Исследование информативности признаков нуклеотидных сайтов при обнаружении однонуклеотидных генетических полиморфизмов

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

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

В работе представлены результаты исследования набора признаков нуклеотидных сайтов при определении генетических полиморфизмов с использованием методов машинного обучения. Исследована информативность признаков в применении к экспериментальным данным геномного секвенирования хромосомы 22 генома человека с добавлением гауссовского шума. Наиболее информативными характеристиками нуклеотидных сайтов являются признаки на основе чисел покрытий референсного нуклеотида и первого по убыванию нереференсного нуклеотида, а также *p*-величина теста биномиального распределения.

Ключевые слова: однонуклеотидный генетический полиморфизм, информативность признаков нуклеотидных сайтов, машинное обучение.

Study of the informativeness of nucleotide site features in the identification of single nucleotide genetic polymorphisms

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 nucleotide site features while identifying genetic polymorphisms using machine learning methods. The informativeness of the features was examined using genomic experimental sequencing data on chromosome 22 of the human genome with the addition of Gaussian noise. The most informative nucleotide site characteristics are features based on the numbers of coverage of the reference nucleotide and the first-in-decreasing non-reference nucleotide, as well as the p-value of the binomial distribution test.

Keywords: single nucleotide genetic polymorphism, informativeness of nucleotide site features, machine learning.

Введение

Методы флуоресцентной спектроскопии, применяемые в полнотранскриптомном секвенировании, позволяют точно установить нуклеотидные последовательности молекул ДНК [1, 2]. В живых организмах различия составов геномов обусловлены генетическими полиморфизмами [3, 4]. Важной задачей является определение сайтов однонуклеотидного генетического полиморфизма (SNP, от англ. single nucleotide polymorphism). Существующие статистические методы идентификации однонуклеотидных полиморфизмов требуют значительных вычислительных ресурсов и сложно применимы при анализе экспериментальных данных с высоким уровнем шума [5]. Применение методов машинного обучения позволяет повысить точность

определения сайтов SNP при увеличении шума в экспериментальных данных геномного секвенирования [6, 7]. При этом важной задачей является исследование информативности характеристик, или признаков, нуклеотидных сайтов при идентификации сайтов SNP с использованием методов машинного обучения [8]. В работе [8] выделены основные признаки нуклеотидных сайтов, информативность которых проанализирована на смоделированных данных. Логическим продолжением работы является исследование информативности признаков нуклеотидных сайтов на экспериментальных данных генома человека, в том числе при увлечении уровня экспериментального шума.

Целью работы является исследование информативности признаков нуклеотидных сайтов при определении генетических полиморфизмов по экспериментальным данным геномного секвенирования с использованием методов машинного обучения. Исследование информативности признаков нуклеотидных сайтов выполнено на данных секвенирования генома человека.

1. Признаки нуклеотидных сайтов

Рассмотрены данные геномного секвенирования нового поколения [4, 9], представленные числами покрытий нуклеотидных сайтов [7, 8]. Полный набор 23 признаков нуклеотидных сайтов включает такие характеристики как количества покрытий нуклеотидов, результаты статистического анализа покрытий, наиболее информативные атрибуты биологических аннотаций [3, 8]. Данный набор признаков может использоваться в общем случае, по результатам проведенного секвенирования некоторой аннотированной нуклеотидной последовательности.

Для применения в методах машинного обучения достаточно ограничиться набором из 14 признаков, таких как: X_1 – число прочтений референсного нуклеотида; $X_2 – X_4$ – отсортированные в порядке убывания числа покрытий нереференсных нуклеотидов; X_5 – энтропия нуклеотидного сайта; X_6 – p-величина теста о значимости энтропии сайта; X_7 – p-величина теста биномиального распределения; X_8 – p-величина точного теста Фишера; X_9 – p-величина теста Пуассона; X_{10} – логарифм вероятности ошибки в прочтениях; X_{11} – дисперсия в позициях прочтений, где встречается вариантный (нереференсный) аллель; X_{12} – средняя позиция в прочте-ниях, где встречается вариантный (нереференсный) аллель; X_{13} – индикатор, указывающий совпадает ли нуклеотид с максимальным количеством покрытий с референсным основанием; X_{14} – индикатор, указывающий на числа почтений для референсного и нереференсных нуклеотидов.

