Simulation Modelling for Machine Learning Identification of Single Nucleotide Polymorphisms in Human Genomes

Mikalai M. Yatskou
Deprt. of Systems Analysis and Computer Modelling
Belarusian State University
Minsk, 220030, Belarus
yatskou@bsu.by

Elizabeth V. Smolyakova
Deprt. of Systems Analysis and Computer Modelling
Belarusian State University
Minsk, 220030, Belarus
smolyakova580@gmail.com

Victor V. Skakun
Deprt. of Systems Analysis and Computer Modelling
Belarusian State University
Minsk, 220030, Belarus
skakun.victor@gmail.com

Vasily V. Grinev
Deprt. of Genetics
Belarusian State University
Minsk, 220030, Belarus
grinev_vv@bsu.by

***Abstract* — An approach for simulation modelling of Single Nucleotide Polymorphisms (SNPs) in DNA sequences is proposed, which implements the generation of random events according to the beta or normal distributions, the parameters of which are estimated from the available experimental data. This approach improves the accuracy of determining SNPs in DNA molecules. The verification of the developed model and analysis methods was carried out on a set of reference data provided by the GIAB consortium. The best results were obtained for the machine learning model of Conditional Inference Trees – the accuracy of the SNP identification by the score F_1 is 82,8 %, which is higher than those obtained by traditional SNP identification methods, such as binomial distribution, entropy-based and Fisher's exact tests.**

Keywords — single nucleotide polymorphism, SNP identification, simulation modelling, machine learning

I. INTRODUCTION

Genetic polymorphism affects the human phenotype and other living organisms [1]. Single nucleotide polymorphisms (SNPs) are one of the most common types of genetic variation in the human genome. Knowledge of the genes involved in cancer development, combined with the ability of gene sequencing and bioinformatics analysis, is an important tool for screening patients at risk and assisting in genetic counseling [2].

Statistical methods of binomial distribution, entropy-based, Fisher's exact tests and machine learning are applied for identifying the SNPs in humans and plants [1, 3, 4]. These methods are quite universal and simple for program implementation, however, are computationally expensive and difficult to be used in the analysis of experimental data with a high noise level and various experimental distortions, which are sources of gaps, repetitions, and other anomalous values often observed in genomic sequencing by the PacBio and Oxford Nanopore technologies [5]. In practical experimental studies, simulation modelling is used to select the most optimal SNP identification algorithm, test competing plans/methods of analysis, and evaluate the performance of specific experimental design for studying biophysical systems [6, 7]. Simulations are critical for testing methods and studying the effects of different phenotypic and genetic architectures of biological traits. Modeled genotypes and phenotypes reflect the intended understanding of the true structure of the phenotype, but do not guarantee the biological correctness of real phenotypes [8]. Simulation modelling is also used to generate training data for machine learning methods to directly identify SNP sites of various organisms from a single sequencing experiment [4]. In this case, the formation of simulated training data can have advantages in terms of accuracy and efficiency in the analysis of experimental data both with a low number of coverages and with gaps due to experimental distortions.

Various approaches to mathematical modelling of genetic polymorphisms, based on accounting the parameters of experimental equipment, the use of probabilistic models and statistical approaches, and auxiliary biological information, are published elsewhere [9, 10]. However, due to complexities in the types of genetic data, modelling methods, evolutionary characteristics, data formats, terminology, and assumptions made in existing software applications, choosing a reliable tool for a particular study could be a resource- and time-consuming process [11]. It should be noted that only few modelling methods use experimental results (or measured parameters) and a complex simulation scheme with covariant noise structure. As the complexity of analysis increases, researchers need sophisticated modelling of realistic genotype and phenotype structures from the measured characteristics of specific experiments. Simulated data from a particular experiment provide more accurate training datasets for machine learning algorithms to identify SNP sites.

This article presents an approach for simulating SNP sites in DNA sequences based on the beta and normal distributions, the parameters of which are determined from the available experimental data. It allows to model the features of specific experiments and form learning datasets for training classification models of machine learning algorithms. The performance of the developed computational algorithms was confirmed in the course of a comparative analysis of the most effective existing algorithms for identifying SNP sites on experimental genomic sequencing data.

II. METHODOLOGY

A. Simulation Modelling of SNPs in DNA sequences

The object (nucleotide sites of sequenced DNA molecules) can be investigated using a natural experiment or simulation modelling [12]. The scheme of study of the object according to experimental data is shown in Fig. 1.

