BBBBBBBBBBBBBBBBB

Modifications of the notion of complexity of infinite words and analysis of the structure of biological sequences (DNA, RNA) through the lens of combinatorics on words

Kirova Valeriya

HSE university

29 июня 2025 г.

Starring

1. Factor complexity function (Morse, Hedlund, 1940)
2. Sturmian words
3. Arithmetical complexity function (Avgustinovich,Fon-Der-Flaas, Frid, 2000)
4. Abelian complexity function (G. Richomme, K. Saari, and L. Q. Zamboni)
5. Phylogenetic analysis of DNA sequences based on k-word and rough set theory (Chun Li ,Yan Yanga, Meiduo J ...)

Introduction

History

The study of combinatorics on words begins in 1906 with the works of the A. Thue, continued by M. Morse and G. Hedlund in 1940. In his work A. Thue established the next question: is there an infinite word over a finite alphabet in which there are no words consisting of two consecutive occurrences of the same factor ? Such words are called cubes.
For example, this would be the word *ababab*.

The factor complexity function (Morse, Hedlund, 1940)

The factor complexity function of a word w is the function $p(n)$ that counts the number of distinct factors of length n from that word.

Example

Let $\Sigma_3 = \{0, 1, 2\}$ and given the word

$$w = 012212111$$

Factors of length 2: 02, 12, 22, 21, 11.
Thus, $p(2) = 5$

1. $p_w(n)$ is a non-decreasing function;
2. $1 \leq p_w(n) \leq q^n$;
3. $\exists C : p_w(n) \leq C$, when w is ultimately periodic [1].

[1] A sequence (a_k) satisfying the condition $a_{k+r} = a_k$ for all sufficiently large values of k and some r is called ultimately periodic with period r.

Theorem (Morse and Hedlund)

Theorem (Morse and Hedlund)

Let w be an infinite word. If there exists n such that $p_w(n) \leq n$, then w is eventually periodic.

This leads to two natural questions:

1. Given a word w — what is its complexity function?
2. Given a function $f(n)$ — does there exist a word with this complexity?

Sturmian words

Infinite word $w = w_0 w_1 w_2 \cdots$ over the alphabet $\{0, 1\}$ is called Sturmian if its factor complexity $p_w(n)$ equals to $n + 1$ for all n.

It is a non-periodically balanced word, i.e. for each value of n, the number of ones in the factors of the word w of length n takes no more than two values - in fact, exactly two adjacent values.

Let u and v be two factors of the word w of the same length. Let $|u|_1$ and $|v|_1$ be the number of ones in the factors u and v, respectively. Then

$$||u|_1 - |v|_1| \leq 1.$$

Sturmian Word is a mechanical word: an infinite word, each symbol of which is given by one of two equalities:
$w_i = \lfloor \alpha(i+1) + \beta \rfloor - \lfloor \alpha i + \beta \rfloor$, or $w_i = \lceil \alpha(i+1) + \beta \rceil - \lceil \alpha i + \beta \rceil$
where $\alpha \in (0, 1)$ is a slope of Sturmian word, and $\beta \in [0, 1)$.

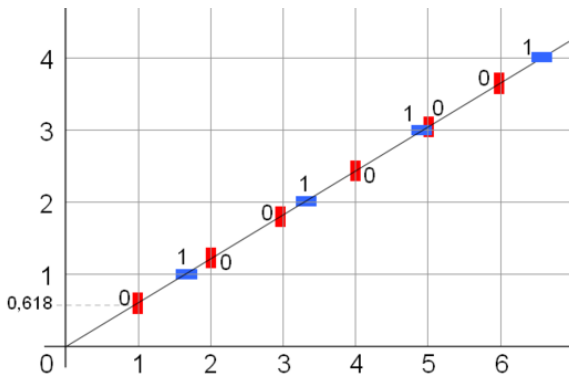
Example

The Fibonacci word is an example of a Sturmian word.

