Идентификация сайтов однонуклеотидного генетического полиморфизма с использованием методов машинного обучения

Н. Н. Яцков, Е. В. Смолякова, К. И. Грудовик, В. В. Скакун, В. В. Гринев

Белорусский государственный университет, Минск, Беларусь,
e-mail: yatskou@bsu.by

В работе представлен алгоритм имитационного моделирования сайтов нуклеотидных вариаций в геномной ДНК. Для идентификации сайтов однонуклеотидного генетического полиморфизма предложено использовать методы машинного обучения, обученные на смоделированных данных. Выполнен сравнительный анализ наиболее эффективных методов машинного обучения и классических алгоритмов идентификации сайтов однонуклеотидного полиморфизма на смоделированных данных.

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

Identification of single nucleotide genetic polymorphisms using machine learning methods

M. M. Yatskou, E. V. Smolyakova, K. I. Grudovik, V. V. Skakun, V. V. Grinev

Belarusian State University, Minsk, Belarus, e-mail: yatskou@bsu.by

The paper presents an algorithm for simulation of nucleotide variations in the genomic DNA. To identify single-nucleotide genetic polymorphisms, it is proposed to use machine learning methods trained on simulated data. A comparative analysis of the most effective classical and machine learning algorithms for identifying single nucleotide polymorphisms was performed on simulated data.

Keywords: single nucleotide genetic polymorphism, simulation modelling, machine learning.

Введение

Методы экспериментальной флуоресцентной спектроскопии используются для изучения оптических свойств молекулярных соединений и находят широкое применение при секвенировании молекул ДНК [1, 2]. Генетические процессы изучаются с помощью экспериментов геномного секвенирования, в результате которых регистрируется информация о составе молекул ДНК и РНК, экспрессия их кодирующих фрагментов [3]. Полное секвенирование генома или секвенирование только функционально значимых регионов 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]. Среди существующих способов определения сайтов SNP следует отметить статистические методы точного теста Фишера, биномиального распределения, на основе энтропии и машинного обучения с использованием нейронных сетей [4, 6, 7]. Методы достаточно универсальны и просты для программной реализации, однако вычислительно затратные и трудно применимы при анализе экспериментальных данных с высоким уровнем

шума и различными экспериментальными искажениями, являющимися источниками пропусков и повторов [3]. В практических экспериментальных исследованиях для выбора наиболее оптимального алгоритма идентификации сайтов, проверки конкурирующих методик анализа и оценки производительности конкретных экспериментальных планов исследования биофизических систем используется имитационное моделирование [8]. Имитационное моделирование также применяется для генерации обучающих данных для методов машинного обучения с целью прямой идентификации сайтов SNP растений по данным отдельного эксперимента секвенирования [9]. В этом случае формирование смоделированных обучающих данных может иметь преимущества по точности и эффективности при анализе экспериментальных данных как с невысоким числом покрытий, так и с наличием пропусков, обусловленных экспериментальными искажениями. Предполагается, что обучение на смоделированных данных конкретного эксперимента по исследованию генома человека позволит повысить точность алгоритмов машинного обучения при идентификации сайтов SNP.

Целью исследования является разработка имитационной модели сайтов нуклеотидной последовательности и сравнительный анализ наиболее эффективных методов машинного обучения и классических алгоритмов идентификации сайтов SNP. Сравнительный анализ алгоритмов идентификации сайтов SNP выполнен на смоделированных данных.

1. Имитационное моделирование сайтов нуклеотидных последовательностей

Имитационное моделирования SNP сайтов производится по экспериментальным данным, в предположении подчинения основных характеристик данных, а именно – числа покрытий, бета- или нормальному законам распределений [10]. Предположим, что сайт j содержит референсное нуклеотидное основание r (нуклеотиды A, C, G или T); $D = \{b_1, b_2, b_3, b_4\}$ – набор из n прочтений (ридов или покрытий) нуклеотидных оснований A, C, G или T, покрывающих сайт j; количества покрытий сайта n, b_1, b_2, b_3, b_4 подчиняются бета- (1) или нормальному (2) законам распределений

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

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

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

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

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 разыгрываются числа покрытий n, b_1, b_2, b_3, b_4 по бета- или нормальному законам распределений в заданном диапазоне $[n_{min}; n_{max}]$. Для

нереференсного сайта j моделируется общее число покрытий n, затем по полученному n генерируется число покрытий для референсного b_R и нереференсного b_{notR} чисел покрытий сайта. Аналогично моделируются покрытия для SNP сайта. Принимается допущение о наличии покрытий не более двух различных нуклеотидных оснований на сайте. Предложенный алгоритм позволяет воспроизводить наборы данных максимально приближенные к экспериментальным условиям, задаваемыми числами покрытий и законами их распределений, количеством полиморфных сайтов. Блок-схема алгоритма моделирования сайтов однонуклеотидного полиморфизма представлена на рисунке.