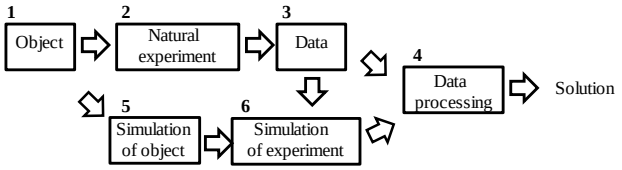

Fig.1. Scheme of the study of sequenced DNA molecules in natural and simulation experiments

In a natural experiment (Fig. 1, block 2), data from the object of study (block 1) are recorded and characterized by the structure of the corresponding genome sequencing experiment. Data processing is carried out in block 6, analyzing the integral characteristics of the data, and in block 4, identifying the SNPs. The choice of data processing methods is determined by the specifics of a certain problem being solved and includes methods and models for finding the required solution. In a simulated or computational experiment (blocks 5 and 6) the same object model is considered as in the real experiment (block 2). The mathematical model of the object under study M can be either parametric (the operator of mathematical transformations F is known up to some parameters A), or non-parametric (a family of operators F is considered, among which the most optimal ones are selected for solving a given problem), and includes a physical model, representing both the object and the experimental sequencing facility (block 2). To describe the behavior of the object in various experiments, it is required to include the output experimental characteristics of the equipment and the recorded data (block 3) in the object of simulation. The concept of an object of simulation includes modelling the behavior of an object under specific experimental conditions (for example, with known distributions and parameters describing the data). Modelling of nucleotide sites based on the estimated characteristics of the experimental data is carried out in block 6. In block 4, data processing is performed, namely, the search for SNP sites using a proper algorithm. The choice of data processing methods is determined by the complexity of real data (a small number of coverages, gaps, duplicates, a high level of experimental noise, etc.). To confirm the validity of simulation models, a comparison of the data characteristics of computational and natural experiments is required. For generative modelling tasks, applied to improve the prediction accuracy of machine learning models, the presence of experimental data might not be necessary.

B. Algorithm for Simulation of SNP Sites

The subsection describes the algorithm for simulating SNP sites, assuming that the main data characteristics, such as the numbers of nucleotide coverages, are of the beta or normal distributions [13], whose parameters are determined from the available experimental data.

Suppose a site j contains the reference nucleotide base r (nucleotides A, C, G, or T); $D = \{b_1, b_2, b_3, b_4\}$ is a set of n reads (coverages) of nucleotide bases A, C, G or T, recorded from sequencing the site j; the numbers of site coverages n, b_1, b_2, b_3, b_4 obey the beta (1) or normal (2) distributions

$$n_b(x,\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)+\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \qquad (1)$$

where β and α (β, α > 0) are some parameters that determine the shape of the distribution curve; Γ is the gamma function;

$$n_g(x,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad (2)$$

where μ and σ are parameters of mathematical mean and standard deviation.

The idea of modelling is to randomly generate N_{SNP} positions of SNP sites in the sequence of the considered molecule S, consisting of N nucleotide sites, for each of which the numbers of coverages n, b_1, b_2, b_3, b_4 are reproduced according to the beta or normal distributions in the defined range $[n_{min}; n_{max}]$. It should be noted that experimental histograms can be considered as distributions (nonparametric method of solution). For a non-reference site j, the total number of coverages n is modeled, then the number of coverages for the reference b_{Ref} and non-reference b_{notRef} nucleotides is generated from the resulting n. Nucleotide coverages for the SNP site are modeled similarly. It is assumed that there are coverages of no more than two different nucleotide bases on the site. The proposed algorithm allows to reproduce datasets as close as possible to experimental conditions, given by the numbers of site coverages and the laws of their distributions, the number of SNP sites. The flow diagram of the algorithm for modeling SNP sites is shown in Fig. 2.

