ВЫСШАЯ ШКОЛА ЭКОНОМИКИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Межвузовская студенческая научная школа-конференция

Информационные технологии и системы. Биоинформатика.

11-18 апреля 2020

Оглавление

11.04.2020

Структура ДНК

Гарницкий Марк Антонович - Распознавание паттернов ассоциации G-квадруплексов и гистоновых меток методами машинного обучения

Автор: Гарницкий Марк Антонович, НИУ ВШЭ ФКН ПМИ, бакалавриат 4 курс.
Научный руководитель: Попцова Мария Сергеевна
В этом проекте предлагается применение сверточных нейронных сетей для обнаружения мотивов в областях пересечения G-квадруплексов и гистоновых меток.
Мы предполагаем, что информация о паттернах ассоциации G-квадруплексов и гистоновых меток может быть получена из фильтров обученной сверточной нейронной сети для бинарной классификации. Положительный класс при этом будет состоять из отрезков последовательности нуклеиновых оснований фиксированной длины, содержащих и G-квадруплекс, и гистоновую метку. В данной работе будут рассмотрены гистоновые метки в тканях сердечно-сосудистой системы человека.
Планируется определение транскрипционных факторов, специфичных для тканей сердечно-сосудистой системы, а также участков регуляции с помощью квадруплексов, статистически значимо ассоциируемыми с гистоновыми метками и обнаруженными транскрипционными факторами.

Терентьева Юлия Андреевна - Анализ консервативных промоторных квадруплексов в геномах мыши и человека

Бекназаров Назар Сохибжонович - DeepZ: подход глубокого обучения для предсказания Z-ДНК

Константиновский Никита Олегович - Применение генеративно-состязательной нейронной сети для предсказания вторичных структур ДНК

Ностаева Арина - G-квадруплексы и эпигеномика

Рябых Григорий Кириллович - Разработка базы данных с веб-интерфейсом для хранения и сравнительного анализа РНК-ДНК интерактома
Рябых Г. К. аспирант 1 года обучения ФББ МГУ

Миронов А.А. проф. ФББ МГУ, д.б.н, д.ф-м.н

Известно большое количество некодирующих РНК: микроРНК, пиРНК, малые ядерные РНК, длинные некодирующие РНК, энхансерные РНК и другие. Они имеют различную клеточную локализацию и играют важную роль почти во всех процессах жизнедеятельности клетки. Например, одна из главных функций длинных некодирующих РНК – это регуляция экспрессии генов на разных уровнях, включая привлечение аппарата транскрипции, посттранскрипционные модификации и эпигенетику. Одни РНК могут взаимодействовать с хроматином «цис» (например, РНК XIST инактивирует Х-хромосому, на которой сама и закодирована [1]), другие (например, MALAT1 и NEAT1 [2]) способны образовывать контакты с соседними хромосомами – «транс».
На сегодняшний день существует несколько методов, с помощью которых можно определить полногеномную локализацию одной РНК на хроматине: ChIRP [3], CHART [4], RAP-DNA [5]. Однако эти методы позволяют анализировать только одну известную РНК за один эксперимент, и, следовательно, они не дают возможности полногеномно посмотреть на взаимодействия всех РНК с ДНК. Однако в 2017 году появились первые работы, которые предлагают методы, позволяющие получить данные обо всех потенциальных РНК-ДНК контактах в клетке: MARGI [6] и GRID-seq [7] и другие.
Данная работа посвящена разработке базы данных, предназначенной для накопления данных РНК-ДНК контактов, их быстрого и удобного анализа. За основу мы взяли колоночно-

