Identification of Target Genetic Variations via the Mathematical Properties of Genetic code

K. Jablonskaitė[1], A. Aldujeli[2], I. Čiapienė[3], U. Meškauskaitė[4], I. Matulevičiūtė[5], V. Tatarūnas[6], M. Landauskas[7]

[1,7]Kaunas University of Technology; [2,3,4,6]Lithuanian University of Health Sciences;
[5]The Hospital of Lithuanian University of Health Sciences
kamilija.jablonskaite@ktu.edu

Motivation

One of the most common types of genetic variation is the single nucleotide polymorphism (SNP). It refers to the presence of a particular nucleotide at a specific position in the genetic code. Such genetic variations are often associated with various phenotypes and are commonly investigated in genome-wide association studies (GWAS) to uncover genotype–phenotype relationships [1]. Genetic variations might span regions of genetic code. This study aims at identifying target genetic variations with respect to mathematical properties of genetic code.

Introduction

Genetic sequence of interest is encoded as integer number sequence comprising of the first 4 natural numbers, each assigned to a particular nucleotide. In this way a gene, chromosome or any other genetic sequence could be further analyzed mathematically. Patterns, variability, statistical distribution, etc. could be assessed using a number of methods. Here Shannon entropy and generalized permutation entropy are computed for moving windows of the encoded sequence. Entropy based features are a proven class of methods for genetic data analysis [2]. For example, differences in entropy were used to detect mutations in the virus DNA [3].

The aforementioned numerical features here are analyzed as a new sequence. Local extreme values (peaks) are identified and corresponding SNPs are selected. These are the target genetic positions to be further analyzed by experts in order to determine possible common aspects.

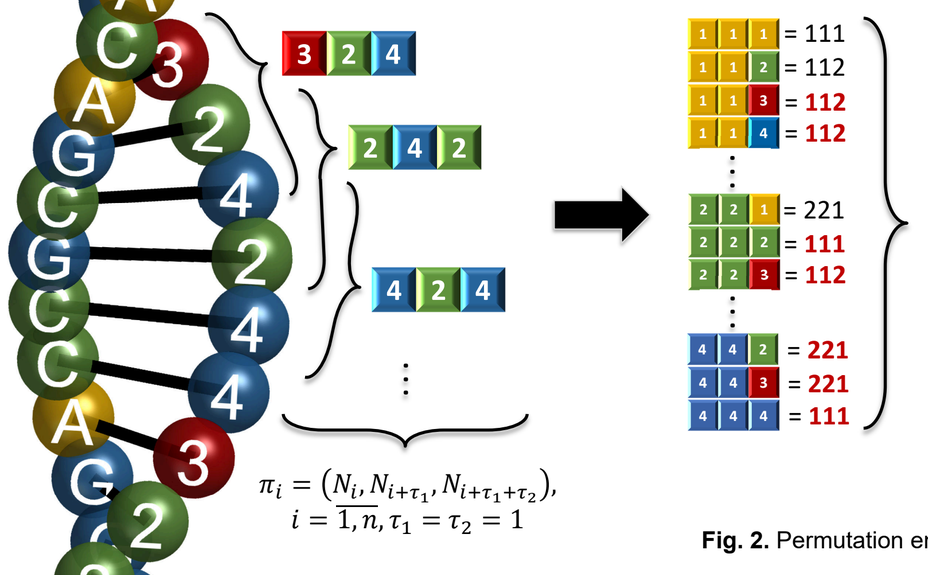
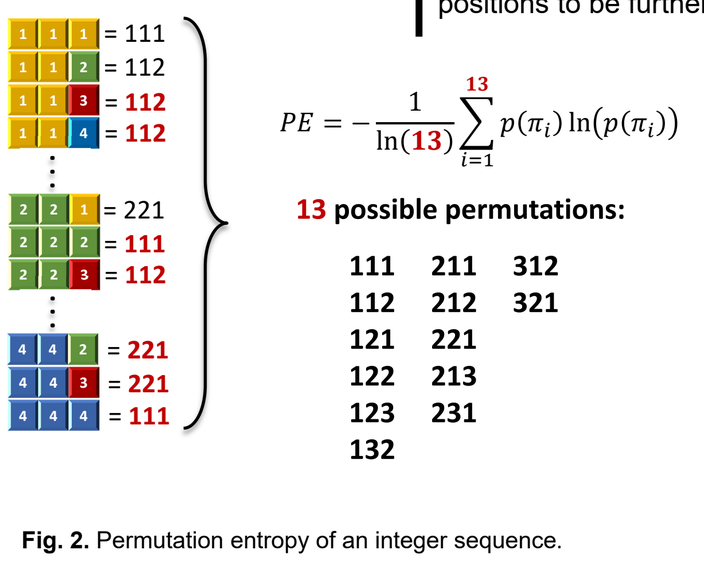


Fig. 1. Encoding of a genetic data into an integer sequence.



