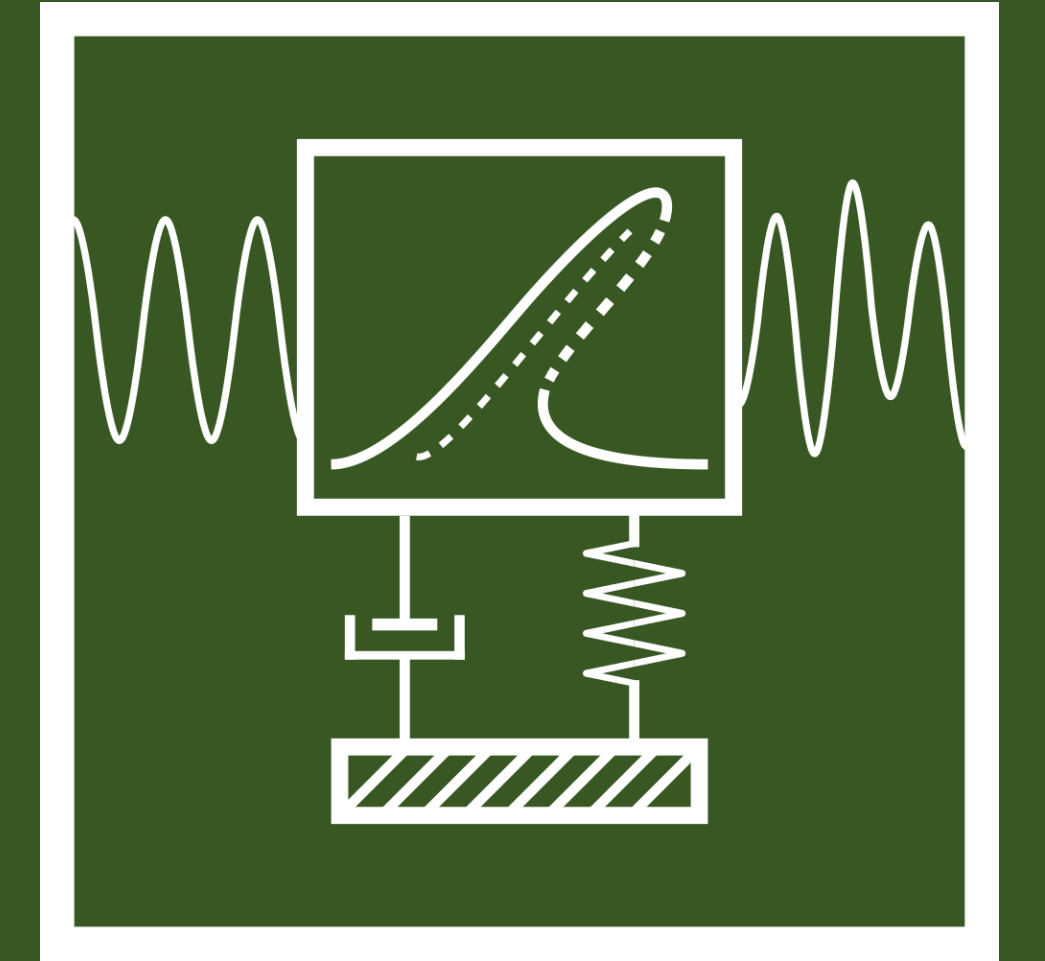


Algebraic Modelling of Bounded Integer Sequences

Kamilija Jablonskaitė, Mantas Landauskas

Department of Mathematical modeling, Kaunas University of Technology, Kaunas, Lithuania

Corresponding author e-mail: mantas.landauskas@ktu.lt



A

Abstract

Bounded integer sequences are often encountered as real-world data in engineering, biology, etc. Modelling of these sequences may be targeted to feature extraction, prediction or generation of high-quality synthetic data. This presentation aims at generating random DNA sequences that would have algebraic properties similar to those in a real sequence (here, EGFR gene [4]). The algebraic approach, which encodes a series of nucleotides in numbers [1,3], is a rather novel way to examine the properties of the genetic code. By constructing Hankel matrices of an encoded genetic sequence with a moving window, a new numerical sequence is formed, one comprising the ranks computed for these Hankel matrices.

