Отчет О.В.Вишневского.

Тема работы.

Разработка компьютерных методов выявления функциональных контекстных сигналов в регуляторных районах генов.

Состав коллектива (на момент выполнения работы):

Олег Владимирович Вишневский. К.б.н. Ведущий программист ИЦиГ (основное место работы), Старший преподаватель НГУ (совмещение). oleg@bionet.nsc.ru

Информация о гранте, гос.задании, программе исследований, ФЦП и т.п. (если есть): номер, название, руководитель, срок выполнения, …

Бюджетный проект ИЦиГ (FWNR-2022-0020 «Системная биология и биоинформатика: реконструкция, анализ и моделирование структурно-функциональной организации и эволюции генных сетей человека, животных, растений и микроорганизмов»).

Научное содержание работы:

Постановка задачи.

ChIP-Seq (иммунопреципитационное секвенирование хроматина) является одним из наиболее эффективных подходов к полногеномному анализу специфических особенностей связывания транскрипционных факторов (ТФ), связывания полимеразы и модификаций гистонов. Выявление этих особенностей позволяет прояснить механизмы, лежащие в основе регуляции транскрипции, и предложить новые подходы к распознаванию и описанию регуляторных областей генов во всем геноме. Для этого необходима разработка компьютерных подходов анализа огромных объемов данных об участках локализации ТФ на ДНК.

Современное состояние проблемы (на момент начала работы).

Большинство доступных в настоящее время инструментов выявления мотивов основаны либо на ранее полученной экспериментальной информации о сайтах связывания ДНК/белок, хранящихся в специализированных базах данных, либо на использовании анализа de novo, основанного на сравнении анализируемых последовательностей и поиске в них относительно схожих участков.

Для обнаружения мотивов de novo были разработаны различные методы. Они включают в себя анализ частот k-меров (k-буквенных подстрок), суффиксных деревьев, поиск наибольших клик в графе, построенном на основе использования расстояния редактирования между k-мерами, подходы локального множественного выравнивания, основанные на жадном алгоритме, алгоритме Expectation - Maximization и стратегии стохастического отбора. Как правило, результаты их работы представляются либо в виде позиционно-зависимых весовых матриц (PSWM), либо в виде олигонуклеотидных мотивов, записанных в 4-(A,T,G, C) или 15-буквенном коде IUPAC (A,T,G,C, R=G/A, Y=T/C, M=A/C, K=G/T, W=A/T, S=G). /C, B=T/G/C, V=A/G/C, H=A/T/C, D=A/T/G, N=A/T/G/C).

Методы обнаружения de novo значимых олигонуклеотидных мотивов достаточно быстры и не требуют ни множественного выравнивания запросных последовательностей, ни экспериментальной информации о точной локализации сайтов связывания транскрипционных факторов. Огромное разнообразие мотивов значительно затрудняет их распознавание. Так, мотивы длины 8 в 15-буквенном коде представлены до $15^8 \sim 2{,}5*10^9$ различными вариантами. Это заставляет исследователей использовать различные эвристические подходы для обнаружения мотивов. Однако эвристические подходы не гарантируют нахождение глобального оптимума, т.е. обнаружение наиболее представленного и значимого мотива. Вот почему должны быть реализованы исчерпывающие алгоритмы обнаружения мотивов. Для этого необходимы высокопроизводительные компьютерные системы, основанные на массовом распараллеливании задач.

Подробное описание работы, включая используемые алгоритмы.

Нами предложен алгоритм расчета представленности мотива **M** длины *l* в выборке **D**, состоящей из N_{seq} последовательностей длины L_{seq}, основанный на оценке соответствия мотива **M** каждой из $N_{seq} * (L_{seq}-l+1)$ позиций выборки **D**. Для этого каждый символ мотива в 15-ти буквенном IUPAC коде записывается в виде целого числа от 1 до 15 (табл. 1а), а каждый нуклеотид выборки анализируемых последовательностей **D** записывается в виде целого числа от 0 до 3 (табл. 1б).

Таблица 1а. Бинарное представление 15-буквенного IUPAC кода для букв мотива.

	A	**T**	**G**	**C**	**R**	**Y**	**M**	**K**	**W**	**S**	**B**	**H**	**V**	**D**	**N**
	A	**T**	**G**	**C**	**G/A**	**T/C**	**A/C**	**G/T**	**A/T**	**C/G**	**!A**	**!G**	**!T**	**!C**	**N**
A	1	0	0	0	1	0	1	0	1	0	0	1	1	1	1
T	0	1	0	0	0	1	0	1	1	0	1	1	0	1	1
G	0	0	1	0	1	0	0	1	0	1	1	0	1	1	1
C	0	0	0	1	0	1	1	0	0	1	1	1	1	0	1
Code	1	2	4	8	5	10	9	6	3	12	14	11	13	7	15

Таблица 1б. Бинарное представление 15-буквенного IUPAC кода для нуклеотидной последовательности.