ориентированную систему управления базами данных (СУБД) ClickHouse, позволяющую выполнять аналитические запросы в режиме реального времени на больших данных.
На данный момент мы можем работать с полногеномными данными РНК-ДНК контактов (данные GRID[7], iMARGI[9], MARGI[6], ChAR-seq[10]; данные, полученные от наших коллабораторов из лаборатории Сергея Владимировича Разина, клеточная линия K562), а также с результатами экспериментов RAP[5] для длинной некодирующей РНК XIST.
Разработанная нами база данных предоставляет пользователю:
● строить интерактивные профили контактов единичной РНК, определенной группы или всех РНК из заданного локуса, по всему геному или выбранному интервалу
● строить тепловую карту распределения контактов единичной РНК, определенной группы или всех РНК из заданного локуса, по всему геному
● строить распределение плотности контактов единичной РНК, определенной группы или всех РНК из заданного локуса, в

зависимости от расстояния между кодирующим соответствующую РНК геном и локусом, с которым она контактирует в эксперименте
● получить список РНК (по которому можно осуществлять поиск и применять различные фильтры) с метаинформацией и исходными/нормированными суммарными значениями их контактируемости в выбранных экспериментах
● получать метаинформацию по выбранным экспериментам, их препроцессингу, количество и долю разных типов РНК, по сравнению с специально собранной аннотацией РНК

Кроме этого мы разрабатываем связанный с базой данных веб-интерфейс, который позволит пользователю формировать треки

контактов из разных экспериментов по одной\группе РНК, которые можно будет визуализировать в Genome Browser или построить интерактивные графики, описанные выше, а также выбирать только те контакты, которые, например, попадают в тела/upstream-, downstream-области всех или конкретных генов или в указанный локус.
В планах у нас:
● закончить сбор необходимых для анализа данных
● доделать и сделать общедоступным веб-интерфейс для базы данных
● сделать возможным выкачивание наших данных, если пользователь захочет провести свою нормировку, фильтрацию или любой другой анализ

Эта работа актуальна, так как аналитической базы данных, содержащей все имеющиеся данные РНК-ДНК взаимодействий, препроцессированные единым образом, с веб-интерфейсом еще не существует, и она позволит глобально взглянуть на данные РНК-ДНК контактов, а также провести масштабный и быстрый анализ некодирующих РНК.

Источники и литература

1) M. D. Simon, S. F. Pinter, R. Fang, K. Sarma, M. Rutenberg-Schoenberg, S. K. Bowman, B. A. Kesner, V. K. Maier, R. E. Kingston, and J. T. Lee, “Highresolution Xist binding maps reveal two-step spreading during X-chromosome inactivation,” Nature, vol. 504, no. 7480, pp. 465–469, 2013.
2)J. A. West, C. P. Davis, H. Sunwoo, M. D. Simon, R. I. Sadreyev, P. I. Wang, M. Y. Tolstorukov, and R. E. Kingston, “The Long Noncoding

RNAs NEAT1 and MALAT1 Bind Active Chromatin Sites,” Mol. Cell, vol. 55, no. 5, pp. 791–802, 2014.

3) Chu, C., Qu, K., Zhong, F.L., Artandi, S.E. & Chang, H.Y. Genomic maps of long noncoding RNA occupancy reveal principles of RNA–chromatin interactions. Mol. Cell 44, 667–678 (2011).
4) Simon, M.D. et al. The genomic binding sites of a noncoding RNA. Proc. Natl. Acad. Sci. USA 108, 20497–20502 (2011).
5) Engreitz, J.M. et al. The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. Science 341, 1237973 (2013).
6) Sridhar B, Rivas-Astroza M, Nguyen TC, Chen W, Yan Z, Cao X, Hebert L, Zhong S. Systematic mapping of RNA-chromatin interactions in vivo. Curr Biol. 2017;27:602–609.
7) Li, X., Zhou, B., Chen, L., Gou, L.-T., Li, H., and Fu, X.-D. (2017). GRID-seq reveals the global RNA-chromatin interactome. Nat Biotechnol.
8) Djebali, S. et al. Landscape of transcription in human cells. Nature 489, 101–108 (2012).
9) Yan, Z. et al. Genome-wide co-localization of RNA-DNA interactions and fusion RNA pairs. PNAS February 19, 2019
10) J. C. Bell, et. al. “Chromatin-associated RNA sequencing (ChAR-seq) maps genome-wide RNA-to-DNA contacts,” Elife, vol. 7, pp. 1–28, 2018

