АНАЛИЗ БОЛЬШИХ ДАННЫХ В БИОИНФОРМАТИКЕ ДЛЯ ОБРАБОТКИ РЕЗУЛЬТАТОВ ПОЛНОГЕНОМНОГО СЕКВЕНИРОВАНИЯ

Р.С. Сергеев
Научный сотрудник лаборатории математической кибернетики ОИПИ НАН Беларуси

Ю.В. Грушецкий
Младший научный сотрудник лаборатории математической кибернетики ОИПИ НАН Беларуси

Объединенный институт проблем информатики НАН Беларуси, Республика Беларусь
E-mail: roma.sergeev@gmail.com

Abstract. This paper is devoted to the key issues associated with analysis of next-generation sequencing data in bioinformatics and retrieving new knowledge from it. Special attention is drawn to genome-wide association studies that allows developing recommendations for optimizing chemical therapy based on identified drug-resistant mutations.

Введение. Стремительное развитие технологий привело к взрывному росту в накоплении данных. Скорость и объем получаемых данных диктуют свои условия, необходимые для соответствующего осмысления получаемой информации. Одним из активных пользователей больших данных стала биоинформатика, которая за последнее десятилетие сделала огромный скачок в своем развитии. В то время как значительная часть ее практических приложений направлена на применение в области персонифицированной медицины и фарминдустрии, среди перспективных направлений следует отметить сельское хозяйство, пищевую промышленность, промышленные биотехнологии и ресурсодобывающие отрасли.

Рост интереса к обработке больших объемов данных в биоинформатике во многом связан с развитием технологий обработки и анализа нуклеиновых кислот, быстрым ростом технических возможностей современных компьютеров и резким падением стоимости получения новой информации. Вместо классических лабораторных экспериментов биологи все чаще используют геномные данные, полученные исследовательскими группами по всему миру. Примерами крупных международных проектов, создающих в настоящее время огромные массивы информации, являются ENCODE (https://www.encodeproject.org/) по расшифровке функций элементов генома человека [1] и «Атлас ракового генома» (http://cancergenome.nih.gov/), целью которого декларируется систематизация данных о генетических мутациях,

приводящих к возникновению рака [2].

Современные методы высокопроизводительного секвенирования позволяют получать большие наборы перекрывающихся прочтений для коротких участков генома, размером от 35 до 3000 пар оснований в зависимости от используемой технологии [3]. В связи с этим требуются огромные емкости для хранения данных. Например, несложные подсчеты показывают, что результаты секвенирования полного генома одного человека могут занимать порядка 200 гигабайт дискового пространства. С другой стороны, без обширного и глубокого статистического анализа, а также наличия требуемых для этого вычислительных ресурсов, исследования в области биоинформатики невозможны в принципе. В правильно организованных проектах постановка эксперимента с самого начала учитывает то, как будут анализироваться его результаты.

Особенности данных в биоинформатике. Данные биоинформатических экспериментов обладают всеми классическими характеристиками, которые определяют «большие данные»: объем (volume), скорость прироста (velocity), многообразие (variety). Однако трудности, с которыми сталкиваются при анализе биоинформатических данных, отличны от других хорошо известных проблем, возникающих при обработке данных физических экспериментов или спутниковых данных. Прежде всего, данные в биоинформатике сильно гетерогенны по природе. Многие аналитические задачи в биоинформатике требуют привлечения информации из нескольких разнородных независимых друг от друга источников для выполнения предсказаний и проверки гипотез. Более того, интересующие исследователя данные могут продуцироваться большим числом организаций, не имея единого формата представления.

Еще одной особенностью биоинформатических данных является их географическая распределенность по всему миру [4]. В то время как часть этих данных может быть передана через Интернет, остальные данные не подлежат передаче из-за их размера, стоимости передачи, наличия законов о неприкосновенности частной жизни и другим причинам. Для преодоления этих проблем разрабатываются методы разбивки и сжатия информации, алгоритмы параллельной обработки, развивается инфраструктура для удаленного анализа данных и обмена результатами.

