Biophysical Reviews (2023) 15:1367–1378
https://doi.org/10.1007/s12551-023-01140-y

REVIEW

Bioinformatics tools for the sequence complexity estimates

Yuriy L. Orlov[1,2,3] · **Nina G. Orlova**[4]

Received: 15 August 2023 / Accepted: 1 September 2023 / Published online: 15 September 2023

Abstract

We review current methods and bioinformatics tools for the text complexity estimates (information and entropy measures). The search DNA regions with extreme statistical characteristics such as low complexity regions are important for biophysical models of chromosome function and gene transcription regulation in genome scale. We discuss the complexity profiling for segmentation and delineation of genome sequences, search for genome repeats and transposable elements, and applications to next-generation sequencing reads. We review the complexity methods and new applications fields: analysis of mutation hot-spots loci, analysis of short sequencing reads with quality control, and alignment-free genome comparisons. The algorithms implementing various numerical measures of text complexity estimates including combinatorial and linguistic measures have been developed before genome sequencing era. The series of tools to estimate sequence complexity use compression approaches, mainly by modification of Lempel–Ziv compression. Most of the tools are available online providing large-scale service for whole genome analysis. Novel machine learning applications for classification of complete genome sequences also include sequence compression and complexity algorithms. We present comparison of the complexity methods on the different sequence sets, the applications for gene transcription regulatory regions analysis. Furthermore, we discuss approaches and application of sequence complexity for proteins. The complexity measures for amino acid sequences could be calculated by the same entropy and compression-based algorithms. But the functional and evolutionary roles of low complexity regions in protein have specific features differing from DNA. The tools for protein sequence complexity aimed for protein structural constraints. It was shown that low complexity regions in protein sequences are conservative in evolution and have important biological and structural functions. Finally, we summarize recent findings in large scale genome complexity comparison and applications for coronavirus genome analysis.

Keywords Bioinformatics · Text complexity · Lempel–Ziv compression · Genetic codes · Sequence information · Entropy · Low complexity regions · Sequencing artefacts · Genomic rearrangement · Alignment-free · Genome comparison · Online tools

Introduction

We rereview current works and bioinformatics tools for DNA text complexity estimates with next applications in short sequence analysis, complete genome studies, and protein annotations. The methods for complexity estimates have been realized in 1990s before the Human genome project (Trifonov 1990; Gusev et al. 1991; Román-Roldán et al. 1998). Information measures and entropies estimates serve as background for biophysical models of genome structure and evolution (Sadovsky et al. 2008). The theory of overlapping genetic codes (protein coding triplets, RNA structure signal, nucleosome positioning codes, and topological chromosome codes) started from the works by E.N.Trifonov (Trifonov 1989, 1990) relies to numerical sequence complexity.

✉ Yuriy L. Orlov
y.orlov@sechenov.ru

1 The Digital Health Institute, I.M. Sechenov First Moscow State Medical University of the Russian Ministry of Health (Sechenov University), Moscow 119991, Russia

2 Institute of Cytology and Genetics SB RAS, 630090 Novosibirsk, Russia

3 Agrarian and Technological Institute, Peoples' Friendship University of Russia, 117198 Moscow, Russia

4 Department of Mathematics, Financial University under the Government of the Russian Federation, Moscow 125167, Russia

With development of the high-throughput sequencing technologies the complexity analysis tools evolved from simple algorithm realization to advanced online programs, large scale genome data processing software (Orlov and Potapov 2004; Kryukov et al. 2020; Agenis-Nevers et al. 2021; Karakatsanis et al. 2021; Zimnyakov et al. 2023; Bello et al. 2023).

In the next sections of this "Introduction," we will discuss applications of the complexity profiling to segmentation and delineation of genome sequences, search for genome repeats and transposable elements, and next-generation sequencing reads. Furthermore, we review the applications of complexity estimates for gene transcription regulatory regions analysis, the alignment-free sequences comparison methods, and the compression-based complexity approaches.

The rest of the review follow standard scheme—methods and algorithms, results, and discussion. The section "Methods and algorithms for DNA sequence complexity" reviews the algorithms for DNA sequence complexity estimates. The "Results" section present comparison of the complexity methods on the different sequence sets, online tools for sequence complexity analysis (summarized in the table), discuss low complexity for protein sequences, and the genome database compression methods based on the complexity. The "Discussion" section summarizes recent findings and actual applications for coronavirus genome analysis.

Complexity for segmentation and delineation of genome sequences

The measures of compositional complexity coming from the statistical physics methods help to find abnormalities in linear genome structure and make corresponding segmentation (Karakatsanis et al. 2021; Bernaola-Galván et al. 2023). Shannon information (Shannon 1948) as the first measure of nucleotide frequencies allows delineation of complexity blocks, coding, and non-coding regions in a sequence (Deng et al. 2012). Shannon information, as well as entropy, could be measured for nucleotides, dinucleotides, and oligonucleotides of any reasonable length (up to 10) in available genome sequences. Despite this measure is easy to count, it can separate real and artificial sequences (Sadovsky et al. 2008). Chang and colleagues have shown that the Shannon information for sequences from complete genomes are much higher than for random sequences of the same size (Chang et al. 2005). This observation raised the problem of artificial sequence generation that resembles properties of real genome sequences (Wang et al. 2020).

Fluctuations in nucleotide frequencies in genome regions allow find heterogeneous sequence regions at varying scale—from short gene regulatory regions (kilobases) to isochores and chromosome segments (megabases) (Bernaola-Galván et al. 2023). Thus, applications of complexity analysis could be listed by sequence size:

1) Short sequences (transcription factor binding sites, promoters, gene regulatory regions, small domains, and microsatellites (Orlov and Potapov 2004; Safronova et al. 2016).
2) Medium size genome regions (genes, patching exon/intron structures, distal gene enhancers) (Abnizova et al. 2007; Deng et al. 2012).
3) Chromosome arms, and complete prokaryotic genomes (Agenis-Nevers et al. 2021; Bonidia et al. 2022; Bernaola-Galván et al. 2023).

The concept of triplet periodicity class and a measure of similarity between such classes were introduced in (Frenkel and Korotkov 2008). Triplet periodicity in DNA is related to coding sequence properties. It could be used to find ORF (open reading frame) shifts (Frenkel and Korotkov 2009).

Suvorova et al. (2014) compared periodicity search methods in DNA sequences. It was shown that combination of spectral methods and information decomposition methos is necessary to define hidden periodicities with high mutation rate. Suvorova and Korotkov (2015) studied triplet periodicity differences inside and between genomes extending the approach discussed by Dios et al. (2014).

Visible elements of low complexity regions in a genome are tandem repeats (Benson 1999; Frenkel et al. 2017). Such tandem (tail-to-tail) repeats are considered as a kind low complexity region. Tandem consist of tens to hundreds of residues of a repeated pattern, such as *atcatcatcatcatc* ("*atc*" repeated). They are classified as mini and microsatellites (Jurka et al. 2007). Molecular mechanism of replication slippage lead reproducing of tandem repeats. Tandem duplications are common for cancers (Li et al. 2020) and may occur in somatic cells at larger scale. Enrichment of head-to-tail somatic segmental tandem duplications in genome defined as the tandem duplicator phenotype also defines lower sequence complexity (Menghi et al. 2018). There are set of computational tools such TRF (Tandem Repeat Finder) (Benson 1999; Frenkel et al. 2017) and ULTRA (Olson and Wheeler 2018) to effectively search for degenerated tandem repeats (Delucchi et al. 2021).

There are set of methods for tandem repeat search—mreps (Kolpakov et al. 2003), TRStalker (Pellegrini et al. 2010), T-REKS (Jorda and Kajava 2009), G-IMEx (Mudunuri et al. 2010). The methods for tandem repeat search are limited due to sensitivity to nucleotide deletions and insertions.

Korotkov et al. (2022) studied triplet and *k*-mer periodicities in relation to genome adaptation. The grouping of bacterial genomes by periodicity in repeat composition was shown. Plant genomes present special case for genome complexity studies. There are abundant repeat elements, transposons and satellites. Korotkov et al. (2021) found highly divergent tandem repeats in the rice genome. Recently, Rudenko and Korotkov (2023) classified tandem repeats (TRs) in the *Capsicum annuum* (pepper plant) genome.

