УДК 57.087.1

ИМИТАЦИОННОЕ МОДЕЛИРОВАНИЕ ОДНОНУКЛЕОТИДНЫХ ГЕНЕТИЧЕСКИХ ПОЛИМОРФИЗМОВ

Н. Н. ЯЦКОВ[1)]***, В. В. АПАНАСОВИЧ***[2)]***, В. В. ГРИНЕВ***[1)]

[1)]*Белорусский государственный университет, пр. Независимости, 4, 220030, г. Минск, Беларусь*
[2)]*Независимый исследователь, г. Минск, Беларусь*

Аннотация. Для идентификации однонуклеотидных полиморфизмов в последовательностях молекул ДНК предложен подход, основанный на имитационном моделировании сайтов отдельных нуклеотидов с использованием генерации случайных событий по бета-распределению или нормальному распределению, параметры которых оцениваются на базе имеющихся экспериментальных данных. Разработанный подход повышает точность определения однонуклеотидных полиморфизмов в молекулах ДНК и позволяет исследовать достоверность результатов отдельных экспериментов и оценить точность параметров, полученных в реальных условиях проведения эксперимента. Имитационная модель и методы анализа верифицированы на наборе данных геномного секвенирования молекул ДНК человека, предоставленных консорциумом GIAB (Genome in a Bottle Consortium). Выполнен сравнительный анализ известных статистических алгоритмов идентификации однонуклеотидных полиморфизмов и методов машинного обучения, параметры которых настраиваются по смоделированным данным геномного секвенирования молекул ДНК человека. Лучшие результаты получены для моделей машинного обучения, у которых точность идентификации сайтов однонуклеотидных полиморфизмов на 2–5 % выше, чем у классических статистических методов.

Ключевые слова: однонуклеотидный генетический полиморфизм; обнаружение однонуклеотидных полиморфизмов; имитационное моделирование; машинное обучение.

Благодарность. Работа выполнена в рамках государственной программы научных исследований «Конвергенция-2025» (грант № 3.04.3.1, № гос. регистрации 20211918).

Образец цитирования:
Яцков НН, Апанасович ВВ, Гринев ВВ. Имитационное моделирование однонуклеотидных генетических полиморфизмов. *Журнал Белорусского государственного университета. Математика. Информатика.* 2024;2:104–112 (на англ.).
EDN: JHAXAE

For citation:
Yatskou MM, Apanasovich VV, Grinev VV. Simulation modelling of single nucleotide genetic polymorphisms. *Journal of the Belarusian State University. Mathematics and Informatics.* 2024;2:104–112.
EDN: JHAXAE

Авторы:
Николай Николаевич Яцков – кандидат физико-математических наук, доцент; заведующий кафедрой системного анализа и компьютерного моделирования факультета радиофизики и компьютерных технологий.
Владимир Владимирович Апанасович – доктор физико-математических наук, профессор; независимый исследователь.
Василий Викторович Гринев – кандидат биологических наук, доцент; доцент кафедры генетики биологического факультета.

Authors:
Mikalai M. Yatskou, PhD (physics and mathematics), docent; head of the department of systems analysis and computer simulation, faculty of radiophysics and computer technologies.
yatskou@bsu.by
Vladimir V. Apanasovich, doctor of science (physics and mathematics), full professor; independent researcher.
apanasovichv@gmail.com
https://orcid.org/0000-0003-4525-4234
Vasily V. Grinev, PhD (biology), docent; associate professor at the department of genetics, faculty of biology.
grinev_vv@bsu.by

SIMULATION MODELLING OF SINGLE NUCLEOTIDE GENETIC POLYMORPHISMS

M. M. YATSKOU[a], *V. V. APANASOVICH*[b], *V. V. GRINEV*[a]

[a]*Belarusian State University, 4 Niezaliezhnasci Avenue, Minsk 220030, Belarus*
[b]*Independent researcher, Minsk, Belarus*

Corresponding author: M. M. Yatskou (yatskou@bsu.by)

Abstract. We propose an approach for the identification of single nucleotide polymorphisms (SNPs) in DNA sequences, based on the simulation modelling of sites of single nucleotides using the generation of random events according to the beta or normal distributions, the parameters of which are estimated from the available experimental data. The developed approach improves the accuracy of determining SNPs in DNA molecules and permits to investigate the reliability of specific experiments as well as to estimate the errors of determination of the parameters obtained in real experimental conditions. The verification of the simulation model and analysis methods is carried out on a set of reference human genomic DNA sequencing data provided by the Genome in a Bottle Consortium. The comparative analysis of the existing statistical SNP identification algorithms and machine learning methods, trained on the simulated data from the genomic sequencing of human DNA molecules, is carried out. The best results are obtained for machine learning models, in which the accuracy of SNP identification is 2–5 % higher than for classical statistical methods.

