Отчет О.В.Вишневского.

Аннотация.

Развитие технологии ChIP-seq произвело революцию в генетическом анализе основных механизмов регуляции транскрипции и привело к накоплению информации об огромном количестве последовательностей ДНК. В настоящее время доступно множество веб-сервисов для обнаружения мотивов de novo в наборах данных, содержащих информацию о связывании ДНК/белков. Огромное разнообразие мотивов усложняет поиск. Чтобы избежать трудностей, исследователи используют различные стохастические подходы. К сожалению, эффективность программ обнаружения мотивов резко снижается с увеличением размера набора запросов. Это приводит к тому, что анализировать удается только небольшую часть последовательностей наиболее значимых пиков ChIP-Seq или приходится сужать область анализа. Таким образом, выявление мотивов в массивных наборах данных остается сложной проблемой. Разработанная нами система Argo_CUDA предназначена для обработки огромных данных ДНК. Это программа для обнаружения вырожденных олигонуклеотидных мотивов фиксированной длины, записанных в 15-буквенном коде IUPAC. Argo_CUDA — это комплексный подход, основанный на высокопроизводительных технологиях графических процессоров. По сравнению с существующими методами выявления мотивов, Argo_CUDA демонстрирует более высокое качество выявления контекстных сигналов на смоделированных наборах. С помощью разработанной нами системы были получены наборы олигонуклеотидных мотивов достоверно перепредставленнные в выборках последовательностей ChIP-Seq. На основе наборов выявленных мотивов строились уравнения множественной регрессии для предсказания высоты ChIP-seq пиков по их нуклеотидным последовательностям. В анализе использованы геномные последовательности ChIP-Seq пиков в районах связывания 8 ТФ (CEBPA, CEBPB, SP1, FOXA2, FOXO1, NFYA, MEF2D, STAT5B). На контрольных выборках была показана достоверная корреляции экспериментально полученных и предсказанных величин значимости ChIP-Seq пиков. Нами показано, что последовательности ChIP-Seq пиков обогащены не одиночными значимыми мотивами, а их наборами, имеющими достоверное сходство с сайтам связывания конкретного целевого ТФ. Кроме того, последовательности ChIP-Seq пиков обогащены специфичными наборами значимых мотивов, достоверно сходных с сайтами связывания других ТФ.

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., н.с. 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 [1]. 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).

К настоящему времени разработано большое количество компьютерных методов идентификации сайтов связывания ТФ в нуклеотидных последовательностях пиков ChIP-seq [2]. Однако до сих не разработаны компьютерные методы предсказания по нуклеотидным последовательностя пиков ChIP-seq их высоты, характеризующей достоверность присутствия в них сайтов связывания ТФ. Такой метод разработан нами в рамках данной работы [3].

Кроме того, 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.

Принципиальная новизна подхода заключалась в следующем. Вместо традиционных моделей для описания сайтов связывания ТФ в виде позиционных весовых матриц или консенсусов, использовалась модель представления сайта в виде набора коротких вырожденных мотивов, записанных в 15-буквенном IUPAC коде. Поиск мотивов, достоверно часто присутствующих в последовательностях ChIP-Seq пиков, осуществлялся с помощью пакета Argo_CUDA [3]. На основе набора выявленных мотивов строилось уравнение множественной регрессии для предсказания высоты ChIP-seq пиков по их нуклеотидным последовательностям. В анализе использованы геномные последовательности ChIP-Seq пиков в районах связывания 8 ТФ (CEBPA, CEBPB, SP1, FOXA2, FOXO1, NFYA, MEF2D, STAT5B), взятые из базы данных CistromeDB [4]. Обучающая выборка для каждого эксперимента ChIP-seq, соответствующего отдельному ТФ, включала 5000 геномных последовательностей (участок [-100,+100] относительно максимума пика ChIP-seq).

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.