Let $\phi(0) = 0, \phi(1) = 01$. Now $\phi(n) = \phi(n-1)\phi(n-2)$

$$0 \longrightarrow 01 \longrightarrow 010 \longrightarrow 01001 \longrightarrow 01001010 \longrightarrow 0100101001001 \longrightarrow \cdots$$

The start of the cutting sequence[2] shown here illustrates the start of the word $w = 0100101001$:



Characterization by a cutting sequence with a line of slope $\alpha = 1/\varphi$ where φ is the golden ratio.

[2] A cutting sequence is a sequence of symbols whose elements correspond to the individual grid lines crossed as a curve crosses a square grid.

The arithmetical complexity of infinite words

The concept of infinite word arithmetical complexity was introduced in:

Avgustinovich S. V. , Fon-Der-Flaass D. G. , Frid A. E.. Arithmetical complexity of infnite words. Proc. Words, Languages and Combinatorics III, 2000. Singapore: World Scientifc, 2003. P. 51-62.

Definition

An arithmetical factor of length n an infinite word w is a finite word of the form

$$w_k w_{k+d} w_{k+2d} \cdots w_{k+nd}$$

with common difference d and starting position k. and denote it by w_d^k

Definition

Arithmetical closure of infinite word w is the set of all arithmetical factors:

$$A_w = \{w_k w_{k+d} w_{k+2d} \cdots w_{k+(n-1)d} | k \geq 0, d > 0\}.$$

Other complexity functions

1. Lie complexity
2. Abelian complexity
3. k-Abelian complexity
4. arithmetical complexity
5. maximal pattern complexity
6. cyclic complexity
7. binomial complexity
8. window complexity,
9. periodicity complexity,
10. palindrome complexity

Abelian complexity (G. Richomme, K. Saari, and L. Q. Zamboni)

Let u be a finite word over an alphabet Σ, and let $a \in \Sigma$, then $|u|_a$ denotes the number of occurrences of the letter a in the word u.

Two finite words u and v are called abelian equivalent ($u \sim_{\text{ab}} v$) if for every letter $a \in \Sigma$ we have:

$$|u|_a = |v|_a.$$

In other words, u and v are permutations of the same multiset of symbols.

Parikh vector of a word v over the alphabet $\Sigma = \{a_1, \ldots, a_k\}$ is the vector:

$$\mathrm{PV}(v) = \left(|v|_{a_1}, \ldots, |v|_{a_k}\right).$$

Clearly, two words are abelian equivalent if and only if they have the same Parikh vector.

Abelian complexity

The abelian complexity of a word w is defined as a function $a_w(n)$, which counts the number of distinct abelian equivalence classes of factors of length n occurring in w.

Lemma

An infinite word w is ultimately periodic if and only if there exists n such that

$$a_w(n) = 1.$$

Therefore, if a word is ultimately periodic, its abelian complexity is bounded.

Sturmian words are aperiodic and satisfy: $a_w(n) = 2 \quad$ for all n, and moreover, this property characterizes them:

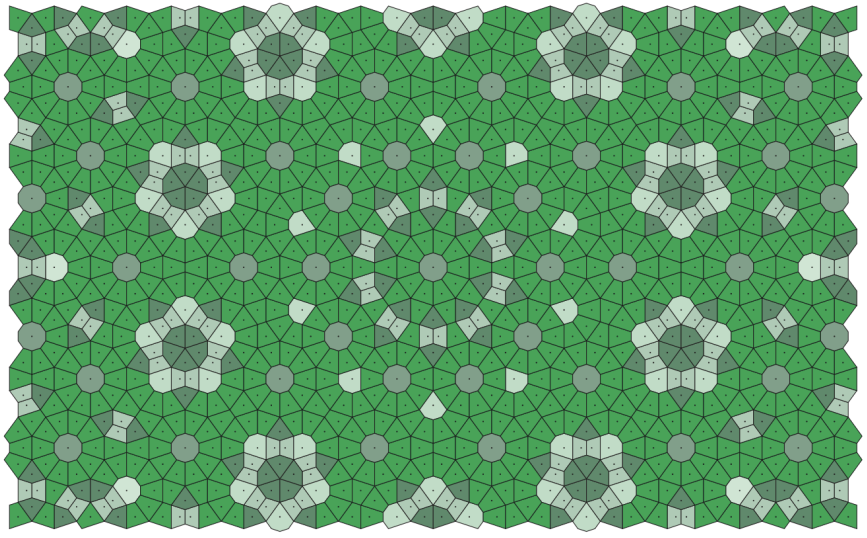
Theorem 2. Let w be an aperiodic infinite word. Then