01

Methods

Let $H_{n \times n}$ be a Hankel matrix and d_n be its determinant. A number sequence $S = \{x_i\}_{i=1}^{\infty}$ is said to have Hankel pseudo rank (H-pseudo rank) $r(S) \in \mathbb{R}$ if $d_r \geq \varepsilon$ but $d_n < \varepsilon$ for all $n > r$. If there is no such number, $r(S) := 1$. If $S = \{0, 0, \dots, 0\}$, $r(S) = 0$. $r(S)$ is used to construct the characteristic equation:

$$\begin{vmatrix} x_1 & x_2 & \dots & x_{n+1} \\ x_2 & x_3 & \dots & x_{n+2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \rho & \dots & \rho^n \end{vmatrix} = 0.$$

The roots of the characteristic equation allows to express each element of S as a weighted sum. In the case of all different roots ρ_k , $k = 1, \dots, r(S)$, any element of S can be expressed (in other words, S can be reconstructed) as follows:

$$x_i = \sum_{k=1}^{r(S)} \mu_k \rho_k^i.$$

Estimating the proposed model

1. A real genetic sequence S is encoded into the sequence of integers comprising of elements from $\mathbb{Z}_4$.
2. Estimate rank $r_1 \in [1; r_{max}]$ for the segment $S_1 = \{x_1, x_2, \dots, x_{2r_{max}+1}\}$. ε is arbitrary set to 2^{-52} .
3. Estimate rank r_2 for the segment $S_{2r_1+2} = \{x_{2r_1+2}, x_{2r_1+3}, \dots, x_{2r_1+2r_{max}}\}$ (non-overlapping approach). Equivalently, fully overlapping approach would check the segment S_2 .
4. The process stops when the sequence S is exhausted and results in the list of subsequence ranks:

$$r_1, r_2, \dots, r_{n_r}.$$

Generating Random Genetic Sequences

A random fragment $\tilde{S}_1 = \{\tilde{x}_k\}_{k=1}^{2r_1+1}$ is obtained:

- $\rho_j = a_j e^{i\phi_j}$, $j = 1, 2, \dots, r_1$, are generated.
 - Here $a_j = 1$ and $\phi_{j-1} = \frac{2\pi(j-1)}{r_1}$ uniformly cover the full angle.
 - Sort into conjugate pairs: $\rho_{1,2} = \pm 1$, $\rho_{3,4} = e^{\pm i\phi_1}$, ..., $\rho_{r_1-1, r_1} = e^{\pm i\phi_{\lfloor \frac{r_1-1}{2} \rfloor}}$.
 - If r_1 is odd, then there are no $\rho_2 = -1$.
- μ_j are generated.
 - If $r_1 = 1$, then $\mu_1 \sim U(0; 4)$, otherwise $\mu_1 \sim U(2; 3)$.
 - Half of the remaining values are taken as the randomly rotated roots of unity. Then, conjugate coefficients are interleaved into them. In the case of even r_1 , 1 or -1 is randomly prepended.
 - The absolute values for μ_j are drawn from $U(0; 1)$ and transformed to ensure that the reconstructed sequence mostly covers the interval $[0.5; 4.5]$.
- $\tilde{x}_k = \sum_{j=1}^{r_1} \mu_j \rho_j^k$. In general, $\tilde{x}_k$ is real, so it is rounded to the closest integer on the interval $[1; 4]$.

The rank r_2 yields the fragment $\tilde{S}_2$ and the process is iterated.

The modelled sequence: $\tilde{S} = \bigcup_{i=1}^{n_r} \tilde{S}_i$.

