Crochemore factorization of infinite words

Jean Berstel, Alessandra Savelli

Basic definition

Crochemore factorization: the c-*factorization* $c(x)$ of a word x is

$$c(x) = (x_1, x_2, \ldots, x_m, x_{m+1}, \ldots)$$

where x_m is the longest prefix of $x_m x_{m+1} \cdots$ occurring twice in $x_1 x_2 \cdots x_m$, or x_m is a letter a if a does not occur in $x_1 \cdots x_{m-1}$.

The c-factorization of $x = ababaab$ is (a, b, aba, ab), since aba occurs twice in $ababa$.

Fibonacci word

The **Fibonacci word** is defined as the limit of the sequence $f_{-1} = b$, $f_0 = a$, and $f_{n+2} = f_{n+1} f_n$:

$$\begin{array}{rcl} f_0 & = & a \\ f_1 & = & ab \\ f_2 & = & aba \\ f_3 & = & abaab \\ & \cdots & \\ \mathbf{f} & = & abaababaaba \cdots = \widetilde{f_0}\widetilde{f_1}\widetilde{f_2}\widetilde{f_3}\cdots \end{array}$$

The c-factorization of **f** is exactly:

$$c(\mathbf{f}) = (a, b, a, aba, baaba, \ldots) = (a, b, a, \widetilde{f_2}, \widetilde{f_3}, \ldots).$$

Fibonacci word

The c-factorization is closely related to two other factorizations of the Fibonacci word.

$$h(\mathbf{f}) = (a, b, a, ab, aba, abaab, \ldots)$$

$$w(\mathbf{f}) = (a, b, aa, bab, aabaa, \ldots)$$

The three factorizations can be visualized through the following scheme:

$$\begin{array}{rl}
h: & \mid a \mid b \mid a \mid a\ b \mid a\ b\ a \mid a\ b\ a\ a\ b \mid \cdots \\
w: & \mid a \mid b \mid a\ a \mid b\ a\ b \mid a\ a\ b\ a\ a \mid \cdots \\
c: & \mid a \mid b \mid a \mid a\ b\ a \mid b\ a\ a\ b\ a \mid \cdots
\end{array}$$

Sturmian words

A **standard Sturmian word** is defined as the limit of

$$s_{-1} = b, s_0 = a, \text{ and } s_n = s_{n-1}^{d_n} s_{n-2},$$

where d_i is a positive integer for all $i > 0$.

Similarly to the Fibonacci word, the Sturmian words have a decomposition in reverse finite words s_n:

$$\mathbf{s} = \widetilde{s_0}^{d_1} \widetilde{s_1}^{d_2} \widetilde{s_2}^{d_3} \cdots .$$

The c-factorization of standard Sturmian words is closely related to that decomposition:

$$c(\mathbf{s}) = (a, a^{d_1-1}, b, a^{d_1} \widetilde{s}_1^{d_2-1}, \widetilde{s}_2^{d_3}, \widetilde{s}_3^{d_4}, \ldots, \widetilde{s}_n^{d_{n+1}}, \ldots).$$

Thue-Morse word

Let τ be the Thue-Morse morphism on a two-letter alphabet defined by $\tau(a) = ab$ and $\tau(b) = ba$, the **Thue-Morse infinite word** $\mathbf{t} = abbabaabbaababba\cdots$ is defined as the limit of the sequence

$$t_0 = a, \quad t_n = \tau(t_{n-1}).$$

Each factor in the c-factorization of $\mathbf{t}$ can be obtained from the previous ones by applying the morphism τ:

$$c(\mathbf{t}) = (c_1, c_2, \ldots), \quad c_{n+2} = \tau(c_n) \text{ for every } n > 6.$$

$$c(\mathbf{t}) = (a, b, b, ab, a, abba, aba, bbabaab, abbaab, babaabbaababba, \ldots)$$

Thue-Morse generalized words

The **Thue-Morse generalized word** on a m-letter alphabet $A = \{a_1, a_2, \ldots, a_m\}$ obtained as the limit of the sequence

$$t_0^{(m)} = a_1, \quad t_{n+1}^{(m)} = \tau_m(t_n^{(m)}),$$

where τ_m is the morphism defined by
$\tau_m(a_i) = a_i a_{i+1} \cdots a_m a_1 \cdots a_{i-1} \quad (i = 1, \ldots, m).$

$$c_{n+2(m-1)}^{(m)} = \tau_m(c_n) \text{ for every } n > m \text{ and } m \geq 3.$$

Example for $m = 3$:

$$c(\mathbf{t}^{(\mathbf{3})}) = (a, b, c, bc, a, ca, b, bcacab, abc, cababc, bca, \ldots)$$

Period doubling sequence

Let δ be the morphism on a two-letter alphabet defined by

$$\delta(a) = ab, \quad \delta(b) = aa.$$

The **period doubling sequence** is the limit of the sequence

$$q_0 = a \text{ and } q_{n+1} = \delta(q_n).$$

Period doubling sequence

Similarly to the case of standard Sturmian words, the period doubling sequence admits the decomposition:

$$\mathbf{q} = \widetilde{q_0}\widetilde{q_1}\widetilde{q_2}\cdots ,$$

and the c-factorization of $\mathbf{q}$ is

$$c(\mathbf{q}) = (a, b, a, aa, ba, baba, aaba, \ldots) = (\widetilde{q_0}, \widetilde{q_0}', \widetilde{q_0}, \widetilde{q_1}', \widetilde{q_1}, \widetilde{q_2}', \widetilde{q_2}, \ldots),$$

where q_n' is q_n with just the last letter changed to its opposite.

Note that $q_{n+1} = q_n q_n'$.

Crochemore vs. Ziv-Lempel

Ziv-Lempel factorization: the z-*factorization* $z(x)$ of a word x is

$$z(x) = (y_1, y_2, \ldots, y_m, y_{m+1}, \ldots)$$

where y_m is the shortest prefix of $y_m y_{m+1} \cdots$ which occurs only once in $y_1 y_2 \cdots y_m$.

The c-factorization of $x = ababaab$ is $(a, b, abaa, b)$.

Crochemore vs. Ziv-Lempel

Facts:

- A Crochemore factor cannot properly include a Ziv-Lempel factor.
- If a Ziv-Lempel factor includes a Crochemore factor, then it ends at most a letter after.

For example, let x be the word $x = aabaaccbaabaabaa$. The c-factorization and the z-factorization of x are:

$$\begin{aligned} c(x) &= (a, a, b, aa, c, c, baa, baabaa) \\ z(x) &= (a, ab, aac, cb, aabaab, aa). \end{aligned}$$

Crochemore vs. Ziv-Lempel

Consider for example the period doubling sequence

$$\mathbf{q} = abaaabababaaabaa \cdots \ .$$

Each Ziv-Lempel factor of $\mathbf{q}$ properly includes a Crochemore factor by ending just a letter before:

$$\begin{array}{ll} z: & |\, a \,|\, b \,|\, a \ a \,|\, a \ b \ a \ b \,|\, a \ b \ a \ a \ a \ b \ a \ a \,|\, \cdots \\ c: & |\, a \,|\, b \,|\, a \,|\, a \ a \,|\, b \ a \,|\, b \ a \ b \ a \,|\, a \ a \ b \ a \,|\, a \ \cdots \end{array}$$