Complexity of words

Zoltán Kása
Babeş-Bolyai University, Cluj
E-mail: `kasa@cs.ubbcluj.ro`
`http://www.cs.ubbcluj.ro/~kasa`

---

finite word over $\mathcal{A}$:

$$w = w_1 w_2 \ldots w_N, \quad w_i \in \mathcal{A} \text{ for } 1 \le i \le N.$$

u *factor* or *subword* of w: $\exists x, y \in \mathcal{A}^*$: $w = xuy$

$F(w)$ the set of all nonempty factors of w
$F_n(w)$ the set of all factors of w of length n

subword complexity of w:

$$f_w(n) = \# F_n(w) \qquad \text{for } 1 \le n \le |w|$$

$w = abaaba$
$F_1(w) = \{a, b\}$, $F_2(w) = \{ab, aa, ba\}$,
$F_3(w) = \{aba, baa, aab\}$

infinite word:

$u = u_0 u_1 u_2 \dots u_n \dots, \quad u_i \in \mathcal{A}$

$f_u(n) = \# F_n(u)$ *subword complexity*

Ex.

1) Fibonacci word: $\sigma(0) = 01$, $\sigma(1) = 0$

0

01

010

01001

01001010 $\quad u_F = \underbrace{01001010}\underbrace{01001}\dots$

$f_{u_F}(n) = n + 1$

2) Power word:

$$u_p = 010011000111 \dots \underbrace{0 \dots 0}_{n}\underbrace{1 \dots 1}_{n} \dots$$

$$f_{u_p}(n) = \frac{n(n+1)}{2} + 1$$

3) Champarnowne word:

$u_C = 0\ 1\ 10\ 11\ 100\ 101\ 110\ 111\ \dots$

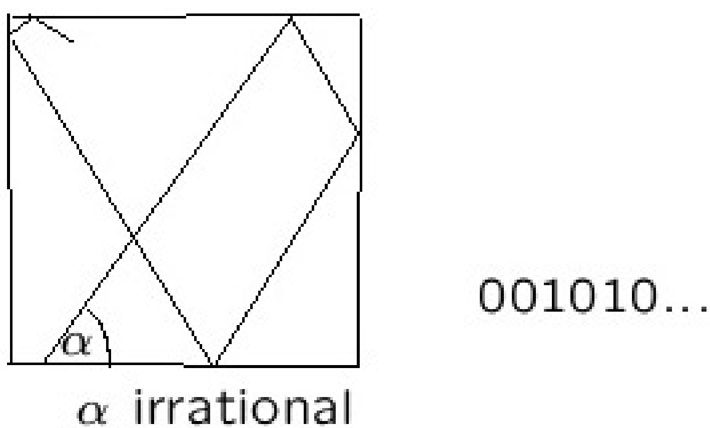
$f_{u_C}(n) = 2^n$

$\underbrace{abc}\underbrace{abc}\underbrace{abc}\ldots\underbrace{abc}\ldots$ *periodic*

$aaaaaba\,\underbrace{abc}\underbrace{abc}\underbrace{abc}\ldots\underbrace{abc}\ldots$ *ultimately periodic*

If $f_u(n) \leq n$ for all $n \geq n_0$ then u is ultimately periodic.

Sturmian words for which $f_u(n) = n + 1$.



3-dimensional case: $n^2 + n + 1$

S. Ferenczi, C. Mauduit: To an infinite word $u = u_0u_1u_2\ldots u_n\ldots$, $u_i \in \{0, 1\}$, we associate a real number $\theta = 0.u_0u_1\ldots u_n\ldots$ in base 2. *If u is Sturmian then θ is a trancendental number.*

Tribonacci word:

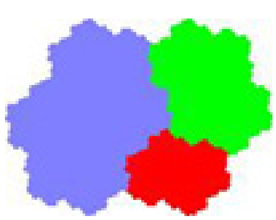
$\sigma(0) = 01, \ \ \sigma(1) = 02, \ \ \sigma(2) = 0$

$$
\begin{array}{l}
0 \\
01 \\
0102 \\
0102010 \\
\underbrace{0102010}\,\underbrace{0102}\,\underbrace{01} \\
\ldots
\end{array}
\qquad
\begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}
$$