2. Методика отбора информативных признаков

Предложена методика отбора информативных признаков, исключающая малоинформативные атрибуты при выполнении набора последовательных условий: значение коэффициента корреляции Пирсона между парами признаков более 0,85; мера характеристик с высокой мультиколлинеарностью *VIF* (Variance Inflation Factor) [10] выше 5; оценка меры взаимной информативности между признаками относительно

целевой переменной менее 0,1; L_1 регуляризация устанавливает коэффициент значимости 0; значимость признака в модели случайного леса, по мере критерия прироста информации на основе индекс Джинни, менее 0,1 от значимости наиболее информативного атрибута; учет признака ведет к приросту точности классификации не более чем на 0,005 [11, 12].

Данная методика позволяет отобрать независимые и информативные признаки при идентификации сайтов SNP с использованием методов машинного обучения.

3. Результаты

Задачей проведения вычислительного эксперимента является исследование информативности признаков и оценка времени работы алгоритмов генерации признаков. В вычислительном эксперименте рассмотрены эталонные данные о хромосоме 22, полученные консорциумом GIAB [13]. Набор данных содержит характеристики 29 633 768 нуклеотидных сайтов, из которых 36 150 истинно идентифицированные сайты SNP.

Разработаны и программно реализованы на языках программирования R и C++ алгоритмы генерации и оценки информативности 14 признаков нуклеотидных сайтов при решении задачи определения сайтов однонуклеотидных полиморфизмов с использованием методов машинного обучения на экспериментальных данных секвенирования хромосомы 22 генома человека. На рис. 1 представлены гистограммы распределений исследуемых признаков и аппроксимирующие их функции.