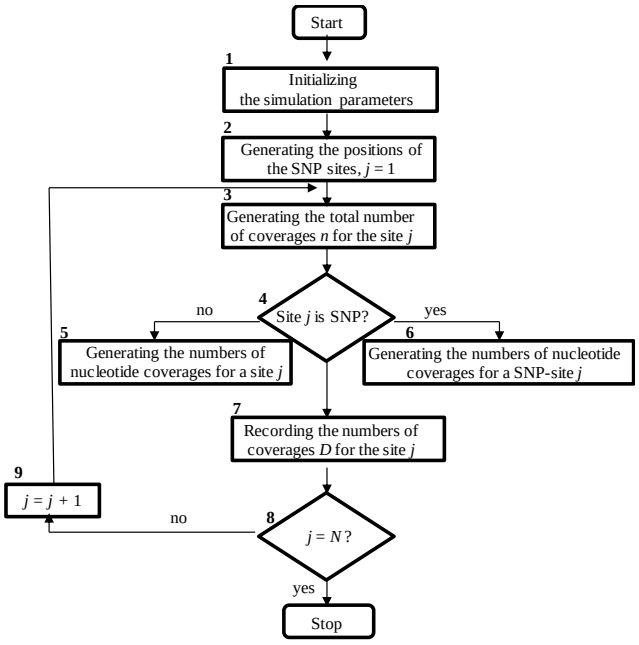

Fig.2. Flow diagram of the algorithm for modelling SNP sites

Algorithm.

Step 1. Initialize the model parameters N, N_{SNP}, n_{min} and n_{max}, α and β (or μ and σ) (Fig. 2, block 1). Parameters α and β (or μ and σ) are given for distributions of the numbers of site nucleotide coverages n, b_1, b_2, b_3, b_4.

Step 2. Generate the SNP site positions $L = \{l_1, l_2, \ldots, l_{NSNP}\}$ in the sequence S according to the uniform discrete distribution in the interval [1; N] (block 2). Set the position index $j = 1$.

Step 3. Gamble the total number of reads n on the current site j as a realization of a random variable of the beta or normal distribution with experimentally extracted parameters (block 3, see subsection IVA).

Step 4. Check if the site j is SNP. Accordingly go to step 5 or 6 (block 4).

Step 5. Generate the numbers of coverages of nucleotide bases b_1, b_2, b_3, b_4 by the beta distribution with experimentally assessed parameters for non-SNP sites (block 5). Go to step 7.

Step 6. Generate the numbers of coverages of nucleotide bases b_1, b_2, b_3, b_4 by the beta distribution with experimentally assessed parameters for SNP sites (block 6).

Step 7. Record the simulated characteristics of the site j to a data file (block 7).

Step 8. Check the termination condition of the simulation algorithm (block 8). If all sites in the sequence are simulated, i.e. $j = N$, then stop the simulation. Otherwise $j = j + 1$ (block 9) and go to step 3.

III. Experimental

Reference data on human chromosome 22, publically available from the GIAB consortium, were taken as an experimental dataset [14]. The choice of GIAB data is due to the fact that today it is the most reliable benchmark data for solving problems related to the study of genomic polymorphism in humans (from the development of new instrumental methods of "wet" biology to the comparison of algorithms for detecting polymorphic sites). The dataset contains characteristics of 29 633 768 nucleotide sites, of which 36 150 are truly identified SNPs. A fragment of the dataset is presented in Table I.

TABLE I. Fragment of the experimental dataset

Chromosome : position	reference	Nucleotide			
		A	C	G	T
chr22:47891620	T	0	0	0	27
chr22:47891621	G	0	0	28	0
chr22:47891622	T	0	0	0	30

IV. Results

We analyze the experimental characteristics of the selected dataset of chromosome 22 in order to determine the distribution law and to estimate its unknown parameters. Then we check the adequacy of the developed mathematical model. Based on the selected sets of experimental data, we conduct a comparative analysis of the most effective SNP identification traditional and machine learning algorithms, trained on simulated data.

A. Analysis of Experimental Characteristics of Genomic Sequencing Datasets

We analyze the histograms of the total number of coverages n, the maximum number of coverages $max\{b_i\}$ and differences between the total and maximum numbers of coverages $m = n - max\{b_i\}$ for non- and SNP sites. Approximation of histograms is performed using the density functions of the beta and normal distributions (the R-functions *dbeta* and *dnorm*). To estimate the parameters of distributions, the optimization method is used (R-function *nls*). The results of histogram approximations are shown in Fig. 3.

The results allow to conclude that the beta distribution is the optimal for the studied integral characteristics of the considered experimental data. The normal distribution is less accurate, but its application might be appropriate to other types of experiments, possibly demonstrating essential normality in data distributions. It should be noted that it is possible to apply in simulation models other types of probability distributions. Promising, in terms of modelling accuracy, but computationally expensive, is the simulation method based on experimental histograms, which can be implemented by modelling a discrete random variable specified by a probability table or by the Neumann method, based on an estimated distribution density function [15].

