Finite-State Randomness

C. S. Calude (UoA)
Joint work with K. Salomaa (Kingston), T. K. Roblot (UoA)

UWO, London Ontario, 2010

Goal

The incomputability of all descriptional complexities is an obstacle towards more “down-to-earth” applications of AIT (e.g. for practical compression).

We develop a version of AIT by replacing Turing machines with finite transducers; the complexity induced is called finite-state complexity.

A transducer is denoted as a triple $T = (Q, q_0, \Delta)$ where Q is the finite set of states, $q_0 \in Q$ is the start state, (all states of Q are considered to be final), and

$$\Delta : Q \times \{0,1\} \to Q \times \{0,1\}^*$$

is the transition function.

The function $\{0,1\}^* \to \{0,1\}^*$ computed by the transducer T is defined by

$$T(\varepsilon) = \varepsilon,\ T(xa) = T(x) \cdot \pi_2(\Delta(\hat{\delta}(q_0, x), a)),$$

for $x \in \{0,1\}^*$, $a \in \{0,1\}$.

Theorem. *The set of all transducers can be enumerated by a regular language. i.e. there is a regular set S_0 such that:*

a) for every $\sigma \in S_0$, $T_\sigma^{S_0}$ is a transducer,

b) for every transducer T one can compute a code $\sigma \in S_0$ such that $T = T_\sigma^{S_0}$.

Theorem. *There is no universal transducer.*

Fix a computable enumeration S of transducers.
A pair (T_σ^S, p), $\sigma \in S$, $p \in \{0,1\}^*$, defines the string x provided

$$T_\sigma^S(p) = x.$$

The *finite-state complexity* (with respect to S) of a string $x \in \{0,1\}^*$ is defined by:

$$C_S(x) = \inf_{\sigma \in S,\, p \in \{0,1\}^*} \Big\{ \mid \sigma \mid + \mid p \mid : T_\sigma^S(p) = x \Big\}.$$

The complexity associated with a transducer T_σ^S is defined by

$$C^S_{T^S_\sigma}(x) = \inf_{p \in \{0,1\}^*} \Big\{ \mid p \mid : T^S_\sigma(p) = x \Big\}.$$

Theorem. For every $\sigma \in S$, $C_S(x) \leq C_{T^S_\sigma}(x) + |\sigma|$, for all x.

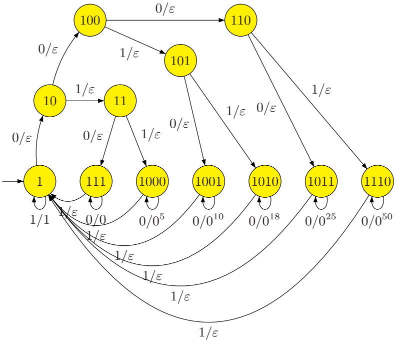
Corollary If $T^S_{\sigma_0}(x) = x$, then $C_S(x) \leq |x| + |\sigma_0|$, for all x. We can take $|\sigma_0| = 8$ for C_{S_0}.

Corollary The complexity C_S is computable.

Conjecture The predicate $C_S(x) \leq n$ is NP-complete.

An example of compressible string

The string $x = 010^2 10^3 1 \cdot \ldots \cdot 0^{99} 10^{100} 1$, of length 5150 can be represented as $x = T_\sigma^{S_0}(p)$ where $|\sigma| = 352$, $|p| \leq 2008$, so $C_{S_0}(x) \leq 2172$.