Источники данных и основные решаемые задачи. Результаты секвенирования геномов живых организмов являются далеко не единственным типом данных в биоинформатике. Можно выделить порядка пяти основных направлений исследований, в которых востребованы методы обработки больших объемов информации [4]:

- анализ данных экспрессии генов;
- обработка результатов секвенирования ДНК, РНК и белковых молекул;
- анализ белок-белковых взаимодействий;
- моделирование и анализ биологических сетей;
- генная онтология.

В задачах анализа экспрессии генов выполняется сравнение уровней экспрессии при различных условиях. Как правило, основным поставщиком данных в подобных экспериментах являются технология профилирования экспрессии на основе ДНК-микрочипов. Данные продуцируются в пространстве ген-образец-время, где в каждой точке (например, в определенные моменты времени) фиксируются уровни экспрессии определенных генов. Такой анализ позволяет идентифицировать гены, работа которых изменена под влиянием патогенов или вирусов, путем сравнения пораженных и здоровых клеток, строить сети совместно экспрессируемых генов и сети регуляции. Быстрый рост объемов данных, продуцируемых ДНК-микрочипами, вызван низкой стоимостью и широким распространением этой технологии.

Обработка результатов секвенирования ДНК, РНК и белковых молекул. Анализ молекул ДНК, РНК и белков необходим для определения их структуры, функций и эволюционных изменений. ДНК-секвенирование широко используется для изучения генетических особенностей, связанных с проявлением определенных фенотипических признаков, идентификации микроорганизмов в образцах, судебно-медицинской экспертизе. По мере роста знаний и достоверной информации, относящейся к технологии полногеномного секвенирования, можно ожидать, что будут раскрыты сравнительные отличия различных штаммов патогенных микроорганизмов в ракурсе вирулентности, токсигенности, резистентности к различным факторам среды, онкогенности и т.п. Методы анализа генетических последовательностей включают алгоритмы выравнивания, сборки геномов, поиска в биологических базах данных. Секвенирование РНК-последовательностей в основном используется в качестве альтернативы микрочипированию, однако может применяться и для решения задач идентификации мутаций, изучения посттранскрипционных механизмов, обнаружения вирусов и экзогенных РНК, определения полиаденилирования.

Анализ данных белок-белковых взаимодействий. Белок-белковые комплексы и изменения в них несут большое количество информации, необходимой для исследования природы различных заболеваниях. Сети белок-белковых взаимодействий изучаются в разных областях биологии и продуцируют огромные объемы данных. Объем, скорость приращения и многообразие данных делают задачи из этой области классическими проблемами «больших данных». Это требует эффективной масштабируемой архитектуры для быстрой и точной генерации белок-белковых комплексов, их проверки и ранжирования.

Моделирование и анализ биологических сетей. В этом случае данные генетики, геномики, метаболомики и протеомики используются для построения сетей, моделирующих взаимодействия, происходящие в клетках живых организмов. С их помощью могут определяться способы регуляции работы генов, исследоваться молекулярные механизмы развития заболеваний, анализироваться связи между продуктами экспрессии генов и фенотипическими признаками для предсказания функций генов, идентифицироваться биомаркеры

заболеваний, выполняться поиск потенциальных мишеней для новых лекарственных препаратов.

Генная онтология. Данные генной онтологии предоставляют собой динамические, структурированные и видонезависимые аннотации генов и генных продуктов, ассоциированные с соответствующими биологическими процессами, клеточными компонентами и молекулярными функциями. Онтология построена по принципу ориентированного ациклического графа, где каждый термин связан с одним или несколькими другими терминами через отношения различных типов, и использует словари для облегчения выполнения запросов на различных уровнях.

Архитектурные решения для анализа больших данных в биоинформатических приложениях. Инструменты, разработанные до начала эпохи «больших данных» не были приспособлены для анализа возрастающих объемов данных биологических экспериментов. Несмотря на то, что кластерные и грид-технологии появились достаточно давно, они предназначались для решения узконаправленных задач, являлись высокозатратными в использовании и требовали высокой квалификации от специалистов. Скачок в развитие методов обработки больших данных связан с появлением технологий облачных вычислений, развитием моделей распределенных вычислений, таких как MapReduce и их открытых реализаций. Решения, где облако используется как для хранения данных, так и для выполнения вычислений, нашли широкое применение в биоинформатике, что стало ответом на специфичные для этой области трудности, связанные с обработкой больших массивов информации. В качестве примеров успешной реализации облачных решений по управлению и анализу данных полногеномного секвенирования можно привести сервисы Gaea и Hecate (http://www.genomics.cn/FlexLab/), разработанные Пекинским институтом геномики на основе программной платформы Hadoop, или сервис Bina (http://www.bina.com/), выросший из исследовательских проектов университетов Стэнфорда и Беркли.