Repeat search in genomes

The complexity measures were used for genome assembly, genome segmentation, search for low complexity regions (Gusev et al. 1991) and repeat masking (Jurka et al. 2007) from the time of first available genome data. RepeatMasker software tool is widely used to identify and mask repetitive genome elements, including low-complexity sequences (Jurka et al. 2007; Tarailo-Graovac and Chen 2009). Interspersed sequence repeats could be treated as low complexity regions of genome. Such repeats play important roles in the evolution, genome variation, and instability, cause the disease. RepeatMasker algorithm searches and classify the repetitive sequences using library of known repeats. RepBase (Jurka et al. 2007) and Dfam (Hubley et al. 2016) databases were most frequently used for masking genome repeats. Due to growth of new sequencing data, especially for non-model species, new msRepDB database became has become more popular for multi-species genome repeat analysis (Liao et al. 2022).

Searching for dispersed repeats in large eukaryotic genomes raises technical and methodical problems. Transposable elements are class of dispersed genome repeats widely represented in mammalian genomes. After insertion in a new position in the genome, Transposable elements accumulate mutations, which complicate their identification and annotation. New Highly Divergent Repeat Search Method suggested by Suvorova et al. (2021) make repeat search more effective than standard RepeatMasker. Recently, Korotkov et al. (2023) presented method for dispersed repeats search in bacterial genomes using an iterative procedure.

Low complexity heterochromatic regions of human genome centromeric such satellite arrays remained not completely sequenced till last year (Nurk et al. 2022). The T2T (Telomere-to-Telomere) Consortium finished complete sequencing of reference human genome (Nurk et al. 2022). New repeat elements in the genome were found previously unknown satellite arrays and mobile elements (Hoyt et al. 2022). Thus, interspersed repeats yet to be found despite detailed previous sequencing and genome assembly. The example of new repeat structure is Short Interrupted Repeat Cassette (SIRC) found in the *A. thaliana* genome (Gorbenko et al. 2023).

Complexity for next-generation sequencing reads

The complexity estimates are important for NGS reads mapping (te Boekhorst et al. 2016; Abnizova et al. 2017) and sequencing error correction. It was shown that the entropy in the sequencing reads not allow accurate mapping onto a reference genome. Moreover, low complexity of the reads relates to technological problem of sequence detection in Illumina sequencing platform (te Boekhorst et al. 2016). Some programs for sequencing error correction may introduce new errors in reads that overlapping low complexity regions. New error correction tool for Illumina sequencing data, BrownieCorrector, specially checks only the reads that overlap with highly repetitive (low complexity) regions in the genome (Heydari et al. 2019). The complexity estimates were used for large plant genome analysis—to analyze the repetitive sequence fraction in wheat (Sergeeva et al. 2014).

Complexity methods for proteins

Low complexity, repetitive protein sequences with a limited amino acid composition are common and important for the protein structure and function (Alba et al. 2002; Ntountoumi et al. 2019; Jarnot et al. 2020; Lee et al. 2022). Intrinsically disordered proteins have lower sequence complexity than ordered proteins, but have unique functions (Uversky 2016). There are set of tools for search of low complexity regions in proteins: SubSeqer (He and Parkinson 2008), 0j.py (Wise 2001), ProBias (Kuznetsov 2008). The Complexity tool has also universal option for amino acid sequences estimates as well as for other alphabets (RNA, grouped amino acids, binary DNA alphabet) (Orlov and Potapov 2004).

Complexity methods for gene regulatory regions analysis

Regulatory regions of gene transcription promoters, transcription factor binding sites, and its cluster also present hierarchical structure to be studied by the complexity methods (Abnizova et al. 2005). The problem is to find signal in gene promoter region and analyze their possible combinations (Vityaev et al. 2001, 2002; Voropaeva et al. 2019). Clusters of different transcription factor binding sites revealed by ChIP-seq technology (Chen et al. 2008) provide data for combinatorial analysis of such regions (Dergilev et al. 2022). It was shown that promoter sequences have varying text complexity, and this feature is statistically significant (Simões et al. 2021). But it is not enough for finding of transcription factor binding sites or weak signal for nucleosome positioning (Orlov et al. 2006a, b; Goh et al. 2010). However, the problem is to reveal the signals in DNA sequence itself, and deal with overrepresentation of the motifs (Abnizova et al. 2005). Note MEME software for analysis of repeated signals, such as transcription factor binding sites, in a sequence set (Tognon et al. 2023). So, the entropy estimates and text linguistic methods could not be used directly for combination of transcription factor elements. Recently, an extension of complexity measure called Abelian complexity was suggested for prediction of gene regulatory regions (Wu et al. 2019).

Analysis of genes and gene regulatory regions raised the challenge of searching for regions with low complexity (Hancock 2002; Wan et al. 2003). It could be used to find borders between coding and non-coding gene regions. Intuitively, the complexity of a symbolic sequence reflects an ability to represent a sequence based on some structural features of this sequence that need a repeated pattern—simple sequence repeats (Hancock 2002), recognizable direct and inverted repeats (Cox and Mirkin 1997). Following (Orlov and Potapov 2004) we note the methods of clusterization of cryptically simple sequences (Alba et al. 2002); evaluation of the alphabet-capacity *l*-gram (combinatorial complexity and linguistic complexity) (Kisliuk et al. 1999; Troyanskaya et al. 2002); complexity measures by Lempel and Ziv (Gusev et al. 1991;1999; Chen et al. 1999; Dai et al. 2013); stochastic complexity (Orlov et al. 2002), and grammatical complexity (Jimenez-Montano et al. 2002).

Alignment-free sequences comparison and visual methods

Visual presentation of DNA sequence in 2D and 3D also gives background for repeat search and new mathematical methods development (Dai et al. 2006; Xie and Mo 2011; Mo et al. 2018). Note chaos game presentation, new approaches such as algebraic biology to find patterns in genome sequences (Petoukhov 2017). We may also refer to these methods as to the methods of extended gene regions analysis.

Alignment-free approaches for sequence comparison assume compression-based sequence analysis. The alignment-free methods may be divided into two groups (Zielezinski et al. 2017): methods based on comparison word frequencies (Provata et al. 2014) and methods that evaluate mutual informational between sequences. In general, alignment-free sequence comparisons used the concepts derived from IT, such as entropy and mutual information (Vinga 2014).

There are also methods that cannot be classified into these groups, including those based on the length of matching words, chaos game representation (Löchel and Heider 2021), iterated maps, as well as graphical representation of DNA sequences, which capture the essence of the base composition in a quantitative manner (de la Fuente et al. 2023).

Compression-based complexity estimates

Discussing the compression-based sequence analysis note the general concept to estimate the complexity of symbolic sequence (text) suggested by Kolmogorov (1965). He proved that there exists an optimal algorithm or binary program p for a binary string s generation. The Kolmogorov complexity, K is the length $|p|$ of a shortest binary program p that computes s in a universal Turing machine and halts (Turing 1936). Complexity $K(s)=|p|$ is the size of compressed storage p—the minimum number of bits required to computationally reproduce the string s. In general, the Kolmogorov complexity is not computable in reasonable time for arbitrary sequence. Various constructive realizations of non-optimal coding have been developed (Lempel and Ziv 1976), including applications for DNA analysis (Gusev et al. 1999; Antão et al. 2018; Li and Vitányi 2019).

The concept of the complexity of a finite symbolic sequence as the compression size was introduced by Lempel and Ziv (Lempel, and Ziv 1976). Initially, this approach was implemented for analyzing DNA by Gusev and coauthors (Gusev et al. 1991; 1999). Based on this approach, we presented the Internet-available tools LZcomposer (http://wwwmgs.bionet.nsc.ru/mgs/programs/lzcomposer/) (Orlov et al 2003) and complexity (Orlov and Potapov 2004). Dai et al. (2013) used Lempel–Ziv decomposition (LZ-words) for sequence comparison without alignment. Note that Lempel–Ziv complexity algorithm could be applied to any sequence of signals (physiological time series) to study repeats and irregularities (Zhang et al. 2016). Thus, the same algorithm and software could be applied for non-DNA arbitrary alphabet. A relative Lempel–Ziv complexity measure was used for alignment-free sequence comparison (Liu et al. 2012). Pirogov et al. (2019) used Lempel–Ziv complexity, and the match complexity measure to analyze the relationship between the complexity and gene function. Enrichment of gene content and development genes in high-complexity genome regions was shown. Hosseini et al. (2020) developed Smash ++ , an alignment-free tool to find and visualize small- and large-scale genomic rearrangements between two DNA sequences. This tool also exploiting a data compression technique to find the rearrangements.

