Программный пакет SNPSimulatoR для моделирования сайтов однонуклеотидного генетического полиморфизма

Н. Н. Яцков, Е. В. Смолякова, В. В. Скакун, В. В. Гринев

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

Разработан R-пакет SNPSimulatoR для моделирования сайтов однонуклеотидного генетического полиморфизма в молекулах ДНК человека. Включает программные средства обработки экспериментальных данных, имитационного моделирования и идентификации сайтов нуклеотидных полиморфизмов с использованием как наиболее эффективных классических алгоритмов, так и машинного обучения, обученных на смоделированных данных. Работоспособность разработанных программных средств подтверждена в ходе сравнительного анализа алгоритмов на примерах экспериментальных данных геномного секвенирования.

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

R-package SNPSimulatoR for modelling single nucleotide genetic polymorphism sites

M. M. Yatskou, E. V. Smolyakova, V. V. Skakun, V. V. Grinev

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

An R-package SNPSimulatoR has been developed for modelling single nucleotide genetic polymorphism sites in human DNA molecules. Includes programming tools for processing experimental data, simulation modelling and identification of nucleotide polymorphism sites using both the most effective classical and machine learning algorithms trained on simulated data. The performance of the developed package was confirmed in a comparative analysis of the algorithms using examples of genomic sequencing experimental data.

Keywords: single nucleotide genetic polymorphism/; R-package; simulation modelling; machine learning.

Введение

Методы флуоресцентной спектроскопии, применяемые в полнотранскриптомном секвенировании, позволяют точно установить нуклеотидные последовательности молекул ДНК [1–3]. В живых организмах состав геномов варьируется. Различия обусловлены генетическими полиморфизмами или вариациями генома. Важной задачей является определение сайтов однонуклеотидного генетического полиморфизма (SNP, от англ. single nucleotide polymorphism) [4, 5]. Имитационное моделирование является стандартным способом оценки методов статистической обработки в ходе анализа данных геномного секвенирования, а также используется при обучении методов классификации с целью прямой идентификации сайтов SNP по данным отдельного эксперимента секвенирования [6, 7]. В литературе представлены различные подходы и программные средства к математическому моделированию сайтов

генетического полиморфизма, основанные на учете параметров экспериментального оборудования, использованию вероятностных моделей и статистических подходов, вспомогательной биологической информации [9, 10]. Основными ограничениями существующих программных пакетов являются: 1) отсутствие возможностей анализа и моделирования особенностей конкретных экспериментов (распределений измеряемых характеристик); 2) необходимость дополнительной реализации проце-дур генерации выборок для обучения классификационных моделей с целью последующей идентификации однонуклеотидных полиморфизмов; 3) отсутствие специализированного R-пакета, интегрирующего основные алгоритмы обработки экспериментальных данных, моделирования и идентификации сайтов SNP, для широкого использования биоинформационным сообществом.

Целью исследования является разработка R-пакета, включающего программные средства обработки экспериментальных данных, имитационного моделирования и идентификации сайтов однонуклеотидных полиморфизмов с использованием как наиболее эффективных классических алгоритмов, так и машинного обучения, обученных на смоделированных данных. Работоспособность разработанных программных средств подтверждена в ходе сравнительного анализа наиболее эффективных алгоритмов идентификации сайтов однонуклеотидного полиморфизма на примерах экспериментальных данных геномного секвенирования.

1. Программный пакет SNPSimulatoR

В программном пакете SNPSimulatoR реализована методика подхода имитационного моделирования и анализа SNP сайтов нуклеотидных последова-тельностей на основе бета- и нормального законов распределений, параметры которых определяются по имеющимся экспериментальным данным. Она позволяет смоделировать особенности конкретных экспериментов и сформировать эталонные выборки для обучения классификационных моделей. Понятие модели объекта включает моделирование поведения объекта в конкретных экспериментальных условиях (например, при известных законах распределений и параметрах, описывающих данные). Выбор методов обработки данных определяется сложностью реальных данных (небольшое число покрытий, пропуски, дубликаты, высокий уровень экспериментального шума и пр.). Для подтверждения адекватности имитационных моделей требуется сравнение характеристик данных вычислительного и реального экспериментов. Для задач генеративного моделирования, применимых с целью повышения точности предсказания моделей машинного обучения, наличие экспериментальных данных не обязательно.

Основные функции пакета и их описание представлены в таблице 1.

2. Экспериментальные данные

В качестве экспериментальных наборов рассмотрены эталонные данные о хромосомах 10 и 22, полученные консорциумом GIAB [11]. Выбор данных GIAB обусловлен тем, что на сегодняшний день это наиболее надежные бенчмарк-данные для

решения задач, связанных с изучением геномного полиморфизма у человека (от разработки новых инструментальных методов «мокрой» биологии до сравнения алгоритмов обнаружения полиморфных сайтов).