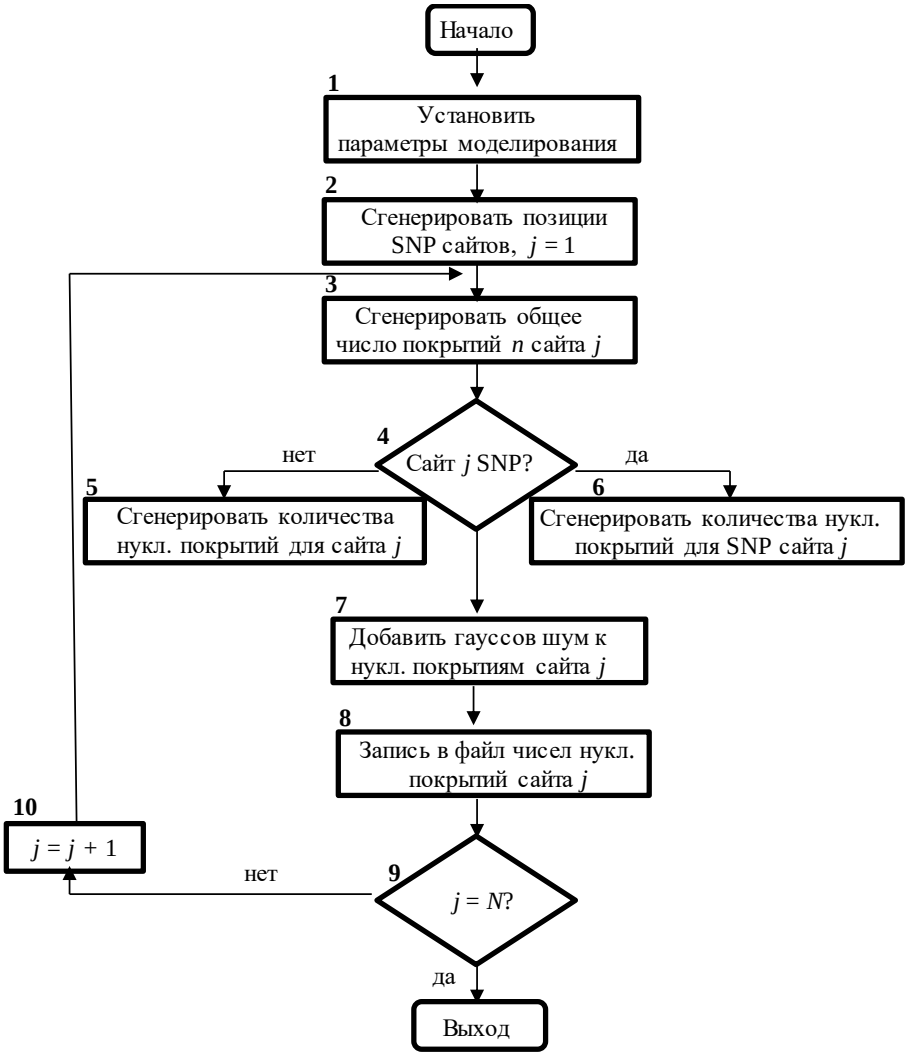

Рис. Блок-схема алгоритма моделирования сайтов однонуклеотидного полиморфизма

2. Алгоритмы машинного обучения

Для применения алгоритмов машинного обучения необходимо сформировать набор признаков нуклеотидных сайтов. Рассмотрены 4 признака: X_1 – число покрытий референсного нуклеотида, X_2 - X_4 – отсортированные в порядке убывания числа покрытий для нереференсных нуклеотидов. Данные нормируются к общему числу покрытий сайта n. Так как в рассматриваемой задаче число признаков невелико (4 признака) и решается задача разделения на два класса (SNP и не SNP сайты), то в

качестве методов машинного обучения предпочтительнее выбрать базовые алгоритмы классификации, такие как на основе деревьев решений – деревьев условного вывода (англ. Conditional Inference Trees – CIT) [11] и классификации и регрессии построением дерева решений (Classification And Regression Tree – CART) [12], опорных векторов с линейной разделяющей функцией (Support Vector Machine – SVM) [13] и экстремального градиентного бустинга (Extreme Gradient Boosting – XGBoost) [14].

3. Организация вычислительного эксперимента

В вычислительном эксперименте модели машинного обучения обучались на отдельно смоделированных наборах данных, а затем проводился сравнительный анализ методов машинного обучения и классических алгоритмов идентификации сайтов SNP на новых сгенерированных наборах данных с различными уровнями добавленного гауссового шума.