Keywords: single nucleotide polymorphism; SNP; SNP identification; simulation modelling; machine learning.

Acknowledgements. This work was carried out in the framework of the state programme of scientific research «Convergence-2025» (grant No. 3.04.3.1, state registration No. 20211918).

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 genome 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 methods are applied for identifying the SNPs in humans and plants [1; 3; 4]. These methods are quite universal and simple for programme implementation, however, they are computationally expensive and difficult to be effectively applied in the analysis of experimental data with a high noise level and various experimental distortions, which are the sources of gaps, repetitions, and other anomalous values [5]. Practical experimental studies use simulation modelling to select a proper SNP identification algorithm, test competing pipelines of analysis, and evaluate the performance of specific experimental designs 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. Modelled 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 in real data 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 coverage (a number of nucleotide reads) 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 methods, and auxiliary biological information, are published elsewhere [9; 10]. However, due to complexities in the types of genetic data, modelling methods, 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 a few modelling methods use experimental results (or measured parameters) and a complex simulation scheme with a 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 us 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 is confirmed in the course of a comparative analysis of the most effective existing statistical algorithms for identifying SNP sites on experimental human genomic sequencing data.

Methodology

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.

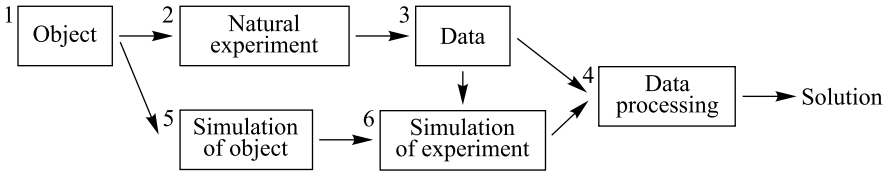

Fig. 1. Scheme of the study of sequenced DNA molecules in natural and simulation experiments

In a natural sequencing experiment (see fig. 1, block 2), data from the object of study (block 1) are recorded. Data processing is carried out in block 6, analysing 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 suitable 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 behaviour 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 behaviour of the object under specific experimental conditions (for example, with known distributions and parameters describing the data). Modelling 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 low coverage, 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.

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, indicating the alignments of the sequenced reads on the reference genome [1]); $D = \{b_1, b_2, b_3, b_4\}$ is a set of n reads (coverage) of nucleotide bases A, C, G or T, recorded from sequencing the site j; the number of coverage n (also the numbers of nucleotide bases b_1, b_2, b_3, b_4, characterising a given nucleotide coverage in the site j) obey the beta (1) or normal (2) distributions:

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

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

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

where μ and σ are parameters of the 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 n, b_1, b_2, b_3, b_4 are reproduced according to the beta or normal distributions in the defined range $[n_{min}; n_{max}]$. For a non-reference site j, the number of coverage n is modelled, then the numbers of coverages for the reference b_{Ref} and non-reference b_{nonRef} nucleotides are generated from the resulting n. Nucleotide coverages for the SNP site are modelled similarly.

It is assumed that there are coverages of no more than two different nucleotide bases on the site. The proposed simulation algorithm reproduces datasets as close as possible to experimental conditions, given by the numbers of nucleotide coverages and the laws of their distributions, the number of SNP sites. The flow diagram of the algorithm for modelling SNP sites is shown in fig. 2.

