УДК 004.94

SAMOVAR: БИБЛИОТЕКА ДЛЯ ГЕНЕРАЦИИ МЕТАГЕНОМОВ С ЗАДАННЫМИ СВОЙСТВАМИ

Смутин Д.В. (Университет ИТМО)
Научный руководитель – Иванов А.Б. (Университет ИТМО)

Введение. Метагеномика - область анализа “омиксных” данных секвенирования ДНК, собранных из окружающей среды. Результатом таких исследований являются описания микробных сообществ. Некоторые методы также позволяют отслеживать генетические сигнатуры многоклеточных организмов, взаимодействующих с ними [1]. Эти данные можно использовать для определения взаимосвязей между членами сообщества, или для предсказания характеристик сообществ по их количественному составу. Полученные данные всегда неполные: геномные базы не содержат информацию обо всех возможных организмах и потому не позволяют точно предсказывать состав микробиомов [2]; так же предсказание сильно затруднено как и на гипервариабельных, и на консенсусных последовательностях [3]. Зачастую объем классифицированных прочтений секвенирований для shotgun-метагеномики не превышает 20%. Несмотря на увеличение количества метагеномных экспериментов, данных о конкретных вариантах сообществ все еще недостаточно. Особенно это актуально для использования метагеномов в алгоритмах машинного обучения [4].

Решить эту проблему может генерация данных *de novo* [5]. Современные программы используют различные bootstrap-подходы. Например, SEQ2MGS [6] анализирует входные FASTQ-файлы и смешивает shotgun reads из разных секвенированных изолятов или spike-изолятов в реальных метагеномах, что позволяет генерировать сложные образцы. BEAR [7] и InSilicoSeq [8] генерируют прочтения с длинами и значениями качества на основе введенных эмпирически вводимых или случайно создаваемых распределений. Однако, такие данные искажают вероятностные пространства. Основной целью текущей работы является совершенствование алгоритмов генерации модельных метагеномов с заданными свойствами.

Основная часть. Данные для тестирования алгоритма были получены через API репозитория GMrepo [9]. Из базы данных загружаются образцы по заданным параметрам (заболевание, тип данных). Для получения данных с нормальным распределением подгоняется прямая и обратная функция. Нормализованные образцы кластеризуются. В пределах каждого кластера между всеми выборками рассчитывается R^2 наилучшей линейной модели и вероятности их сосуществования. Созданные таблицы сохраняются в базе данных.
Затем начинается генерация новых образцов с использованием полученной базы. Пользователь может задать значение обилия для одного из видов. На каждой итерации случайным образом на основании матрицы вероятностей сосуществования определяется видовой состав в генерируемом образце. С использованием матрицы R^2 для всех видов данного кластера, присутствующих в сгенерированной выборке, строится ориентированный граф генерации, начиная с инициализированных видов. На каждой итерации обхода в глубину ищется максимальное значение между видами, присутствующими и отсутствующими в графе, и добавляется новая зависимость. На основе полученного графа прогнозируются значения для всех видов, присутствующих в сгенерированной выборке.
На следующем этапе ищется максимальное значение R^2 между средними значениями предсказанных и необработанных кластеров. Наиболее точно предсказанное обилие вида инициализирует обработку следующего кластера. Алгоритм повторяется до тех пор, пока не будет определено присутствие, обилие и вероятность всех видов во всех образцах.
Для оценки эффективности предсказаний модели оценивался R^2 сгенерированных численностей по отношению к наиболее близким образцам из обучающей выборки, а также f1-значение и аккуратность предсказания на основании PCA. На основании 10000 сгенерированных образцов, из модели, обученной на 200 пробах, аккуратность предсказания видового состава составила 0,86, а f1-значение - 0,92. R^2 для предсказания значений в целом

составляет -0,03, а R^2 для предсказания соотношения внутри кластеров - 0,51. Тестовые и сгенерированные данные объединяются в кластеры в 92% случаев.
Образцы здоровой микробиоты представляют собой гетерогенные выборки, и для получения более точных результатов их следует разбить на более мелкие кластеры, для каждого из которых матрицы сходства должны быть подсчитаны отдельно. Для решения этой проблемы можно реализовать пошаговый k-мерный анализ для разных видов и образцов.
Полученная модель характеризуется высокой воспроизводимостью. Возможной проблемой может стать предсказание видов, дивергентно представленных в разных образцах. Ранее такие проблемы были частично решены с помощью алгоритмов машинного обучения без учителя [7]. Генераторы численностей, использовавшиеся ранее, еще предстоит сравнить с результатами алгоритма. В отличие от реализаций bootstrap-подхода [6,8], R^2 полученной модели близок к нулю. Даже для модели, ориентированной на взаимодействие между видами, эта величина слишком мала. Эта проблема решаема путем генерации большего количества образцов, но тогда вернутся и некоторые проблемы bootstrap-подхода, а предсказательная сила в плане выявления новых возможных связей в метагеноме снизится.

Выводы. Таким образом, генерируемые данные могут использоваться в различных исследованиях микробных сообществ. Созданный алгоритм и ему подобные могут стать ключом к решению проблемы неполноты данных в метагеномике. Пакет и документация доступны на github.com/dsmutin/samovar.

Список использованных источников:

1. D. Garfias-Gallegos и др., «Metagenomics Bioinformatic Pipeline», Methods Mol. Biol. Clifton NJ, т. 2512, сс. 153–179, 2022, doi: 10.1007/978-1-0716-2429-6_10.
2. A. Sczyrba и др., «Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software», Nat. Methods, т. 14, вып. 11, сс. 1063–1071, ноя. 2017, doi: 10.1038/nmeth.4458.
3. S. H. Ye, K. J. Siddle, D. J. Park, и P. C. Sabeti, «Benchmarking Metagenomics Tools for Taxonomic Classification», Cell, т. 178, вып. 4, сс. 779–794, авг. 2019, doi: 10.1016/j.cell.2019.07.010.
4. N. Vaulin, A. Chechenina, A. Ivanov, и V. Ulyantsev, «SAMOVAR: a pipeline for in silico generation of metagenomic communities with given properties», Природа, т. 443, сс. 27–28, 2023.
5. N. Derus, «Omics Data Generators.», searchRxiv, т. 2023, с. 20230256160, янв. 2023, doi: 10.1079/searchRxiv.2023.00257.
6. P.-J. Van Camp и A. Porollo, «SEQ2MGS: an effective tool for generating realistic artificial metagenomes from the existing sequencing data», NAR Genomics Bioinforma., т. 4, вып. 3, с. lqac050, июл. 2022, doi: 10.1093/nargab/lqac050.
7. S. Johnson, B. Trost, J. R. Long, V. Pittet, и A. Kusalik, «A better sequence-read simulator program for metagenomics», BMC Bioinformatics, т. 15, вып. S9, с. S14, сен. 2014, doi: 10.1186/1471-2105-15-S9-S14.
8. H. Gourlé, O. Karlsson-Lindsjö, J. Hayer, и E. Bongcam-Rudloff, «Simulating Illumina metagenomic data with InSilicoSeq», Bioinforma. Oxf. Engl., т. 35, вып. 3, сс. 521–522, фев. 2019, doi: 10.1093/bioinformatics/bty630.
9. D. Dai и др., «GMrepo v2: a curated human gut microbiome database with special focus on disease markers and cross-dataset comparison», Nucleic Acids Res., т. 50, вып. D1, сс. D777–D784, янв. 2022, doi: 10.1093/nar/gkab1019.