	A	**T**	**G**	**C**
Code	0	1	2	3

В этом случае, соответствие между мотивом **M** длины *l* и районом [i;i+*l*] анализируемой последовательности, записанной 4-х буквенном коде, может быть оценено с помощью

операции побитового сдвига вправо. При этом если буквы в позиции мотива **M** и анализируемой нуклеотидной последовательности соответствуют друг другу, то побитовый сдвиг вправо бинарного представления символа мотива **M** (табл. 1а) на число, соответствующее бинарному представлению нуклеотида (табл. 1б) выдаст 1, в противном случае – 0. Таким образом, если все символы мотива и сравниваемого участка последовательности соответствуют друг другу, произведение результатов побитового сдвига для всех позиций будет равным 1. Подобный подход позволяет существенно ускорить оценку соответствия мотива и нуклеотидной последовательности.

Для оценки представленности всех 15^l возможных мотивов, все рассматриваемые мотивы разбиваются на группы равные количеству потоков в потоковом блоке, а каждый потоковый блок обрабатывает свою нуклеотидную последовательность. При этом все нуклеотидные последовательности анализируемой выборки **D** размещались в текстурной памяти, что позволило существенно ускорить доступ к этим последовательностям. Последовательность, с которой работает блок, копировалась в разделяемую память, поскольку доступ к ней существенно быстрее, чем к глобальной памяти.

Загрузка последовательностей из текстурной памяти производится всеми потоками блока. Размер разделяемой памяти на мультипроцессор ограничивает длину последовательностей в ~14000 нуклеотидов, что достаточно для решения большинства задач по анализу регуляторных районов генов. Сократить количество итераций обращения к текстурной памяти можно за счет использования упакованных типов данных. В нашем случае вместо char (один 8-ми битный символ) использовался uchar4 (четыре 8-ми битных символа). То есть, например, для загрузки одной последовательности длины L=2000 нуклеотидов 512 потоками нам потребуется четыре итерации обращений к текстурной памяти для char и только одна для uchar4.

Затем каждый поток в блоке проверяет встречаемость одного мотива в одной последовательности нуклеотидов и запоминает результат в глобальной памяти. В случае использования запакованных типов данных (uchar4) каждый поток может обрабатывать одновременно четыре последовательности. После этого запускается другое ядро на GPU, которое вычисляет встречаемость обработанной порции мотивов во всех последовательностях нуклеотидов. В то время, пока на GPU идет обработка мотивов, на CPU готовится следующая порция мотивов и процесс повторяется.

После того как процесс расчета представленности проведен для всего множества мотивов, производится оценка значимости полученных мотивов согласно биномиальному

критерию и расчет их представленности в выборке случайных последовательностей. Случайная выборка генерировалась с частотами нуклеотидов, соответствующими частотам нуклеотидов в анализируемой выборке. Мотивы, не удовлетворяющие граничным критериям, удалялись из рассмотрения, а среди оставшихся мотивов выбирался наиболее значимый. Позиции этого мотива маскировались в выборке анализируемых последовательностей, и процесс оценки значимости оставшихся мотивов производился заново. Затем, среди найденных мотивов выявлялся следующий по значимости мотив, производилась маскировка позиций его расположения в выборке последовательностей, и цикл поиска значимых мотивов повторялся до тех пор, пока в анализе оставались мотивы, удовлетворяющие граничным критериям.

Предложенный алгоритм был реализован в виде компьютерной программы на языке CUDA. Программа может работать в операционных системах Windows и Linux и позволяет оценивать представленность в заданной выборке нуклеотидных последовательностей всех вырожденных олигонуклеотидных мотивов длины 8, записанных в 15-ти буквенном IUPAC коде. Программа обладает интерфейсом, в котором пользователь может задать границы окна в анализируемой выборке, граничный уровень значимости и представленности в выборке случайных последовательностей. Можно указать такие параметры случайной выборки как количество последовательностей в ней и необходимость использования частот нуклеотидов, характерных для анализируемой выборки последовательностей. Поиск может проводиться как в прямой цепи ДНК, так и в комплементарной. На вход программы подается выборка нуклеотидных последовательностей записанных в FASTA формате. На выходе - набор полученных вырожденных олигонуклеотидных мотивов удовлетворяющих заданным критериям.

