А.Н. Панкратов, М.И. Пятков, Р.К. Тетуев

ВЕБ-СЕРВИСЫ ДЛЯ РЕШЕНИЯ ЗАДАЧ РАСПОЗНАВАНИЯ ПОВТОРЯЮЩИХСЯ СТРУКТУР В НУКЛЕОТИДНЫХ И АМИНОКИСЛОТНЫХ ПОСЛЕДОВАТЕЛЬНОСТЯХ*

ИМПБ РАН – филиал ИПМ им. М.В. Келдыша РАН, г. Пущино,
pan@impb.ru

Математические проблемы поиска повторов

Многие проблемы биоинформатики связаны с задачей определения близости структур. Для первичных структур биополимеров мерой близости является схожесть последовательностей при выравнивании. Для пространственных структур понятие близости тесно связано с классификацией структур. Нахождение и сравнение неточных повторов является одной из базовых задач биоинформатики.

Однако при этом задача решается на основе упрощений и допущений. Обычно кандидаты на повторы в геномах находятся на основе некоторых затравок, или якорей, в последовательности и уточняются на основе выравнивания методом динамического программирования. Такой подход называется локальным выравниванием последовательностей. Его результаты могут различаться в зависимости от используемых якорей и параметров выравнивания.

Поиск неточных повторов в белковых последовательностях также зависит от определения повтора методом динамического программирования. При этом сложность и неоднозначность выравнивания в этом случае возрастает в связи с усложнением алфавита. Всевозможные редукции алфавита приводят к еще большему количеству допущений и параметров в задаче.

При этом настоящей проблемой является множественное выравнивание повторов, когда необходимо выработать критерий того, что некоторое число кандидатов на повтор становятся одним классом. Известно, что любая метрика не в состоянии сама по себе обеспечить решение этой задачи, поскольку пороговое решающее правило не удовлетворяет условию транзитивности устанавливаемого отношения эквивалентности между повторами. То есть если объект А похож на Б, а Б похож на В, то правило не гарантирует, что А похож на В.

Альтернативный подход к задаче поиска повторов

В ходе многолетних исследований коллектива авторов был предложен и развит альтернативный подход к нахождению кандидатов на повторы с помощью сравнения последовательностей целыми блоками [1]. Для этого было предложено переводить символьную последова-

* Работа выполняется при поддержке Российского фонда фундаментальных исследований, проект № 15-29-07063.

тельность в пучок линейно независимых кривых, Фурье-спектры которых могут быть использованы при оценке расстояния между повторами.

Предложенный метод свободен от предопределенных якорей последовательности, а редакционное расстояние заменяется интегральной оценкой отклонения сравниваемых фрагментов кривых.

Потенциально сильная сторона развиваемого метода – это возможность распознавания тандемных повторов, что является одной из задач множественного выравнивания повторов. На точечной матрице, отражающей сравнение всех попарно блоков последовательности, тандемные повторы могут быть представлены в виде совершенного (заполненного) квадрата. Это отражает тот факт, что можно корректно с точки зрения решающего правила установить отношение эквивалентности между неточными повторами.

Упрощение, введенное спектральным подходом, обусловлено введением такого параметра как масштаб. За счет этого параметра сложность задачи может быть существенно снижена. На каждом конкретном масштабе сложность квадратично зависит от длины последовательности, в то же время было показано, что за счет масштаба сложность может быть снижена до линейной [2].

Роль высокопроизводительных вычислений

Роль технологий высокопроизводительных вычислений в данной задаче состоит в том, чтобы решать задачу в исходной постановке, не вводя упрощенных версий алгоритмов. Только таким образом можно исследовать повторы на грани применимости алгоритмов.

Опыт построения таких алгоритмов свидетельствовал о том, что одна из стратегий – это отказ от хранения промежуточных данных в пользу повторяющихся вычислений с открывающейся возможностью распараллеливания вычислений и минимизации использования памяти. Например, не хранить вычисленные значения элементов некоторой матрицы, а иметь способ вычисления элементов матрицы. Как показала практика, этот принцип работает как для векторно-параллельных вычислительных систем с общей памятью, так и для систем с распределенной памятью.