Козюлина Светлана Вадимовна - Создание инструмента для визуализации взаимодействий в молекулярной биологии на основе графического редактора Inkscape

Структура хроматина и эпигеномика

Храмеева Екатерина Евгеньевна - Анализ данных Hi-C и других омиксных типов данных для изучения трехмерной организации хроматина, и не только

Черкасов Александр Вадимович – Исследование пространственной структуры хроматина немодельных живых организмов

Смирнов Дмитрий Николаевич - Topologically Associating Domain optimal set prediction using Armatus software

Максимов Владислав Сергеевич - Анализ и восстановление пространственных взаимодействий ДНК индивидуальных клеток

Максимов Владислав Сергеевич, студент 5 курса ФББ МГУ
Куратор: Галицына Александра Алексеевна
Сколковский Институт Наук и Технологий, ИППИ РАН, ИБГ РАН

Метод Hi-C позволяет получить информацию о взаимодействии каждого локуса со всеми остальными при исследовании контактов генома [1]. Компартменты открытого и закрытого хроматина обычно ищут по данным Hi-C методом разложения по собственным векторам [2]. Для выявления скрытых геномных характеристик, в частности, компартментов хроматина, использовалось представление Hi-C в виде графа контактов и последующим преобразованием в векторное пространство [3].
Однако для частного случая Hi-C индивидуальных клеток (scHi-C) традиционный подход к поиску компартментов невозможен из-за разреженности данных [4]. Также нерешенной проблемой в области обработки scHi-C является фильтрация шума и дополнение контактов. Теоретически обе эти задачи могут быть решены с помощью графовых автокодировщиков[5]. Задачей данного исследования является тестирование этого подхода для взаимодействия хроматина индивидуальных клеток.
В данной работе мы использовали данные Hi-C одиночных ядер (snHi-C) ооцитов мышей с количеством контактов на клетку порядка 10 6 [4]. Пространственные взаимодействия представили в виде графа, в

котором вершины - участки генома, а ребра - контакты между ними. Каждую хромосому мы по отдельности отобразили в эмбединговое пространство с помощью вариационного графового автокодировщика с двумя слоями [6]. Далее мы считали, что контакт между двумя участками генома существует, если их сходство в эмбединговом пространстве выше порога. Все вычисления проводились на облачных ресурсах сервера Google Colab.
Средняя точность восстановления контактов на тестовом наборе данных составила порядка 85%. Получили из эмбедингового пространства дополненную карту взаимодействий.
1. Schmitt, A., Hu, M. & Ren, B. Genome-wide mapping and analysis of chromosome architecture. Nat Rev Mol Cell Biol 17, 743–755 (2016)
2. Imakaev, Maxim, et al. "Iterative correction of Hi-C data reveals hallmarks of chromosome organization." Nature methods 9.10 (2012)
3. Ashoor, H. et al. Graph embedding and unsupervised learning predict genomic sub-compartments from HiC chromatin interaction data. Nat Commun 11, 1173 (2020)
4. Flyamer, I., Gassler, J., Imakaev, M. et al. Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. Nature 544, 110–114 (2017)
5. Spinelli, Indro, et al. "Efficient data augmentation using graph imputation neural networks." arXiv preprint arXiv:1906.08502 (2019)
6. Kipf, Thomas N., and Max Welling.Variational graph auto-encoders. arXiv preprint arXiv:1611.07308 (2016)

Жегалова Ирина Владимировна - Особенности петель хроматина Dictyostelium discoideum

И. Жегалова 1 , О. Цой 2 , С. Стариков 3 , А. Савостьянов 4 , А. Галицына 3 , С.
Ульянов 1 , Е. Храмеева 3,5
1- МГУ им. М.В. Ломоносова
2 - Технический университет Мюнхена
3 - Сколковский институт науки и технологий
4 - Высшая школа экономики
5 - Институт проблем передачи информации им. А.А. Харкевича
e.mail: ir.zhegalova@fbb.msu.ru