$$PE = -\frac{1}{\ln(13)}\sum_{i=1}^{13} p(\pi_i)\ln\left(p(\pi_i)\right)$$

13 possible permutations:

111	211	312
112	212	321
121	221	
122	213	
123	231	
132		

Fig. 2. Permutation entropy of an integer sequence.

$$E = -\frac{1}{\ln(4)}\sum_{i=1}^{4} p_i \ln(p_i)$$

$p_i = P(N = N_i)$, $N_i \epsilon \{1,2,3,4\}$

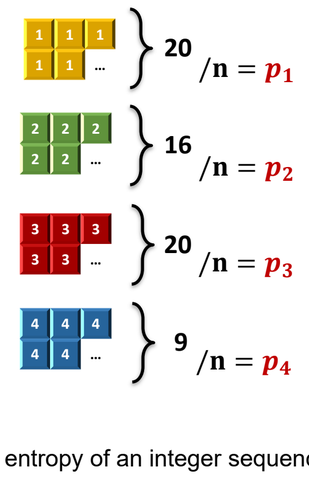


Fig. 3. Shannon entropy of an integer sequence.

Numerical analysis

Genetic data (complete human genome) is encoded into an integer sequence $N_1, N_2, \ldots$. Extreme positions (peaks) of the numerical features of the sequence are merged with SNP positions. Each peak P_k in the feature sequence is assigned a prominence p_k and width w_k. Prominence is the minimum vertical distance a feature must descend (ascend) on either side of the peak before raising (falling) back to a level higher (lower) than the peak (or reaching an endpoint). Lastly, sorting by $\sigma(\beta_k)$ is performed and top prominent and variable positions are picked as targets. $\sigma(\beta_k)$ is the standard deviation of methylation values at genomic position k in the Age-Methylation dataset. The most prominent positions are showed in Table 1.

Shannon entropy (E) describes the amount of disorder in the nucleotide sequence. Lower values of E suggest one or few nucleotides are more common in the sequence than others. Here, E_i at the genomic position i is computed for nucleotide sequence $\{N_j\}_{j=i-32}^{i+32}$.

Since 3 nucleotides code an amino acid, a list of 3D vectors (triplets) are considered: $\{(N_{j-1}, N_j, N_{j+1})\}_{j=i-33}^{i+33}$. PE_i is computed out of 65 triplets, 5 triplets on average for each possible permutation (13 in total, including ties). Permutation entropy (PE) describes how uniform is the distribution of nucleotide triplets. The higher the PE, the more uniform is the distribution of triplets. Conversely, low PE suggests one or a few triplets are more common than others.

Tab. 1. Top prominent (min and max) target positions k with respect to $\boldsymbol{\sigma(\beta_k)}$.

k	rsID	Feature value	w_k	p_k	min (-1) max (1)	$\sigma(\beta_k)$	CHR	Feature
45791224	1006754171	0,8394	94,79	0,16	-1	0,4191	19	E
102439421	996308310	0,9909	211,25	0,10	1	0,3969	4	E
134268627	938787444	0,8701	27,78	0,12	-1	0,3599	9	PE
2980648	186737398	0.9777	84.41	0,10	1	0,3309	7	PE

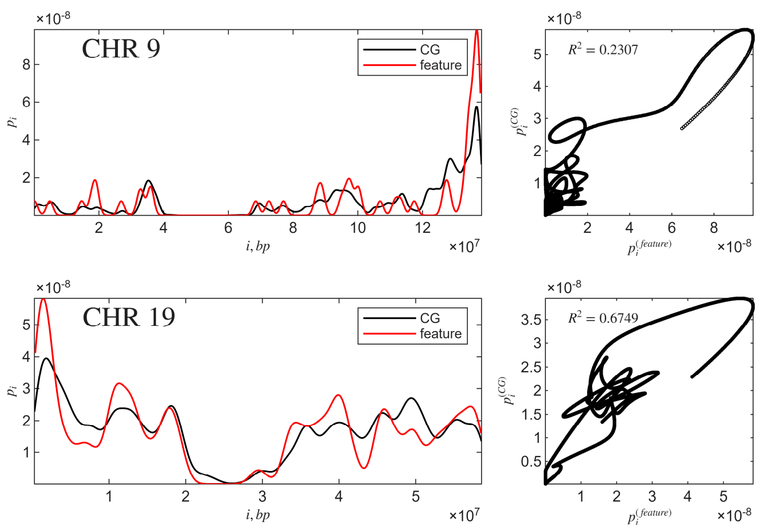


Fig. 7. Kernel density estimates of genomic positions i being a CG island location (in black) or having a peak in a numerical feature (in red).