Methods and algorithms for DNA sequence complexity

General classification of complexity approaches

We overview approaches for sequence complexity measurement in the general scheme (Fig. 1). DNA sequence complexity as well as protein sequence complexity methods could be broadly classified into large groups—entropy-based and compression-based methods.

Entropy based methods of complexity estimates include word frequency (linguistic) approaches, Shannon entropy and its variants. Compression based methods include modifications of Lempel–Ziv compression scheme, could be applied for repeat search in genomes (direct, inverted, generated) and alignment-free genome comparisons. Spatial (3D) visualization of linear sequence, fractal presentation,

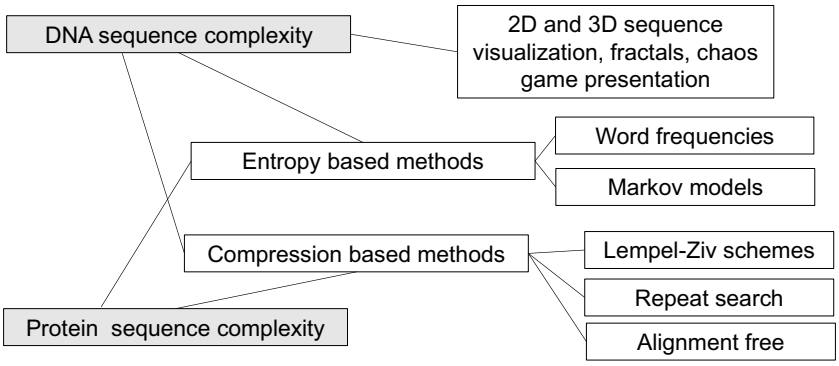

Fig. 1 Classification of the methods for sequence complexity analysis

sequence polarization, and other techniques could be used for genome analysis. Basically, protein sequence complexity estimates use entropy approaches, but may refer to all the methods for DNA complexity.

Algorithms for DNA sequence complexity estimates

Several estimates of complexity were incorporated to the complexity tool to compare different approaches (Orlov and Potapov 2004). It includes frequency of nucleotide content (Wootton and Federhen 1996), entropy estimates, and linguistic complexity (Gabrielian and Bolshoy 1999; Troyanskaya et al. 2002). By applying *l*-gram trees for the sequence representation in the complexity software the operation time for computation was optimized. We further refer to (Orlov and Potapov 2004) for the examples and details of sequence complexity algorithms.

Since main complexity algorithms were first published in 2000s, the software implementations differ in sequence size to be processed, optimization, and the applications areas. Though novel additions such as long-range correlations (Abnizova et al. 2007) and polarization coding (Zimnyakov et al. 2023) differ from information-based and compression-based techniques.

The Hurst exponent estimate for long-range correlation was added to the software to measure dependencies in DNA sequences (Orlov et al. 2006a, b). It was shown that the complexity of introns and regulatory regions is lower than that of coding regions, while Hurst exponent is larger due to long-range correlation between transcription factor binding sites (Abnizova et al. 2007). Promoter sequences have lower complexity than protein coding regions. Long-range correlation analysis tool was implemented in CorGen software (as http://corgen.molgen.mpg.de) (Messer and Arndt 2006). The examples of sequence complexity for transcription factor binding sites were considered (Orlov et al. 2006a, b). It was noted that the DNA sequence of transcript factor binding sites have in average lower complexity than protein coding regions.

Naumenko et al. (2018) used complexity estimates to reveal artefacts in short sequencing read mapping on a chromosome (aligner artefacts) (te Boekhorst et al. 2016; Naumenko et al. 2018; Subkhankulova et al. 2021).

Nucleotide sequences containing human mutation sites are associated with varying sequence complexity related to mutagenesis mechanism (Chuzhanova et al. 2003). Complexity estimates for the analysis of mutation sites (SNP containing regions) confirmed presence of low complexity regions at the flanking sites of mutation/polymorphism position (Safronova et al. 2015; 2016).

Nucleotide sequences forming non-B DNA structures (not double DNA helix) have repeated patterns as palindromes that can form hairpins, cruciform or triplexes. To analyze the sequences potentially forming non-B DNA structures the NeSSie tool was presented (Berselli et al. 2018).

Gene expression regulation studies (Orlov and Baranova 2020; Voropaeva et al. 2019) give broad field for application of application of information and entropy measures. Nucleotide sequences containing binding sites of many protein transcription factors have symmetrical structure due to contacts with protein dimers. Thus, due to presence of the repeated elements the transcription factor bindings sites have lower sequence complexity. Transcription factor binding sites have been catalogued in the databases such as TRRD, TRANSFAC, and JASPAR (Heinemeyer et al. 1998; Sandelin et al. 2004) filled by data high-throughput sequencing technologies (ChIP-seq, ATAC-seq and related approaches) (Chen et al. 2008). The clusters of transcription factor (TF) binding sites in genomes were constructed based on ChIP-seq data (Dergilev et al. 2016, 2022). The effect of lower complexity estimates was shown for longer gene regulatory regions (with multiple TF binding) in plants (Dergilev et al. 2022). Recently, cooperative effect of transcription factor binding in sequential and spatial proximity was shown using chromatin conformation capture data (Vadnala et al. 2023), thus extending the theory of chromosome and high order regulatory codes.

Note recent work original by Zimnyakov et al. (2023) on revealing and visualization of sequence dependencies using

so called polarization coding. It refers to 2D sequence visualization for pattern search (Dai et al. 2006). The polarization coding presents nucleotide sequence in a two-dimensional phase screen, where each element corresponds to a specific nucleotide. The polarization-based technique was shown on the model data from a comparative analysis of the spike protein gene sequences of the SARS-CoV-2 virus complementing other complexity-based works on this topic (Akbari Rokn Abadi et al. 2023).

Results

Comparison of the methods and analysis of sequence sets

The complexity estimates for different classes of genome sequences—exons, introns, regulatory nucleotide sequences—confirmed general theory about overlapping genetic codes suggested by E.N.Trifonov (Trifonov 1989, 1990). The more genetics messages have the sequence, the higher its complexity is. It was demonstrated that the complexity of exons is, on average, higher, whereas that of introns is lower (Orlov and Potapov 2004). The alteration in the local complexity for splicing sites was shown using sequences the SpliceDB database (Burset et al. 2001). The splicing sites in eukaryotic genes have intermediate place between protein coding regions (exons) and non-coding regions (introns). The average complexity of protein coding genes is higher than for non-coding, as it might be expected (short sequence repeats, simple repeats and polytracks (like AAAA…, TTTTT…) are common for introns making lower complexity). So, the average complexity of splicing sites sequence has corresponding intermediate place between complexity exons and introns (Orlov and Potapov 2004).

The complexity estimates were used for large eukaryotic genome analysis—in plants. To analyze the repetitive sequence fraction in plant genomes such as wheat Sergeeva et al. (2014) used RepeatMasker program (http://www.repeatmasker.org/) and calculated GC content and total content of satellites, simple repeats, and low complexity regions (Orlov et al. 2006a, b).

To search for satellite repeat structures such as simple sequence repeats (GAA)n(GGA)m and telomeric (TTTAGGG)n satellites and inverted repeats in chromosome 5B 454 sequences, the authors used a custom script based on the Lempel–Ziv approach, which identified the perfect satellite repeat tracts and inverted, as well as direct, repeat structures in DNA sequences (Sergeeva et al. 2014). This program works with different types of repeats and does not require sequence alignment. The algorithm implements structural and comparative analysis of the significant amount of collections of DNA fragments of moderate length, developed at the Sobolev Institute of Mathematics (Gusev et al. 1999; Orlov et al. 2003). Analysis of genome inversions in plant as mutual genome comparison method was presented for rice genome (Suvorova et al. 2021; Zhou et al. 2023). The lack of consensus concerning the biological meaning of entropy and complexity of genomes and the different ways to assess these data hamper conclusions concerning what are the causes of genomic entropy variation among species (Simões et al. 2021).

Online tools for sequence complexity analysis

There are novel online tools for the sequence complexity analysis and information processing that were not widely published, being presented at the conferences, such as ICGenomics (Orlov et al. 2020). Table 1 shows online tools for text complexity estimates for DNA and proteins.