Дихтиостелиум – одноклеточная почвенная амёба, при определённых условиях образующая многоклеточные плодовые тела сложного строения. Его хроматин образует петли, но не образует топологически ассоциированных доменов, свойственных геномам млекопитающих [1]. Чтобы исследовать пространственные изменения генома в течение жизненного цикла дихтиостелиума, был проведен эксперимент Hi-C с двумя биологическими репликами в разные моменты времени (0, 2, 5, 8 часов). Также для соответствующих временных точек были сделаны RNA-seq и ATAC-seq. Размер бина для всех экспериментов после обработки - 2 килобазы.
Все данные прошли предобработку, включая фильтрацию по качеству и нормализацию при необходимости для данных Hi-C. Аннотация петель была выполнена двумя различными алгоритмами, разработанными сотрудниками лаборатории, а петли при 0 ч. имели также ручную аннотацию. Дифференциальная экспрессия генов была проанализирована с помощью алгоритма edgeR GLM. Гены с четырехкратным изменением уровня экспрессии между стадиями и с FDR <0,05 были выбраны как дифференциально экспрессирующиеся.
Количество кластеров экспрессии было идентифицировано с помощью t-SNE. При перплексии 50 шесть крупных кластеров экспрессии

было определено. Большие скопления были разделены так, что в итоге получилось десять кластеров. Каждый кластер был аннотирован с использованием Gene Onthology. Были отмечены изменения уровней экспрессии для кластеров в различных временных точках.
Касательно изучений изменений, происходящих на границе петель, как было отмечено ранее [2], транскрипция генов симметрично направлена в сторону границ петли. Кроме того, было обнаружено, что плотность кодирующих последовательностей, а также уровень экспрессии в бине, содержащем границу петли резко возрастает. Корреляции между экспрессией генов в петлях и количеством контактов не было обнаружено, однако, наблюдается тенденция падения insulation score на границе. GC-

состав на границе петли резко возрастает, обнаружены крупные (~24 пн) полиадениновые повторы. Кроме того, в открытом хроматине около границ петель обнаружены мотивы, аннотированные для других организмов как имеющие функции поддержания пространственной структуры. Дальнейшие исследования требуются для изучения функций факторов, мотивы которых обнаружены на границах петель у дихтиостелиума, а также прочих функциональных и структурных особенностей петель хроматина. Выполненные исследования поддержаны грантом [3].

1- Dixon J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions //Nature. – 2012. – Т. 485. – №. 7398. – С. 376.
2- Tsoy O. et al. The chromatin structure of Dictyostelium discoideum //2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). – IEEE, 2018. – С. 2492-2492.
3- Грант Российского научного фонда № 19-74-00112.

Галицына Александра Алексеевна - Структура хроматина: видовое разнообразие и развитие

Быков Николай Сергеевич - HiChew: консольная программа для поиска ТАДов и их кластеризации в эмбриогенезе

Плискин Александр Маркович - Предсказание петель в хроматине Dictyostelium discoideum

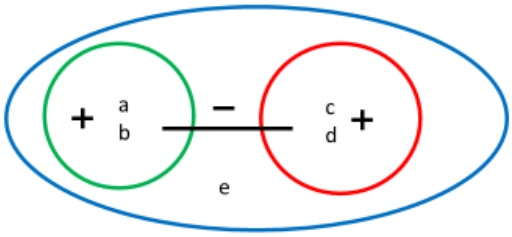
Алишев Наиль Анварович - Generation of Hi-C maps from DNA sequence data using Deep Learning

Федоров Александр Николаевич - Приложение машинного обучения к проблеме гетерогенности ChIP-seq данных

