«Сейфуллин оқулары – 18: « Жастар және ғылым – болашаққа көзқарас» халықаралық ғылыми -практикалық конференция материалдары = Материалы международной научно-практической конференции «Сейфуллинские чтения – 18: « Молодежь и наука – взгляд в будущее» - 2022.- Т.I, Ч.IV. - С. 66-68

БИОИНФОРМАТИЧЕСКИЕ АЛГОРИТМЫ ДЛЯ ИДЕНТИФИКАЦИИ ТАНДЕМНЫХ ПОВТОРОВ В БАКТЕРИАЛЬНЫХ ГЕНОМОВ ПРИ ГЕНОТИПИРОВАНИИ

В.А. Шевцов, докторант 1 курса
Нур-Султан, Казахский агротехнический университет им. С.Сейфуллина

Введение

В настоящее время определение нуклеотидной последовательности геномов является основной технологией в биологических исследованиях. Еще 20 лет назад изучение геномных данных казалась дорогой и трудно решаемой задачей именно в области определения нуклеотидной последовательности. Первичная стоимость генома человека оценивался в 3 миллиарда долларов[1]. Прогресс в развитии секвенаторов нового поколения(NGS) мгновенно изменили ситуацию, что уже сейчас позволяет получать геном человека за 1000 долларов США[2], а стоимость малых геномов бактерий вирусов колеблется в диапазоне от 500 до 30 долларов США за штамм[3].Снижение стоимости геномных исследований привело к бурному развитию данной области, и уже реализуются грандиозные проекты, нацеленные на геномное секвенирование 5000 насекомых[4], десятьтысяч геномов позвоночных, миллионы микроорганизмов и др. Эксперты предупредили,что обработка генетических данных скоро превысит вычислительные потребности TwitteriиYouTube[5].

Процессинг данных генотипирования микроорганизмов полученных в «догеномный» период с новыми геномными данными стало настоящей проблемой. Данная пропасть возникла в связи с различиями протоколов изучения ДНК микроорганизмов на протяжении развития технологий генотипирования. Генотипирование на основании фрагментации ДНК с разделением фрагментов в pulsed-field gel electrophoresis (PFGE), Random Amplified Polymorphic DNA (RAPD), Restriction Fragment Length Polymorphism (RFLP) сменились на более простые, точные и легко воспроизводимые методы как мультилокусный анализ вариабельных тандемных повторов (MLVA) и мультилокусное сиквенс типирование

(MLST). В настоящее время существует переходный период, когда для полноценной эпидемиологии необходимо использовать данные, полученные по всему миру различными методами генотипирования, а также при ретроспективном анализе ранее полученных результатов генотипирования циркулирующих штаммов. В переходный период многие лаборатории вынуждены параллельно использовать полногеномное секвенирование, PFGE and MLVA для ряда патогенов, чтобы проводить полноценный эпидемиологический контроль. В связи с этим развитие методологии in silico генотипирования на полногеномных данных является актуальной задачей, особенно для патогенов, для которых MLVA и MLST рассматривался в качестве злотого стандарта генотипирования включая Brucella spp, Bacillus anthracis, Yersenia pestis, Francisella tularensis and Neisseria meningitidis.

Биоинформатические алгоритмы

Технологии сбора биологических данных становятся все более дешевыми и эффективными, такие как автоматические секвенаторы генома, что дает начало новой эре больших данных в биоинформатике. Таким образом, из-за увеличения объема данных в биоинформатике, еще до этапа анализа возникают непосредственные проблемы, связанные с хранением и обменом данными между группами, генерирующими данные, и группами анализа. Такой огромный объем данных далеко за пределами вычислительных возможностей в большинстве биомедицинских исследовательских лабораторий. Чтобы преодолеть эту практическую трудность при создании сервера для хранения данных и обширных вычислений в отдельных лабораториях, многие компании, такие как Amazon, разработали платформу облачных вычислений и предоставили сервис, позволяющий исследователям использовать свои серверы по мере необходимости, или многие исследовательские институты создали кластерную вычислительную платформу для их дочерние лаборатории.