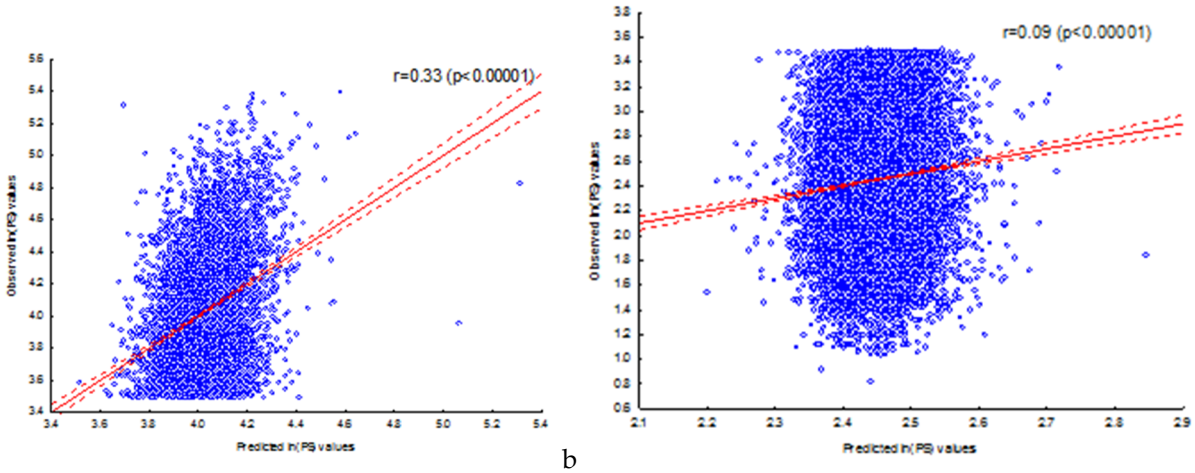
Предложенный алгоритм был реализован в виде компьютерной программы на языке CUDA. Программа может работать в операционных системах Windows и Linux и позволяет оценивать представленность в заданной выборке нуклеотидных последовательностей всех вырожденных олигонуклеотидных мотивов длины 8, записанных 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.

Для каждого из рассмотренных ТФ была выявлена группа вырожденных мотивов длиной 8 п.о., записанных в 15-буквенном IUPAC коде, достоверно часто ($p< 10^{-2}$) присутствующих в последовательностях его ChIP-Seq пиков. Количество таких значимых мотивов для различных ТФ варьировало от 270 до 151. Для каждого из 8 ТФ на основе набора выявленных мотивов было построено уравнение множественной регрессии, которое предсказывало высоту пика ChIP-Seq по наличию значимых мотивов, присутствующих в его нуклеотидных последовательностях. На контрольных выборках нуклеотидных последовательностей, не включенных в построение регрессионных моделей, была предсказана высота пиков ChIP-Seq для перечисленных выше ТФ. Далее было проведено сравнение теоретически рассчитанных и экспериментальных высот этих пиков. На всех 8-и контрольных выборках была показана достоверная корреляции экспериментально полученных и предсказанных величин (см. пример на рисунке 1). Новизна полученных результатов заключается в следующем: (1) впервые показано, что: а) последовательности ChIP-Seq пиков обогащены не одиночными значимыми мотивами, а их наборами, имеющими достоверное сходство с сайтам связывания конкретного целевого ТФ, кроме того (б) последовательности ChIP-Seq пиков обогащены специфичными наборами значимых мотивов, достоверно сходных с сайтами связывания других ТФ и (2) компьютерный метод предсказания высоты пиков ChIP-Seq разработан нами впервые и не имеет аналогов.



a b

Рисунок 1. Зависимость наблюдаемого значения логарифмированной величины значимости ChIP-seq пиков в обучающей (a) и контрольной (b) выборках FOXA2 от

ожидаемого значения. Ожидаемое значение значимости пиков было предсказано с использованием модели множественной регрессии на основе присутствия в них 202 значимых IUPAC мотивов, ранее выявленных в обучающей выборке FOXA2. Сплошные и пунктирные линии представляют линию регрессии и границы ее 95% доверительного интервала. r — коэффициент линейной корреляции, p — его статистическая значимость.

1. Bailey, T.; Krajewski, P.; Ladunga, I.; Lefebvre, C.; Li, Q.; Liu, T.; Madrigal, P.; Taslim, C.; Zhang, J. Practical guidelines for the comprehensive analysis of ChIP-seq data. PLoS computational biology 2013, 9, e1003326. https://doi.org/10.1371/journal.pcbi.1003326

2. 36. Heinz, S.; Benner, C.; Spann, N.; Bertolino, E.; Lin, Y. C.; Laslo, P.; Cheng, J. X.; Murre, C.; Singh, H.; Glass, C. K. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Molecular cell 2010, 38, 576–589. https://doi.org/10.1016/j.molcel.2010.05.004
3. Vishnevsky OV, Bocharnikov AV, Ignatieva EV. Peak Scores Significantly Depend on the Relationships between Contextual Signals in ChIP-Seq Peaks. Int J Mol Sci. 2024;25(2):1011. Published 2024 Jan 13. doi:10.3390/ijms25021011
4. Zheng, R.; Wan, C.; Mei, S.; Qin, Q.; Wu, Q.; Sun, H.; Chen, C. H.; Brown, M.; Zhang, X.; Meyer, C. A.; Liu, X. S. Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. Nucleic acids research 2019, 47, D729–D735. https://doi.org/10.1093/nar/gky1094

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. Vishnevsky OV, Bocharnikov AV, Ignatieva EV. Peak Scores Significantly Depend on the Relationships between Contextual Signals in ChIP-Seq Peaks. Int J Mol Sci. 2024;25(2):1011. Published 2024 Jan 13. doi:10.3390/ijms25021011