Транскриптомы и РНК

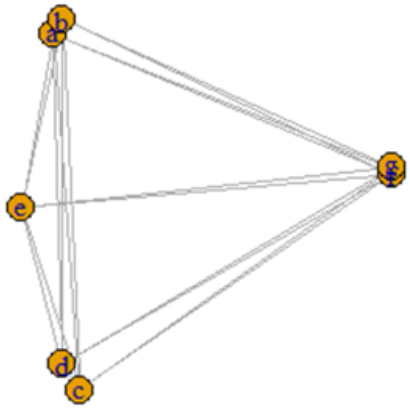
Самосюк Алексей Владимирович - Быстрая пермутационная оценка геной ко-экспрессии с использованием метода случайной проекции на данных одноклеточного секвенирования

- Differentially expressed genes are positively co-localized in each cluster and mutually negative between each other



	a	b	c	d	e	f	g
a	+	+	-	-	+	-	-
b	+	+	-	-	+	-	-
c	-	-	+	+	+	-	-
d	-	-	+	+	+	-	-
e	+	+	+	+	+	-	-
f	-	-	-	-	-	+	+
g	-	-	-	-	-	+	+

	1	2	3	4	5	6
a	1	1				
b	1	1				
c			1	1		
d			1	1		
e	1	1	1	1		
f					1	1
g					1	1



Методы построения сети ко-экспрессии для данных одноклеточного секвенирования корректно работают для генов с высокими уровнями экспрессии (основной фенотип) но показывают нестабильный результат для сильно гомогенных клеточных состояний когда отличия определяются небольшим набором из 100-200 генов, не являющихся сильноэкспрессирующимися

Мы предлагаем метод пермутационной оценки ко-экспрессии, позволяющий быстро находить такие генные наборы в полу-автоматическом режиме без использования приорных знаний о датасете

Шаймарданов Абусаид Муратович - Investigation of the Evolution of Gene Expression in the Brain Tissue of Primates

Муртазалиева Халимат Асадулаевна - cCFE: a database of chemical compound-based direct reprogramming and differentiation experiments

Сизых Алексей Дмитриевич - cCFE: a database of chemical compound-based direct reprogramming and differentiation experiments

Плетенев Илья Андреевич - ПЦР-дубликаты в РНК-сек: удалить нельзя оставить

Бобровский Даниил Максимович - Изменения транскриптома, метаболома и липидома в прелимбической коре мозга макак-резус под влиянием антидепрессанта флуоксетин

Камышева Анна Леонидовна - Изучение эволюции липидома мозга человека: выявление человеко-специфичных липидов

Озерова Александра Михайловна - Особенности индивидуального развития насекомых с полным превращением на уровне экспрессии генов

Валяева Анна Александровна - Биоинформатический анализ полу-экстрагируемых РНК

Валяева Анна Александровна, МГУ им. Ломоносова; Шеваль Евгений Валерьевич, Миронов Андрей Александрович

Стандартные методы экстракции РНК не позволяют выделить всю клеточную РНК: определенные фракции РНК остаются в связанном состоянии и вместе с белками осаждаются из раствора либо по другим причинам плохо детектируются. Усиленный дополнительной механической гомогенизацией метод экстракции РНК позволяет обогатить выделяемую фракцию так называемыми полу-экстрагируемыми РНК, которые могут быть как белок-кодирующими, так и некодирующими РНК, в том числе архитектурными РНК (например, NEAT1).
В данной работе мы провели эксперимент по выделению РНК стандартным методом и методом с дополнительной гомогенизацией клеточного лизата в шприце (использовалась клеточная линия HeLa). Выделение РНК проводилось на колонках и с помощью Тризола, затем было проведено секвенирование образцов. Для поиска РНК, проявляющих свойство полу-экстрагируемости, был адаптирован пайплайн для анализа