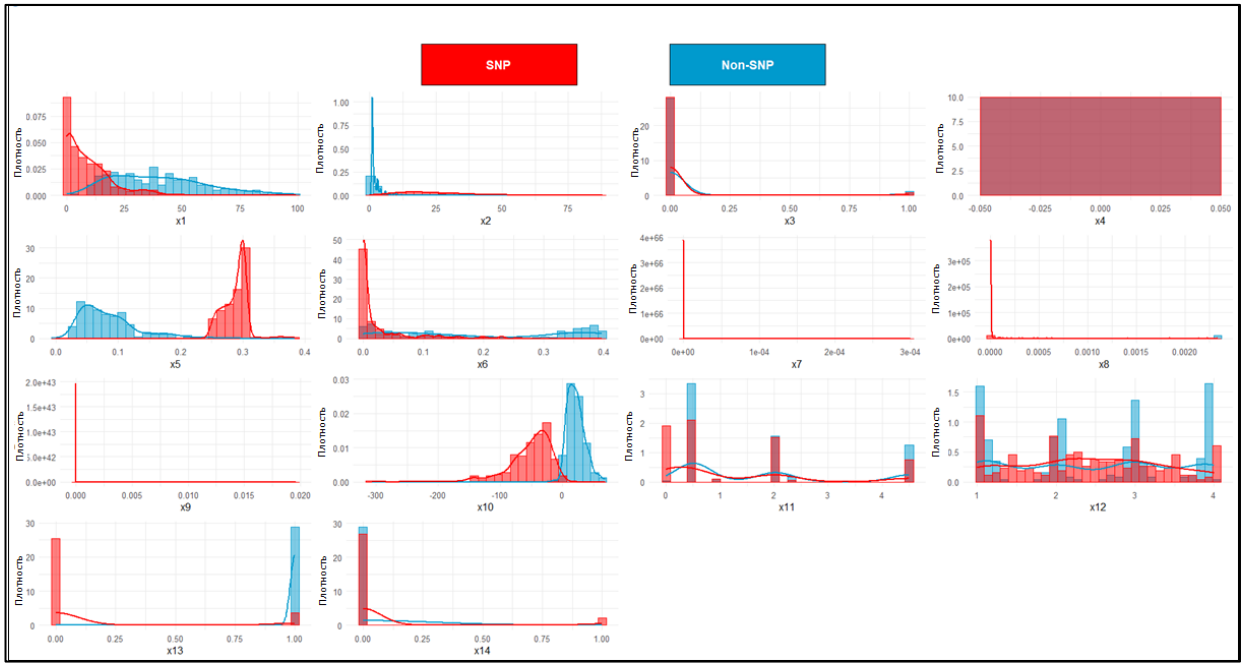

Рис. 1. Гистограммы 14 признаков и аппроксимирующие их функции для SNP и не SNP сайтов

В соответствии с разработанной методикой отбора, как сильно коррелированные, исключены признаки X_5, X_6, X_9, и X_{10}, по мере взаимной информативности – X_3,

X_4 и X_{14}, по L_1 регуляризации – X_{13}, по критерию важности в методе случайного леса – X_{12}, в результате выполнения алгоритма обратного отбора признаков с 5-ти кратной перекрёстной проверкой – X_1, X_2, X_8 и X_{11}. В результате отобран признак X_7 – p-величина теста биномиального распределения.

Важным свойством реализованных алгоритмов является их шумоустойчивость [5, 14]. Проведен вычислительный эксперимент с добавлением гауссового шума. Для исследования методики отбора признаков на зашумленном наборе данных в каждый нуклеотидный канал добавлен небольшой аддитивный гауссовский шум с параметрами $m = 0$ и $\sigma = 5$. В результате применения разработанной методики отбора информативных признаков выделены 2 признака – числа покрытий референсного нуклеотида X_1 и первого по убыванию нереференсного нуклеотида X_2. Диаграмма расположения нуклеотидных сайтов в пространке признаков X_1 и X_2 представлена на рис. 2.