Модели CIT (функция *ctree* R-пакета *party*), CART (*rpart* R-пакета *rpart*), SVM (функция *svm* R-пакета *e1071*) и XGBoost (функция *xgboost* R-пакета *xgboost*) были обучены на смоделированных данных по бета-закону распределения без добавления аддитивного гауссового шума на выборке из 40 000 сайтов, из которых 20 000 – SNP.

Для всестороннего исследования алгоритмов идентификации сайтов SNP смоделированы данные с учетом добавления гауссового шума с параметрами $\mu = 0$ и $\sigma_l = q_l \cdot b_l$, l =1-4 (индексы советуют нуклеотидам A, C, G и T),

$$b_l^* = b_l + z \cdot \sigma_l, \qquad (3)$$

где z – реализация стандартизированной нормальной случайной величины, $q_l > 0$. Варьирование параметром σ_l 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. Исследовались две группы наборов данных: 1) значения параметров для референсных и нереференсных каналов сайта q_R и q_{notR} полагались равными и варьировалось от 0 до 0,6 – таким образом воспроизводятся условия возрастающего шума во всех каналах регистрации числа покрытий. 2) параметр $q_R = 0$, а q_{notR} варьировался от 0,5 до 2,0 – моделируются условия возрастающего шума в нереференсном канале регистрации числа покрытий. Смоделированы наборы данных по 20 000 сайтов, в каждом из которых случайно сгенерированы 20 сайтов SNP. Число наборов данных для каждой комбинации параметров – 3.

4. Результаты

Проведено исследование алгоритмов машинного обучения и двух наиболее эффективных классических алгоритмов идентификации сайтов – тестов на основе биномиального распределения и энтропии [4, 6]. Разработана эффективная программная реализация теста биномиального распределения (ТБР), особенностью которой является автоматизация выбора порогового значения при идентификации сайтов SNP. Предложено в качестве порогового значения вероятностей использовать величину 10^{-k}, где k – среднее число покрытий сайтов, оцененное по смоделированной или экспериментальной выборке. В качестве энтропийного теста (ЭТ) используется

опубликованная программная реализация [6]. Пороговые значения при идентификации сайтов SNP: энтропия $E > 0.21$ и p-величина < 0.5. Эффективность алгоритмов оценена с помощью мер точности *Precision*, чувствительности *Recall* и счета F_1, характеризующих свойства алгоритмов не включать ложно положительные события (*Precision*, неверно классифицированные сайты как SNP), включать истинно положительные события (*Recall*, верно классифицированные сайты SNP) и их комбинированного вклада F_1) [15]. Результаты анализа смоделированных наборов данных представлены в таблице.

Точность алгоритмов идентификации сайтов SNP по мере F_1

q_R; q_{notR}	F_1, %					
	ТБР	ЭТ	CIT	CART	SVM	XGBoost
0; 0	91,9 (0,8)	97,6(1,4)	**100 (0)**	**100 (0)**	**100 (0)**	**100 (0)**
0,2; 0,2	91,8 (2,1)	96,0 (0,8)	**99,0 (0,8)**	98,3 (0,9)	98,3 (0,9)	98,3 (0,9)
0,4; 0,4	84,1 (1,8)	82,3 (3,2)	47,4 (0,5)	**88,8 (3,1)**	2,6 (3,0)	44,9 (1,5)
0,6; 0,6	**82,7 (4,5)**	79,3 (2,3)	19,0 (1,1)	81,1 (1,7)	1,6 (1,6)	17,6 (1,3)
0,0; 0,5	87,0 (3,0)	92,2 (1,9)	**92,6 (2,4)**	91,8 (1,7)	92,0 (2,8)	91,8 (1,7)
0,0; 1,5	75,8 (2,9)	77,1 (3,4)	85,6 (2,8)	**88,2 (4,0)**	**88,2 (4,0)**	**88,2 (4,0)**
0,0; 1,5	72,6 (2,5)	74,7 (2,1)	87,0 (3,5)	92,1 (2,7)	**92,3 (1,6)**	92,1 (2,7)
0,0; 2,0	56,7 (0,6)	59,5 (1,6)	72,9 (1,7)	**88,9 (2,0)**	85,6 (2,8)	**88,9 (2,0)**

Примечание. В скобках указана стандартная ошибка среднего.