Table 1 Existing tools for complexity estimates

Sequence type	Tool name	URL	Reference
Protein	RES repeatability scanner	http://cbdm-01.zdv.uni-mainz.de/~munoz/res/	(Kamel et al. 2019)
Protein	0j.py	https://doi.org/10.1093/bioinformatics/17.suppl_1.S288	(Wise 2001)
Proteins and DNA	fLPS 2.0 (find low probability subsequences)	https://biology.mcgill.ca/faculty/harrison/flps.html	(Harrison 2017)
Proteins and DNA	Complexity	http://wwwmgs.bionet.nsc.ru/mgs/programs/low_complexity/	(Orlov and Potapov 2004)
Proteins and DNA	CLC (Local complexity)	https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/750/index.php	(Wootton and Federhen 1993)
Protein	ProBias	http://lcg.rit.albany.edu/ProBias/	(Kuznetsov 2008)
Protein	Subseqer	https://www.compsysbio.org/subseqer	(He and Parkinson 2008)
DNA	Macle	http://guanine.evolbio.mpg.de/complexity/	(Pirogov et al. 2019)
Protein	PlaToLoCo	http://platoloco.aei.polsl.pl	(Jarnot et al. 2020)
DNA	CorGen	http://corgen.molgen.mpg.de/	(Messer and Arndt 2006)

Some online tools developed early are no longer accessible or have no web-version to be included to Table 1. In CLC Genomics Workbench it is possible to calculate local complexity for both DNA and protein sequences (CLC bio was a bioinformatics software company that developed the software suite. It was subsequently purchased by QIAGEN (https://digitalinsights.qiagen.com/products-overview/discovery-insights-portfolio/analysis-and-visualization/qiagen-clc-genomics-workbench/). The local complexity realizes measure of the diversity in the amino acid composition (Wootton and Federhen 1993).

The AC tool was created for compression of amino acid sequences (Hosseini et al. 2019). New version of his protein sequence compression tool, AC2, was proved to be more effective for specialized compression purposes (Silva et al. 2021). The methods to detect such regions in protein include classical entropy, SEG (Wootton and Federhen 1993) measure, and other tools such as LCR-eXXXplorer (Kirmitzoglou and Promponas 2015), see also Table 1.

Low complexity analysis tools for protein sequences

Low complexity regions of proteins are abundant in proteomes (Lee et al. 2022). Low complexity regions in the proteins have important functional roles (Alba et al. 2002). They are highly conserved (Ntountoumi et al. 2019). Contrary to a widespread belief based on older and not complete data, low complexity regions have a significant, persistent, and highly conserved presence in many prokaryotes. Their specific amino acid content is linked to proteins with certain molecular functions, such as the binding of RNA, ions and polysaccharides (Jarnot et al. 2020). Jarnot et al. (2022) show that existing methods for protein similarity search need improvements to count low complexity regions. Li and Kahveci (2006) defined new complexity measures to compute the complexity of a sequence based on a given scoring matrix, such as BLOSUM 62.

Database compression and large scale analysis

Compression of genomes data for all the studied species is important by technical reasons. Growth of NGS (Next Generation Sequencing) data challenge analysis of multiple sequences and effective database storage (Agenis-Nevers et al. 2021). Due to multiple sequence repeats this task could be solved using operational compression algorithms (variants of Lempel–Ziv compression). Mapping, quality control, and redundancy removal are related to sequence complexity. Comparison of existing tools Mardre (Expósito et al. 2017), Bioseqzip (Urgese et al. 2020), and others to process such sequencing database by algorithmic complexity was presented by (de Oliveira Veras 2021). Overall, the operational complexity could be used for duplicate removal and effective NGS sequencing database processing. Future application here are for cloud computing. Although the complexity of algorithms is not a new subject, there is a lack of materials within the area. Note first works by Gusev et al. (1999) for optimization of Lempel-Zive algorithm for DNA compression in terms of operation time. Performance of order $O(n \log n)$ and even $O(\log \log n)$ was shown (de Oliveira Veras 2021).

Mutual information measures (relative information could be used for complete genomes comparison, alignment free method (Veluchamyet al. 2021). At the same time, it could be used for compression of genome databases. The iDoComp tool may compress an individual genome using a reference data (Ochoa et al. 2015).

Discussion

Previously, the complexity estimates were applied for analysis of decompositions of several complete bacterial genomes and fragments of eukaryotic chromosomes (Orlov et al. 2003). The complexity of sequences containing introns and regulatory regions is less than that of coding regions (Orlov and Potapov 2004). This observation is also valid in eukaryotes by estimating complexity of gene regions using several other complexity measures (Orlov et al. 2006a, b). Modern works on complexity use large scale calculations, present online tools for convenient and reproducible complexity analysis ((Jarnot et al. 2020; Pirogov et al. 2019). Recent studies described low complexity regions and compressed hundreds of complete genome sequences (Agenis-Nevers et al. 2021; Munagala et al. 2022).

The problem of long-range correlations in genome could be also studied by sequence complexity methods. Experimental data on 3D genomics show presence of topologically associated domains (TAD) in eukaryotes (Li et al. 2012; Kulakova et al. 2017). The DNA sequences from such chromosome regions are close in 3D cell space giving new point of view for long range genome correlations and topological code (Vadnala et al. 2023). Such topologically proximal sequences should be analyzed by complexity methods (Kulakova et al. 2015). Such TAD in chromosomes present higher level of sequence constraints (chromosome come) after triple coding, gene regulation signals, and nucleosome positioning. Thus, the theory of multiple genetic codes (Trifonov 1990) could be extended to new signals and repetitive sequence elements.

Repetitive sequences in annotated non-coding RNAs (ncRNAs) are found to constitute functional components that perform specific biological functions (Zeng et al. 2022). Complexity measures are applied to novel data such as ncRNAs (miRNA, siRNA, tsRNA, circRNA, lncRNA) in plants (Chao et al. 2022).

Alignment-free technique for sequence analysis is wide area for complexity algorithms. Information theory and data compression algorithms provide mathematical and computational tools to capture essential patterns in biological sequences (Bonidia et al. 2022). Recently, Munagala et al. (2022) investigated the use of compression-complexity based distance measures for analyzing genomic sequences. The proposed distance measure is used to successfully reproduce the phylogenetic trees for a mammalian dataset consisting of eight species clusters, a set of coronaviruses. *k*-mer and physic-chemical properties of nucleotides were recently used for SARS-CoV-2 genomes classification (Akbari Rokn Abadi et al. 2023).

Inversion index (number of genome inversion) was used for rice genome structure analysis (*Oryza sativa*) (Zhou et al. 2023). The authors used 73 genomes of rice (*O. sativa*) and the genomes of wild relative species to build a pan-genome inversion index for the reference genome sequence. Detailed analyses of these inversions show evidence of their effects on gene expression, recombination rate, and linkage disequilibrium. Complex plant genomes became object of the hidden periodicity search (Suvorova et al. 2021). Recombinations and inversions in plant genomes could be revealed by sequence compression technique (Chao et al. 2023).

Scaling of computations in comparative genomics demand new algorithm development. Bello et al. (2023) used text compression to accelerate algorithms for Hidden Markov Models. Their work provides an efficient approach to big data computations with HMM using compression measures.

Overall, information theory is widely used for model development and data analysis for a variety of biologically derived data types ranging from molecular, sequence and phenotypic data in genomics and genetics to gene expression, protein and spectral data in transcriptomics, proteomics and metabolomics, respectively (Chanda et al. 2020; Bartal and Jagodnik 2022). Sequence compression of whole genomes became routine procedure to be standardized in benchmark test—Sequence Compression Benchmark database (Kryukov et al. 2020). Novel machine learning applications for classification of complete genome sequences also include sequence compression and complexity algorithms (Silva et al. 2021; Akbari Rokn Abadi et al. 2023). We have reviewed here text complexity applications starting from basic definitions and algorithms to the online applications and approaches for large-scale NGS analysis and machine learning techniques. The complexity analysis, sequence compression, and information-based methods gave raise to new findings and challenges in molecular biophysics such as coronavirus genome studies (Munagala et al. 2022).

We conclude by mentioning new application areas of sequence complexity estimates in Big Data and Machine Learning methods. Statistical estimates of sequence complexity and periodicities patterns could be using in Machine Leaning application as input parameters that might be even not interpretable (Silva et al. 2021; Munagala et al. 2022; Balcı et al. 2023). There are new AI solutions in bioinformatics that use additional sequence features and statistics as input parameters (Expósito et al. 2017; Penzar et al. 2023).

Acknowledgements The authors are grateful to the Committee of the Russian Biophysicists Society, to A. I. Dergilev and A. V. Mitina for technical help.

Author contribution All the authors conceptualized and outlined the review.

Funding The publication was prepared with the support of the RUDN University Strategic Academic Leadership Program (Y. O.)

Data availability Not applicable.