$$a_w(n) = 2 \text{ for all } n \geq 1 \quad \iff \quad w \text{ is Sturmian.}$$

Examples of apllications of combinatorics on words

Quasicrystals: algebraic, combinatorial and geometrical aspects Edita Pelantová , Zuzana Masáková

The paper presents mathematical models of quasicrystals with particular attention given to cut-and-project sets. There exists a general family of sets that are known to have quasicrystalline properties: the so-called cut and project sets, here abreviated to CP sets. For the description of their properties we use the methods of combinatorics on words. The construction of a one-dimensional set of CP is illustrated:



Phylogenetic analysis of DNA sequences based on k-word and rough set theory (Chun Li ,Yan Yanga, Meiduo J ...)

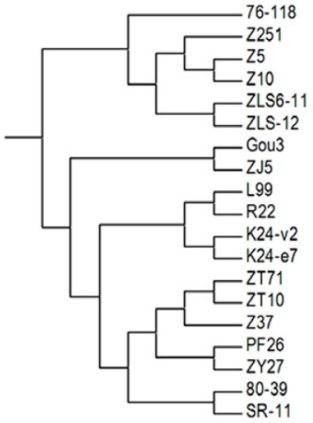
Most methods for constructing phylogenetic trees require prior sequence alignment because:

- It allows for the comparison of homologous regions (sequences inherited from a common ancestor).
- It enables the calculation of evolutionary distances (e.g., the number of mutations between sequences).

However, alignment has significant limitations:

- Computational complexity
- Ambiguity in alignment criteria
- Genomic rearrangements
- Differences in sequence lengths

Phylogenetic analysis of DNA sequences based on k-word and rough set theory (Chun Li ,Yan Yanga, Meiduo J ...)



The phylogenetic tree of 19Hantaviruses.

How can biological sequences be characterized and compared while avoiding multiple alignment?

Some methods shift the focus from analyzing individual nucleotides or amino acids to studying sequence composition. In these approaches, each sequence is represented as a vector whose components are derived from k-words (substrings of length k)

Complete word-based vector

Let S be a binary $(0,1)$-sequence of length m. The count of a word w of length k (called a k-word) in S, denoted by $c(w)$, is the number of occurrences of w in the sequence S.

Since S contains $m-k+1$ overlapping k-words, the frequency of occurrence of w in S is defined as:

$$f(w) = \frac{c(w)}{m-k+1}.$$

Once the frequencies of all $n = 2^k$ possible k-words (or k-mers) are known, we can construct the frequency vector:

$$F_k = \left(f(w_{k,1}), f(w_{k,2}), \ldots, f(w_{k,n})\right).$$

If some frequencies are equal, the corresponding k-words are sorted in lexicographic order. Therefore, in the sorted frequency vector F_s, the following relations hold:

$$f(w_{k,i_1}) \leq f(w_{k,i_2}) \leq \cdots \leq f(w_{k,i_n}).$$

Thus, each frequency $f(w)$ is assigned a unique position in the sorted vector F_s, denoted as $g(w)$. By combining this positional information with the frequency itself, we form the vector:

$$V_{FP} = \left(g(w_{k,1}) \cdot e^{f(w_{k,1})},\ g(w_{k,2}) \cdot e^{f(w_{k,2})},\ \ldots,\ g(w_{k,n}) \cdot e^{f(w_{k,n})} \right).$$