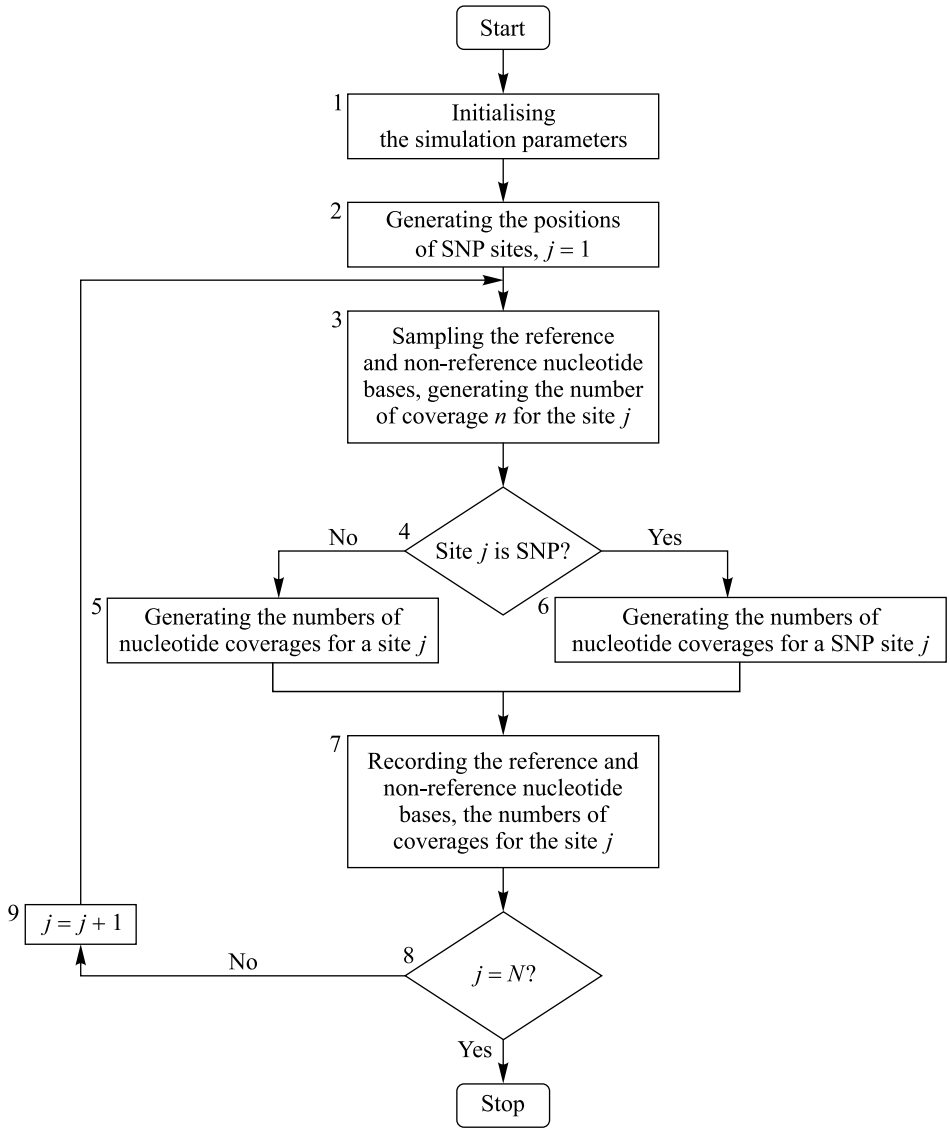

Fig. 2. Flow diagram of the algorithm for modelling SNP sites

Algorithm

Step 1: initialise the model parameters N, N_{SNP}, n_{min} and n_{max}, α and β (or μ and σ) (see fig. 2, block 1). Experimentally extracted sets of the parameters α and β (or μ and σ) of the beta (or normal) distribution for SNP and non-SNP sites are given for simulating the numbers 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: sample the reference and non-reference nucleotide bases, gamble the total number of reads n on the current site j as a realisation of a random variable of the beta or normal distribution with experimentally extracted parameters (block 3).

Step 4: check if the site j is SNP, i. e. $j \in L$ (block 4). Accordingly go to step 5 or 6.

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 nucleotide coverages 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.

Machine learning methods. To apply the machine learning algorithms [14; 15], it is necessary to form a set of features characterising a nucleotide site. It was decided to use four features: X_1 is the number of coverage of the reference nucleotide; $X_2 - X_4$ are the numbers of coverages for non-reference nucleotides sorted in descending order. The data are normalised to the number of coverage n.

Taking into account the limited number of four features, and the binary classification problem (SNP and non-SNP site classes) to be solved, it is preferable to test the basic machine learning methods [16], such as conditional inference trees (CIT) [17], classification and regression tree (CART) [18] and support vector machines with a linear separating function (SVM) [19].

