Имитационное моделирование сайтов однонуклеотидных генетических полиморфизмов в молекулах ДНК

Д. Д. Сарнацкий, Н. Н. Яцков, В. В. Гринев

Белорусский государственный университет, Минск, Беларусь,
e-mail: denisiussarnatski@gmail.com

В работе представлены результаты исследования алгоритмов имитационного моделирования сайтов однонуклеотидных генетических полиморфизмов в нуклеотидных последовательностях молекул ДНК с учетом параметрического и непараметрического способов определения функций распределений вероятностей при воспроизведении чисел покрытий нуклеотидных сайтов. Сравнительный анализ алгоритмов имитационного моделирования выполнен на экспериментальных данных генома человека.

Ключевые слова: молекулы ДНК, однонуклеотидный генетический полиморфизм, имитационное моделирование.

Simulation modelling of single nucleotide genetic polymorphism sites in DNA molecules

D. D. Sarnatski, M. M. Yatskou, V. V. Grinev

Belarusian State University, Minsk, Belarus, e-mail: denisiussarnatski@gmail.com

This paper presents the results of a study of simulation algorithms for the modelling of single nucleotide polymorphism sites in DNA nucleotide sequences, taking into account parametric and nonparametric methods for determining probability distribution functions for reproducing nucleotide site coverage numbers. A comparative analysis of the simulation algorithms was performed using experimental data from the human genome.

Keywords: DNA molecules, single nucleotide genetic polymorphism, simulation modelling.

Введение

Методы экспериментальной флуоресцентной спектроскопии используются при секвенировании молекул ДНК [1-3]. Секвенирование генома человека позволяет одновременно идентифицировать множество сайтов однонуклеотидного генетического полиморфизма (SNP, от англ. single nucleotide polymorphism), имеющих диагностическую или прогностическую значимость в отношении многих заболеваний человека [4, 5]. Имитационное моделирование сайтов SNP позволяет предсказывать и изучать влияние генетических вариаций на фенотипы, заболевания и ответы на лекарственные препараты. Использование статистических оценок экспериментальных распределений в имитационных моделях генетических вариаций позволяет моделировать данные, близкие к экспериментальным [6]. В работах [6, 7] представлены имитационные модели, в основе которых лежит использование заданных функций распределений вероятностей (параметрический способ моделирования). Основным ограничением моделей является понижение точности модели-

рования при несоответствии подобранных распределений экспериментальным характеристикам регистрируемых прочтений нуклеотидных сайтов. Возможным способом устранения ограничения является учет экспериментальных гистограмм чисел покрытий нуклеотидных сайтов в качестве законов распределений (непараметрический способ моделирования) [8].

Целью данной работы является разработка и исследование имитационных моделей сайтов однонуклеотидных генетических полиморфизмов в нуклеотидных последовательностях молекул ДНК с учетом параметрического и непараметрического способов определения функций распределений вероятностей при воспроизведении чисел покрытий нуклеотидных сайтов. Сравнительный анализ алгоритмов имитационного моделирования выполнен на экспериментальных данных генома человека.

1. Имитационное моделирование сайтов нуклеотидных последовательностей

Имитационное моделирование нуклеотидных последовательностей молекул ДНК производится по экспериментальным данным, в предположении подчинения основных характеристик данных, а именно – чисел покрытий нуклеотидных сайтов, некоторым законам распределений вероятностей [6, 7]. Идея моделирования состоит в случайной генерации N_{SNP} позиций SNP сайтов в последовательности рассматриваемой молекулы S, состоящей из N нуклеотидных сайтов, для каждого из которых числа покрытий в референсном канале R и трех нереференсных N_1, N_2, N_3 воспроизводятся с учетом корреляционных взаимосвязей между каналами в соответствии с заданным законом распределения. Блок-схема алгоритма моделирования сайтов нуклеотидных последовательностей молекул ДНК с учетом однонуклеотидных полиморфизмов представлена на рис. 1.