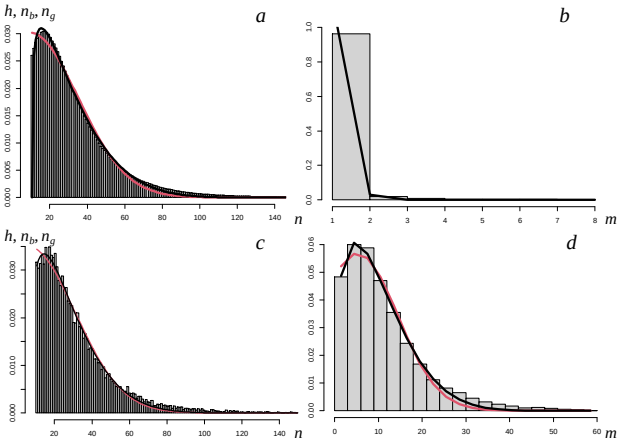

Fig.3. Normalised histograms h of the total number of coverages n (*a*, *c*) and the differences between the total and maximum numbers of coverages m (*b*, *d*) for non-SNP (*a*, *b*) and SNP (*c*, *d*) sites. Approximations are made by the density functions of the beta n_b (black) and normal n_g (red) distributions; parameter estimates are *a*: α = 1,57 (std. error = 0,02), β = 7,9 (0,2), and μ = 9,2 (1,1), σ = 25,9 (0,7); *b*: α = 0,5 (0,05), β = 20 (2), *c*: α = 1,45 (0,02), β = 8,4 (0,2), and μ = 5,8 (1,6), σ = 25,2 (0,8); *d*: α = 1,71 (0,05), β = 7,7 (0,3), and μ = 5,3 (0,6), σ = 9,2 (0,6)

The experimental estimates of the distribution parameters are used in the simulation model. A fragment of the simulated dataset is presented in Table II.

TABLE II. Fragment of the simulated dataset

Chromosome : position	reference	Nucleotide			
		A	C	G	T
chrX:1	G	0	0	33	0
chrX:2	C	0	14	0	0
chrX:3	T	0	0	0	20

B. Program Development of Algorithms

In the course of the work, R-functions were developed that implement various stages of simulation modelling and data analysis algorithms. It is proposed to integrate the developed functions into a dedicated R-package that can be used to model synthetic datasets, according to a concrete experiment, in order to comprehensively test and select the best algorithms for identifying SNP sites, as well as for generative data modelling to train identification algorithms based on machine learning methods.

As a test of the validity of the developed model, we use visual inspection of the plots of simulated and experimentally verified histograms for the number of site coverages n and the accuracy of restoring the modeled parameters when estimating the distribution parameters. We simulated a sequence of 10 000 sites with the parameters of the beta and normal distributions, reconstructed from experimental data,

and approximated the histograms using the beta and normal distributions. Model parameters were estimated using R-functions *dbeta* and *dnorm*. The histograms were successfully approximated by given density functions (Fig. 4). The parameters of the simulation models fall within 95 % confidence intervals of parameter estimates, which support correctness of the developed simulation model, namely, that the procedures for modelling the numbers of site coverages according to the beta and normal distributions are correct.

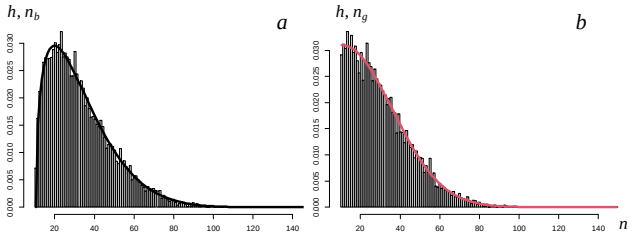

Fig.4. Normalised histograms h of the total number of coverages n in datasets modeled with experimentally estimated parameters of the beta (*a*) and normal (*b*) distributions. Approximations are made by the density functions of the beta n_b (black) and normal n_g (red) distributions; parameter estimates are *a*: α = 1,50 (std. error = = 0,02), β = 7,6 (0,2); *b*: μ = 10,4 (0,9), σ = 25,2 (0,6)

C. Comparative Analysis of SNP Identification Algorithms

We performed the comparative analysis of the most effective existing SNP identification algorithms, such as binomial distribution, entropy-based and Fisher's exact tests, with some fundamental machine learning techniques trained on simulated datasets. We have developed an efficient software implementation of the binomial distribution test, the feature of which is the automation of the selection of a threshold value when identifying SNP sites. It is proposed to use the value 10^{-k} as a threshold value of probabilities, where k is the average number of site coverages estimated from the experimental dataset. As Fisher's exact test, a modification of the algorithm from the R-package *Rsubread* is considered [16]. Our program implementation is used as an entropy-based test [17]. Thresholds in identifying SNP sites are: the entropy $E > 0{,}21$ and the p-value $< 0{,}5$.