Evaluating SNP identification algorithms. The performance of the algorithms is evaluated using the standard classification measures for unbalanced classes, such as precision (P), recall (R) and score F_1, characterising the properties of the algorithms accept false positive (FP), non-SNPs as SNPs (precision) and false negative (FN), SNPs as non-SNPs (recall), events and their combined contribution F_1 (equations (3)–(5), where TP is true positive) [20]:

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{3}$$

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{4}$$

$$F_1 = \frac{2}{\mathrm{P}^{-1} + \mathrm{R}^{-1}}. \tag{5}$$

Programme development of algorithms. In the course of the work, R-functions are developed that implement various stages of simulation modelling and SNP identification 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 models. The statistical analyses were conducted using the R-functions *dbeta*, *dnorm*, *nls*, *ctree*, *rpart* and *svm* [21].

Results

Experimental data. Reference data on human chromosomes 10 and 22, publicly available from the Genome in a Bottle Consortium (GIAB), are taken as experimental datasets [22]. 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 on chromosomes 22, used for recovering the experimental parameters for simulation modelling, 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 1.

Table 1

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

Organisation of a computational experiment. We analysed the experimental characteristics of the selected dataset of chromosome 22 in order to determine the distribution laws of the nucleotide coverages and to estimate their unknown parameters. Then we checked the adequacy of the developed mathematical model. The machine learning models were trained on specially simulated datasets, generated with the estimated experimental parameters of data distributions on the chromosome 22. Based on the selected sets of experimental data on chromosomes 10 and 22, we conducted a comparative analysis of the most effective existing (traditional or classical) statistical SNP identification and machine learning algorithms, trained on simulated data.

Analysis of the experimental characteristics of genomic sequencing datasets. We analysed the histograms of the number of coverage n, the maximum number of nucleotide coverages $\max\{b_i\}$ and residuals between the total and maximum numbers of coverages $m = n - \max\{b_i\}$ for non-SNP and SNP sites. Approximations of histograms were performed using the density functions of the beta and normal distributions (the R-functions *dbeta* and *dnorm*). To estimate the parameters of distributions, the nonlinear least-squares method was used (the R-function *nls*). The results of histogram approximations are shown in fig. 3.