Code availability The reviewed software applications are available online (See Table 1).

Declarations

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Conflict of interest The authors declare no competing interests.

References

Abnizova I, te Boekhorst R, Walter K, Gilks WR (2005) Some statistical properties of regulatory DNA sequences, and their use in predicting regulatory regions in the Drosophila genome: the fluffy-tail test. BMC Bioinformatics 6:109. https://doi.org/10.1186/1471-2105-6-109

Abnizova I, Walter K, Te Boekhorst R, Elgar G, Gilks WR (2007) Statistical information characterization of conserved non-coding elements in vertebrates. J Bioinform Comput Biol 5(2B):533–547. https://doi.org/10.1142/s0219720007002898

Abnizova I, te Boekhorst R, Orlov Y (2017) Computational errors and biases of short read next generation sequencing. J Proteom Bioinform 10:1–17. https://doi.org/10.4172/jpb.1000420

Agenis-Nevers M, Bokde ND, Yaseen ZM, Shende MK (2021) An empirical estimation for time and memory algorithm complexities: newly developed R package. Multimed Tools Appl 80(2):2997–3015. https://doi.org/10.1007/s11042-020-09471-8

Akbari Rokn Abadi S, Mohammadi A, Koohi S (2023) A new profiling approach for DNA sequences based on the nucleotides' physicochemical features for accurate analysis of SARS-CoV-2 genomes. BMC Genomics 24(1):266. https://doi.org/10.1186/s12864-023-09373-7

Alba MM, Laskowski RA, Hancock JM (2002) Detecting cryptically simple protein sequences using the SIMPLE algorithm. Bioinformatics 18:672–678. https://doi.org/10.1093/bioinformatics/18.5.672

Antão R, Mota A, Tenreiro Machado JA (2018) Kolmogorov complexity as a data similarity metric: application in mitochondrial DNA. Nonlinear Dyn 93(3):1059–1071. https://doi.org/10.1007/s11071-018-4245-7

Balcı AT, Ebeid MM, Benos PV, Kostka D, Chikina M (2023) An intrinsically interpretable neural network architecture for sequence-to-function learning. Bioinformatics 39(39 Suppl 1):i413–i422. https://doi.org/10.1093/bioinformatics/btad271

Bartal A, Jagodnik KM (2022) Progress in and opportunities for applying information theory to computational biology and bioinformatics. Entropy (basel) 24(7):925. https://doi.org/10.3390/e24070925

Bello L, Wiedenhöft J, Schliep A (2023) Compressed computations using wavelets for hidden Markov models with continuous observations. PLoS One 18(6):e0286074. https://doi.org/10.1371/journal.pone.0286074

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27(2):573–580. https://doi.org/10.1093/nar/27.2.573

Bernaola-Galván P, Carpena P, Gómez-Martín C, Oliver JL (2023) Compositional structure of the genome: a review. Biology (basel) 12(6):849. https://doi.org/10.3390/biology12060849

Berselli M, Lavezzo E, Toppo S (2018) NeSSie: a tool for the identification of approximate DNA sequence symmetries. Bioinformatics 34(14):2503–2505. https://doi.org/10.1093/bioinformatics/bty142

Bonidia RP, Avila Santos AP, de Almeida BLS, Stadler PF, Nunes da Rocha U, Sanches DS, de Carvalho ACPLF (2022) Information theory for biological sequence classification: a novel feature extraction technique based on Tsallis entropy. Entropy (basel) 24(10):1398. https://doi.org/10.3390/e24101398

Burset M, Seledtsov IA, Solovyev VV (2001) SpliceDB: database of canonical and non-canonical mammalian splice sites. Nucleic Acids Res 29:255–259. https://doi.org/10.1093/nar/29.1.255

Chanda P, Costa E, Hu J, Sukumar S, Van Hemert J, Walia R (2020) Information theory in computational biology: where we stand today. Entropy 22(6):627. https://doi.org/10.3390/e22060627

Chang CH, Hsieh LC, Chen TY, Chen HD, Luo L, Lee HC (2005) Shannon information in complete genomes. J Bioinform Comput Biol 3(3):587–608. https://doi.org/10.1142/s0219720005001181

Chao H, Hu Y, Zhao L et al (2022) Biogenesis, functions, interactions, and resources of non-coding RNAs in plants. Int J Mol Sci 23(7):3695. https://doi.org/10.3390/ijms23073695

Chao H, Zhang S, Hu Y, Ni Q, Xin S, Zhao L, Ivanisenko VA, Orlov YL, Chen M (2023) Integrating omics databases for enhanced crop breeding. J Integr Bioinform. https://doi.org/10.1515/jib-2023-0012. (Online ahead of print)

Chen X, Kwong S, Li MA (1999) Compression algorithm for DNA sequences and its applications in genome comparison. Genome Inform Ser Workshop Genome Inform 10:51–61. https://doi.org/10.11234/gi1990.10.51

Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB et al (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell 133:1106–1117. https://doi.org/10.1016/j.cell.2008.04.043

Chuzhanova NA, Anassis EJ, Ball E, Krawczak M, Cooper DN (2003) Meta-analysis of indels causing human genetic disease: mechanisms of mutagenesis and the role of local DNA sequence complexity. Hum Mutation 21:28–44. https://doi.org/10.1002/humu.10146

Cox R, Mirkin SM (1997) Characteristic enrichment of DNA repeats in different genomes. Proc Natl Acad Sci USA 94:5237–5242. https://doi.org/10.1073/pnas.94.10.5237

Dai Q, Liu X, Wang T (2006) A novel 2D graphical representation of DNA sequences and its application. J Mol Graph Model 25(3):340–344. https://doi.org/10.1016/j.jmgm.2005.12.004

Dai Q, Yan Z, Shi Z, Liu X, Yao Y, He P (2013) Study of LZ-word distribution and its application for sequence comparison. J Theor Biol 336:52–60. https://doi.org/10.1016/j.jtbi.2013.07.008

de la Fuente R, Díaz-Villanueva W, Arnau V, Moya A (2023) Genomic signature in evolutionary biology: a review. Biology (basel) 12(2):322. https://doi.org/10.3390/biology12020322

de Oliveira Veras AA (2021) Complexity analysis of algorithms: a case study on bioinformatics tools. World J Biol Biotechnol 6(3): 11-13. https://doi.org/10.33865/wjb.006.03.0445. Available at: <https://sciplatform.com/index.php/wjb/article/view/445> (Date accessed: 31 Aug. 2023)

Delucchi M, Näf P, Bliven S, Anisimova M (2021) TRAL 2.0: tandem repeat detection with circular profile hidden Markov models and evolutionary aligner. Front Bioinform 1:691865. https://doi.org/10.3389/fbinf.2021.691865

Deng S, Shi Y, Yuan L, Li Y, Ding G (2012) Detecting the borders between coding and non-coding DNA regions in prokaryotes based on recursive segmentation and nucleotide doublets statistics. BMC Genomics 13(Suppl 8):S19. https://doi.org/10.1186/1471-2164-13-S8-S19

Dergilev AI, Spitsina AM, Chadaeva IV, Svichkarev AV, Naumenko FM, Kulakova EV et al (2016) Computer analysis of colocalization of the TFs' binding sites in the genome according to the ChIP-seq data. Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding 20(6):770–778. https://doi.org/10.18699/VJ16.194. (In Russian).