Таблица 1

Функции пакета SNPSimulatoR

R-функция	Описание	Результат
CreateHistogram1.R CreateHistogram2.R CreateHistogram2.R	Построение гистограмм чисел покрытий нуклеотидных сайтов. Аппроксимация гистограмм с помощью бета- и нормального законов распределений, оценка параметров распределений	Оцененные параметры бета- и нормального распределений
CreateHistogramSNVs.R	Построение гистограмм чисел покрытий SNP-сайтов. Аппроксимация гистограмм с помощью бета- и нормального законов распределений, оценка параметров распределений	Оцененные параметры бета- и нормального распределений
GenerateSNPSeqBeta.R	Имитационное моделирование сайтов и чисел их покрытий в нуклеотидной последовательности с использованием бета-распределения	Смоделированные данные, представленные символами референсного нуклеотида и числами покрытий сайта в каналах нуклеотидов A, C, G и T
GenerateSNPSeqBetaGauss-Noise.R	Имитационное моделирование сайтов и чисел их покрытий в нуклеотидной последовательности с использованием бета-распределения и добавлением гауссового шума	Смоделированные данные, представленные символами референсного нуклеотида и числами покрытий сайта в каналах нуклеотидов A, C, G и T
GenerateSNPSeqGauss.R	Имитационное моделирование сайтов и чисел их покрытий в нуклеотидной последовательности с использованием нормального распределения	Смоделированные данные, представленные символами референсного нуклеотида и числами покрытий сайта в каналах нуклеотидов A, C, G и T
GenerateSNPSqGaussGausNoise.R	Имитационное моделирование сайтов и чисел их покрытий в нуклеотидной последовательности с использованием нормального распределения и добавлением гауссового шума	Смоделированные данные, представленные символами референсного нуклеотида и числами покрытий сайта в каналах нуклеотидов A, C, G и T
TestSimModel.R	Построение гистограмм чисел покрытий нуклеотидных сайтов по смоделированным данным. Аппроксимация гистограмм с помощью бета- и нормального законов распределений, оценка параметров распределений. Применение критериев оценки адекватностей имитационных моделей	Оцененные параметры бета- и нормального распределений. Выводы об адекватности имитационных моделей.

Окончание табл. 1

R-функция	Описание	Результат
BinRatioTest.R	Программная реализация теста биномиального распределения.	Список SNP сайтов, *p*-величины
EntropyTest.R	Программная реализация энтропийного теста	Список SNP сайтов, оценки энтропии *E* и *p*-величины
TestAnalysis.R	Анализ смоделированных или экспериментальных данных с использованием тестов биномиального распределения и энтропийного	Списки SNP сайтов
VectorizeCites.R	Векторизация экспериментальных данных для обучения моделей классификации	Векторизированные наборы данных
TestMLonEDchr22.R	Создание и обучение моделей классификации на эксприментальных данных	Модели машинного обучения
TestMLonSimData.R	Создание и обучение моделей классификации на смоделированных данных	Модели машинного обучения
Classif_ctree_SimData.rds	Модель идентификации SNP сайтов на основе алгоритма CIT, обученная на смоделированных данных	Список SNP сайтов
Classif_rpart_SimData.rds	Модель идентификации SNP сайтов на основе алгоритма CART, обученная на смоделированных данных	Список SNP сайтов
Classif_SVM_SimData.rds	Модель идентификации SNP сайтов на основе алгоритма SVM, обученная на смоделированных данных	Список SNP сайтов
Classif_xgboost_SimData.rds	Модель идентификации SNP сайтов на основе алгоритма XGBoost, обученная на смоделированных данных	Список SNP сайтов

3. Результаты

Бета-распределение наиболее оптимально подходит для исследуемых интегральных характеристик рассматриваемых экспериментальных наборов. Нормальное распределение – менее точное, однако его применение допустимо к другим типам экспериментов, возможно демонстрирующим гауссовость данных. Полученные экспериментальные оценки параметров распределений используются в имитационных моделях для генерации обучающих данных.

Проведено исследование наиболее эффективных алгоритмов идентификации сайтов – тестов на основе биномиального распределения (ТБР), энтропии (ЭТ) и модифицированного точного теста Фишера (МТТФ), базовых методов машинного обу-

чения – деревьев условного вывода (англ. Conditional Inference Trees – CIT), классификации и регрессии построением дерева решений (Classification And Regression Tree – CART), опорных векторов с линейной разделяющей функцией (Support Vector Machine – SVM), обученных на имитационно смоделированных данных [12–14]. Эффективность алгоритмов оценена с помощью меры точности F_1 [15]. Результаты идентификации сайтов SNP для 9 наборов по 20 000 сайтов, считанных с позиций 12, 60, 84, 108×10^6 в хромосоме 10 и позиций 3, 9, 15, 21, и 27×10^6 в хромосоме 22, представлены в таблицах 2 и 3.