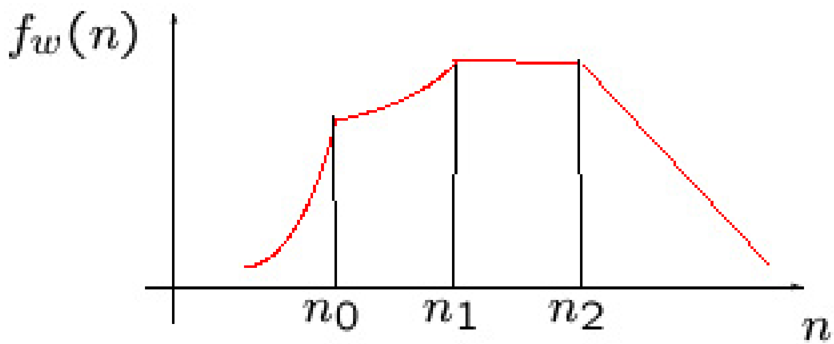
$$\begin{vmatrix} 1-x & 1 & 0 \\ 1 & -x & 1 \\ 1 & 0 & -x \end{vmatrix} = 0$$

characteristic polynomial, roots: $\beta > 1$, $\alpha, \overline{\alpha}$ complex

$$\mathcal{R} = \Big\{ \sum_{i \geq 0} \varepsilon_i \, \alpha^i ; \ \varepsilon_i = 0, 1; \ \varepsilon_i \varepsilon_{i+1} \varepsilon_{i+2} = 0 \Big\} \subset \mathbf{C}.$$



n	1	2	3	4	5
11111	1	1	1	1	1
11112	2	2	2	2	1
21122	2	4	3	2	1
21211	2	3	3	2	1
22112	2	4	3	2	1
22211	2	3	3	2	1



maximal complexity:

$C(w) = \max\{f_w(n) \mid n \geq 1\}$

global maximal complexity in $\mathcal{A}^N$:

$K(N) = \max\{C(w) \mid w \in \mathcal{A}^N\}$

$R(N) = \{i \in \overline{1, N} \mid \exists w \in \mathcal{A}^N : f_w(i) = K(N)\}$

$M(N)$: the number of words in $\mathcal{A}^N$ with maximal complexity equal to the global maximal complexity

N	$K(N)$	$R(N)$	$M(N)$
1	1	1	2
2	2	1	2
3	2	1, 2	6
4	3	2	8
5	4	2	4
6	4	2, 3	36
7	5	3	42
8	6	3	48
9	7	3	40
10	8	3	16
11	8	3, 4	558
12	9	4	718
13	10	4	854
14	11	4	920
15	12	4	956
16	13	4	960
17	14	4	912
18	15	4	704
19	16	4	256
20	16	4, 5	79006

Theorem 1.
If $\#\mathcal{A} = q$ and $q^k + k \le N \le q^{k+1} + k$ then $K(N) = N - k$.

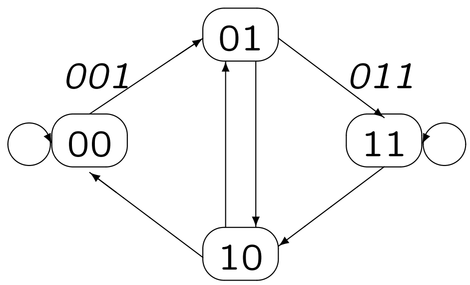
Theorem 2.
If $\#\mathcal{A} = q$ and $q^k + k < N < q^{k+1} + k + 1$ then $R(N) = \{k+1\}$;
if $N = q^k + k$ then $R(N) = \{k, k+1\}$.

De Bruijn graphs

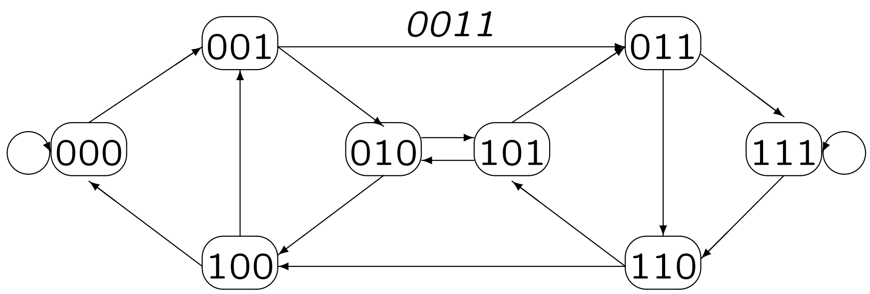
For a q-letter alphabet $\mathcal{A}$ the de Bruijn graph is defined as:

$$B(q,k) = (V(q,k), E(q,k))$$

with $V(q,k) = \mathcal{A}^k$ as the set of vertices, and $E(q,k) = \mathcal{A}^{k+1}$ as the set of directed arcs. There is an arc from $x_1x_2\ldots x_k$ to $y_1y_2\ldots y_k$ if $x_2x_3\ldots x_k = y_1y_2\ldots y_{k-1}$, and this arc is denoted by $x_1x_2\ldots x_k y_k$.