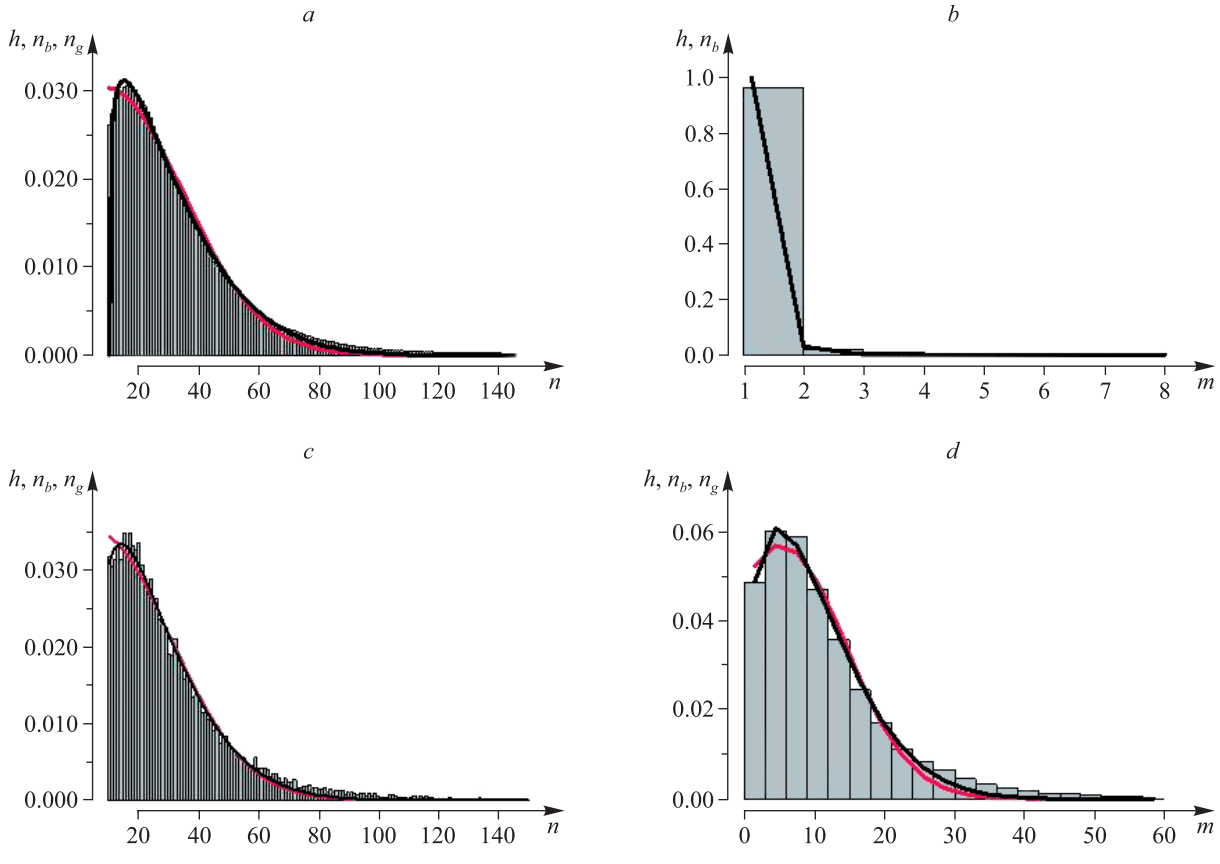

Fig. 3. Normalised histograms h of the number of coverage n (*a*, *c*) and the residuals between the total and maximum numbers of coverages m (*b*, *d*) for non-SNP (*a*, *b*) and SNP (*c*, *d*) sites. Approximations are made by the density functions of the beta n_b (black) and normal n_g (red) distributions; parameter estimates are $\alpha = 1.57$ (standard error is equal 0.02), $\beta = 7.9$ (0.2), and $\mu = 9.2$ (1.1), $\sigma = 25.9$ (0.7) for the fragment *a*; $\alpha = 0.5$ (standard error is equal 0.05), $\beta = 20$ (2) for the fragment *b*; $\alpha = 1.45$ (standard error is equal 0.02), $\beta = 8.4$ (0.2), and $\mu = 5.8$ (1.6), $\sigma = 25.2$ (0.8) for the fragment *c*; $\alpha = 1.71$ (standard error is equal 0.05), $\beta = 7.7$ (0.3), and $\mu = 5.3$ (0.6), $\sigma = 9.2$ (0.6) for the fragment *d*

Our results suggest that the beta distribution is appropriate for the studied integral characteristics of the considered experimental data. The normal distribution is less accurate, but its application might be valid 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.

The experimental estimates of the distribution parameters are further used in the simulation model to generate training data for machine learning methods. A fragment of the simulated dataset is presented in table 2.

Table 2

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

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 coverage n and the accuracy of restoring the modelled 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 the experimental data, and approximated the histograms using the beta and normal distributions. Model parameters were estimated using the R-functions *dbeta* and *dnorm*. The histograms were successfully fitted by the given density functions (fig. 4). The parameters of the simulation models are within 95 % confidence intervals of the parameter estimations, which supports the 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.

