ИССЛЕДОВАНИЕ ИНФОРМАТИВНОСТИ ПРИЗНАКОВ НУКЛЕОТИДНЫХ САЙТОВ ПРИ ОПРЕДЕЛЕНИИ ГЕНЕТИЧЕСКИХ ПОЛИМОРФИЗМОВ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Сарнацкий Д. Д., Яцков Н. Н., Гринев В. В.
Кафедра системного анализа и компьютерного моделирования, кафедра генетики,
Белорусский государственный университет
Минск, Республика Беларусь
E-mail: denisiussarnatski@gmail.com, {Yatskou, Grinev_vv}@bsu.by

В работе рассмотрены способы генерации признаков нуклеотидных сайтов при определении генетических полиморфизмов с использованием методов машинного обучения. Исследована информативность признаков в применении к смоделированным данным геномного секвенирования с добавлением гауссовского шума. Наиболее информативными и независимыми характеристиками нуклеотидных сайтов являются признаки на основе логарифма вероятности ошибки в ридах, p-величине теста Пуассона, чисел покрытий референсного и первого нереференсного каналов.

Введение

Однонуклеотидные полиморфизмы (SNP – от англ. Single Nucleotide Polymorphism) являются одним из наиболее распространенных типов генетических вариаций в геноме человека. Существующие статистические методы идентификации однонуклеотидных полиморфизмов [1] требуют значительных вычислительных ресурсов и сложно применимы при анализе экспериментальных данных с высоким уровнем шума [2].

Применение методов имитационного моделирования и машинного обучения позволяет повысить точность определения сайтов SNP при увеличении шума в экспериментальных данных геномного секвенирования. Логическим продолжением работы является повышение степени детализации моделируемых процессов в имитационной модели [3] и исследование информативности харкетристик или признаков нуклеотидных сайтов при решении задачи определения сайтов SNP с использованием методов машинного обучения.

Целью данной работы является выделение и исследвоание эффективности признаков нуклеотидных сайтов при определении генетических полиморфизмов с использованием методов машинного обучения. Расммотрены 23 признака, характеризующие данные геномного секвенирования нового поколения [1], представленные числом покрытий нуклеотидных сайтов. Исследование информативности признаков нуклеотидных сайтов выполнено на примере анализа геномных данных человека с использованием методов машинного обучения.

I. Признаки нуклеотидных сайтов

В таблице 1 представлен фрагмент набора экспериментальных данных , содержащий числа покрытий нуклеотидных сайтов (колонки 3–7) и тип референсного нуклеотида (колонка 2).

Таблица 1 – Фрагмент набора данных секвенирования хромосомы 22 человека

	Расположение	Референсное значение	A	C	G	T
0	chr22:16050343	T	0	2	0	25
1	chr22:42771714	A	25	0	0	0
2	chr22:42771715	T	0	21	0	3

Выделены 23 статистических признака, включающие исходные числа покрытий, результаты ключевых классических критериев, эмпирические признаки сайтов SNP и др.: X_1 – нормированное число покрытий референсного нуклеотида; $X_2 - X_4$ – отсортированные в порядке убывания нормированные числа покрытий для нереференсных нуклеотидов; X_5 – энтропия сайта [1]; X_6 – p-величина энтропии сайта [1]; X_7 – p-величина теста биномиального распределения [1]; X_8 – p-величина точного теста Фишера [1]; X_9 – p-величина точного теста Пуассона [1]; X_{10} – логарифм вероятности ошибки в ридах [4]; X_{11} – вариация позиции чтения [4]; X_{12} – средняя позиция чтения [4]; X_{13} – бинарный признак, указывающий совпадает ли нуклеотид с максимальным кол-вом покрытий с референсным (эмпирический признак); X_{14} – бинарный признак, указывающий равняется ли приблизительно кол-во ридов для референсного и нереференсных нуклеотидов (эмпирический признак); X_{15} – баланс аллелей [4]; X_{16} – качество рядом расположенных нуклеотидов [4]; X_{17} – кол-во повторений динуклеотидов [4]; X_{18} – средняя позиция чтения [4]; X_{19} – направленность стренда [4]; X_{20} – суммарное расхождение площадей [4]; X_{21} – разнообразие нуклеотидов [4]; X_{22} – кол-во несоответствий ридов [4]; X_{23} – кол-во последовательных повторов одного нуклеотида [4].

Признаки $X_1 - X_{14}$ могут быть выделены из исходных частотных таблиц вида таблицы 1. Назовем их дифференциальными признаками, их удобно применять для методов машинного обучения. Для выделения признаков $X_{15} - X_{23}$ нужна дополнительная аннотирующая информация [5] и выделяются они с учетом состояния смежных сайтов. Назовем их интегральными признаками. Полный набор признаков имеет смысл применять в основном для обучения нейросетей.