$S = 10110110111000000111111001001011011001101111110001000000110$

and let $k = 3$. Then there are $2^3 = 8$ possible 3-words in total:

$$000,\ 001,\ 010,\ 011,\ 100,\ 101,\ 110,\ 111.$$

the frequency vector:

$$F_k = (f(000), f(001), f(010), f(011), f(100), f(101), f(110), f(111)) =$$

$$= (0.1250,\ 0.0795,\ 0.1136,\ 0.1364,\ 0.0909,\ 0.1705,\ 0.1364,\ 0.1477)\,.$$

Sorting the entries in ascending order gives the sorted frequency vector:

$$F_s = (0.0795,\ 0.0909,\ 0.1136,\ 0.1250,\ 0.1364,\ 0.1364,\ 0.1477,\ 0.1705)\,.$$

Therefore, the positional ranks (lex order for ties) are:

$$(g(000), g(001), g(010), g(011), g(100), g(101), g(110), g(111)) =$$

$$= (4,\ 1,\ 3,\ 5,\ 2,\ 8,\ 6,\ 7).$$

Thus, combining positional and frequency information, we get:

$$V_{FP} = (4.5326,\ 1.0828,\ 3.3610,\ 5.7305,\ 2.1903,\ 9.4867,\ 6.8766,\ 8.1144)\,.$$

DNA bases — A, G, C, and T — can be classified according to various properties:

- By chemical structure: purines $R = \{A, G\}$ and pyrimidines $Y = \{C, T\}$;
- By functional group: amino group $M = \{A, C\}$, keto group $K = \{G, T\}$;
- By the number of hydrogen bonds: weak $W = \{A, T\}$, strong $S = \{G, C\}$.

Assign value 1 to bases from the classes R, M, and W, and value 0 to bases from Y, K, and S, respectively. Thus, a primary DNA sequence is transformed into three binary $(0, 1)$-sequences:

- (R, Y)-characteristic,
- (M, K)-characteristic,
- (W, S)-characteristic.

For each of these, we compute a frequency-position vector:

$$V_{FP}^{(1)},\ V_{FP}^{(2)},\ V_{FP}^{(3)},$$

and the final combined vector:

$$v = \left(V_{FP}^{(1)},\ V_{FP}^{(2)},\ V_{FP}^{(3)}\right).$$

- Goal: Extract more information from biological sequences.
- Approach: Uses three binary representations of a DNA sequence, constructing a full feature vector of dimension 3×2^k for any given k.
- Feature selection: Since not all k-mers contribute equally to evolutionary distance, the most informative k-mers (carrying maximal evolutionary signal) are selected.
- Dimensionality reduction: A compact feature vector is built exclusively from these k-mers, drastically reducing dimensionality while preserving sequence information.

Feature selection scheme based on rough set theory

Rough Set Theory (RST), proposed by Pawlak in the 1980s, is an effective mathematical tool for studying intelligent systems described by imprecise, uncertain or vague information. Currently, rough set theory is actively used for feature reduction and attribute selection.

An information system is defined as a quadruple $S = \langle U, A, V, f \rangle$, where:

- $U = \{x_1, x_2, \ldots, x_n\}$ is a finite non-empty set of objects (universe);
- A is a set of attributes that describe the objects;
- V is the domain of attribute values;
- f is an information function such that $\forall a \in A, \forall x \in U : f(x, a) \in V$.

If the set of attributes A is divided into condition attributes C and decision attributes D, then the system $S = \langle U, C \cup D, V, f \rangle$ is called a decision system (or a decision table).

Let $S = \langle U, C \cup D, V, f \rangle$ be a decision system. For an attribute $c_k \in C$ and objects $x, y \in U$, we say that they are equivalent with respect to c_k if

$$f(x, c_k) = f(y, c_k).$$

The equivalence class of x with respect to c_k is denoted by:

$$E[x]_{c_k} = \{y \in U \mid f(y, c_k) = f(x, c_k)\}.$$

If x and y are equivalent with respect to c_k but $f(x, D) = f(y, D)$, they are called consistent with respect to c_k; otherwise, they are inconsistent with respect to c_k.