02

Results

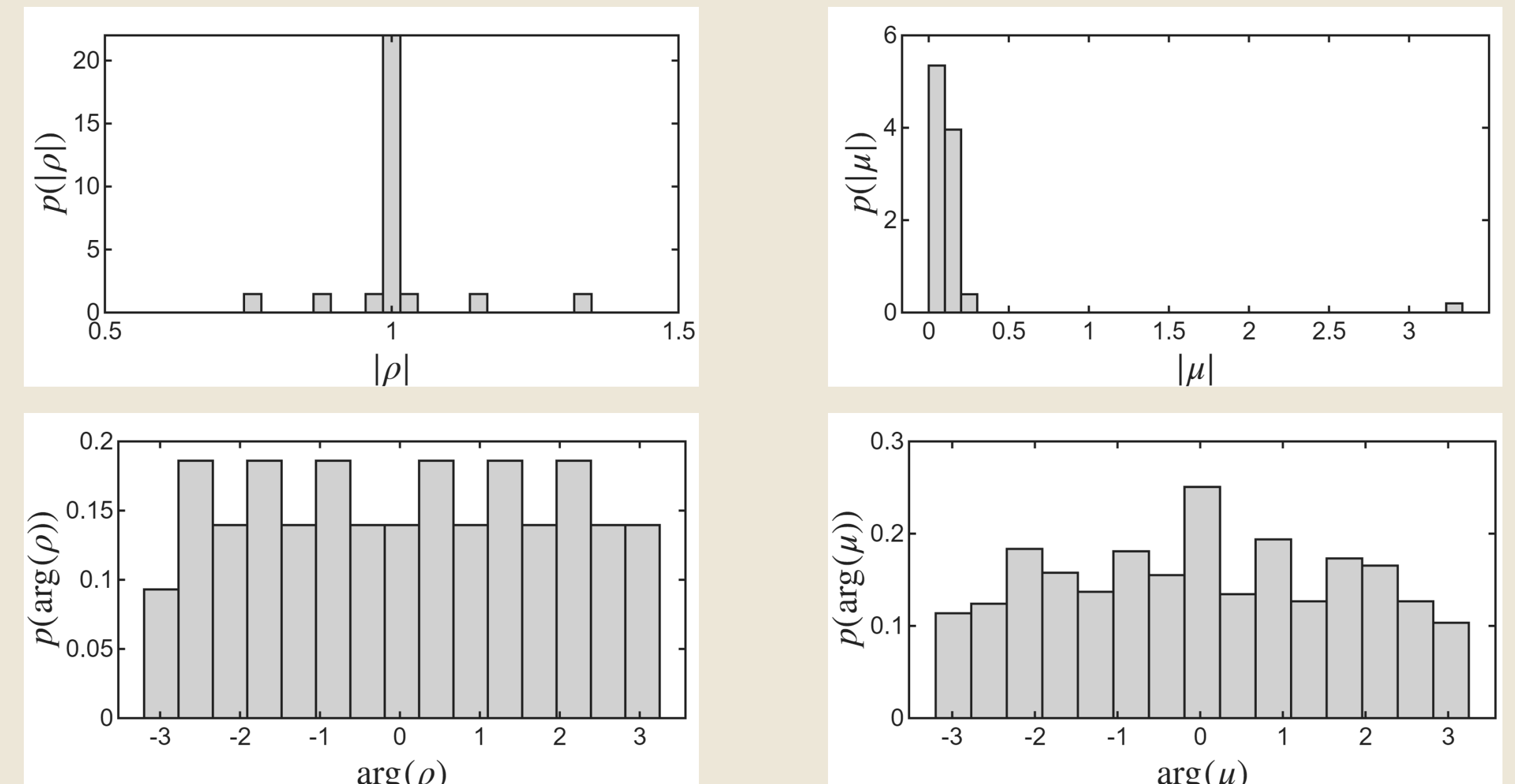


Fig 1. Distribution of modulus and argument for characteristic roots ρ_j (left) as well as for coefficients μ_j (right), $j = 1, 2, \dots, r(S)$, of numerically encoded EGFR gene when $r(S) = 50$.

