

ИМИТАЦИОННАЯ МОДЕЛЬ ГЕНЕРАЦИИ САЙТОВ ОДНОНУКЛЕОТИДНОГО ПОЛИМОРФИЗМА В МОЛЕКУЛАХ ДНК ЧЕЛОВЕКА

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

Белорусский государственный университет, Минск, Беларусь
E-mail: sarnacki@bsu.by;

В работе предложена имитационная модель генерации сайтов однонуклеотидного полиморфизма в молекулах ДНК человека для обработки экспериментальных данных геномного секвенирования. В модели имитируются числа нуклеотидных прочтений в четырех каналах системы регистрации, учтены статистические линейные взаимосвязи между каналами и реализовано добавление гауссового шума для тестирования и сравнительного анализа алгоритмов определения сайтов полиморфизма.

Ключевые слова: однонуклеотидный полиморфизм; имитационное моделирование.

ВВЕДЕНИЕ

В живых организмах состав геномов варьирует. Различия называются генетическим полиморфизмом и одной из наиболее частых форм такой вариации является однонуклеотидный полиморфизм (SNP, от англ. Single Nucleotide Polymorphism). Генетические полиморфизмы влияют на фенотипы и заболевания [1]. 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 [2]. В работе [2] представлена имитационная модель, основанная на бета-распределении. Основным ограничением модели является отсутствие воспроизведения числа покрытий в четырех каналах системы регистрации прочтений нуклеотидных сайтов.

В связи с этим, целью настоящей работы стала разработка имитационной модели для воспроизведения числа покрытий в четырех каналах регистрации нуклеотидных прочтений сайтов с учетом статистических взаимосвязей между каналами.

МЕТОДОЛОГИЯ ИССЛЕДОВАНИЯ

Предположим, что сайт j содержит референсное нуклеотидное основание r (нуклеотиды А, С, G или Т); $D = \{R, N_1, N_2, N_3\}$ – набор из n прочтений (ридов) нуклеотидных оснований А, С, G или Т, покрывающих сайт j , получаемых при геномном секвенировании.

Идея алгоритма имитационного моделирования заключается в случайной генерации N_{SNP} позиций SNP-сайтов в последовательности рассматриваемой молекулы S , состоящей из N нуклеотидных сайтов, для каждого из которых сумма чисел покрытий сайта n , 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 (формулы (1)-(4)).

Бета-распределение

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

где β и α ($\beta, \alpha > 0$) – произвольные фиксированные параметры, задающие форму кривой распределения; Γ – гамма-функция.

Нормальное распределение

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

где μ и σ – параметры, математическое ожидание и среднеквадратическое отклонение.

Распределение Вейбулла

$$n_w(x, k, \lambda) = \frac{k}{\lambda} \left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^k}, \quad (3)$$

где k и λ ($k, \lambda > 0$) – параметры, задающие форму кривой распределения.

Гамма-распределение

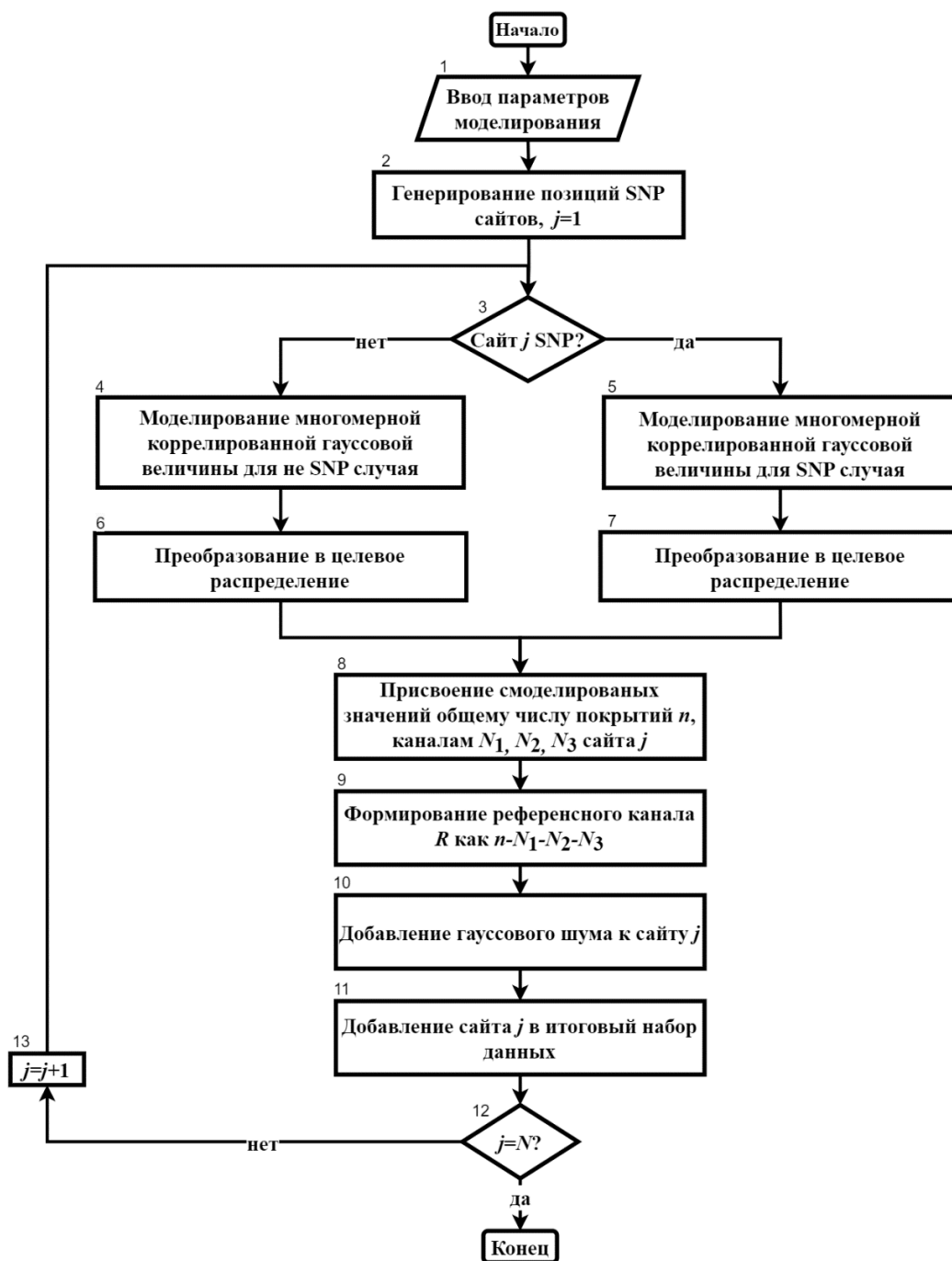
$$n_\Gamma(x, k, \theta) = x^{k-1} \frac{\exp(-\frac{x}{\theta})}{\Gamma(k)\theta^k}, \quad (4)$$