Таблица 2

Точность по мере F_1 алгоритмов идентификации сайтов SNP в хромосоме 10

i[1]	F_1, %					
	ТБР	ЭТ	МТТФ	CIT	CART	SVM
12	88,9	100	97,4	100	94,8	97,4
60	96,8	94,1	96,9	100	98,4	98,4
84	90,3	97,0	96,9	95.4	90,0	90,0
108	100	96,9	96,8	100	98,4	98,4
с.з.[2]	94,0	97,0	97,0	**98,9**	95,4	96,1

Примечания. [1] i, $\times10^6$ – номер первой позиции набора из 20 000 сайтов в хромосоме 10; [2] с.з. – среднее значение.

Таблица 3

Точность по мере F_1 алгоритмов идентификации сайтов SNP в хромосоме 22

i[1]	F_1, %					
	ТБР	ЭТ	МТТФ	CIT	CART	SVM
3	15,4	17,1	11,8	22,2	21,1	20,0
9	97,2	97,3	94,3	98,6	95,8	95,8
15	86,7	95,7	90,6	98,5	90,3	92,1
21	90,3	82,9	91,4	97,1	87,5	90,9
27	92,7	88,9	97,5	97,6	95,0	97,6
с.з.[2]	76,5	76,4	77,1	**82,8**	77,9	79,3

Точность идентификации сайтов SNP по мере F_1 на 2–5% выше у метода на основе деревьев решений CIT, чем у сравниваемых методов. Прочие модели машинного обучения и классические тесты идентификации имеют сопоставимую точность.

Дополнительно были исследованы модели машинного обучения, обученные на экспериментальных данных для хромосомы 22 (сформирована выборка из 72 261 сайтов, из которых 36 150 SNP, остальные – случайно выбранные не SNP). Точность классификации не превышала 60–70% на смоделированных и экспериментальных данных. Невысокая точность может быть обусловлена двумя факторами: 1) возможно «неполноценной» обучающей выборкой; 2) предположительно имитационная модель действительно точнее формирует обучающие данные, фокусируясь на воспроизведении важных/первостепенных источников информации и не учитывает второстепенные сигналы, присутствующие в реальных данных.

4. Заключение

Разработан программный пакет SNPSimulatoR для моделирования и анализа сайтов нуклеотидных последовательностей по экспериментальным наборам данных, основанный на генерации случайных событий бета- или нормального законов распределений, параметры которых оцениваются по имеющимся экспериментальным данным. Проверка работоспособности разработанных моделей и методов пакета произведена на примерах наборов эталонных данных о хромосомах 10 и 22 человека. Выполнен сравнительный анализ наиболее эффективных алгоритмов идентификации сайтов однонуклеотидного полиморфизма. Наивысшая точность идентификации по мере F_1 получена для модели классификации деревьев условного вывода.

Пакет может использоваться для моделирования синтетических данных, по экспериментальным данным или самостоятельно, с целью всестороннего тестирования и выбора наилучших алгоритмов идентификации сайтов SNP, а также для генеративного моделирования данных с целью обучения алгоритмов идентификации на основе методов машинного обучения (нейронных и байесовых сетей, ансамблевых алгоритмов и пр.).

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. *Su Z.* HAPGEN2: simulation of multiple disease SNPs / Z. Su, J. Marchini, P. Donnelly // Bioinformatics, 2011. Vol. 27(16), P. 2304–2305.
7. 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).
8. DHOEM: a statistical simulation software for simulating new markers in real SNP marker data / L. Jacquin [et al.] // BMC Bioinformatics. 2015. Vol. 16:404.
9. A comparison of gene region simulation methods/ A.E. Hendricks [et al.] // PLoS One, 2012. vol. 7(7) : e40925.
10. Genetic Simulation Resources: a website for the registration and discovery of genetic data simulators / B. Peng [et al.]// Bioinformatics, 2013 vol. 29(8). P. 1101–1102.
11. An open resource for accurately benchmarking small variant and reference calls / J. M. Zook [et al.] // Nat. Biotechnol. 2019. Vol. 37(5). P. 561–566, May 2019.
12. 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.
13. 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.
14. *Hastie T.* The Elements of Statistical Learning. Data Mining, Inference, and Prediction. / T. Hastie, R. Tibshirani, J. Friedman. 2nd ed. New York : Springer, 2009.
15. *Murphy K. P.* Probabilistic Machine Learning, London : The MIT Press, 2022.