Роль облачных вычислений

При разработке специализированного программного обеспечения, предназначенного для определенного круга задач и пользователей, встает необходимость выбора платформы для разработки приложения. Технологии облачных вычислений являются наиболее перспективными как с точки зрения масштабируемости и развития разрабатываемых приложений, так и удобства предоставления доступа для пользователя.

Предоставление программного обеспечения пользователям в виде вебсервисов является одним из распространенных и перспективных видов облач-

ных технологий. При этом разработчики сервисов сталкиваются с необходимостью разработки сложных программно-аппаратных комплексов, поддерживающих эти технологии. С развитием сервиса встает необходимость расширения его возможностей и привлечения большего числа пользователей.

В то время пока каждый пользователь обдумывает свою задачу, его компьютер зачастую простаивает. Это машинное время естественным образом через виртуальную машину браузера может быть отдано в помощь серверу или другим клиентам. Таким образом может быть построена распределенная вычислительная система, составленная из пользователей сервиса. Этот путь может быть выбран при построении бесплатных сервисов для проведения научных расчетов. Проблемы реализации таких распределенных вычислительных систем с недавнего времени стали активно обсуждаться в научной литературе [5–8].

Например, в области биоинформатики научные сервисы в сети Интернет зачастую являются необходимым подспорьем в работе биолога. Однако создание таких сервисов, зачастую связанных с большими объемами информации, требует специальных организационных условий. Не решая эту задачу в целом, можно рассмотреть построение вычислительно мощного, но дешевого с аппаратной точки зрения сервиса.

Предложим следующую упрощенную вычислительную архитектуру, состоящую из сервера и двух видов клиентов: ведущий (master), мастер-клиент, и ведомый (slave), служебный клиент. На практике предполагается, что два вида клиента совмещаются в одном с двумя ролями. Роли этих трех агентов вычислительной сети строго разделены.

Ведущий клиент играет роль главного приложения в системе: запускает процесс вычислений по заданному параллельному алгоритму над данными, которые также находятся под его управлением. Ведущий клиент разделяет задачу на автономные порции вычислений, формирует задания по ним, отправляет задания на сервер и принимает решенные задания с сервера до тех пор, пока задача не будет решена.

Сервер представляет собой диспетчер пакетов, сформированных ведущими клиентами, ничего не знает о решаемой задаче, принимает пакеты с заданиями, отдает на вычисления ведомым клиентам, принимает обратно, возвращает выполненные пакеты ведущему клиенту.

Ведомый клиент – это фоновый вычислительный процесс на виртуальной машине браузера, находящегося на связи с веб-сервером.

Таким образом, данная архитектура естественным образом объединяет пользователей ресурса и обобществляет их вычислительные мощности для решения вычислительных задач, используя заданное программное обеспечение. Каждый клиент выступает одновременно мастер-клиентом, под управлением пользователя инициируя расчеты, и служебным клиентом, выполняющим теневые задания от других пользователей.

Результаты

В рамках парадигмы минимизации памяти было построено несколько алгоритмов:

1) кэш-оптимальный векторный алгоритм вычисления коэффициентов разложения с вычислением матрицы ортогональных многочленов по требованию;
2) распределенный алгоритм поиска повторов по точечной матрице с отложенным вычислением элементов матрицы по требованию;
3) распределенный алгоритм глобального выравнивания последовательностей, минимизирующий использование памяти [4].

В качестве тестовой задачи для организации распределенной системы и проверки рабочей гипотезы о принципиальной возможности построения подобной вычислительной сети из клиентов-браузеров нами выбрана классическая задача биоинформатики – глобальное попарное выравнивание генетических последовательностей. Этот алгоритм был реализован в виде иерархии усложняющихся веб-приложений на языке JavaScript.

На первом этапе разработки алгоритма был создан высокопроизводительный однопоточный алгоритм. Особенностью этой версии стало достижение максимальной универсальности алгоритма, а также использование таблиц поиска для построения последовательного кода без условных операторов. На втором этапе разработки алгоритма было проведено его распараллеливание с помощью веб-worker’ов. На третьем этапе порции вычислений, производимые веб-worker’ами, отправляются в распределенную систему вычислений, организованную на сервере.

В результате построен сервис (http://sbars.impb.ru/aligner.html), превосходящий аналоги по своим возможностям и востребованный при проведении научных исследований.