Предложенный алгоритм позволяет воспроизводить наборы данных максимально приближенные к экспериментальным условиям, задаваемыми числами покрытий и законами их распределений, количеством полиморфных сайтов. Повысить точность моделирования можно с помощью подбора наиболее точного закона распределения или использования экспериментальных гистограмм характеристик нуклеотидных сайтов в качестве законов распределений.

В работе рассмотрены три подхода имитационного моделирования – с использованием параметрического и непараметрического способов определения функций распределений, в том числе статистического метода бутстрапирования, основанного на построении эмпирических функций распределений по (бутстра-повским) выборкам из анализируемых данных [9]. Для параметрического имитационного моделирования выбраны нормальное, бета-, гамма-, Вейбула и полиномиальное распределения, параметры которых оцениваются по экспериментальным данным. Для параметрического имитационного моделирования рассмотрены экспериментальные гистограммы, представляющие собой эмпирические функции распределения числе нуклеотидах покрытий в качестве оценок неизвестных распределений.

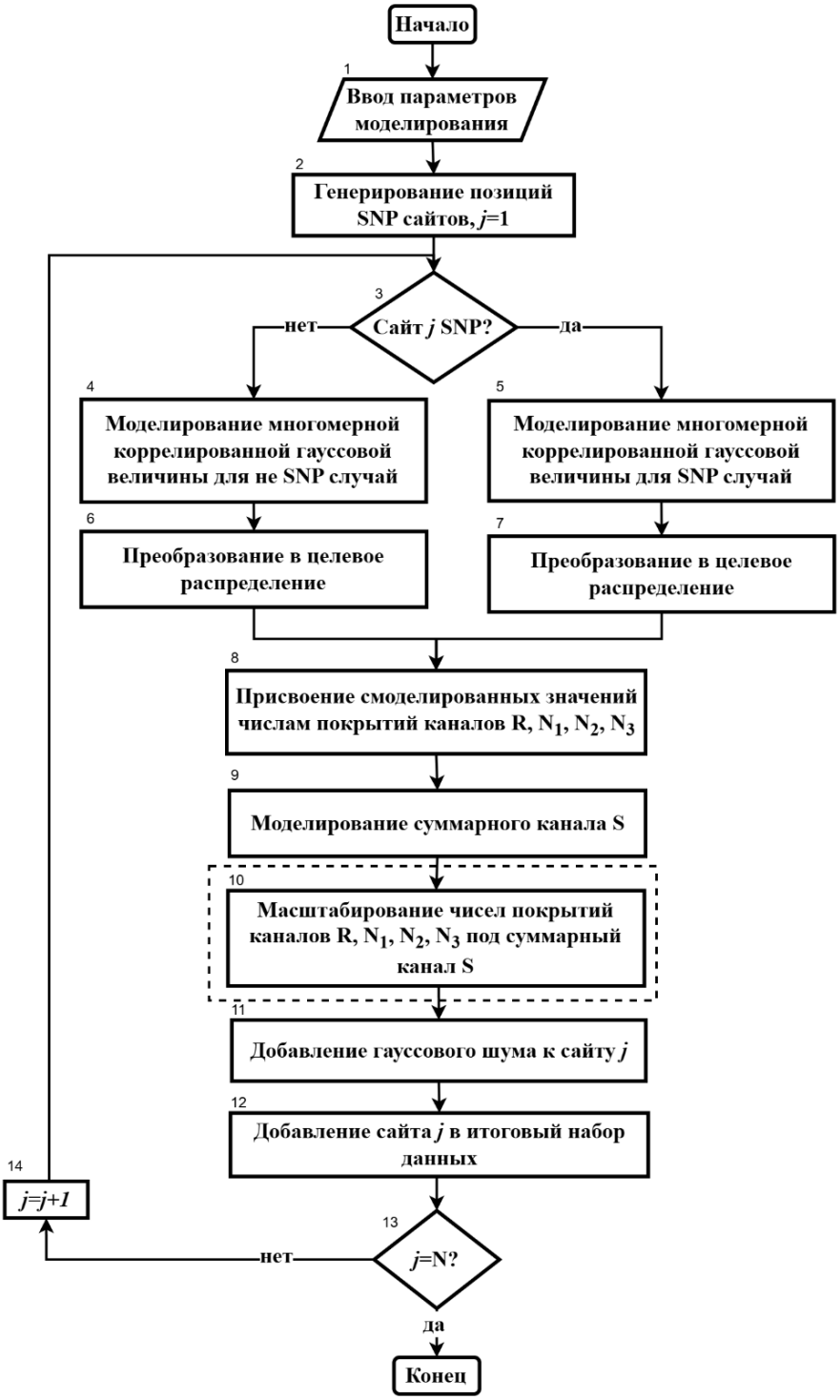

Рис. 1. Блок-схема алгоритма моделирования сайтов нуклеотидных последовательностей молекул ДНК с учетом однонуклеотидных полиморфизмов

2. Организация вычислительного эксперимента

В вычислительном эксперименте рассмотрены эталонные данные о хромосоме 22, полученные консорциумом GIAB [10]. Набор данных содержит характеристики 29 633 768 нуклеотидных сайтов, из которых 36 150 истинно идентифицированные

сайты SNP. При параметрическом моделировании выполняется анализ экспериментальных характеристик набора данных хромосомы 22 с целью определения наиболее точного закона распределения и оценки его параметров, при непараметрическом – построение экспериментальных гистограмм чисел нуклеотидных покрытий, при бутстрапировнии – построение эмпирических функций распределений по бутстраповским выборкам. Смоделированы наборы данных по 40 000 сайтов, в каждом из которых 20 000 сайтов SNP.