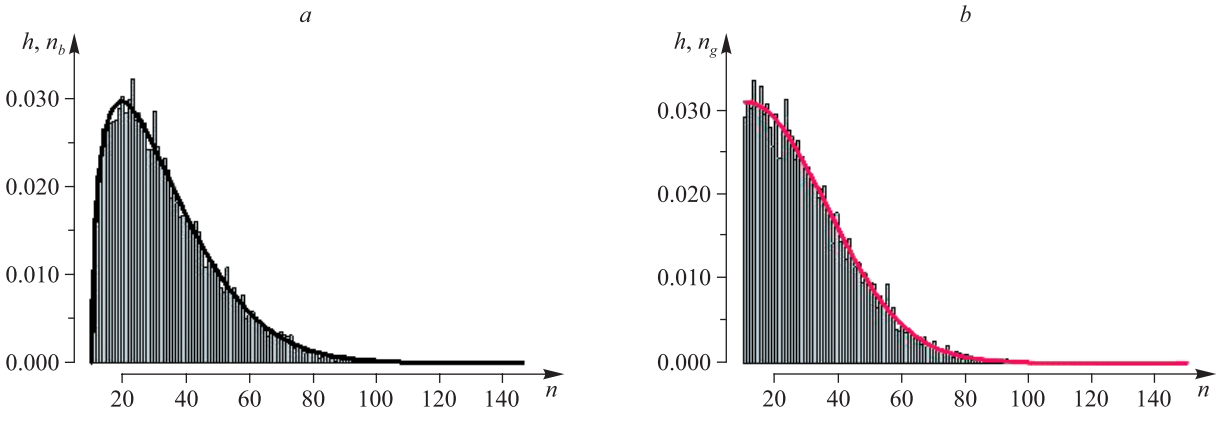

Fig. 4. Normalised histograms h of the number of coverage n in datasets modelled with the experimentally estimated parameters of the beta (*a*) and normal (*b*) distributions. Approximations are made by the density functions of the beta n_b (black) and normal n_g (red) distributions; parameter estimations are $\alpha = 1.50$ (standard error is equal 0.02), $\beta = 7.6$ (0.2) for the fragment *a*; $\mu = 10.4$ (standard error is equal 0.9), $\sigma = 25.2$ (0.6) for the fragment *b*

Comparative analysis of SNP identification algorithms. We performed the comparative analysis of the most effective existing statistical SNP identification algorithms, such as binomial distribution test (BDT), entropy-based test (EBT) and Fisher's exact test (FET), with some fundamental machine learning techniques trained on simulated datasets. An efficient software implementation of BDT was developed, a 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 coverage estimated from the simulated or experimental dataset. As FET, a modification of the algorithm from the R-package *Rsubread* is considered [23]. Our programme implementation of EBT [24] is taken, where thresholds in identifying SNP sites are the entropy E which is more than 0.21 and the p-value which is less than 0.5.

The machine learning methods of CIT (the R-function *ctree* of the package *party*), CART (the R-function *rpart* of the package *rpart*) and SVM (the R-function *svm* of the package *e1071*) were trained on synthetic data simulated with the beta distribution. A training dataset contained 40 000 nucleotide sites, of which 20 000 were SNPs.

Based on the nine selected sets of experimental data on chromosomes 10 and 22, we conducted a comparative analysis of the most effective SNP identification and machine learning algorithms, trained on simulated data. The results of SNP identification at nine datasets of 20 000 sites (per each set), starting from site positions $12 \cdot 10^6$, $60 \cdot 10^6$, $84 \cdot 10^6$, $108 \cdot 10^6$ on chromosome 10 and from site positions $3 \cdot 10^6$, $9 \cdot 10^6$, $15 \cdot 10^6$, $21 \cdot 10^6$, $27 \cdot 10^6$ on chromosome 22, are collected in tables 3 and 4.

Table 3

SNP identification algorithms efficiency by the score F_1 on chromosome 10

Start position	F_1, %					
	BDT	EBT	FET	CIT	CART	SVM
$12 \cdot 10^6$	88.9	100	97.4	100	94.8	97.4
$60 \cdot 10^6$	96.8	94.1	96.9	100	98.4	98.4
$84 \cdot 10^6$	90.3	97.0	96.9	95.4	90.0	90.0
$108 \cdot 10^6$	100	96.9	96.8	100	98.4	98.4
Mean	*94.0*	*97.0*	*97.0*	*98.9*	*95.4*	*96.1*

Table 4

SNP identification algorithms efficiency by the score F_1 on chromosome 22

Start position	F_1, %					
	BDT	EBT	FET	CIT	CART	SVM
$3 \cdot 10^6$	15.4	17.1	11.8	22.2	21.1	20.0
$9 \cdot 10^6$	97.2	97.3	94.3	98.6	95.8	95.8
$15 \cdot 10^6$	86.7	95.7	90.6	98.5	90.3	92.1
$21 \cdot 10^6$	90.3	82.9	91.4	97.1	87.5	90.9
$27 \cdot 10^6$	92.7	88.9	97.5	97.6	95.0	97.6
Mean	*76.5*	*76.4*	*77.1*	*82.8*	*77.9*	*79.3*

Chromosome 10. The highest mean score F_1 was obtained for the CIT machine learning model, and the lowest score F_1 was obtained for the classical BDT method. The accuracy by the score F_1 of machine learning methods slightly exceeds that of classical methods. The mean score F_1 of the best machine learning model CIT does not significantly differ from the value of the FET method (the *p*-value of the two sample paired Student's *t*-test for not equal variances is 0.19).