Dergilev AI, Orlova NG, Dobrovolskaya OB, Orlov YL (2022) Statistical estimates of multiple transcription factors binding in the model plant genomes based on ChIP-seq data. J Integr Bioinform 19(1):20200036. https://doi.org/10.1515/jib-2020-0036

Dios F, Barturen G, Lebron R, Rueda A, Hackenberg ML, Oliver JL (2014) DNA clustering and genome complexity. Comput Biol Chem 53(PA):71–78. https://doi.org/10.1016/j.compbiolchem.2014.08.011

Expósito RR, Veiga J, González-Domínguez J, Touriño J (2017) Mardre: efficient mapreduce-based removal of duplicate DNA reads in the cloud. Bioinformatics 33(17):2762–2764. https://doi.org/10.1093/bioinformatics/btx307

Frenkel FE, Korotkova MA, Korotkov EV (2017) Database of periodic DNA regions in major genomes. BioMed Res Int 2017:7949287, 9. https://doi.org/10.1155/2017/7949287

Frenkel FE, Korotkov EV (2008) Classification analysis of triplet periodicity in protein-coding regions of genes. Gene 421(1–2):52–60. https://doi.org/10.1016/j.gene.2008.06.012

Frenkel FE, Korotkov EV (2009) Using triplet periodicity of nucleotide sequences for finding potential reading frame shifts in genes. DNA Res 16(2):105–114. https://doi.org/10.1093/dnares/dsp002

Gabrielian A, Bolshoy A (1999) Sequence complexity and DNA curvature. Comput Chem 23:263–274. https://doi.org/10.1016/S0097-8485(99)00007-8

Goh WS, Orlov Y, Li J, Clarke ND (2010) Blurring of high-resolution data shows that the effect of intrinsic nucleosome occupancy on transcription factor binding is mostly regional, not local. PLoS Comput Biol 6(1):e1000649

Gorbenko IV, Petrushin IS, Shcherban AB, Orlov YL, Konstantinov YM (2023) Short interrupted repeat cassette (SIRC)—novel type of repetitive DNA element found in *Arabidopsis thaliana*. Int J Mol Sci 24(13):11116. https://doi.org/10.3390/ijms241311116

Gusev VD, Kulichkov VA, Chupakhina OM (1991) Complexity analysis of genomes. I. Complexity and classification methods of detected structural regularities. Mol Biol (mosk) 25:825–834

Gusev VD, Nemytikova LA, Chuzhanova NA (1999) On the complexity measures of genetic sequences. Bioinformatics 15:994–999. https://doi.org/10.1093/bioinformatics/15.12.994

Hancock JM (2002) Genome size and the accumulation of simple sequence repeats: implications of new data from genome sequencing projects. Genetica 115:93–103. https://doi.org/10.1023/A:1016028332006

Harrison PM (2017) fLPS: fast discovery of compositional biases for the protein universe. BMC Bioinformatics 18(1):476. https://doi.org/10.1186/s12859-017-1906-3

He D, Parkinson J (2008) SubSeqer: a graph-based approach for the detection and identification of repetitive elements in low-complexity sequences. Bioinformatics 24(7):1016–1017. https://doi.org/10.1093/bioinformatics/btn073

Heinemeyer T, Wingender E, Reuter I, Hermjakob H, Kel AE, Kel OV, Ignatieva EV, Ananko EA, Podkolodnaya OA, Kolpakov FA, Podkolodny NL, Kolchanov NA (1998) Databases on transcriptional regulation: TRANSFAC. TRRD and COMPEL Nucleic Acids Res 26(1):362–367. https://doi.org/10.1093/nar/26.1.362

Heydari M, Miclotte G, Van de Peer Y, Fostier J (2019) Illumina error correction near highly repetitive DNA regions improves de novo genome assembly. BMC Bioinformatics 20(1):298. https://doi.org/10.1186/s12859-019-2906-2

Hosseini M, Pratas D, Pinho AJ (2019) AC: a compression tool for amino acid sequences. Interdiscip Sci 11(1):68–76. https://doi.org/10.1007/s12539-019-00322-1

Hosseini M, Pratas D, Morgenstern B, Pinho AJ (2020) Smash++: an alignment-free and memory-efficient tool to find genomic rearrangements. Gigascience 9(5):giaa048. https://doi.org/10.1093/gigascience/giaa048

Hoyt SJ, Storer JM, Hartley GA et al (2022) From telomere to telomere: the transcriptional and epigenetic state of human repeat elements. Science 376(6588):eabk3112. https://doi.org/10.1126/science.abk3112

Hubley R, Finn RD, Clements J et al (2016) (2016) The Dfam database of repetitive DNA families. Nucleic Acids Res 44(D1):D81–D89. https://doi.org/10.1093/nar/gkv1272

Jarnot P, Ziemska-Legiecka J, Dobson L et al (2020) PlaToLoCo: the first web meta-server for visualization and annotation of low complexity regions in proteins. Nucleic Acids Res 48(W1):W77–W84. https://doi.org/10.1093/nar/gkaa339

Jarnot P, Ziemska-Legiecka J, Grynberg M, Gruca A (2022) Insights from analyses of low complexity regions with canonical methods for protein sequence comparison. Brief Bioinform 23(5):bbac299. https://doi.org/10.1093/bib/bbac299

Jimenez-Montano MA, Ebeling W, Pohl T, Rapp PE (2002) Entropy and complexity of finite sequences as fluctuating quantities. Biosystems 64:23–32

Jorda J, Kajava AV (2009) T-REKS: identification of Tandem REpeats in sequences with a K-means based algorithm. Bioinformatics 25(20):2632–2638. https://doi.org/10.1093/bioinformatics/btp482

Jurka J, Kapitonov VV, Kohany O, Jurka MV (2007) Repetitive sequences in complex genomes: structure and evolution. Annu Rev Genomics Hum Genet 8:241–259. https://doi.org/10.1146/annurev.genom.8.080706.092416

Kamel M, Mier P, Tari A, Andrade-Navarro MA (2019) Repeatability in protein sequences. J Struct Biol 208(2):86–91. https://doi.org/10.1016/j.jsb.2019.08.003

Karakatsanis LP, Pavlos EG, Tsoulouhas G, Stamokostas GL, Mosbruger T, Duke JL, Pavlos GP, Monos DS (2021) Spatial constrains and information content of sub-genomic regions of the human genome. iScience 24(2):102048. https://doi.org/10.1016/j.isci.2021.102048

Kirmitzoglou I, Promponas VJ (2015) LCR-eXXXplorer: a web platform to search, visualize and share data for low complexity regions in protein sequences. Bioinformatics 31(13):2208–2210. https://doi.org/10.1093/bioinformatics/btv115

Kisliuk OS, Borovina TA, Nazipova NN (1999) Otsenka izbytochnosti geneticheskikh tekstov s pomoshch'iu vysokochastotnoĭ komponenty grafa l-grammnogo razlozheniia [Evaluation of genetic test redundancy using a high-frequency component of the l-gram graph]. Biofizika 44(4):639–648 (in Russian)

Kolmogorov AN (1965) Three approaches to definition of information quantity. Probl Peredachi Inf 1:3–11 (in Russian)

Kolpakov R, Bana G, Kucherov G (2003) mreps: efficient and flexible detection of tandem repeats in DNA. Nucleic Acids Res 31(13):3672–3678. https://doi.org/10.1093/nar/gkg617

Korotkov EV, Kamionskya AM, Korotkova MA (2021) Detection of highly divergent tandem repeats in the rice genome. Genes 12(4):473. https://doi.org/10.3390/genes12040473

Korotkov E, Zaytsev K, Fedorov A (2022) Use of 6 nucleotide length words to study the complexity of gene sequences from different organisms. Entropy 24(5):632. https://doi.org/10.3390/e24050632

Korotkov E, Suvorova Y, Kostenko D, Korotkova M (2023) Search for dispersed repeats in bacterial genomes using an iterative procedure. Int J Mol Sci 24(13):10964. https://doi.org/10.3390/ijms241310964

Kryukov K, Ueda MT, Nakagawa S, Imanishi T (2020) Sequence Compression Benchmark (SCB) database-a comprehensive evaluation of reference-free compressors for FASTA-formatted sequences. Gigascience 9(7):giaa072. https://doi.org/10.1093/gigascience/giaa072

Kulakova EV, Spitsina AM, Orlova NG, Dergilev AI, Svichkarev AV, Safronova NS et al (2015) Supercomputer analysis of genomics and transcriptomics data revealed by high-throughput DNA sequencing. Program Syst: Theory Appl 6(2):129–148. https://doi.org/10.25209/2079-3316-2015-6-2-129-148. (in Russian)

Kulakova EV, Spitsina AM, Bogomolov AG, Orlova NG, Dergilev AI, Chadaeva IV et al (2017) Program for analysis of genome distribution of chromosome contacts in cell nucleus by the data obtained using ChIA-PET and Hi-C technologies. Program Syst: Theory Appl 8:219–142. https://doi.org/10.25209/2079-3316-2017-8-1-219-242. (in Russian)

Kuznetsov IB (2008) ProBias: a web-server for the identification of user-specified types of compositionally biased segments in protein sequences. Bioinformatics 24(13):1534–1535. https://doi.org/10.1093/bioinformatics/btn233

Lee B, JaberiLashkari N, Calo E (2022) A unified view of low complexity regions (LCRs) across species. Elife 11:e77058. https://doi.org/10.7554/eLife.77058

Lempel A, Ziv J (1976) On the complexity of finite sequences. IEEE Trans Inf Theory IT 22:75–81