To apply machine learning algorithms, it is necessary to form a set of features charactering a nucleotide site. It was decided to use 4 features: X_1 – the number of coverages of the reference nucleotide, X_2 - X_4 – the numbers of coverages for non-reference nucleotides sorted in descending order. The data are normalized to the total number of site coverages n.

Taking into account the limited number of 4 features, and the binary classification problem (SNP and non-SNP site classes) to be solved, it is optimal and effective to test decision trees as machine learning methods [18]. For example – the algorithms of Conditional Inference Trees [19] and CART [20]. Conditional Inference Trees (the function *ctree* of the package *party*) and CART (the function *rpart* of the package *rpart*) were trained on simulated data, generated with the beta distribution. Dataset consisted of 40 000 sites, of which 20 000 were SNPs.

The performance of the algorithms is evaluated using the standard classification measures for unbalanced classes, such as *Precision*, *Recall* and *score* F_1, characterizing the properties of the algorithms accept false positive (non-SNPs as SNPs, *Precision*) and false negative (SNPs as non-SNPs, *Recall*) events and their combined contribution F_1 [21]. The results of SNP identification for 5 datasets of 20 000 sites, starting from positions 3, 9, 15, 21, and 27×10^6 on chromosome 22, are collected in Table III.

TABLE III. SNP IDENTIFICATION ALGORITHMS EFFICINCY BY THE SCORE F_1

Start cite, ×10⁶	**F_1,%**				
	Entropy-based test	**Binomial distribution test**	**Fisher' s exact test**	**Conditional Inference Trees**	**CART**
3	17,1	15,4	11,8	22,2	21,1
9	97,3	97,2	94,3	98,6	95,8
15	95,7	86,7	90,6	98,5	90,3
21	82,9	90,3	91,4	97,1	87,5
27	88,9	92,7	97,5	97,6	95,0
Mean	**76,4**	**76,5**	**77,1**	**82,8**	**77,9**

The mean accuracy of SNP identification in terms of the score F_1 is higher for decision tree-based methods than for classical statistical methods. Conditional Inference Trees model shows the highest accuracy – 82,8 %. The binomial distribution, entropy-based, and Fisher's exact tests have similar mean accuracy 76,4 – 77,1 %.

Additionally, we investigated Conditional Inference Trees and CART models trained on experimental datasets, sampled from the chromosome 22 data. A typical dataset of 72 261 sites was considered, of which 36 150 were SNPs and the rest – randomly selected non-SNPs. The classification accuracy F_1 did not exceed 60-70 % on simulated and experimental data. The poor classification may be due to some reasons, for example, a possibly inferior training dataset or, perhaps, the simulation model is indeed better at forming the training datasets by focusing on reproducing the important/primary sources of information in the data and not taking into account the minor/secondary signals present in the real data.

V. CONCLUSIONS

An approach for simulation modelling of SNPs in DNA sequences is developed, which is based on the generation of random events according to the beta or normal distribution, the parameters of which are estimated from experimental data. This approach increases the accuracy of determining SNPs in genomic sequencing data. The verification of the developed model and the analysis algorithms was done on the example of a large experimental dataset. The comparative analysis of efficient existing statistical SNP identification algorithms and two selected machine learning models trained on synthetic data was carried out. The best results were obtained for machine learning models − the accuracy of SNP identification by the score F_1 is higher for the trained on simulated data Conditional Inference Trees and CART than those for the methods of binomial distribution, entropy-based and Fisher's exact tests.

REFERENCES

[1] W.K. Sung, Algorithms for Next Generation Sequencing, 1st ed., Chapman & Hall/CRC, 2017.

[2] M. Kappelmann-Fenzl, Ed., Next Generation Sequencing and Data Analysis, Springer, Cham, 2021.

[3] X.L. Wu, J. Xu, G. Feng, G.R. Wiggans, J.F. Taylor, J. He, C. Qian, J. Qiu, B. Simpson, J. Walker, S. Bauck, “Optimal design of low-density SNP arrays for genomic prediction: algorithm and applications”, PLoS One, vol. 11(9):e0161719, September 2016.