Chromosome 22. The highest mean score F_1 was obtained for the CIT machine learning model, the lowest score F_1 was obtained for the classical EBT method. The accuracy by the score F_1 exceeds that of classical methods. The mean score F_1 of the best machine learning model CIT is statistically significantly higher than that of the best classical FET method (the *p*-value of the two sample paired Student's *t*-test for equal variances is 0.03). A significant increase in accuracy may be due to training the model on experimental data for chromosome 22.

Additionally, we investigated the CIT and CART methods trained on experimental datasets, sampled from the experimental data on chromosome 22. A typical dataset of 72 261 sites was considered, 36 150 of which were SNPs and the rest randomly selected non-SNPs. The classification accuracy by the score F_1 did not exceed 60–70 % on the 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.

These results let us conclude that for real experimental data it is preferable to use machine learning models, trained on simulated data. The mean accuracy of SNP identification in terms of the score F_1 is 2–5 % higher for the machine learning models, in particular decision tree-based, than for classical statistical methods. The CIT model shows the highest accuracy, while BDT, EBT, FET have similar mean accuracy.

Conclusions

An approach for simulation modelling of SNPs in DNA sequences has been 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, and it applies machine learning methods trained on simulated data to identify the single nucleotide genetic polymorphism sites. This approach has some distinct advantages, namely, it permits: a) to achieve the higher accuracy of determining SNPs in genomic sequencing data; b) to simulate data closely reproducing the real experimental conditions in order to study the reliability of specific experiments and assess the accuracy of the results obtained under the observed experimental conditions; c) to generate synthetics data for training machine learning methods and subsequently create the classification models of machine learning algorithms to identify the SNPs in specific experimental datasets; d) to generate datasets for testing and comparing available SNP identification methods to analyse real data obtained for specific experimental conditions. The verification of the developed simulation model and the analysis algorithms is realised on the examples of large humane chromosome sequencing datasets. The comparative analysis of efficient existing statistical SNP identification algorithms of BDT, EBT and FET and machine learning methods of CIT, CART and SVM, trained on synthetic data, is carried out. The best results are obtained for machine learning models, namely, the accuracy of SNP identification by the score F_1 is 2–5 % higher for the trained on simulated data CIT than those for the methods of BDT, EBT and FET.

References

1. Sung WK. *Algorithms for next-generation sequencing*. 1st edition. New York: Chapman & Hall/CRC; 2017. 364 p. DOI: 10.1201/9781315374352.

2. Kappelmann-Fenzl M, editor. *Next generation sequencing and data analysis*. 1st edition. Cham: Springer; 2021. 218 p. DOI: 10.1007/978-3-030-62490-3.

3. Wu XL, Xu J, Feng G, Wiggans GR, Taylor JF, He J, et al. Optimal design of low-density SNP arrays for genomic prediction: algorithm and applications. *PLoS ONE*. 2016;11(9):e0161719. DOI: 10.1371/journal.pone.0161719.

4. Korani W, Clevenger JP, Chu Y, Ozias-Akins P. Machine learning as an effective method for identifying true single nucleotide polymorphisms in polyploid plants. *Plant Genome*. 2019;12(1):180023. DOI: 10.3835/plantgenome2018.05.0023.

5. Masoudi-Nejad A, Narimani Z, Hosseinkhan N. *Next generation sequencing and sequence assembly. Methodologies and algorithms*. 1st edition. New York: Springer; 2013. 86 p. DOI: 10.1007/978-1-4614-7726-6.

6. Su Z, Marchini J, Donnelly P. HAPGEN2: simulations of multiple disease SNPs. *Bioinformatics*. 2011;27(16):2304–2305. DOI: 10.1093/bioinformatics/btr341.

7. Oh JH, Deasy JO. SITDEM: a simulation tool for disease/endpoint models of association studies based on single nucleotide polymorphism genotypes. *Computers in Biology and Medicine*. 2014;45:136–142. DOI: 10.1016/j.compbiomed.2013.11.021.

8. Meyer HV, Birney E. PhenotypeSimulator: a comprehensive framework for simulating multi-trait, multi-locus genotype to phenotype relationships. *Bioinformatics*. 2018;34(17):2951–2956. DOI: 10.1093/bioinformatics/bty197.