$B(2,2)$

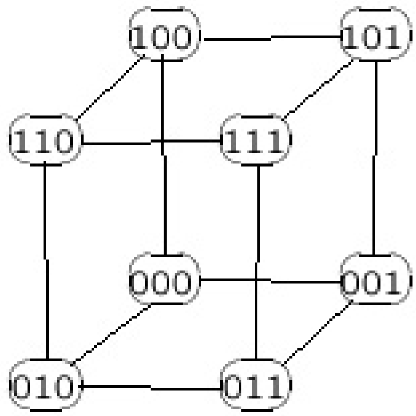


$B(2,3)$

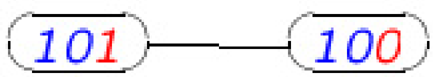
path 001, 011, 111, 110 $\Longrightarrow$ word *001110*
Hamilton path:
000, 001, 011, 111, 110, 101, 010, 100 $\Longrightarrow$ word *0001110100*

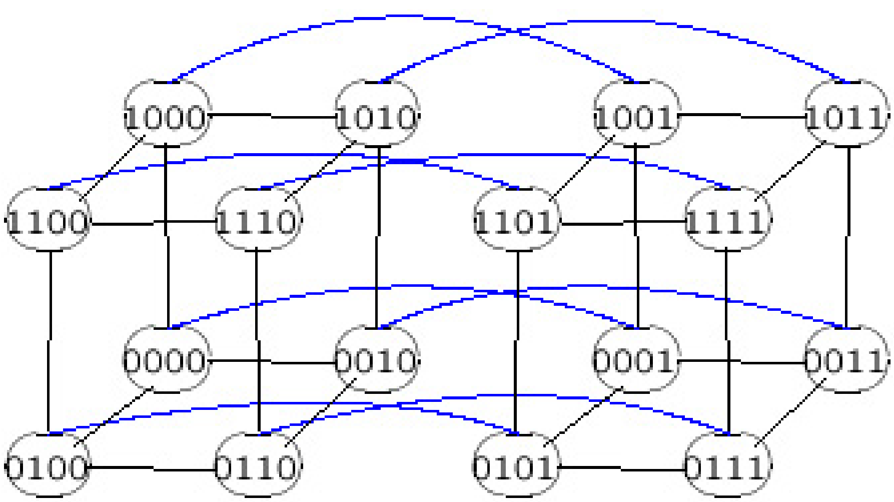
Every maximal length path in the graph $B(q,k)$ (which is a Hamiltonian one) corresponds to a de Bruijn word.

undirected de Bruijn graph as *network model*



3-dimensional hypercube





4-dimensional hypercube

Theorem 3.
If $\#\mathcal{A} = q$ and $q^k + k \le N \le q^{k+1} + k$ then $M(N)$ is equal to the number of different paths of length $N - k - 1$ in the de Bruijn graph $B(q, k + 1)$.

Theorem 4.
If $N = 2^k + k - 1$ then $M(N) = 2^{2^{k-1}}$.

The number of distinct Hamiltonian cycles in the de Bruijn graph $B(2, k)$ is equal to $2^{2^{k-1}-k}$. With each vertex of a Hamiltonian cycle a de Bruijn word (containing all the factors of length k) begins, which has maximal complexity, so $M(n) = 2^k \cdot 2^{2^{k-1}-k}$.

