УДК 001.53:[575.112::575.224.22]

СРАВНИТЕЛЬНЫЙ АНАЛИЗ АЛГОРИТМОВ ОБНАРУЖЕНИЯ САЙТОВ ОДНОНУКЛЕОТИДНЫХ ВАРИАЦИЙ

Я.В. Шинкевич[1], Н.Н. Яцков[2], И.С. Трусов[1], И.Н. Ильюшенок[1], В.В. Скакун[2], В.В. Гринев[1]

[1]Белорусский государственный университет, Кафедра генетики, пр. Независимости, д. 4, 220050, г. Минск, Беларусь, grinev_vv@bsu.by
[2]Белорусский государственный университет, Кафедра системного анализа и компьютерного моделирования, пр. Независимости, д. 4, 220050, г. Минск, Беларусь

В работе апробировано применение нейронной сети по типу многослойного персептрона для обнаружения сайтов однонуклеотидного полиморфизма у человека. Использованная нейронная сеть сравнима по результативности с точным тестом Фишера и превосходит биномиальный тест отношения правдоподобия по точности.

***Ключевые слова*:** геномное секвенирование; однонуклеотидный полиморфизм; алгоритмы идентификации.

COMPARATIVE ANALYSIS OF ALGORITHMS FOR DETECTION OF SINGLE NUCLEOTIDE VARIATIONS

Y.V. Shynkevich[1], M.M. Yatskou[2], I.S. Trusau[1], I.M. Ilyushonak[1], V.V. Skakun[2], V.V. Grinev[1]

[1]Department of Genetics, Belarusian State University, Nezavisimosti Avenue-4, 220050, Minsk, Belarus, grinev_vv@bsu.by
[2]Department of System Analysis and Computer Modeling, Belarusian State University, Nezavisimosti Avenue-4, 220050, Minsk, Belarus.
Corresponding authors: grinev_vv@bsu.by

In this work, we tested the applicability of perceptron-like neural network to detect sites of single nucleotide polymorphism in human genome. The used neural network is comparable in performance to Fisher's exact test and outperforms the binomial likelihood ratio test in precision.

Keywords: genomic sequencing; single nucleotide polymorphism; calling algorithms.

Введение

Геном человека содержит множество изменчивых (полиморфных) однонуклеотидных сайтов, варианты которых встречаются в популяциях людей с частотой более 1% и именуются сайтами однонуклеотидных вариаций (SNVs, от англ. single nucleotide variations). Сюда также следует добавить точечные мутации, встречающиеся с частотой менее 1%, и мы получим очень высокую изменчивость нашего генома, влияющую на

транскрипцию генов, процессинг их первичных РНК и структуру кодируемых белков.

Поскольку для многих полиморфных сайтов нашего генома обнаружены ассоциации с фенотипом (физическими и умственными способностями человека, предрасположенностью к заболеваниям, характером ответа на лекарства, продолжительностью жизни и т.д.), то в настоящее время особое внимание уделяется идентификации таких сайтов в индивидуальных геномах по данным полногеномного, экзомного или прицельного секвенирования. Такого рода задача имеет ряд классических решений, основанных, например, на использовании точного теста Фишера. Однако, исходя из структуры геномных данных, вполне возможно применение и нетривиальных подходов, в частности, нейронных сетей, хотя такие подходы и не фигурируют в научной литературе. В связи с этим мы нацелились на проверку потенциала нейронных сетей в обнаружении SNVs и сравнение такого подхода с несколькими наиболее мощными классическими алгоритмами.

1. Методология исследования

В работе были использованы эталонные данные, полученные консорциумом GIAB [1]. Выбор именно этих данных был обусловлен тем, что на сегодняшний день это наиболее надежные бенчмарк-данные для решения любых задач, связанных с изучением геномного полиморфизма у человека (от разработки новых инструментальных методов «мокрой» биологии до сравнения алгоритмов обнаружения полиморфных сайтов).