Li X, Kahveci T (2006) A novel algorithm for identifying low-complexity regions in a protein sequence. Bioinformatics 22(24):2980–2987. https://doi.org/10.1093/bioinformatics/btl495

Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, Poh HM, Goh Y, Lim J, Zhang J, Sim HS, Peh SQ, Mulawadi FH, Ong CT, Orlov YL, Hong S, Zhang Z, Landt S, Raha D, Euskirchen G, Wei CL, Ge W, Wang H, Davis C, Fisher-Aylor KI, Mortazavi A, Gerstein M, Gingeras T, Wold B, Sun Y, Fullwood MJ, Cheung E, Liu E, Sung WK, Snyder M, Ruan Y (2012) Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell 148(1–2):84–98. https://doi.org/10.1016/j.cell.2011.12.014

Li Y, Roberts ND, Wala JA, Shapira O, Schumacher SE, Kumar K, Khurana E, Waszak S, Korbel JO, Haber JE, Imielinski M, PCAWG Structural Variation Working Group, Weischenfeldt J, Beroukhim R, Campbell PJ, PCAWG Consortium (2020) Patterns

of somatic structural variation in human cancer genomes. Nature 578(7793):112–121. https://doi.org/10.1038/s41586-019-1913-9
Li M, Vitányi P (2019) An introduction to Kolmogorov complexity and its applications. Texts in Computer Science. Springer Cham.p 834. https://doi.org/10.1007/978-3-030-11298-1
Liao X, Hu K, Salhi A, Zou Y, Wang J, Gao X (2022) msRepDB: a comprehensive repetitive sequence database of over 80 000 species. Nucleic Acids Res 50(D1):D236–D245. https://doi.org/10.1093/nar/gkab1089
Liu L, Li D, Bai F (2012) A relative Lempel-Ziv complexity: application to comparing biological sequences. Chem Phys Lett 530:107–112. https://doi.org/10.1016/j.cplett.2012.01.061
Löchel HF, Heider D (2021) Chaos game representation and its applications in bioinformatics. Comput Struct Biotechnol J 19:6263–6271. https://doi.org/10.1016/j.csbj.2021.11.008
Menghi F, Barthel FP, Yadav V et al (2018) The tandem duplicator phenotype is a prevalent genome-wide cancer configuration driven by distinct gene mutations. Cancer Cell 34(2):197-210.e5. https://doi.org/10.1016/j.ccell.2018.06.008
Messer PW, Arndt PF (2006) CorGen–measuring and generating long-range correlations for DNA sequence analysis. Nucleic Acids Res 34:W692–W695. https://doi.org/10.1093/nar/gkl234
Mo Z, Zhu W, Sun Y et al (2018) One novel representation of DNA sequence based on the global and local position information. Sci Rep 8(1):7592. https://doi.org/10.1038/s41598-018-26005-3
Mudunuri SB, Kumar P, Rao AA, Pallamsetty S, Nagarajaram HA (2010) G-IMEx: a comprehensive software tool for detection of microsatellites from genome sequences. Bioinformation 5:221–223. https://doi.org/10.6026/97320630005221
Munagala NVTS, Amanchi PK, Balasubramanian K, Panicker A, Nagaraj N (2022) Compression-complexity measures for analysis and classification of coronaviruses. Entropy (basel) 25(1):81. https://doi.org/10.3390/e25010081
Naumenko FM, Abnizova II, Beka N, Genaev MA, Orlov YL (2018) Novel read density distribution score shows possible aligner artefacts, when mapping a single chromosome. BMC Genomics 19(Suppl 3):92. https://doi.org/10.1186/s12864-018-4475-6
Ntountoumi C, Vlastaridis P, Mossialos D, Stathopoulos C, Iliopoulos I, Promponas V, Oliver SG, Amoutzias GD (2019) Low complexity regions in the proteins of prokaryotes perform important functional roles and are highly conserved. Nucleic Acids Res 47(19):9998–10009. https://doi.org/10.1093/nar/gkz730
Nurk S, Koren S, Rhie A et al (2022) The complete sequence of a human genome. Science 376(6588):44–53. https://doi.org/10.1126/science.abj6987
Ochoa I, Hernaez M, Weissman T (2015) iDoComp: a compression scheme for assembled genomes. Bioinformatics 31(5):626–633. https://doi.org/10.1093/bioinformatics/btu698
Olson D, Wheeler T (2018) ULTRA: a model based tool to detect tandem repeats. ACM BCB 2018:37–46. https://doi.org/10.1145/3233547.3233604
Orlov YL, Baranova AV (2020) Editorial: bioinformatics of genome regulation and systems biology. Front Genet 11:625. https://doi.org/10.3389/fgene.2020.00625
Orlov YL, Potapov VN (2004) Complexity: an internet resource for analysis of DNA sequence complexity. Nucleic Acids Res 32:W628-633. https://doi.org/10.1093/nar/gkh466
Orlov YL, Filippov VP, Potapov VN, Kolchanov NA (2002) Construction of stochastic context trees for genetic texts. In Silico Biol 2(3):233–247
Orlov YL, Gusev VD, Miroshnichenko LA (2003) LZcomposer: decomposition of genomic sequences by repeat fragments. Biofizika 48(1):7–16
Orlov IuL, Levitskiĭ VG, Smirnova OG, Podkolodnaia OA, Khlebodarova TM, Kolchanov NA (2006a) Statistical analysis of DNA sequences containing nucleosome positioning sites. Biofizika 51(4):608–614 (In Russian)
Orlov YL, Te Boekhorst R, Abnizova II (2006b) Statistical measures of the structure of genomic sequences: entropy, complexity, and position information. J Bioinform Comput Biol 4(2):523–536. https://doi.org/10.1142/s0219720006001801
Orlov YL, Bragin AO, Babenko RO, Dresvyannikova AE, Kovalev SS, Shaderkin IA et al (2020) Integrated computer analysis of genomic sequencing data based on ICGenomics tool. In: Hu, Z., Petoukhov, S., He, M. (eds) Advances in intelligent systems, computer science and digital economics. CSDEIS 2019. Advances in Intelligent Systems and Computing, vol 1127. Springer, Cham. https://doi.org/10.1007/978-3-030-39216-1_15
Pellegrini M, Renda ME, Vecchio A (2010) TRStalker: an efficient heuristic for finding fuzzy tandem repeats. Bioinformatics 26:i358–i366. https://doi.org/10.1093/bioinformatics/btq209
Penzar D, Nogina D, Noskova E, Zinkevich A, Meshcheryakov G, Lando A, Rafi AM, de Boer C, Kulakovskiy IV (2023) LegNet: a best-in-class deep learning model for short DNA regulatory regions. Bioinformatics. 39(8):btad457. https://doi.org/10.1093/bioinformatics/btad457
Petoukhov SV (2017) Genetic coding and united-hypercomplex systems in the models of algebraic biology. Biosystems 158:31–46. https://doi.org/10.1016/j.biosystems.2017.05.002
Pirogov A, Pfaffelhuber P, Börsch-Haubold A, Haubold B (2019) High-complexity regions in mammalian genomes are enriched for developmental genes. *Bioinformatics* 35(11): 1813–1819. https://doi.org/10.1093/bioinformatics/bty922
Provata A, Nicolis C, Nicolis G (2014) Complexity measures for the evolutionary categorization of organisms. Computational Biology and Chemistry 53(Part A):5–14. https://doi.org/10.1016/j.compbiolchem.2014.08.004
Román-Roldán R, Bernaola-Galván P, Oliver J (1998) Sequence compositional complexity of DNA through an entropic segmentation method. Phys Rev Lett 80(6):1344–1347. https://doi.org/10.1103/PhysRevLett.80.1344
Rudenko V, Korotkov E (2023) Detection of tandem repeats in the Capsicum annuum genome. DNA Res 30(3):dsad007. https://doi.org/10.1093/dnares/dsad007
Sadovsky MG, Putintseva JA, Shchepanovsky AS (2008) Genes, information and sense: complexity and knowledge retrieval. Theory Biosci 127(2):69–78. https://doi.org/10.1007/s12064-008-0032-1
Safronova NS, Babenko VN, Orlov YL (2015) 117 analysis of SNP containing sites in human genome using text complexity estimates. J Biomol Struct Dyn 33(sup 1):73–74. https://doi.org/10.1080/07391102.2015.1032750
Safronova NS, Ponomarenko MP, Abnizova II, Orlova GV, Chadaeva IV, Orlov YL (2016) Flanking monomer repeats determine decreased context complexity of single nucleotide polymorphism sites in the human genome. Rus J Genet: Appl Res 6:809–815. https://doi.org/10.1134/S2079059716070121
Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res 32(Database issue):D91-4. https://doi.org/10.1093/nar/gkh012
Sergeeva EM, Afonnikov DA, Koltunova MK, Gusev VD, Miroshnichenko LA, Vrána J et al (2014) Common wheat chromosome 5B composition analysis using low-coverage 454 sequencing. Plant Genome 7:plantgenome2013.10.0031. https://doi.org/10.3835/plantgenome2013.10.0031
Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27, pt I, 379–423; pt II, 623–656
Silva M, Pratas D, Pinho AJ (2021) AC2: an efficient protein sequence compression tool using artificial neural networks and Cache-Hash Models. Entropy (basel) 23(5):530. https://doi.org/10.3390/e23050530