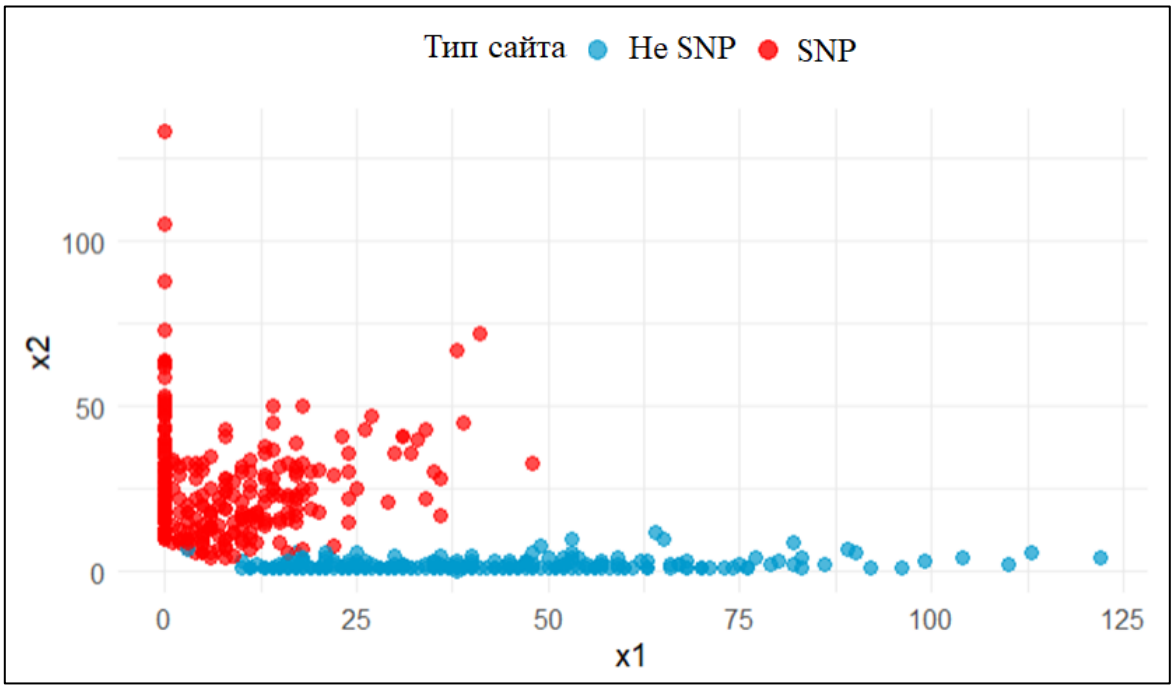

Рис. 2. Диаграмма расположения нуклеотидных сайтов в пространке признаков X_1 и X_2

Одним из критериев оценки эффективности разработанных алгоритмов является время их работы. Вычисления проведены на ПК, оснащенным 12-ядерным процессором Ryzen 9 5900X. Реализованы распараллеленные варианты вычислительных процедур – с использованием R пакетов parallel, Rcpp и RcppParallel. Преимущество распараллеленной RcppParallel-реализации над распараллеленной R-реализацией составляет порядка 10 раз. Время работы алгоритмов генерации признаков сопоставимо с временем работы классических статистических методов идентификации нуклеотидных сайтов.

4. Заключение

Разработаны и программно реализованы алгоритмы генерации 14 признаков нуклеотидных сайтов. Проведено исследование информативности признаков нуклеотидных сайтов при определении генетических полиморфизмов с использованием методов машинного обучения на экспериментальных данных секвенирования хромосомы 22 генома человека с учетом добавления гауссовского шума. Определены наиболее информативные признаки – числа покрытий референсного нуклеотида и первого по убыванию нереференсного нуклеотида, *p*-величина теста биномиального распределения.

Программные реализации алгоритмов выделения признаков нуклеотидных сайтов распараллелены. Преимущество распараллеленной RcppParallel-реализации над распараллеленной R-реализацией составляет порядка 10 раз.

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. *Sung W.-K.* Algorithms for Next-Generation sequencing. 1st ed. Chapman & Hall/CRC, 2017.
4. *Kappelmann-Fenzl M., ed.* Next Generation Sequencing and Data Analysis. Cham : Springer, 2021.
5. 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.
6. 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.
7. *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.
8. *Сарнацкий Д. Д.* Исследование информативности признаков нуклеотидных сайтов при определении генетических полиморфизмов с использованием методов машинного обучения / Д. Д. Сарнацкий, Н. Н. Яцков, В. В. Гринев // Информационные технологии и системы 2024 (ITS 2024): материалы международной научной конференции, Минск, 20 ноября 2024 г. – Минск : БГУИР, 2024. С. 69-70.
9. *Masoudi-Nejad A.* Next Generation Sequencing and Sequence Assembly. Methodologies and Algorithms / A. Masoudi-Nejad, Z. Narimani, N. Hosseinkhan. – New York : Springer, 2013.
10. *Marquardt D. W.* Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation // Technometrics.1970. Vol. 12(3). P. 591–612.
11. *Hastie T.* The Elements of Statistical Learning. Data Mining, Inference, and Prediction. 2nd ed. / T. Hastie, R. Tibshirani, J. Friedman. – New York : Springer, 2009.
12. *Murphy K. P.* Probabilistic Machine Learning, London : The MIT Press, 2022.
13. An open resource for accurately benchmarking small variant and reference calls / J. M. Zook [et al] // Nature Biotechnology. 2019. Vol. 37(5). P. 561–566.
14. *Яцков Н. Н.* Программный комплекс для имитационного моделирования сайтов однонуклеотидного генетического полиморфизма / Н. Н. Яцков, [и др.] // Информатика. 2025. Т. 22(2). С. 81–94.