Для сравнения разработанных имитационных моделей используются графики соответствия смоделированных и экспериментально верифицированных плотностей распределений чисел покрытий нуклеотидных сайтов и точность восстановления коэффициентов линейных корреляций между каналами покрытий нуклеотидных сайтов.

3. Результаты

На языке программирования R реализованы алгоритмы имитационных моделей нуклеотидных последовательностей молекул ДНК с учетом параметрического и непараметрического способов определения функций распределений вероятностей при воспроизведении чисел покрытий нуклеотидных SNP и не SNP сайтов. Среди параметрических распределений наиболее точно аппроксимируют экспериментальные данные хромосомы 22 полиномиальное распределение (полином 12-й степени). Проведено исследование реализованных алгоритмов имитационного моделирования. Графики смоделированных и экспериментальных гистограмм чисел покрытий нуклеотидных сайтов представлены на рис. 2.

Наиболее точные воспроизведения исследуемых характеристик моделируемых покрытий сайтов в различных нуклеотидных каналах и линейных статистических взаимосвязей между каналами получены для метода бутстрапирования. Основными недостатками данного способа моделирования являются невозможность генерации наборов данных при исследовании специальных экспериментальных условий, труднореализуемых на практике, и необходимость использования полного набора экспериментальных данных.

К преимуществам параметрического моделирования можно отнести шумоустойчивость, интерпретируемость, вычислительную эффективность и удобство настройки параметров моделирования, в том числе при исследовании специальных экспериментальных условий. Недостатки – сложность учета и невысокая точность воспроизведения статистических взаимосвязей между нуклеотидными каналами.

При непараметрическом моделировании важно отметить довольно точное воспроизведение покрытий отдельных каналов, в том числе лучшее, в сравнении с параметрическим моделированием, генерирование статистических корреляций между нуклеотидными каналами. Основные недостатки – неточное воспроизведение суммарного канала для SNP сайтов, степенная вычислительная эффективность, отсутствие шумоустойчивости и необходимость использования полного набора экспериментальных данных.