A set $P \subseteq U$ is called consistent with respect to c_k if any two objects $x, y \in P$ are consistent with respect to c_k.

A set $P \subseteq U$ is called inconsistent with respect to c_k if:

- all objects $x, y \in P$ are equivalent with respect to c_k,
- but there exist $x_0, y_0 \in P$ that are inconsistent with respect to c_k.

A decision system S is said to be consistent with respect to c_k if every equivalence class with respect to c_k is consistent.

Let $S = \langle U, C \cup D, V, f \rangle$ be a system consistent with respect to c_k, and let $x, y \in U$ such that $f(x, D) \neq f(y, D)$ (then necessarily $f(x, c_k) \neq f(y, c_k)$).
Then there exists an inflection point between x and y, whose value is defined as:

$$f_{\mathrm{IP}}(x, y, c_k) = \frac{1}{2}\left[f(x, c_k) + f(y, c_k)\right]$$

The corresponding inflection ratio is given by:

$$R_{\mathrm{IP}}(x, y, c_k) = \frac{|f(x, c_k) - f(y, c_k)|}{f_{\mathrm{IP}}(x, y, c_k)} \times 100\%$$

Now suppose that a certain order is defined on $U = \{x_1, x_2, \ldots, x_n\}$ and the objects are sorted:

$$x_1 \leq x_2 \leq \cdots \leq x_n.$$

Then it is possible to find the inflection points between any adjacent objects and compute the corresponding R_{IP}. The smallest among them is called the minimal inflection ratio with respect to c_k and is denoted by:

$$R_{\mathrm{mi}}(c_k)$$

Algorithm for computing $R_{\mathrm{mi}}(c_k)$:

Input: decision table $S = \langle U, C \cup D, V, f \rangle$;
Output: $R_{\mathrm{mi}}(c_k)$ for $c_k \in C$

1. Sort the objects $\{x_1, \ldots, x_n\}$ in ascending order of $f(x_i, c_k)$. Rename them as $\{y_1, \ldots, y_n\}$ such that $f(y_i, c_k) \leq f(y_{i+1}, c_k)$.
2. Partition the set $\{y_1, \ldots, y_n\}$ into equivalence classes $E_1, \ldots, E_m$ with respect to c_k.
3. For each class E_i, if it is inconsistent, redefine the decision values $f(y, D)$ as the union of all decision labels in that class. After this, the table becomes consistent with respect to c_k.
4. In the resulting table, find all adjacent pairs (y_j, y_{j+1}) such that $f(y_j, D) \neq f(y_{j+1}, D)$, compute R_{IP} for them, and select the minimum. This is $R_{\mathrm{mi}}(c_k)$.

By repeating steps 1–4 for each attribute c_k, one can compute $R_{\mathrm{mi}}(c_k)$ for all attributes.

Definition 8.

Definition

Let $S = \langle U, C \cup D \rangle$ be a decision system. The significance of each condition attribute $c_k \in C$ is defined as

$$SIG(c_k, C, D) = R_{mi}(c_k) - \lambda,$$

where $\lambda \geq 0$ is an equilibrium constant.

Definition

Let $S = \langle U, C \cup D \rangle$ be a decision system. If $SIG(c_k, C, D) > 0$, the condition attribute c_k is said to be important.

The greater the value of $SIG(c_k, C, D)$, the more important the attribute c_k is. Therefore, in this work, we use the proposed function $SIG(c_k, C, D)$ to evaluate the significance of the condition attribute c_k. By identifying important attributes, one can obtain a feature vector for DNA sequences with significantly reduced dimensionality.