Полученные результаты.

Программа Argo_CUDA была разработана для обнаружения мотивов de novo в массивных данных ДНК. Это исчерпывающий подход, который не так быстр, как уже существующие эвристические подходы. Однако он позволяет находить в массивных наборах последовательностей достаточно сильно вырожденные и плохо представленные мотивы. Благодаря использованию высокопроизводительных технологий графического процессора анализ реальных наборов данных ChIP-Seq может быть выполнен за разумное время. Argo_CUDA показывает хорошее качество поиска на смоделированных наборах данных.

Анализ последовательностей Foxa2 ChIP-Seq выявил как мотивы, соответствующие сайтам связывания транскрипционного фактора Foxa2, так и сайты связывания возможных кофакторов.

Иллюстрации, визуализация результатов (опционально).

Эффект от использования кластера в достижении целей работы.

Все основные расчеты проводились на суперкомпьютерном кластере НГУ.

Перечень публикаций, содержащих результаты работы.

1. O.V. Vishnevsky, A.V. Bocharnikov, N.A. Kolchanov ARGO_CUDA: Exhaustive GPU based approach for motif discovery in large DNA datasets. //Journal of Bioinformatics and Computation Biology, 2018, 16(1), Epub 2017 Dec 10. [23 pages].
2. O.V. Vishnevsky, A.V. Bocharnikov, N.A. Kolchanov ARGO_CEL: GPU Based Approach For Potential Composite Elements Discovery In Large DNA Datasets. The 3rd International Symposium “Mathematical modeling and high performance computing in bioinformatics, biomedicine and biotechnology” (MM-HPC-BBB-2018) p.71
3. А. В. Бочарников, Е. В. Игнатьева, О. В. Вишневский. Использование графических ускорителей для выявления функциональных сигналов в регуляторных районах дифференциально экспрессирующихся генов AGRP нейронов гипоталамуса мыши в ответ на голодание // Вестник СибГУТИ. 2019. № 3. С.36-44.
4. Бочарников А.В., Игнатьева Е.В., Вишневский О.В. Использование графических ускорителей для выявления функциональных сигналов в регуляторных районах генов прокариот //В сборнике: Наукоемкое программное обеспечение. труды семинара . 2019. С. 59-67.
5. O. Vishnevsky, A. Bocharnikov and N. Kolchanov, "GPU Based Composite Elements Discovery In Large DNA Datasets" 2020 Cognitive Sciences, Genomics and Bioinformatics (CSGB), Novosibirsk, Russia, 2020, pp. 135-138, doi: 10.1109/CSGB51356.2020.9214777.
6. BGRS/SB-2020: 12th International Multiconference “Bioinformatics of Genome Regulation and Structure/Systems Biology”, 06-10 July 2020, Novosibirsk, Russia.
7. O. V. Vishnevsky and A.V. Bocharnikov, New motif discovery approach, Marchuk Scientific Readings-2021: Abstracts of the International conference Marchuk, October 4–8, 2021. 174 p.
8. Rasskazov D, Chadaeva I, Sharypova E, Zolotareva K, Khandaev B, Ponomarenko P, Podkolodnyy N, Tverdokhleb N, Vishnevsky O, Bogomolov A, Podkolodnaya O, Savinkova L, Zemlyanskaya E, Golubyatnikov V, Kolchanov N, Ponomarenko M. Plant_SNP_TATA_Z-tester: a Web service that unequivocally estimates the impact of proximal promoter mutations on plant gene expression. *Int J Mol Sci.* 2022, 23: 8684. DOI: 10.3390/ijms23158684
9. Вишневский О.В., Чадаева И.В., Шарыпова Е.Б., Хандаев Б.М., Золотарева К.А., Казачек А.В., Пономаренко П.М., Подколодный Н.Л., Рассказов Д.А., Богомолов А.Г., Подколодная О.А., Савинкова Л.К., Землянская Е.В., Пономаренко М.П. Промоторы генов, кодирующих β-амилазу, альбумин и глобулин пищевых растений в сравнении с непищевыми, характеризуются более низкой аффинностью к TATA-связывающему белку: in silico анализ. *Вавиловский журнал генетики и селекции*. 2022;26(8):798-805. DOI 10.18699/VJGB-22-96

10. Вишневский О.В., Ворожейкин П.С., Титов И.И. Контекстные сигналы в митохондри-альных микроРНК млекопитающих. *Вавиловский журнал генетики и селекции*. 2022;26(8):819-825. DOI 10.18699/VJGB-22-99.