С другой стороны, большинство биологических или клинических исследователей до сих пор не знакомы с вычислительными подходами для работы с этими данными, а биоинформатика является чужой территорией. Недавние эксперименты на основе микрочипов или секвенирования часто генерируют существенно большие данные и являются более широко применимыми, чем раньше , и проблема может возникнуть, когда большинство биомедицинских исследователей имеют очень ограниченные возможности для проведения анализа таких больших данных с использованием соответствующих инструментов, которые могут быть полностью поняты другими[6].

В современной биоинформатике стало необходимстьюпроводить генетическую идентификацию и выявление уникальных характеристик организма, что показывает необходимость развития биоинформатики. В настоящее время система генотипирования построена на получении

полногеномных данных, при этом программное обеспечение в данном направлении в основном нацелены на выявление однонуклеотидных палеморфизмов в кодирующей и некодирующей части и упускает важный элемент генетического материала бактерий такой как тандемные повторы(рисунок 1).

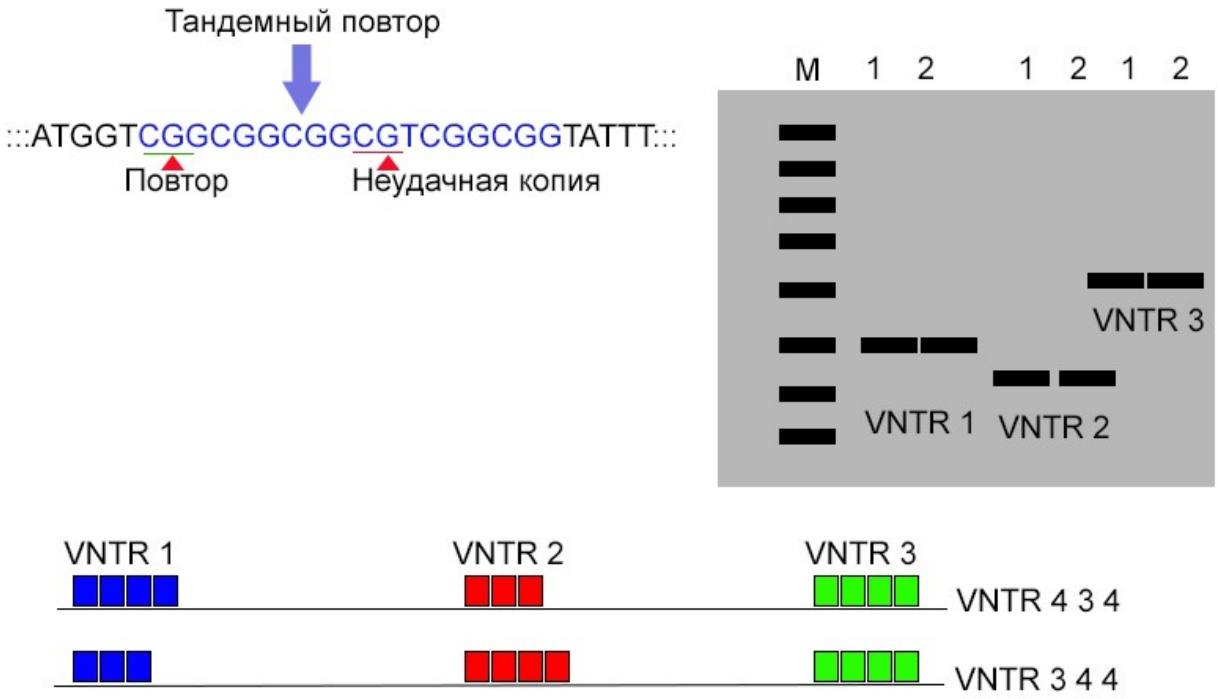

Рисунок 1 - Представление тандемных повторах в ридах(сырые данные)