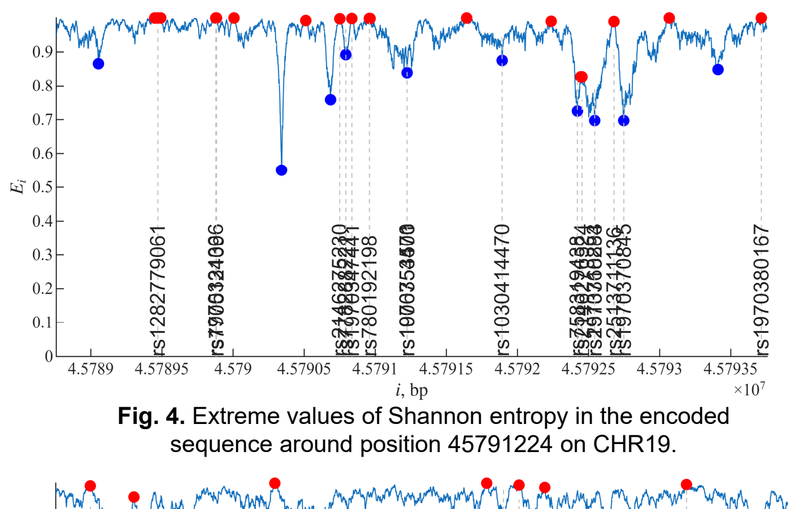


Fig. 4. Extreme values of Shannon entropy in the encoded sequence around position 45791224 on CHR19.

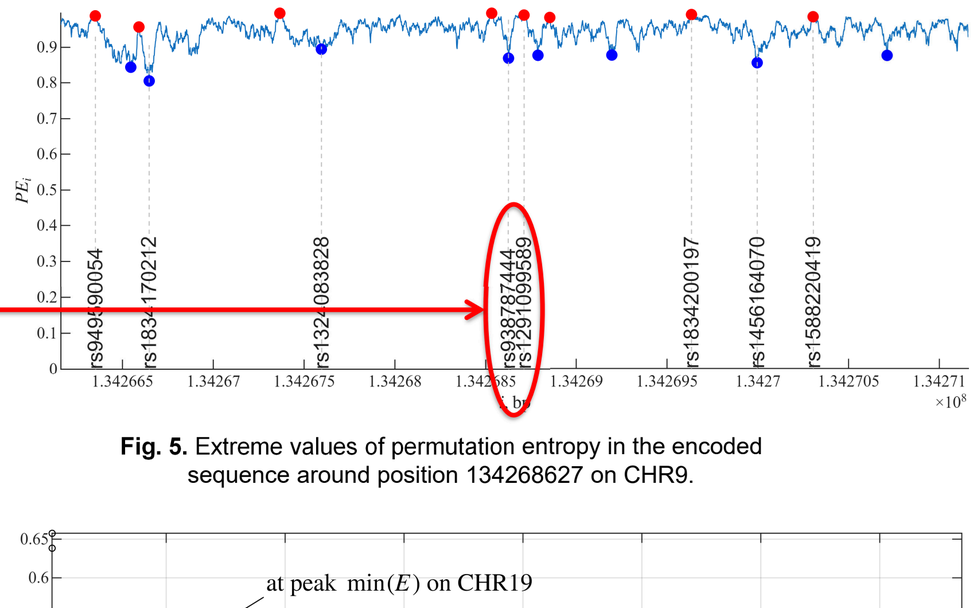


Fig. 5. Extreme values of permutation entropy in the encoded sequence around position 134268627 on CHR9.

Fig. 6. A subset of the most prominent target positions in the context of top 1% $\sigma(\beta)$ in the methylation dataset.

Discussion

Experiments with live tissues and further genomic analysis are costly to perform. Thus, it is important to select target genomic positions carefully. Our approach is to pick positions with extreme values of E and PE. If 10 targets are needed, 5-7 could be picked with respect to highest variation in methylation (for epigenetics studies), the remaining 3-5 could be corresponding to extreme features.

Minimum peak prominence was set to 0.1 (10% of a feature range). This resulted in 2956 extreme feature values across the human genome, i.e. only around 0.000095% of all positions. There are peaks with prominences exceeding 0.9 but the most interesting fact is that such a handful set of identified positions include 20 which are among the top 1% most variable methylations in the methylation dataset (Fig. 6).

Variability in methylation (or other biological property in general) is an additional criterion for identification of target positions since it creates an opportunity to distinguish between cell types or their pathologies. Next steps will include a more detailed analysis of the identified target positions and their use as features of a genetic code.

Acknowledgements

This research was supported by the Research and Innovation Funds of Kaunas University of Technology and Lithuanian University of Health Sciences (MEGRAC, Grant No. INP2025/7).

References

[1] **Ding, X., & Guo, X**. (2018). A survey of SNP data analysis. Big Data Mining and Analytics, 1(3), 173-190.

[2] **Orlov, Y. L., & Orlova, N. G**. (2023). Bioinformatics tools for the sequence complexity estimates. Biophysical reviews, 15(5), 1367-1378.

[3] **Vopson, M. M., & Robson, S. C**. (2021). A new method to study genome mutations using the information entropy. Physica A: Statistical Mechanics and its Applications, 584, 126383.