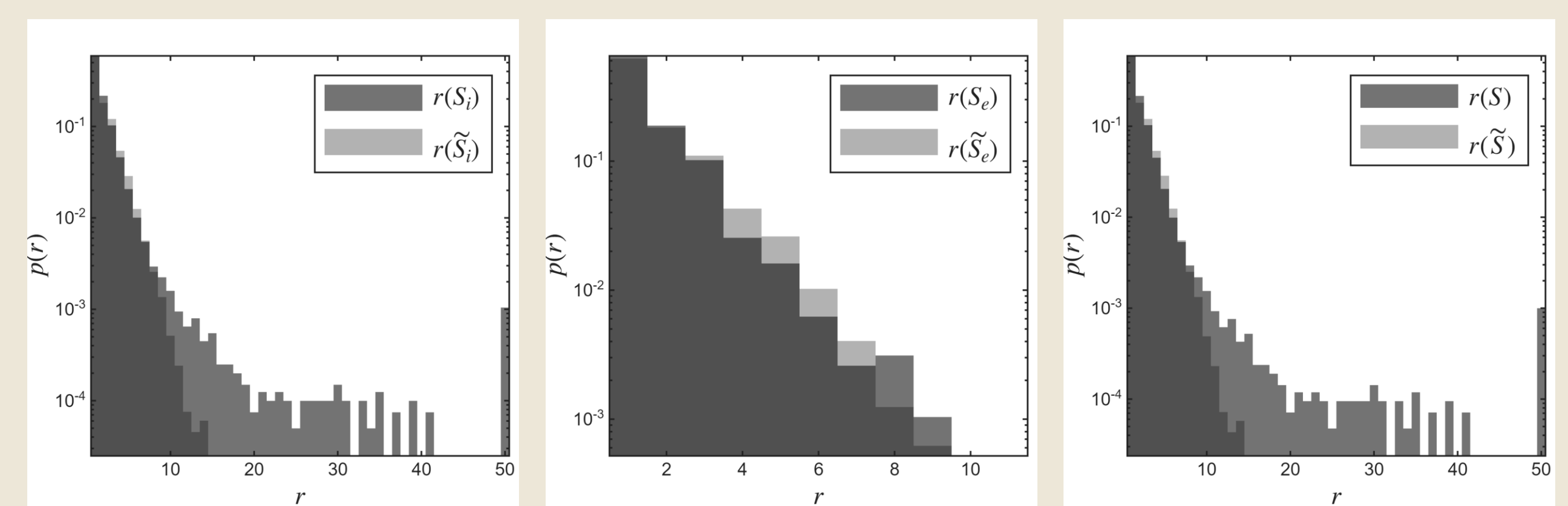


Fig 2. Comparison between empirical distributions of H-pseudo ranks for EGFR gene (segments do not include lower rank subsequences). Overall distribution (left), exon (middle) and intron (right) regions. The darkest gray regions represent overlaps of the original distribution $p(S)$ and the one of the generated sequence: $p(\tilde{S})$.

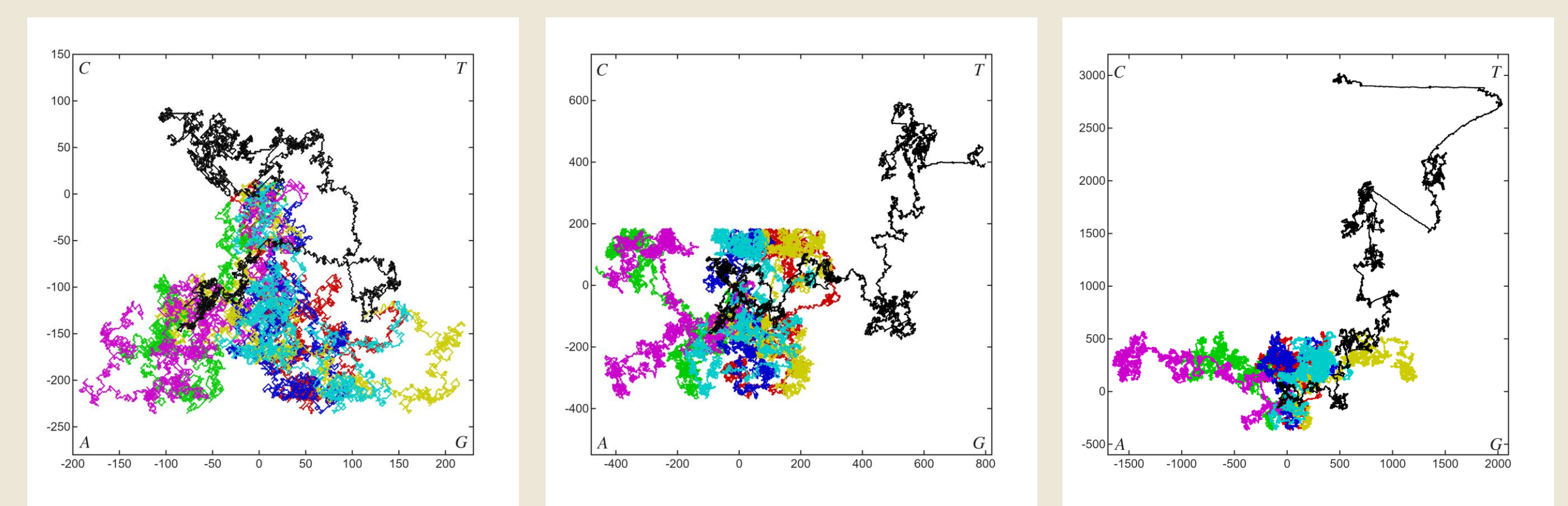


Fig 3. Track visualization of EGFR gene (in black) and randomly generated nucleotide sequences (in colors). The first 10 000 nucleotides (left), the first 50 000 (middle) and the full gene (right).

03

Conclusions

- Random nucleotide sequence generation is convenient for constructing synthetic genetic code which differs from the original sequence in actual nucleotides yet has similar algebraic complexity. The model was able to preserve an alternating pattern between a random walk-like track and local scattering of the sequence.
- The highest ranks are found only in intron regions. Maximal set ranks are obtained at certain genomic positions and also indicate that a number of segments might not have any algebraic relations.

References

- [1] Nair, A.S., Sreenadhan, S.P.: A coding measure scheme employing electron-ion interaction pseudopotential (eiip). *Bioinformatics* 1(6), 197 (2006)
- [2] Dolgin, E.: The most popular genes in the human genome. *Nature* 551(7681), 427–432 (2017)
- [3] Hebert, P.D., Cywinska, A., Ball, S.L., DeWaard, J.R.: Biological identifications through dna barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 270(1512), 313–321 (2003)
- [4] National Center for Biotechnology Information: EGFR epidermal growth factor receptor [Homo sapiens (human)]. Gene ID: 1956 (2025), www.ncbi.nlm.nih.gov, updated November 25, 2025. Accessed December 16, 2025