[4] W. Korani, J.P. Clevenger, Y. Chu, P. Ozias-Akins, “Machine learning as an effective method for identifying true single nucleotide polymorphisms in polyploid plants”, Plant Genome, vol. 12(1), March 2019.

[5] A. Masoudi-Nejad, Z. Narimani, N. Hosseinkhan, Next Generation Sequencing and Sequence Assembly. Methodologies and Algorithms, Springer, New York, 2013.

[6] Z. Su, J. Marchini, P. Donnelly, “HAPGEN2: simulations of multiple disease SNPs”, Bioinformatics, vol. 27(16), pp. 2304-2305, August 2011.

[7] J.H. Oh, J.O. Deasy, “SITDEM: a simulation tool for disease/endpoint models of association studies based on single nucleotide polymorphism genotypes”, Comput. Biol. Med., vol. 45, pp. 136-42, February 2014.

[8] H.V. Meyer, E. Birney, “PhenotypeSimulator: A comprehensive framework for simulating multi-trait, multi-locus genotype to phenotype relationships”, Bioinformatics, vol. 34(17), pp. 2951-2956, September 2018.

[9] A.E. Hendricks, J. Dupuis, M. Gupta, M.W. Logue, K.L. Lunetta, “A comparison of gene region simulation methods”, PLoS One, vol. 7(7):e40925, 2012.

[10] B. Peng, H.S. Chen, L.E. Mechanic, B. Racine, J. Clarke, L. Clarke, E. Gillanders, E.J. Feuer, “Genetic Simulation Resources: a website for the registration and discovery of genetic data simulators”, Bioinformatics, vol . 29(8), pp. 1101-1102, April 2013.

[11] B. Peng, H.S. Chen, L.E. Mechanic, B. Racine, J. Clarke, E. Gillanders, E.J. Feuer, “Genetic data simulators and their applications: an overview”, Genet. Epidemiol., vol. 39(1), pp. 2-10, January 2015.

[12] M.M. Yatskou, V.V. Apanasovich, “Simulation modeling and machine learning platform for processing fluorescence spectroscopy data”, Communications in Computer and Information Science, vol. 1562, pp. 178-190, 2022.

[13] L. Jacquin, T.V. Cao, C. Grenier, N. Ahmadi, “DHOEM: a statistical simulation software for simulating new markers in real SNP marker data”, BMC Bioinformatics, vol. 16:404, December 2015.

[14] J.M. Zook, J. McDaniel, N.D. Olson, J. Wagner, H. Parikh, H. Heaton, S.A. Irvine, L. Trigg, R. Truty, C.Y. McLean, F.M. De La Vega, C. Xiao, S. Sherry, M. Salit, “An open resource for accurately benchmarking small variant and reference calls”, Nat. Biotechnol., vol. 37(5), pp. 561-566, May 2019.

[15] MM Yatskou, Computer Simulation of Energy Relaxation and Transport in Organized Porphyrin Systems, Wageningen University, The Netherlands, 2001.

[16] Y. Liao, G.K. Smyth, W. Shi, “The R Package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads”, Nucleic Acids Research, vol. 47: e47, 2019.

[17] M.M. Yatskou, E.V. Smolyakova, V.V. Skakun, V.V. Grinev, “Entropy-based detection of single-nucleotide genetic polymorphism sites”, Proceedings of the 7th Intern. scientific-practical. conf. “Applied Problems of Optics, Informatics, Radiophysics and Condensed Matter Physics”, May 18–19, 2023, Minsk:Institute of Applied. physical problems for them. AN Sevchenko BGU, pp. 191–193 (in Russian).

[18] V.V. Grinev, M.M. Yatskou, V.V. Skakun, M. Chepeleva, P.V. Nazarov, “ORFhunteR: An accurate approach to the automatic identification and annotation of open reading frames in human mRNA molecules”, Software Impacts, vol. 12:100268, 2022.

[19] T. Hothorn, K. Hornik, A. Zeileis, “Unbiased recursive partitioning: a conditional inference framework”, Journal of Computational and Graphical Statistics, vol. 15(3), pp. 651–674, 2006.

[20] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, 1st ed., Wadsworth, 1984.

[21] K.P. Murphy, Probabilistic Machine Learning, The MIT Press, London, 2022.