В зависимости от поставленной задачи и существующих зависимостей между данными, большинство решений для крупных биоинформатических сервисов используют один из традиционных подходов к организации распределенных вычислений: модель MapReduce, отказоустойчивая архитектура на основе анализа графов (например, GraphLab), либо модель потоковой обработки данных (streaming graph architecture) с распределенной памятью и обменом сообщениями с помощью MPI.

Статистический анализ и машинное обучение. Обучение с учителем, обучение без учителя и гибридные подходы к машинному обучению являются основными методами в описательной и прогностической аналитике «больших данных».

Методы обучения с учителем предполагают наличие обучающей выборки, где для каждого наблюдения определены метки классов, и полученные знания используются для классификации новых наблюдений. Эти методы устойчивы к

зашумленности наборов данных и подходят для анализа данных, полученных из разных источников. Особенностью данных биологических экспериментов является то, что априорное знание, которое закладывается в обучающую выборку или метод в качестве «золотого стандарта» и должно охватывать суть вопроса, зачастую бывают неполным. К часто используемым методам обучения с учителем можно отнести метод опорных векторов, случайный лес и различные методы для решения задач регрессии.

Обучение без учителя, с другой стороны, не требуют наличия априорного знания или размеченной обучающей выборки. Эти методы ставят целью кластеризацию и определение структуры данных, например, выявление шаблонов в экспрессии генов при некотором заболевании. Методы обучения без учителя чувствительны к искажающим факторам, таким как групповой эффект, если анализируются данные нескольких экспериментов. В связи с этим их применяют для анализа однородных данных либо выполняют предобработку данных для получения достоверных результатов.

Гибридные методы, такие как глубинное обучение, позволяют обеспечить высокую точность предсказаний при наличии большой выборки данных. Методы глубинного обучения используют иерархическое представление данных для классификации и пытаются моделировать высокоуровневые абстракции в данных, используя алгоритмы обучения с учителем либо обучения без учителя, чтобы обучаться на каждом из уровней абстракции. Благодаря экспоненциальному росту данных этот класс методов стал особо популярным в последние годы.

Тем не менее, не все классические методы машинного обучения могут быть использованы для обработки больших объемов информации без доработки. Можно выделить несколько свойств, которыми должен обладать подходящий метод.

1 Масштабируемость к объему данных: метод должен обрабатывать большие порции данных с низкой пространственной сложностью и низкими накладными расходами на чтение данных с диска.

2 Робастность к скорости прироста: метод должен иметь низкую временную сложность и быть способным обрабатывать потоки данных в реальном времени без заметного падения производительности.

3 Открытость к многообразию: данные могут быть полуструктурированными или неструктурированными по своей природе. Однако традиционные методы машинного обучения предполагают обработку данных, которые генерируются одним источником и имеют фиксированную схему (упорядоченный набор признаков и отношений между ними). Для анализа «больших данных» метод должен быть способен обрабатывать данный из нескольких источников с различными схемами.

4 Учитывается инкрементальность данных: классические методы машинного обучения, как правило, оперируют со всем набором данных сразу, не учиты-

вая ситуаций, когда набор данных растет динамически со временем. Методы машинного обучения должны учитывать неравномерное приращение данных во времени и обрабатывать такие данные с минимальными затратами без компромиссов по качеству.

5 Распределенность: метод должен позволять распределенную обработку частей данных и объединение результатов. Это особенно важно для работы с источниками больших данных, распределенных по всему миру, что часто встречается в биоинформатических приложениях.

Распределенные версии некоторых методов машинного обучения реализованы в библиотеках, адаптированных для использования в условиях обработки больших объемов информации, наиболее известными из которых являются Apache Mahout [5] и MLlib [6].