1. *Панкратов, А.Н.* Поиск протяженных повторов в геномах на основе спектрально-аналитического метода [Текст] / А.Н. Панкратов, М.И. Пятков, Р.К. Тетуев, Н.Н. Назипова, Ф.Ф. Дедус // Математическая биология и биоинформатика. – 2012. – Т. 7, № 2. – С. 476–492.
2. *Pyatkov, M.I.* SBARS: fast creation of dotplots for DNA sequences on different scales using GA-, GC-content / M.I. Pyatkov, A.N. Pankratov // Bioinformatics. – 2014. –Vol. 30, № 12. – P. 1765–1766.
3. *Панкратов, А.Н.* Спектрально-аналитический метод распознавания неточных повторов в символьных последовательностях [Текст] / А.Н. Панкратов, Р.К. Тетуев, М.И. Пятков, В.П. Тойгильдин, Н.Н. Попова // Труды Института системного программирования РАН. – 2015. – Т. 27, Вып. 6. – С. 335–344.
4. *Тетуев, Р.К.* Параллельный алгоритм глобального выравнивания протяжённых аминокислотных и нуклеотидных последовательностей / Р.К. Тетуев, М.И. Пятков, А.Н. Панкратов // Математическая биология и биоинформатика. – 2017. – 12(1). – P. 137–150.

5. Pan, Yao & White, Jules & Sun, Yu & Gray, Jeff. (2017). Gray Computing: A Framework for Computing with Background JavaScript Tasks // IEEE Transactions on Software Engineering. P. 1-1. 10.1109/TSE.2017.2772812.
6. *Zorrilla, M.* SaW: Video Analysis in Social Media with Web-Based Mobile Grid Computing / M. Zorrilla, J. Flórez, A. Lafuente, A. Martin, J. Montalbán, I.G. Olaizola, I. Tamayo // Mobile Computing IEEE Transactions on. – 2018. – Vol. 17. – P. 1442–1455. ISSN 1536-1233.
7. *Cushing, R.* Distributed Computing on an Ensemble of Browsers / R. Cushing, G.H.H. Putra, S. Koulouzis, A. Belloum, M. Bubak and C. de Laat // In IEEE Internet Computing, Sept.-Oct. 2013. – Vol. 17, No. 5. – P. 54–61. – DOI: 10.1109/MIC.2013.3.
8. *Fabisiak, T.* Browser-Based Harnessing Of Voluntary Computational Power / T. Fabisiak, A. Danilecki // Foundations Of Computing And Decision Sciences. – 2017. – Vol. 42, No. 1. – DOI: 10.1515/fcds-2017-0001.

А.А. Прохоров, Н.О. Пересторонин

ПОДХОД К ИСПОЛЬЗОВАНИЮ ОБЛАЧНЫХ ВЫЧИСЛЕНИЙ ДЛЯ РЕШЕНИЯ КРУПНОМАСШТАБНЫХ НАУЧНЫХ И ИНЖЕНЕРНЫХ ЗАДАЧ*

ООО «ДАТАДВАНС», г. Москва,
alexander.prokhorov@datadvance.net

В решении большинства актуальных научных и инженерных задач применяются вычислительные эксперименты, реализация которых подразумевает совместное использование ряда специализированных вычислительных средств, передачу данных между ними, и зачастую организацию сложной последовательности расчетов. Возможность автоматизировать такие вычислительные эксперименты и повторно использовать ранее полученные результаты является важным фактором, влияющим на производительность труда исследователей и инженеров в широком спектре предметных областей.

Распространенным подходом для достижения необходимого уровня автоматизации является использование интеграционной программной платформы. Такие системы, как правило, обладают графическим пользовательским интерфейсом, где визуально задается описание процесса расчета (который часто представляется в виде графа), и предоставляют среду исполнения этих процессов. Последовательность вычислений представляется в виде потока работ – набора взаимосвязанных задач с четко определенным порядком исполнения [1]. Таким образом, на базе интеграционной платформы пользователем создается некая расчетная схема, управляющая исполнением компонентов вычислительного эксперимента (отдельных расчетных средств) и передачей данных между

* Работа выполнена при финансовой поддержке РФФИ в рамках научного проекта № 15-29-07043. Условия для выполнения работ по проекту предоставлены ООО «ДАТАДВАНС».