Для достижения своей цели мы воспользовались результатами секвенирования генома HG001 – одного из семи геномов GIAB. В общей сложности 1268026804 ридов, полученных при секвенировании этого генома, были картированы относительно эталонной сборки GRCh38 генома человека и собраны в BAM файл объемом 101,42 Гбайт. Кроме того, для генома HG001 мы получили эталонные наборы SNVs, сгенерированные с помощью двух подходов (10X Genomics и benchmark) [1]. Эти наборы были использованы нами для оценки чувствительности и специфичности наших алгоритмов.

2. Результаты и их обсуждение

В данном исследовании мы провели сравнение трех алгоритмов идентификации сайтов SNVs по данным геномного секвенирования: точный тест Фишера, биномиальный тест и нейронная сеть. Сравнение проводилось по чувствительности (recall или sensitivity) и точности (precision), которые рассчитывались по следующим формулам:

$$\text{recall} = \text{SNVs}^{(\text{empir.true})}/\text{SNVs}^{(\text{ref})}, \quad (1)$$

$$\text{precision} = \text{SNVs}^{(\text{empir.true})}/\text{SNVs}^{(\text{empir.all})}, \quad (2)$$

где $\text{SNVs}^{(\text{ref})}$ – это количество сайтов SNVs в эталонном наборе,
$\text{SNVs}^{(\text{empir.true})}$ – это количество истинных сайтов SNVs, обнаруженных тестируемым алгоритмом,
$\text{SNVs}^{(\text{empir.all})}$ – это количество всех сайтов SNVs (как истинных, так и ложных), обнаруженных тестируемым алгоритмом.

Точный тест Фишера, как известно, работает с таблицами сопряженности признаков. Построение таких таблиц для идентификации сайтов SNVs имеет свои особенности. Первая пара сопряженных значений получается путем подсчета количества ридов, подтверждающих полиморфизм, и общего количества ридов, покрывающих данный сайт. Вторая же пара сопряженных значений отражает локальный шум, возникающий при секвенировании, главными источниками которого являются инструментальные ошибки самого секвенирования и ошибки картирования. Она получается путем подсчета количества нуклеотидов, не совпадающих с эталонной последовательностью, и количества нуклеотидов, полностью тождественных эталону, среди ридов, картированных в геномном окне размером 100 нуклеотидов.

Эталонный набор	Чувствительность	Точность
10X Genomics	0,762	0,888
benchmark	0,896	0,796

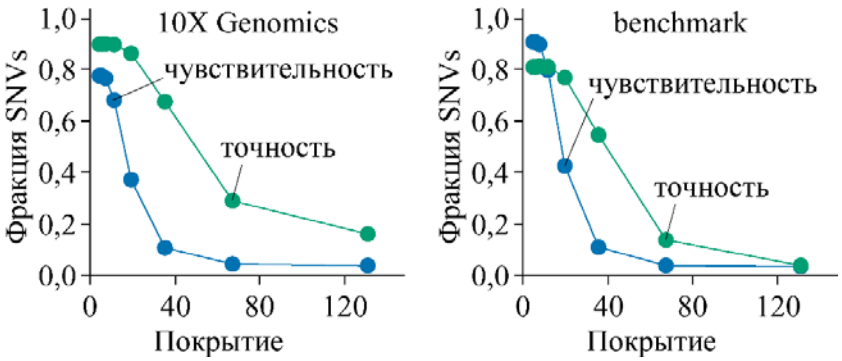

Рисунок 1 – Идентификация SNVs с помощью точного теста Фишера

В своих исследованиях мы воспользовались функцией exactSNP() из пакета Rsubread [2], где имплементирован быстрый алгоритм поиска сайтов SNVs с помощью точного теста Фишера. Как видно на рисунке 1, минимальная чувствительность этого алгоритма равна 76,2% (при сравнении с эталонным набором 10X Genomics), а минимальная точность – 79,6% (с

эталонным набором benchmark).

Однако очевидна также зависимость результатов этого теста от глубины секвенирования (покрытия эталонной последовательности ридами): чем выше порог покрытия, тем ниже чувствительность и точность. И если падение чувствительности является ожидаемым, то падение точности, по-видимому, обусловлено искусственным изменением локального шума из-за фильтрации геномных данных.