Применение методологии «больших данных» для анализа лекарственно-устойчивого туберкулеза. Если к геномным данным добавить оцифрованные данные о пациенте, историю болезней, результаты анализов на чувствительность к лекарственным препаратам, данные томографии и т.п., то вместе с полнотой сведений существенно возрастает и их объем, что требует использования подходов из области «больших данных» для хранения и обработка такого массива информации. Белорусский портал о туберкулезе (www.tuberculosis.by) является уникальным ресурсом, пользователи которого имеют доступ к самому большому набору реальных случаев заболеваний туберкулезом, а также возможность ознакомиться с клиническими данными, изучить файлы с компьютерной томографией и получить полные геномы возбудителя туберкулеза по ряду случаев [7].

В рамках настоящего исследования [8] по анализу геномов лекарственно устойчивого туберкулеза была сформирована выборка из 144 организмов, включающая 17.65% штаммов микобактерий туберкулеза, чувствительных ко всем противотуберкулезным препаратам, 2.94% монорезистентных штаммов и 79.41% штаммов с множественной лекарственной устойчивостью. Все образцы M.tuberculosis были получены из клинических проб и секвенированы на платформе Illumina HiSeq2000 с длиной прочтений в 101 пару оснований при среднем покрытии 140x. Для каждого образца было выполнено четыре прогона секвенатора и получены наборы коротких парных прочтений, доступные в виде SRA-архивов размером 50 – 600 Мб, в зависимости от библиотеки, использованной для секвенирования ДНК. В ходе сборки геномов короткие прочтения картировались на референсный штамм H37Rv, после чего было выполнено выделение и аннотирование геномных вариаций.

Технологии изучения полных геномов тесно связаны с исследованиями, направленными на поиск ассоциаций генетического набора организма и его фенотипа. Под множественными маркерами лекарственной устойчивости будем понимать некоторую комбинацию однонуклеотидных полиморфизмов, которая наилучшим образом обуславливает наличие или отсутствие резистентности к исследуемому препарату. Для поиска таких комбинаций на этапе предобработки

были выполнены процедуры по очистке и фильтрации данных, проведен сравнительный анализ геномов и исследована популяционная структура образцов. Последующие шаги были направлены на исследование зависимостей между геномными полиморфизмами и результатами фенотипических тестов на резистентность (рисунок 1).

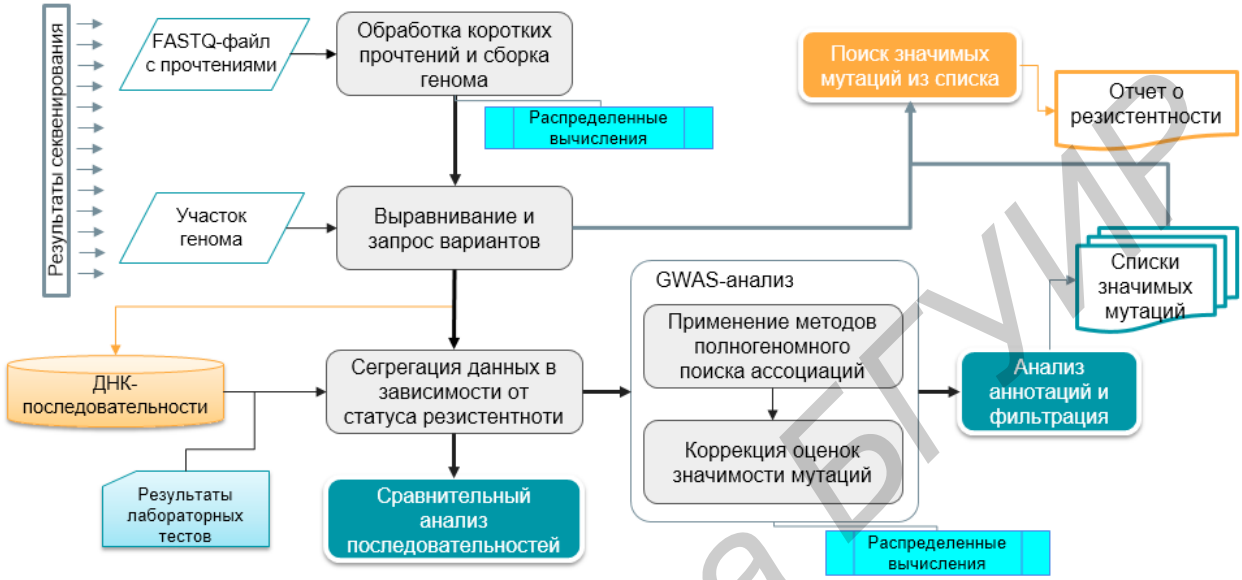

Рис. 1. Схема сервиса по анализу геномных данных микроорганизмов для поиска маркеров лекарственной устойчивости