II. Описание вычислительного эксперимента

Задачей проведения вычислительного эксперимента является исследование: i) информативности признаков; ii) точности методов машинного обучения, обученнных на выделенных признаках; iii) времени работы алгоритмов.

Информативность признаков оценивается по методу случайного леса с критерием прироста информативности (критерием расщепления узлов) индекс Джинни [6]:

$$Gini(Q) = 1 - \sum_{i=1}^{n} p_i^2$$

где Q — результирующее множество, n — число классов в нем, p_i — вероятность i-го класса.

III. Результаты

Рассмотрены дифференциальные признаки $X_1 - X_{14}$ для решения задачи определения сайтов однонуклеотидных полиморфизмов с использованием методов машинного обучения.

На рисунке 1 представлена столбчатая диаграмма важности признаков, оцененных на смоделированных данных с добавлением стандартного гауссовского шума.

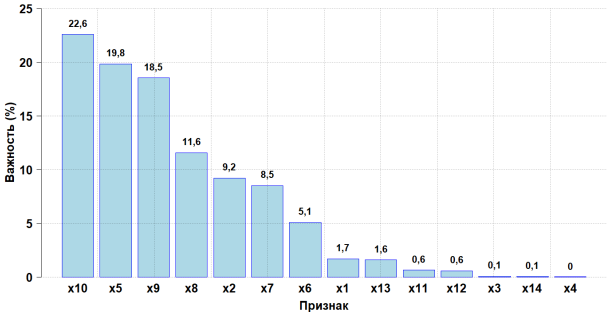

Рис. 1 – Оцененная важность признаков по методу случайного леса с критерием прироста информативности индекс Джинни

Наиболее информативными признаками являются: X_{10} – логарифм вероятности ошибки в ридах; X_9 – p-величина теста Пуассона; X_5 – энтропия сайта. Однако здесь не учтена корреляция между признаками, которая может достигать порядка 60%.

После сравнения ансамбля моделей случайного леса, обученных на разных комбинациях признаков, наибольшую точность в 98,26% имеет модель, обученная на следующих 4-х признаках: X_{10} – логарифм вероятности ошибки в ридах; X_9 – p-величина теста Пуассона; X_1 – относительное число покрытий референсного нуклеотида; X_2 – относительное число покрытий 1-го нереференсного канала.

Они являются уравновешенными и логичными: присутствует вероятность ошибки в ридах, результат одного из статистических тестов, а также числа покрытий в референсном и первом нереференсном каналах.

Сравнение времени выделения признаков на 1000 исходных сайтов выполнено на персональном компьютере, обладающим 12 ядерным процессором AMD Ryzen 5900X (3.7 GHz). В однопоточной реализации на языке R усредненное время выделения простейших признаков $X_1 - X_4$ занимает 0,39 с, отобранных признаков X_1,X_2,X_9,X_{10} – 1,6 с, полного набора признаков $X_1 - X_{14}$ – 38 с.

IV. Заключение

Для решения задачи определения сайтов однонуклеотидного полиморфизма с использованием методов интеллектуального анализа данных в общем случае рассмотрены 23 признака. Они разделены на дифференциальные, используемые для методов машинного обучения и интегральные, применимые для методов с учетом порядка структуры генома.

Определены наиболее информативные и независимые признаки для задачи идентификации сайтов SNP с использованием методов машинного обучения. Это признаки X_{10} – логарифм вероятности ошибки в ридах; X_9 - p–величина теста Пуассона; X_1 – относительное число покрытий референсного нуклеотида; X_2 – относительное число покрытий 1-го нереференсного канала.

Недостатком предложенных признаков является существенное возрастание времени их выделения. Время работы алгоритма с новыми признаками возрастает примерно в 4 раза.

1. Sung, W.-K. Algorithms for next-generation sequencing / Wing-Kin Sung // Chapman & Hall/CRC Comput Biol Series. – 2017. – P. 175-185.
2. Oh, J. H. SITDEM: A simulation tool for disease/endpoint models of association studies based on single nucleotide polymorphism genotypes / J. H. Oh, J. O. Deasy // Comput Biol Med, Volume 45. – 2014. – P. 136-142.
3. Имитационная модель генерации сайтов однонуклеотидного полиморфизма в молекулах ДНК человека / Д. Д. Сарнацкий [и др.] // CTDA'24. – 2024. – Материалы. – С. 265-268.
4. A support vector machine for identification of single-nucleotide polymorphisms from next-generation sequencing data / B. D. O'Fallon [et al.] // Bioinformatics, Volume 29. – 2013. – P. 1361–1366.
5. Sequence Alignment/Map Format Specification [Electronic recourse] / – Mode of access: `https://samtools.github.io/hts-specs/SAMv1.pdf`. – Date of access: 24.10.2024.
6. О.В. Классические методы машинного обучения / А.В. Кугаевских [и др.] // Университет ИТМО. – 2022. – С. 36-40, 42-45.