Вторым алгоритмом был биномиальный тест отношения правдоподобия. Для постановки этого теста мы воспользовались функционалом двух R-пакетов VariantTools [3] и gmapR [4]. Результаты, полученные нами с помощью биномиального теста для двух эталонных наборов данных, неоднозначны. Как видно из таблицы 1, чувствительность этого теста может достигать почти 98%, что является самым высоким значением среди всех изученных алгоритмов. Однако специфичность данного теста не превышает 30%, что является самым низким показателем среди всех алгоритмов. При этом следует отметить, что применение многокомпонентных правил фильтрации (частота ошибок секвенирования, глубина покрытия ридами, частота альтернативной аллели и т.д.) как исходных данных, так и обнаруженных первичных сайтов SNVs не улучшают ситуацию, а наоборот, только снижают целевые показатели.

Наконец, для идентификации сайтов SNVs нами также был апробирован подход, основанный на нейронных сетях. Мы использовали сети по типу многослойного персептрона с тремя последовательными слоями (входной, скрытый, выходной) размерностями 9, 300 и 100, соответственно, между которыми находились слои Batch Normalization, а функцией активации выходного слоя была функция sigmoid.

Таблица 1 – Результативность применения биномиального теста для идентификации сайтов SNVs.

Эталонный набор	Чувствительность	Точность
10X Genomics	0,874	0,296
benchmark	0,979	0,262

Исходные данные, представленные матрицей частот нуклеотидов по каждой из позиций эталонного генома, были разделены на обучающую и тестовую выборку в соотношении 80 к 20, соответственно, и трансформи-

рованы согласно специфике подачи данных в нейросеть. Кроме того, так как исходные данные имеют сильный дисбаланс классов (с явным преобладанием сайтов, не являющихся SNVs), то к обучающей выборке был применен upsampling, а в функции ошибки был использован эквивалентный пенализирующий коэффициент weight decay.

Как видно из таблицы 2, полученной для эталонного набора 10X Genomics, с повышением количества сайтов SNVs в обучающей выборке модель работает все с большей точностью, но падает чувствительность. Компромисными вариантами являются коэффициенты upsampling и weight decay, равные 0,05 и 0,1: чувствительность в этом случае достигает 92%, а точность 83%. Аналогичные результаты были получены и для эталонного набора benchmark.

Таблица 2 – Результативность применения нейронной сети для идентификации сайтов SNVs.

Пенализирующий коэффициент	Чувствительность	Точность
0,01	0,95	0,72
0,05	0,92	0,81
0,10	0,88	0,83
0,15	0,84	0,84
0,20	0,74	0,85
0,30	0,59	0,86

Таким образом, все три алгоритма обнаружения SNVs, прошедшие испытание в представленном исследовании, дают схожие результаты по чувствительности, но различаются по точности с минимумом для биномиального теста. Существенно, что использованная нами нейронная сеть сравнима по результативности с точным тестом Фишера и превосходит биномиальный тест отношения правдоподобия по точности. Следовательно, наши результаты открывают перспективу для дальнейшего развития алгоритмов анализа геномного полиморфизма человека с использованием нейронных сетей, в частности, повышения их чувствительности и точности.

Библиографические ссылки

1. Zook J.M., McDaniel J., Olson N.D., Wagner J., Parikh H., Heaton H., Irvine S.A., Trigg L., Truty R., McLean C.Y., De La Vega F.M., Xiao C., Sherry S., Salit M. An open resource for accurately benchmarking small variant and reference calls // Nature Biotechnology. 2019. № 37(5). P. 561–566. DOI: 10.1038/s41587-019-0074-6.
2. Liao Y., Smyth G.K., Shi W. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads // Nucleic Acids Research. 2019. № 47(8). P. e47. DOI: 10.1093/nar/gkz114.
3. Lawrence M., Degenhardt J., Gentleman R. VariantTools: Tools for exploratory analysis of variant calls. R package version 1.38.0.
4. Barr C., Wu T., Lawrence M. gmapR: An R interface to the GMAP/GSNAP/GSTRUCT suite. R package version 1.38.0.