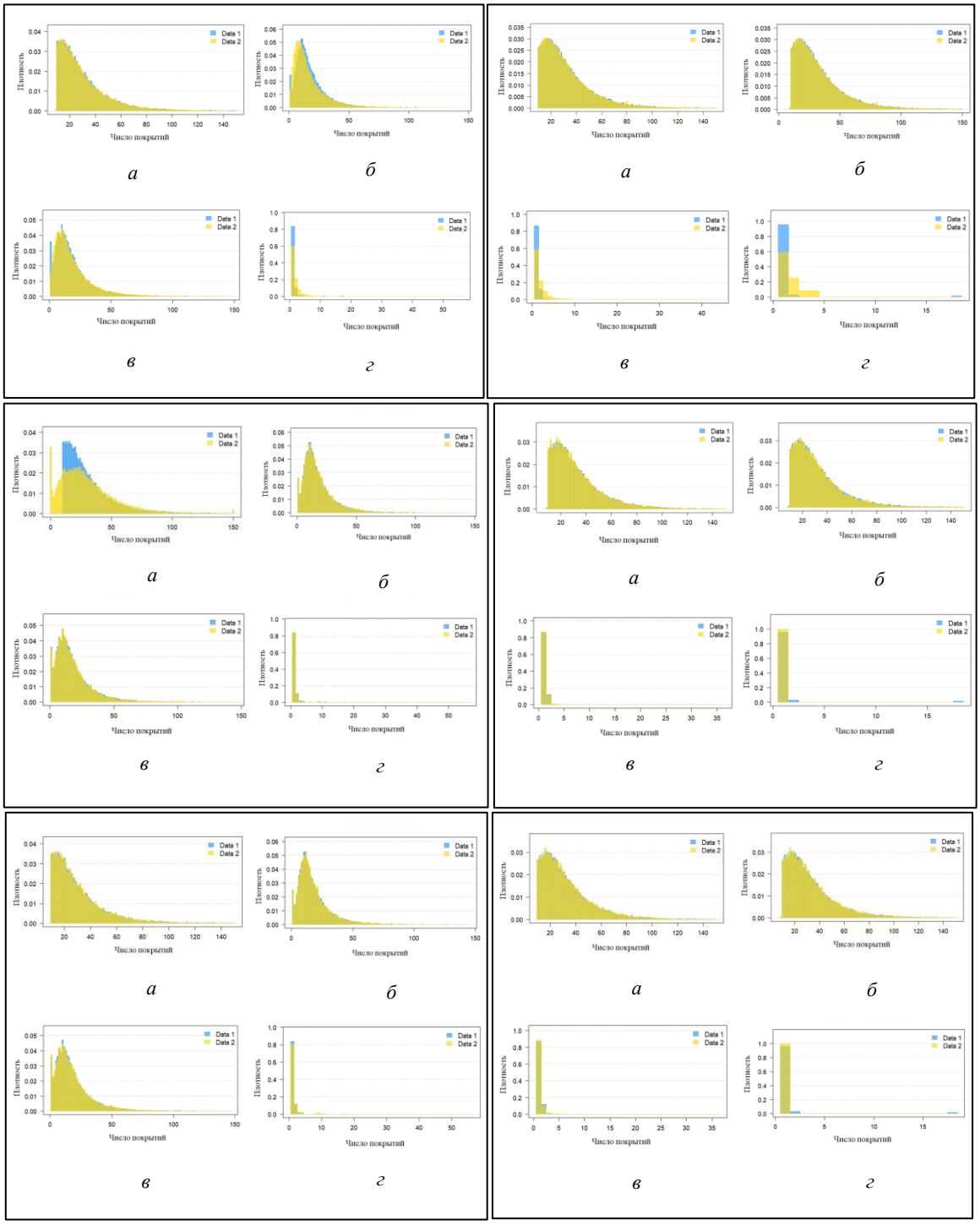

Рис. 2. Гистограммы экспериментальных (Data1) и смоделированных (Data2) данных для SNP (колонка слева) и не SNP (колонка справа) сайтов. *Ряды рисунков 1-3* – результаты параметрического, непараметрического и бутстрап-моделирования соответственно;
а – суммарный канал; *б* – референсный канал; *в* – 1-й нереференсный канал;
г – 2-й нереференсный канал