где k и θ ($k, \theta > 0$) – параметры, задающие форму кривой распределения.

Для экспериментальных данных характерно преобладание нулевых значений в нереференсных каналах. Предложено использовать при моделировании вероятности нулевых значений p_1 для канала N_1 и условные вероятности $p_2(0|N_1)$ и $p_3(0|N_2)$ для каналов N_2 и N_3 . Если $\sum_{i=1}^3 N_i \geq n$, то $R = 0$.

Для моделирования линейно-зависимых гауссовых случайных величин используется метод Халецкого с последующим преобразованием к заданному распределению. Реализовано добавление гауссового шума для тестирования и сравнительного анализа алгоритмов определения сайтов однонуклеотидного полиморфизма.

Блок-схема алгоритма имитационного моделирования однонуклеотидных полиморфизмов представлена на рисунке.



Блок-схема алгоритма моделирования

РЕЗУЛЬТАТЫ

Рассмотрены экспериментальные данные 22-ой хромосомы генома HG001 [2]. Удалены нуклеотидные сайты с числом покрытий меньше 10. Гистограммы чисел покрытий n , R , N_1 , N_2 , N_3 , построенные по экспериментальным данным, аппроксимированы функциями плотностей нормального, бета-, Вейбулла и гамма-распределений. Наилучшие результаты получены

для гамма-распределения. Полученные экспериментальные оценки параметров распределений используются в имитационном моделировании. В качестве проверки адекватности разработанной модели рассмотрены диаграммы соответствия смоделированных и экспериментально верифицированных плотностей распределений для чисел покрытий n , N_1 , N_2 , N_3 и точность восстановления моделируемых параметров при оценке параметров распределений.

Параметры имитационных моделей попадают в 95% доверительные интервалы оценок параметров, что позволяет предположить об адекватности имитационных моделей, а именно – о корректности процедур моделирования числа покрытий сайтов в четырех каналах по нормальному, бета, Вейбулла и гамма законам распределений.

Для исследования алгоритмов идентификации сайтов SNP смоделированы данные с учетом добавления гауссового шума с параметром σ . Варьирование параметром σ изменяет уровень экспериментального шума, а именно – регулирует информативность полезного сигнала, что позволяет всесторонне исследовать эффективность разрабатываемых или выбранных алгоритмов идентификации сайтов SNP и воссоздавать специальные экспериментальные условия.

ЗАКЛЮЧЕНИЕ

Разработана имитационная модель для воспроизведения чисел покрытий в четырех каналах регистрации нуклеотидных прочтений сайтов с учетом статистических взаимосвязей между каналами. Наивысшая точность моделирования числа покрытий установлена для Гамма распределения. Предложенная имитационная модель учитывает статистические линейные взаимосвязи между каналами регистрации числа прочтений, что позволяет более точно исследовать характеристики нуклеотидных сайтов. При сравнительном анализе методов обнаружения сайтов однонуклеотидных полиморфизмов добавление гауссового шума позволяет учесть инструментальную ошибку секвенирующего устройства и повысить точность имитационного моделирования.

БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. Sung W.K. Algorithms for next-generation sequencing. // Chapman & Hall/CRC Computational Biology Series. 2017. P. 175-180.
2. Yatskou M.M., Smolyakova E.V., Skakun V.V., Grinev V.V. Simulation modelling for machine learning identification of single nucleotide polymorphisms in human genomes. // Pattern Recognition and Information Processing (PRIP'2023). Artificial Universe: New Horisont: Proceedings of the 16th International Conference, October 17–19, 2023, Belarus, Minsk / Belarusian State University: eds. A. Nedzved. A. Belotserkovsky. Minsk: BSU, 2023. P. 49-53.