В случае регуляризованной логистической регрессии в качестве регрессоров послужили все имеющиеся полиморфизмы, значимыми признавались мутации, для которых коэффициенты регрессии имели наибольшее отклонение от нуля. По данным перекрестной проверки эта модель имела наилучшие показатели, однако по итогам анализа аннотаций выяснилось, что в результирующее множество включается большое число шумовых признаков.

Линейная смешанная модель позволяет учесть скрытые взаимосвязи за счет включения в регрессионную модель вектора случайных эффектов, рассчитанных по матрице родства микроорганизмов. Как и в модели логистической регрессии, исследуемые полиморфизмы рассматривались в качестве регрессоров, а значимость коэффициентов определялась с помощью теста отношения правдоподобия.

Хорошим переборным методом поиска наиболее вероятных моделей регрессии и идентификации значимых полиморфизмов послужил алгоритм ориентированного на моду стохастического поиска на множестве иерархических лог-линейных моделей с числом переменных от двух до пяти. Благодаря этому методу удалось проанализировать не только влияние отдельных геномных маркеров, но и их взаимодействие. Процедура на основе этого метода позволила

ранжировать и сравнивать предложенные модели данных в соответствии с их апостериорными вероятностями и использовать байесовское усреднение, чтобы оценить значимость отдельных полиморфизмов по нескольким моделям, которые наилучшим образом согласуются с данными.

Важно отметить, что большинство из выявленных маркеров лекарственной устойчивости входит в состав современных тест-систем молекулярной диагностики MTBDRplus/MTBDRsl (HAIN-тест) и GeneXpert MTB/RIF, что является аргументом в пользу корректности полученных результатов.

Резюме. В заключение отметим основную проблему, связанную с анализом больших данных в биоинформатике. Главным критерием доказательности в биологии является прямой эксперимент, однако на основе анализа данных можно делать предсказания, экспериментальная проверка которых сильно ограничена естественным образом: ценой эксперимента, временем его проведения, технической осуществимостью. В результате экспериментальная валидация не успевает за предсказаниями, сделанными на основе анализа данных. Однако стоит надеяться, что совершенствование технологий сравнения больших последовательностей и работы с большими данными, приведет к тому, что выводы анализа перестанут восприниматься как предсказания, а начнут восприниматься как достоверные факты. Для этого нужно набирать статистику и увеличивать статистическую значимость этих предсказаний, то есть нужно еще больше данных.

Литература

[1]. ENCODE Project Consortium et al. The ENCODE (ENCyclopedia of DNA elements) project //Science. – 2004. – №. 5696(306). – С. 636-640.

[2]. Weinstein J. N. et al. The cancer genome atlas pan-cancer analysis project // Nature genetics. – 2013. – №. 10(45). – С. 1113-1120.

[3]. Liu L. et al. Comparison of next-generation sequencing systems //BioMed Research International. – 2012.

[4]. Kashyap H. et al. Big Data Analytics in Bioinformatics: A Machine Learning Perspective //arXiv preprint arXiv:1506.05101. – 2015.

[5]. Anil R., Dunning T., Friedman E. Mahout in action. – Shelter Island : Manning, 2011. – С. 145-183.

[6]. Meng X. et al. Mllib: Machine learning in apache spark //arXiv preprint arXiv:1505.06807. – 2015.

[7]. Кириченко В.В. и др. Белорусский портал о туберкулезе: уникальная база данных в открытом доступе // Медицинская панорама. - 2015. - № 9. - С. 75-78.

[8]. Сергеев Р.С. и др. Алгоритмы поиска мутаций лекарственной устойчивости в геномах микобактерий туберкулеза // Информатика. - 2016. - № 1(49). - С. 75-91.