To estimate the evolutionary distance between two sequences i and j, the following formula is used:

$$D(i,j) = \frac{1}{2}\left(1 - \cos(v_i, v_j)\right) \cdot d(v_i, v_j)$$

where:

- v_i, v_j are the feature vectors for sequences i and j,
- $\cos(v_i, v_j)$ is the cosine of the angle between the vectors,
- $d(v_i, v_j)$ is the Euclidean distance between them.

Based on the resulting distance matrix, a phylogenetic tree is constructed using the UPGMA method.

To estimate the evolutionary distance between two sequences i and j, the following formula is used:

$$D(i,j) = \frac{1}{2}\left(1 - \cos(v_i, v_j)\right) \cdot d(v_i, v_j)$$

where:

- v_i, v_j are the feature vectors for sequences i and j,
- $\cos(v_i, v_j)$ is the cosine of the angle between the vectors,
- $d(v_i, v_j)$ is the Euclidean distance between them.

Based on the resulting distance matrix, a phylogenetic tree is constructed using the UPGMA method.

Testing the Algorithm on Hantavirus Genomes

The training set consists of 19 hantavirus strains. Hantavirus (HV) is a negative-sense RNA virus from the Bunyaviridae family. For a given value of k, we construct a full vector of size 3×2^k based on k-mers for each of the 19 HV sequences, and then compute the minimal inflection ratio and the corresponding significance of each k-mer.

It was found that for $k = 11$ and $\lambda = 0.17\%$, all 19 HV strains can be correctly classified. Moreover, the resulting phylogenetic tree aligns well with the accepted taxonomy.

The total number of selected significant k-mers was 869, meaning that out of $3 \times 2^{11} = 6144$ possible components, 869 of the most informative were selected. These selected k-mers were used to construct the feature vector for each DNA sequence.

Testing the Algorithm on Hantavirus Genomes

The training set consists of 19 hantavirus strains. Hantavirus (HV) is a negative-sense RNA virus from the Bunyaviridae family. For a given value of k, we construct a full vector of size 3×2^k based on k-mers for each of the 19 HV sequences, and then compute the minimal inflection ratio and the corresponding significance of each k-mer.

It was found that for $k = 11$ and $\lambda = 0.17\%$, all 19 HV strains can be correctly classified. Moreover, the resulting phylogenetic tree aligns well with the accepted taxonomy.

The total number of selected significant k-mers was 869, meaning that out of $3 \times 2^{11} = 6144$ possible components, 869 of the most informative were selected. These selected k-mers were used to construct the feature vector for each DNA sequence.

References

- Avgustinovich S. V. , Fon-Der-Flaass D. G. , Frid A. E.. Arithmetical complexity of infnite words. Proc. Words, Languages and Combinatorics III, 2000. Singapore: World Scientifc, 2003. P. 51-62.
- Avgustinovich S., Cassaigne J., Frid A. Sequences of low arithmetical complexity // Theoret. Inform. Appl. 40 (2006) 569–582
- Berstel J., P. Séébold. Sturmian words, in: M. Lothaire, Algebraic Combinatorics on Words . Cambridge University Press, 2002. P. 40-97.
- Hedlund G.A. , M. Morse. Symbolic dynamics. Amer. J. Math, 1938, 815-866.
- Thue A., Uber unendliche Zeichenreihen. Norske Vid. Skrifter I Mat. Nat. Kl., Christiania 1906. V. 7 P. 1 22. V. 10. P. 1-67.

References

- G. Richomme, K. Saari, L. Zamboni, Abelian complexity of minimal subshifts. J. Lond. Math. Soc. (2) 83 (2011), no. 1, 79–95.
- Edita Pelantova , Zuzana Masakova. Quasicrystals: algebraic, combinatorial and geometrical aspects
- Chun Li, Yan Yang, Meiduo Jia, Yingying Zhang, Xiaoqing Yu, Changzhong Wang, Phylogenetic analysis of DNA sequences based on k-word and rough set theory, Physica A: Statistical Mechanics and its Applications, Volume 398, 2014,