Computational results

x	$C_{S_0}(x)$	(σ, p)	x	$C_{S_0}(x)$	(σ, p)
ε	4	(0000,ε)	00000	11	(000110,11111)
0	7	(000110,1)	00001	13	(01000110,11110)
00	8	(000110,11)	00010	13	(01000110,11101)
01	9	(00011100,1)	00011	13	(01000110,11100)
000	9	(000110,111)	00100	13	(01000110,11011)
001	11	(01000110,110)	00101	13	(01000110,11010)
010	11	(01000110,101)	00110	13	(01000110,11001)
011	11	(01000110,100)	00111	13	(01000110,11000)
0000	10	(000110,1111)	01000	13	(01000110,10111)
0001	12	(01000110,1110)	01001	13	(01000110,10110)
0010	12	(01000110,1101)	01010	13	(01000110,10101)
0011	12	(01000110,1100)	01011	13	(01000110,10100)
0100	12	(01000110,1011)	01100	13	(01000110,10011)
0101	10	(00011100,11)	01101	13	(01000110,10010)
0110	12	(01000110,1001)	01110	13	(01000110,10001)
0111	12	(01000110,1000)	01111	13	(01010110,10000)

Table: Finite-state complexity and minimal descriptions (for S_0) of all strings in lexicographic order from ε to 01111.

Theorem For $n \geq 1$, $C(0^n) \in \Theta(\sqrt{n})$.

Corollary For all x, $C(x) \geq 2 \cdot \lfloor \sqrt{|x|} \rfloor$.

Corollary There is no constant c such that for all strings x, $C(xx) \leq C(x) + c$.

Proposition There exists c such that for all $u \in \{0,1\}^*$ and n,

$$C(u^n) \leq 2 \cdot (\lfloor \sqrt{n} \rfloor + 1) \cdot |u| + 2\sqrt{n} + c.$$

Corollary We have: $C(0^n 1^n) \in \Theta(\sqrt{n})$.

A string x is *finite-state i–incompressible* if $C(x) > |x| - i$.

Fact Finite-state incompressible strings of any length exist.

A binary *de Bruijn word* of order $r \geq 1$ is a binary string w of length $2^r + r - 1$ such that any binary string of length r occurs as a substring of w (exactly once); de Bruijn words of any order exist.

Theorem There is a constant d such that for any $r \geq 1$ there exist strings w of length $2^r + r - 1$ with an explicit construction such that

$$C(w) \geq \frac{d \cdot |w|}{\log(|w|)}.$$

Conjecture de Bruijn words are finite-state incompressible.

Theorem For any $n > 0$ there exists a string $x_n \in \{0,1\}^*$ such that whenever $C_{S_0}(x_n) = |\sigma| + |p|$, the transducer $T_\sigma^{S_0}$ has more than n states.

By $L_{\leq m}^{S_0}$ ($L_{=m}^{S_0}$) we denote the language of all strings for which a minimal description uses a transducer with at most (exactly) $m \geq 1$ states.

Corollary For any $n \geq 1$, there exists effectively $k_n \geq 1$ such that $L_{\leq n}^{S_0} \subset L_{\leq n+k_n}^{S_0}$.

Conjecture $L^{S_0}_{\leq n} \subset L^{S_0}_{\leq n+1}$, for all $n \geq 1$.

Problem What is asymptotically the length of the shortest words in $L^{S_0}_{=n}$ as a function of n?

Problem Are there strings with two minimal descriptions for which the respective transducers have different numbers of states?

Problem How "robust" is C_S when S varies?

Problem What are the relations between finite-state complexities for different size alphabets?

Conjecture
Finite-state infinite sequences (defined via C) are exactly the Borel-normal sequences.

References

1. C. S. Calude, K. Salomaa, T. K. Roblot. Finite-state complexity and the size of transducers, in I. McQuillan and G. Pighizzini (eds.). *12th International Workshop on Descriptional Complexity of Formal Systems (DCFS 2010)* EPTCS 26, 2010.
2. C. S. Calude, K. Salomaa, T. K. Roblot. Finite-state complexity and randomness, in F. Ferreira, H. Guerra, E. Majordomo, J. Rasga (eds). *Programs, Proofs, Processes, 6th Conference on Computability in Europe, CiE 2010, Abstract and Handbook Booklet*, Ponta Delgada, Azores, Portugal, University of Azores, 2010, 73–82.