

Proof. If $w <_i v$, then $w <_u v$ ($w = v$ or $l(w) < l(v)$) and $w <_h v$. \square

Final comment. For every string x we can construct a context (u, v) such that uxv is t -random, whereas there exist strings y and z such that uy (respectively, zv) are not t -random, for all strings u and v .

References

- [1] C. Calude, Topological size of sets of partial recursive functions, *Z. Math. Logik Grundlag. Math.* **32** (1982) 81–88.
- [2] C. Calude, *Theories of Computational Complexity* (North-Holland, Amsterdam, 1988).
- [3] C. Calude and I. Chitescu, Random strings according to A.N. Kolmogorov. Classical approach, *Found. Control Engrg.* **7** (1982) 73–85.
- [4] C. Calude and I. Chitescu, Qualitative properties of P. Martin-Löf random sequences, *Bolletino Un. Mat. Ital. B* **3** (1989) 229–240.

- [5] C. Calude, I. Chitescu and L. Staiger, P. Martin-Löf tests: representability and embeddability, *Rev. Roumaine Math. Pures Appl.* **30** (1985) 719–732.
- [6] A. Ehrenfeucht and G. Rozenberg, Each regular code is included in a maximal regular code, *RAIRO Theor. Inform. Appl.* **20** (1985) 89–96.
- [7] J.P. Jones and Y.V. Matijasevič, Register machine proof of the theorem on exponential diophantine representation of enumerable sets, *J. Symbolic Logic* **49** (1984) 818–829.
- [8] H. Jürgensen and S.S. Yu, Relations on free monoids, their independent sets and codes, Report No. 257, Dept. of Computer Science, Univ. of Western Ontario, 1989.
- [9] J. Kelley, *General Topology* (Van Nostrand, Princeton, NJ, 1968).
- [10] A.N. Kolmogorov, Three definitions for defining the concept of information quantity, *Problemy Peredachi Informatsii* **1** (1965) 3–11.
- [11] M. Li and P.M. Vitanyi, Kolmogorov complexity and its applications, in: J. van Leeuwen, ed., *Handbook of Theoretical Computer Science, Vol. A* (Elsevier, Amsterdam, The MIT Press, Cambridge, MA, 1990) 189–254.
- [12] P. Martin-Löf, The definition of random sequences, *Inform. and Control* **9** (1966) 602–619.
- [13] M. Zimand, On the topological size of random strings, *Z. Math. Logik Grundlag. Math.* **32** (1986) 81–88.
- [14] A. Zvonkin and L. Levin, The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms. *Uspekhi Mat. Nauk* **156** (1970) 85–127.