Набор данных 1. На данных без добавления нормального шума наивысшая точность по мере F_1 (100 %) характерна для методов машинного обучения. Точность ЭТ (97,6 %) выше чем у ТБР (91,9 %). При увеличении шума в данных с q_l = 0,2 до 0,6 точность алгоритмов ТБР, ЭТ и CART понижается до 80–82 %, а для алгоритмов машинного обучения CIT, SVM и XGBoost – до 18 % и ниже. Наименьшая точность, при увеличении шума с 0,4 и выше, наблюдается у алгоритма SVM (1,6–2,6 %).

Набор данных 2. При увеличении шума в нереференсном канале с q_{notR} = 0,5 до 2,0 точность классических алгоритмов значительно снижается до 57-60 %, алгоритмов машинного обучения до 73–89 %. Наименьшая точность среди методов классификации, при увеличении шума с 1,5 и выше, у алгоритма CIT (73 %).

Полученные результаты позволяют сделать вывод о том, что для не зашумленных данных предпочтительнее использовать алгоритмы машинного обучения. При равномерном зашумлении данных в каналах сайта – классические алгоритмы и CART, при зашумлении нереференсных каналов – алгоритмы машинного обучения. Невысокая точность классификации для алгоритма CIT объясняется ухудшением статистических свойств рассматриваемых выборок, что критично для статистических алгоритмов.

5. Заключение

Для идентификации сайтов однонуклеотидного генетического полиморфизма предложено использовать методы машинного обучения, обученные на смоделированных данных. Разработан алгоритм имитационного моделирования SNP сайтов в последовательности молекулы ДНК по экспериментальным наборам данных, осно-

ванный на генерации случайных событий по бета- или нормальному законам распределений, параметры которых оцениваются по экспериментальным данным. Выполнен сравнительный анализ наиболее эффективных методов машинного обучения и классических алгоритмов идентификации сайтов SNP. На примерах данных без добавления шума наилучшими методами являются машинного обучения – CIT, CART, SVM и XGBoost, с увеличением шума – ТБР, ЭТ и CART. При добавлении шума в нерефернсных каналах наилучшими методами являются машинного обучения – CART, SVM и XGBoost. Проведенное исследование позволяет сделать вывод об успешном использовании имитационного моделирования при обучении предсказательных моделей машинного обучения с целью идентификации сайтов SNP. Наиболее оптимальным методом идентификации однонуклеотидного генетического полиморфизма, при различных уровнях экспериментального шума в смоделированных наборах, данных является алгоритм CART.

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. Methodologies and Algorithms / A. Masoudi-Nejad, Z. Narimani, N. Hosseinkhan, 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. Обнаружение сайтов однонуклеотидного генетического полиморфизма на основе энтропии / Н. Н. Яцков [и др.] // Прикладные проблемы оптики, информатики, радиофизики и физики конденсированного состояния : материалы седьмой Междунар. науч.-практ. конф., 18–19 мая 2023 г., Минск. – Минск: Ин-т прикл. физ. проблем им. А. Н. Севченко БГУ, 2023. – С. 191–193.
7. Сравнительный анализ алгоритмов обнаружения сайтов однонуклеотидных вариаций / Я. В. Шинкевич [и др.] // Информационные системы и технологии = Information Systems and Technologies [Электронный ресурс] : материалы междунар. науч. конгресса по информатике. Ч. 2, Респ. Беларусь, Минск, 27–28 окт. 2022 г. / Белорус. гос. ун-т ; редкол.: С. В. Абламейко (гл. ред.) [и др.]. – Минск : БГУ, 2022. С. 61–66.
8. *Su Z.* HAPGEN2: simulation of multiple disease SNPs / Z. Su, J. Marchini, P. Donnelly // Bioinformatics, 2011. Vol. 27(16), P. 2304-2305.
9. Machine learning as an effective method for identifying true single nucleotide polymorphisms in polyploid plants / W. Korani [et al.] // Plant Genome. 2019. Vol. 12(1).
10. DHOEM: a statistical simulation software for simulating new markers in real SNP marker data / L. Jacquin [et al.] // BMC Bioinformatics. 2015. Vol. 16:404.
11. *Hothorn T.* Unbiased Recursive Partitioning: A Conditional Inference Framework / T. Hothorn, K. Hornik, A. Zeileis // Journal of Computational and Graphical Statistics, 2006 Vol. 15(3). P. 651–674.
12. Classification and Regression Trees./ L. Breiman [et al.]. 1st ed. Wadsworth, 1984.
13. *Vapnik V. N.* The Nature of Statistical Leaning Theory. 2nd ed., New York : Springer-Verlag, 2000.
14. *Hastie T.* The Elements of Statistical Learning. Data Mining, Inference, and Prediction. / T. Hastie, R. Tibshirani, J. Friedman J. 2nd ed. New York : Springer, 2009.
15. *Murphy K. P.* Probabilistic Machine Learning, London : The MIT Press, 2022.