Несмотря на то, что повторы не несут смысловой информации, они необходимы для изменения уровня экспрессии генов и могут быть использованы при генотипировании в оссобенности для бактерий с особой высоко консервативными геномами, таких как бруцелл, сибирской язвы (*Yerseniapestis*) для которых существует конвенция по имеющимся повторам в геноме этих организмов и является практически золотым стандартом. В связи с этим, основной целью данной работы является разработка информационного алгоритма для выявлении и идентификации количества тандемных повторов бактериальных геномов. Были разработаны скрипты на Python[7, 8], предназначенный для анализа нескольких локусов VNTR (переменное количество тандемных повторов), которые выполняет ПЦР in silico для извлечения последовательностей тандемных повторов из отправленных файлов fasta.и вызовает аллели VNTR.Скрпты используют список праймеров для восстановления последовательностей из VNTR. Данный скрипт был протестирован на полногеномных данных (сырых данных полученных с секвенатора) для 6 штаммов *Brucellaabortus*всего в геноме*Brucella* идентифицированно 88 тандемных повторов, все они были идентифицировани

при помощи разработанного алгоритма, при этом время затратность на биоинформатический анализ составлял на сервере всего 15 минут, при этом анализировалоссь примерно 600000 ридов на кажый штамм, что позволяет получить результат на основании которого можно делать генетическое заключение о вариабельности того или иного образца.Таким образом разработанные скрипты позволяют проводить *insilico*MLVA генотипирование бактерий с использованием первичных сырых данных поступающих с секвенаторов нового поколения. Кроме того, результаты, полученные с использованием разработанного алгоритма достоверные и могут быть проверены методом визуализации и сравнения последовательностей в сортированных ридах. Разработанные скрипты показывают высокую достоверность, а также наибольшую работоспособность, что проявлялась в установленных аллелях в локусах.

Заключение

Биоинформатика нашла широкое применение в молекулярной эпидемиологии, позволяя медицинским работникам устанавливать источники инфицирования, определять фенотипические особенности в краткосрочный период времени, отслеживать филогенетическую историю проявляя свет на ранее неизвестные факты об инфекция существовавших тысячелетия назад. Самым дискриминационным методом в молекулярной эпидемиологии является анализ полно геномных данных. Преимущество использования полно геномных данных является возможность филогенетического анализа и выявления фенотипических признаков на основании аннотации геномов. Недостатком метода является высокая стоимость получения полно геномных данных, необходимость наличия профессионального биоинформатика и больших аналитических мощностей.

Списоклитературы

1. Roberts, L.J.S., *Human Genome: Questions of Cost: At a recent workshop scientists tried to tally the costs of the genome project; they came up with a hefty ballpark figure in the low billions.* 1987. 237(4821): p. 1411-1412.
2. Schwarze, K., et al., *The complete costs of genome sequencing: a microcosting study in cancer and rare diseases from a single center in the United Kingdom.* 2020. 22(1): p. 85-94.
3. Turenne, C.Y., et al., *Rapid identification of bacteria from positive blood cultures by fluorescence-based PCR–single-strand conformation polymorphism analysis of the 16S rRNA gene.* 2000. 38(2): p. 513-520.
4. Behura, S.K. and D.W.J.B.R. Severson, *Codon usage bias: causative factors, quantification methods and genome-wide patterns: with emphasis on insect genomes.* 2013. 88(1): p. 49-61.
5. Check Hayden, E.J.N.N., *Genome researchers raise alarm over big data.* 2015.

6. Park, P.J.J.N.r.g., *ChIP–seq: advantages and challenges of a maturing technology.* 2009. 10(10): p. 669-680.
7. Shevtsov V. Search for primers in reads. URL: https://github.com/Vladislav-Shevtsov/search-primers-in-reads (датаобращения: 20.03.22)
8. D.Christiany MLVA_finder. URL: https://github.com/i2bc/MLVA_finder (датаобращения: 20.03.22)