4. Заключение

Реализованы эффективные алгоритмы имитационного моделирования сайтов однонуклеотидных генетических полиморфизмов в нуклеотидных последовательностях молекул ДНК с учетом параметрического и непараметрического способов определения функций распределений вероятностей при воспроизведении чисел покрытий нуклеотидных сайтов. Проверка адекватности разработанных моделей и сравнительный анализ точности моделирования выполнены на примере набора эталонных данных о хромосоме 22, полученных консорциумом GIAB. Наиболее точно воспроизводящим экспериментальные гистограммы чисел покрытий сайтов и статистические взаимосвязи между нуклеотидными каналами является метод бутстрапирования.

Разработанные алгоритмы имитационного моделирования могут использоваться для генерации синтетических данных по экспериментальным данным или самостоятельно с целью всестороннего тестирования и выбора наилучших алгоритмов идентификации сайтов однонуклеотидных полиморфизмов, а также для генеративного моделирования данных с целью обучения алгоритмов идентификации на основе методов машинного обучения [11, 12].

Библиографические ссылки

1. *Lakowicz J. R.* Principles of Fluorescence Spectroscopy. 3rd ed. New York : Springer, 2006.
2. *Demchenko A. P.* Introduction to Fluorescence Sensing. Volume 1: Materials and Devices. 3rd ed. Cham, Switzerland : Springer, 2020.
3. *Masoudi-Nejad A.* Next Generation Sequencing and Sequence Assembly. / A. Masoudi-Nejad, Z. Narimani, N. Hosseinkhan // Methodologies and Algorithms, New York : Springer, 2013.
4. *Sung W.-K.* Algorithms for Next-Generation sequencing. 1st ed. Chapman & Hall/CRC, 2017.
5. *Kappelmann-Fenzl M., ed.* Next Generation Sequencing and Data Analysis. Cham : Springer, 2021.
6. *Yatskou M. M.* Simulation modelling of single nucleotide genetic polymorphisms / M. M. Yatskou , V. V. Apanasovich , V. V. Grinev // Journal of the Belarusian State University. Mathematics and Informatics. 2024. Vol. 2. P. 104–112.
7. *Сарнацкий Д. Д.* Имитационная модель генерации сайтов однонуклеотидного полиморфизма в молекулах ДНК человека / Д. Д. Сарнацкий, Н. Н. Яцков, В. В. Гринев // Компьютерные технологии и анализ данных (CTDA’2024) : материалы IV Междунар. науч.-практ. конф., Минск, 25–26 апр. 2024 г. – Минск : БГУ, 2024. С. 265–268.
8. Simulation modelling for machine learning identification of single nucleotide polymorphisms in human genomes / M. M. Yatskou [et al] // Pattern Recognition and Information Processing (PRIP’2023) : Proc. of the 16th Intern. Conf., Minsk, 17–19 Oct. 2023. – Minsk : BSU, 2023. P. 49–53.
9. *Efron B.* Bootstrap methods: Another look at the jackknife // The Annals of Statistics. 1979. Vol. 7(1). P. 1–26.
10. An open resource for accurately benchmarking small variant and reference calls / N. D. Olson [et al] // Nature Biotechnology. 2019. Vol. 37(5). P. 561–566.
11. *Яцков, Н. Н.* Генеративное имитационное моделирование сложных биофизических систем / Н. Н. Яцков, В. В. Апанасович, В. Н. Яцков // Компьютерные технологии и анализ данных (CTDA’2024): материалы IV Междунар. науч.-практ. конф., Минск, 25–26 апр. 2024 г. – Минск : БГУ, 2024. С. 211–214.
12. Identification of single nucleotide genetic polymorphism sites using machine learning methods / M. M. Yatskou [et al] // Advances in Transdisciplinary Engineering. 2023. Vol. 42. P. 1031–1037.