9. Hendricks AE, Dupuis J, Gupta M, Logue MW, Lunetta KL. A comparison of gene region simulation methods. *PLoS ONE*. 2012;7(7):e40925. DOI: 10.1371/journal.pone.0040925.

10. Peng B, Chen HS, Mechanic LE, Racine B, Clarke J, Clarke L, et al. Genetic Simulation Resources: a website for the registration and discovery of genetic data simulators. *Bioinformatics*. 2013;29(8):1101–1102. DOI: 10.1093/bioinformatics/btt094.

11. Peng B, Chen HS, Mechanic LE, Racine B, Clarke J, Gillanders E, et al. Genetic data simulators and their applications: an overview. *Genetic Epidemiology*. 2015;39(1):2–10. DOI: 10.1002/gepi.21876.

12. Yatskou MM, Apanasovich VV. Simulation modelling and machine learning platform for processing fluorescence spectroscopy data. In: Tuzikov AV, Belotserkovsky AM, Lukashevich MM, editors. *Pattern Recognition and Information Processing. PRIP-2021*. Cham: Springer; 2022. p. 178–190 (Communications in computer and information science; volume 1562). DOI: 10.1007/978-3-030-98883-8_13.

13. Jacquin L, Cao TV, Grenier C, Ahmadi N. DHOEM: a statistical simulation software for simulating new markers in real SNP marker data. *BMC Bioinformatics*. 2015;16:404. DOI: 10.1186/s12859-015-0830-7.

14. Volkau AU, Yatskou MM, Grinev VV. Selecting informative features of human gene exons. *Journal of the Belarusian State University. Mathematics and Informatics*. 2019;1:77–89. Russian. DOI: 10.33581/2520-6508-2019-1-77-89.

15. Xu Silun, Skakun VV. Comparative analysis of deep learning neural networks for the segmentation of cancer cell nuclei on immunohistochemical fluorescent images. *Journal of the Belarusian State University. Mathematics and Informatics*. 2024;1:59–70. Russian. EDN: TOOSJI.

16. Grinev VV, Yatskou MM, Skakun VV, Chepeleva MV, Nazarov PV. ORFhunteR: an accurate approach to the automatic identification and annotation of open reading frames in human mRNA molecules. *Software Impacts*. 2022;12:100268. DOI: 10.1016/j.simpa.2022.100268.

17. Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: a conditional inference framework. *Journal of Computational and Graphical Statistics*. 2006;15(3):651–674. DOI: 10.1198/106186006X133933.

18. Breiman L, Friedman J, Olshen R, Stone C. *Classification and regression trees*. 1st edition. Wadsworth: Wadsworth International Group; 1984. 358 p.

19. Vapnik VN. *The nature of statistical leaning theory*. 2nd edition. New York: Springer; 2000. 314 p. DOI: 10.1007/978-1-4757-3264-1.

20. Murphy KP. *Probabilistic machine learning* [Internet]. London: The MIT Press; 2022. 864 p. Available from: https://mitpress.mit.edu/9780262369305/probabilistic-machine-learning.

21. R Core Team. R: a language and environment for statistical computing. R foundation for statistical computing [Internet]. Vienna: [s. n.]; 2021. Available from: https://www.R-project.org.

22. Zook JM, McDaniel J, Olson ND, Wagner J, Parikh H, Heaton H, et al. An open resource for accurately benchmarking small variant and reference calls. *Nature Biotechnology*. 2019;37(5):561–566. DOI: 10.1038/s41587-019-0074-6.

23. Liao Y, Smyth GK, Shi W. The R-package *Rsubread* is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Research*. 2019;47(8):e47. DOI: 10.1093/nar/gkz114.

24. Yatskou MM, Smolyakova EV, Skakun VV, Grinev VV. Entropy-based detection of single-nucleotide genetic polymorphism sites. In: A. N. Sevchenko Institute of Applied Physical Problems of Belarusian State University. *Proceedings of the 7th International scientific-practical conference «Applied problems of optics, informatics, radiophysics and condensed matter physics»; 2023 May 18–19; Minsk, Belarus*. Minsk: Belarusian State University; 2023. p. 191–193. Russian.

Received 22.01.2024 / revised 01.07.2024 / accepted 01.07.2024.