Generalization:
If $N = q^k + k - 1$ then $M(N) = (q!)^{q^{k-1}}$.

total complexity: $\mathbf{K}(w) = \sum_{i=1}^{|w|} f_w(i)$

- $C \neq 1, 2, 4$ then $\exists$ a nontrivial w such that $\mathbf{K}(w) = C$
- $C \neq 1, 2, 4, 6, 10, 18, 22$ then $\exists$ a nontrivial $w \in \mathcal{A}^*$, $\#\mathcal{A} = 2$, such that $\mathbf{K}(w) = C$

$|w| = 5$

C	5	6	7	8	9	10	11	12	13	14	15
$f_5(C)$	5	0	0	0	60	0	200	400	1140	1200	120

$$f_k(C) \neq 0 \text{ for all } C \text{ with } b_k \leq C \leq \frac{k(k+1)}{2}$$

$$k = \frac{\ell(\ell+1)}{2} + 2 + i,\ \ell \geq 2,\ 0 \leq i \leq \ell \Rightarrow$$

$$b_k = \frac{\ell(\ell^2-1)}{2} + 3\ell + 2 + i(\ell+1)$$

$k = 5 = \frac{2 \cdot 3}{2} + 2 + 0$ so $\ell = 2, i = 0$, then $b_5 = \frac{2 \cdot 5}{2} + 3 \cdot 2 + 2 + 0 = 11$.

F. Levé, P. Séébold, 2000

$$\mathbf{K}(w) = \sum_{k=1}^{|w|} f_w(k)$$

$$\mathbf{K}_u^+(n) = \max_i \mathbf{K}(u_i u_{i+1} \ldots u_{i+n-1})$$

$$\mathbf{C}(w) = \max_{k=1}^{|w|} f_w(k)$$

$$\mathbf{C}_u^+(n) = \max_i \mathbf{C}(u_i u_{i+1} \ldots u_{i+n-1})$$

u non ultimately periodic:

$$f_u(n) \geq n + 1$$

$$\mathbf{C}_u^+(n) \geq \left\lfloor \frac{n}{2} \right\rfloor + 1$$

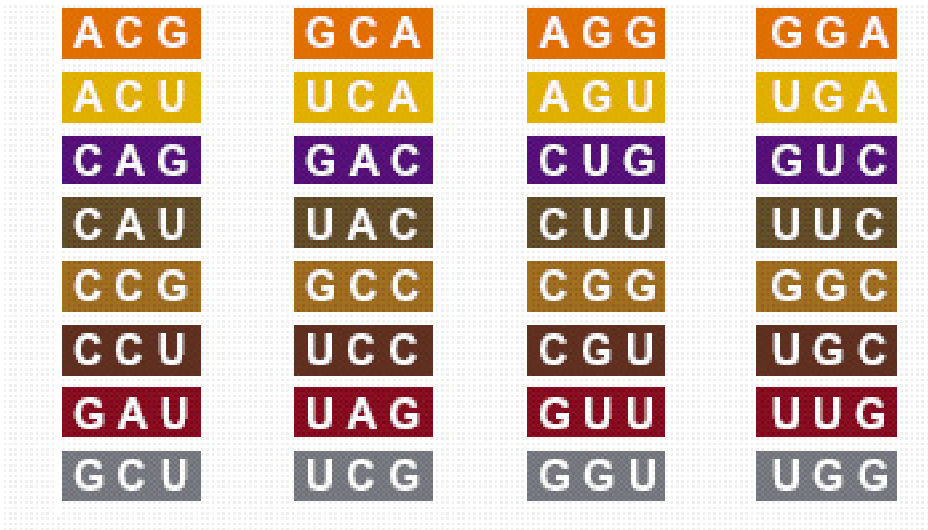
$$\mathbf{K}_u^+(n) \geq \left\lfloor \frac{n^2}{4} + n \right\rfloor$$

u Sturmian:

$$f_u(n) = n + 1$$

$$\mathbf{C}_u^+(n) = \left\lfloor \frac{n}{2} \right\rfloor + 1$$

$$\mathbf{K}_u^+(n) = \left\lfloor \frac{n^2}{4} + n \right\rfloor$$



genetic code

4 nucleotide bases

DNA

A (adenine), *T* (thymine), *G* (guanine), *C* (cytosine)

RNA

A (adenine), *U* (uracil), *G* (guanine), *C* (cytosine)

. . . UGUCGUAAG. . .   UGU, GUC, UCG, . . .

(B. Hayes: *The invention of the genetic code*,
American Scientist, 1998, vol. 86 no. 1

right special subword which can be continued in more than one way

010010100100101001010...
010 right special subword: *0100*, *0101*
100 is not right special: *1001*

(left, right, bi-) special subwords play an important role in DNA and RNA molecules

010 bispecial
0100, *0101*
0010, *1010*

---

pattern recognition, string matching

E-mail: kasa@cs.ubbcluj.ro
http://www.cs.ubbcluj.ro/~kasa