Simões RP, Wolf IR, Correa BA, Valente GT (2021) Uncovering patterns of the evolution of genomic sequence entropy and complexity. Mol Genet Genomics 296(2):289–298. https://doi.org/10.1007/s00438-020-01729-y

Subkhankulova T, Naumenko F, Tolmachov OE, Orlov YL (2021) Novel ChIP-seq simulating program with superior versatility: isChIP. Brief Bioinform 22(4):bbaa352. https://doi.org/10.1093/bib/bbaa352

Suvorova Y, Korotkov E (2015) Study of triplet periodicity differences inside and between genomes. Stat Appl Genet Mol Biol 14(2):113–123. https://doi.org/10.1515/sagmb-2013-0063

Suvorova YM, Korotkova MA, Korotkov EV (2014) Comparative analysis of periodicity search methods in DNA sequences. Comput Biol Chem 53(PA):43–48. https://doi.org/10.1016/j.compbiolchem.2014.08.008

Suvorova YM, Kamionskaya AM, Korotkov EV (2021) Search for SINE repeats in the rice genome using correlation-based position weight matrices. BMC Bioinformatics 22:42. https://doi.org/10.1186/s12859-021-03977-0

Tarailo-Graovac M, Chen N (2009) Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinform Chapter 4:4.10.1-4.10.14. https://doi.org/10.1002/0471250953.bi0410s25

te Boekhorst R, Naumenko FM, Orlova NG, Galieva ER, Spitsina AM, Chadaeva IV, Orlov YL, Abnizova II (2016) Computational problems of analysis of short next generation sequencing reads. Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed 20(6):746–755. https://doi.org/10.18699/VJ16.191

Tognon M, Giugno R, Pinello L (2023) A survey on algorithms to characterize transcription factor binding sites. Brief Bioinform. 24(3):bbad156. https://doi.org/10.1093/bib/bbad156

Trifonov EN (1989) The multiple codes of nucleotide sequences. Bull Math Biol. 51(4):417-32. https://doi.org/10.1007/BF02460081

Trifonov EN (1990) Making sense of the human genome. In Sarma RH, Sarma MH (Eds), *Structure & Methods* Adenine Press, Albany. Vol. 1: 69–77

Troyanskaya OG, Arbell O, Koren Y, Landau GM, Bolshoy A (2002) Sequence complexity profiles of prokaryotic genomic sequences: a fast algorithm for calculating linguistic complexity. Bioinformatics 18:679–688

Turing A (1936) On computable numbers, with an application to the Entscheidungsproblem. Proc Lond Math Soc 42(2):230–265

Urgese G, Parisi E, Scicolone O, Di Cataldo S, Ficarra E (2020) Bioseqzip: a collapser of ngs redundant reads for the optimization of sequence analysis. Bioinformatics 36(9):2705–2711. https://doi.org/10.1093/bioinformatics/btaa051

Uversky VN (2016) Paradoxes and wonders of intrinsic disorder: complexity of simplicity. Intrinsically Disord Proteins 4(1):e1135015. https://doi.org/10.1080/21690707.2015.1135015

Vadnala RN, Hannenhalli S, Narlikar L, Siddharthan R (2023) Transcription factors organize into functional groups on the linear genome and in 3D chromatin. Heliyon 9(8):e18211. https://doi.org/10.1016/j.heliyon.2023.e18211

Veluchamy A, Mehta P, Srividhya KV et al (2021) Information theoretic perspective on genome clustering. Saudi J Biol Sci 28(3):1867–1889. https://doi.org/10.1016/j.sjbs.2020.12.039

Vinga S (2014) Information theory applications for biological sequence analysis. Brief Bioinform 15(3):376–389. https://doi.org/10.1093/bib/bbt068

Vityaev EE, Orlov YL, Vishnevsky OV, Belenok AS, Kolchanov NA (2001) Computer system "Gene Discovery" to search for patterns in eukaryotic regulatory nucleotide sequences. Mol Biol (mosk) 35:810–817. https://doi.org/10.1023/A:1013273932056(inRussian)

Vityaev EE, Orlov YL, Vishnevsky OV, Pozdnyakov MA, Kolchanov NA (2002) Computer system "Gene Discovery" for promoter structure analysis. In Silico Biol 2:257–262

Voropaeva EN, Pospelova TI, Voevoda MI, Maksimov VN, Orlov YL, Seregina OB (2019) Clinical aspects of TP53 gene inactivation in diffuse large B-cell lymphoma. BMC Med Genomics 12(Suppl 2):35. https://doi.org/10.1186/s12920-019-0484-9

Wan H, Li L, Federhen S, Wootton JC (2003) Discovering simple regions in biological sequences associated with scoring schemes. J Comput Biol 10:171–185. https://doi.org/10.1089/106652703321825955

Wang Z, Wang Y, Fuhrman JA, Sun F, Zhu S (2020) Assessment of metagenomic assemblers based on hybrid reads of real and simulated metagenomic sequences. Brief Bioinform 21(3):777–790. https://doi.org/10.1093/bib/bbz025

Wise MJ (2001) 0j.py: a software tool for low complexity proteins and protein domains. Bioinformatics 17:S288–S295. https://doi.org/10.1093/bioinformatics/17.suppl_1.S288

Wootton JC, Federhen S (1993) Statistics of local complexity in amino acid sequences and sequence databases. Comput Chem 17(2):149–163. https://doi.org/10.1016/0097-8485(93)85006-X

Wootton JC, Federhen S (1996) Analysis of compositionally biased regions in sequence databases. Methods Enzymol 266:554–571. https://doi.org/10.1016/S0076-6879(96)66035-2

Wu C, Chen J, Liu Y, Hu X (2019) Improved prediction of regulatory element using hybrid Abelian complexity features with DNA sequences. Int J Mol Sci 20(7):1704. https://doi.org/10.3390/ijms20071704333

Xie G, Mo Z (2011) Three 3D graphical representations of DNA primary sequences based on the classifications of DNA bases and their applications. J Theor Biol 269(1):123–130. https://doi.org/10.1016/j.jtbi.2010.10.018

Zeng C, Takeda A, Sekine K, Osato N, Fukunaga T, Hamada M (2022) Bioinformatics approaches for determining the functional impact of repetitive elements on non-coding RNAs. In: Parrish NF, Iwasaki YW (eds) piRNA Methods in Molecular Biology, vol 2509. Humana, New York, NY

Zhang Y, Wei S, Liu H, Zhao L, Liu C (2016) A novel encoding Lempel-Ziv complexity algorithm for quantifying the irregularity of physiological time series. Comput Methods Programs Biomed 133:7–15. https://doi.org/10.1016/j.cmpb.2016.05.010

Zhou Y, Yu Z, Chebotarov D, Chougule K, Lu Z, Rivera LF, Kathiresan N, Al-Bader N, Mohammed N, Alsantely A, Mussurova S, Santos J, Thimma M, Troukhan M, Fornasiero A, Green CD, Copetti D, Kudrna D, Llaca V, Lorieux M, Zuccolo A, Ware D, McNally K, Zhang J, Wing RA (2023) Pan-genome inversion index reveals evolutionary insights into the subpopulation structure of Asian rice. Nat Commun 14(1):1567. https://doi.org/10.1038/s41467-023-37004-y

Zielezinski A, Vinga S, Almeida J, Karlowski WM (2017) Alignment-free sequence comparison: benefits, applications, and tools. Genome Biol 18(1):186. https://doi.org/10.1186/s13059-017-1319-7

Zimnyakov D, Alonova M, Skripal A, Dobdin S, Feodorova V (2023) Quantification of the diversity in gene structures using the principles of polarization mapping. Curr Issues Mol Biol 45(2):1720–1740. https://doi.org/10.3390/cimb45020111