дифференциально экспрессируемых генов. Полученные риды были картированы на референсный геном человека GRCh38 с помощью программы HISAT2 и уникально картированные на гены (а также на экзоны и интроны) риды были подсчитаны с помощью программы HTSeq count. Анализ дифференциальной экстракции РНК (по аналогии с дифференциальной экспрессией) был произведен с помощью пакета R DESeq2.
Проведенный анализ показал, что РНК 1083 генов проявляют свойство полу-экстрагируемости (ДЭ, padj<0.05, FC>1.5). В 1.5 и более раз эффективнее при усиленной экстракции выделились РНК 536 белок-кодирующих генов и 497 длинных некодирующих РНК (в том числе NEAT1 и MALAT1). GO анализ, однако, не позволил выделить какие-то закономерности в функциональных ролях продуктов найденных белок-кодирующих РНК. Вероятно, свойство полу-экстрагируемости связано с какими-то особенностями молекул РНК, а не с тем, что эти молекулы кодируют.
Нами было замечено, что суммарная длина интронных участков у ДЭ генов больше, чем у неДЭ. Поэтому далее мы попытались проанализировать, в каких случаях полуэкстрагируемость определялась интронами, в каких экзонами. Для этого был проведен анализ дифференциальной экспрессии/экстракции отдельно по ридам, картирующимся на интроны или экзоны генов, в результате которого были получены соответствующие списки ДЭ генов. Оказалось, что в большинстве случаев полуэкстрагируемость генов связана с полуэкстрагируемостью именно интронов (634 гена отличались по покрытию интронов, FC>1.5).
В настоящее время разрабатываем альтернативный пайплайн поиска полу-экстрагируемых РНК, который бы разрешил проблему множественного картирования ридов и позволил бы, например, проанализировать влияние повторов на свойство полу-экстрагируемости РНК.

Менделевич Ася Владимировна - Unexpected variability of allelic imbalance estimates from bulk and scRNA sequencing
Asia Mendelevich. Skolkovo Institute of Science and Technology, Moscow, Russia. Svetlana Vinogradova, Saumya Gupta, Andrey A. Mironov, Shamil Sunyaev, Alexander A. Gimelbrant

RNA sequencing and other experimental methods that produce large amounts of data are increasingly dominant in molecular biology. However, the noise properties of these techniques have not been fully understood.
We assessed the reproducibility of allele-specific expression measurements by conducting replicate sequencing experiments from the same RNA sample. Surprisingly, variation in the estimates of allelic imbalance (AI) between technical replicates was up to 7-fold higher than expected from commonly applied noise models.

We show that AI overdispersion varies substantially between replicates and between experimental series, appears to arise during the construction of sequencing libraries, and can be measured by comparing technical replicates. We demonstrate that compensation for AI overdispersion greatly reduces technical variation and enables reliable differential analysis of allele-specific expression across samples and across experiments. Conversely, not taking AI overdispersion into account can lead to a substantial number of false positives in analysis of allele-specific gene expression.

Транкова Наталья Аркадьевна - Малые бактериальные некодирующие РНК как антисенс-регуляторы экспрессии генов

Дрождев Алексей Иванович - Анализ экспрессионных данных Escherichia coli при помощи нейросетей
ФББ МГУ им. Ломоносова. Червонцева Зоя Сергеевна

Автоэнкодеры представляют собой тип нейросетей, которые должны на выходе восстановить данные, поданные на вход, однако в структуре сети существует ограничение, не позволяющее ей просто запомнить входные значения, и вынуждающие её каким-то образом структурировать их и искать закономерности. Данный подход был применен для анализа экспрессий генов Escherichia coli, была проверена гипотеза, что нейросеть сможет таким образом усвоить оперонную структуру. Ещё одна разновидность автоэнкодеров, concrete autoencoder, позволяет найти наиболее важные входные параметры, по значениям которых можно восстановить значения остальных параметров с хорошей точностью. При помощи concrete autoencoder были найдены наборы генов, по экспрессиям которых лучше всего восстанавливаются экспрессии остальных, и проведен их анализ.

Червонцева Зоя Сергеевна - Роль вторичной структуры мРНК бактерий в процессе трансляции

Григорашвили Елизавета Игоревна - Применение нейронных сетей для нахождения признаков, важных для фолдинга РНК

Григорашвили Елизавета Игоревна, Сколтех, Червонцева Зоя Сергеевна
Современные алгоритмы для предсказания вторичной структуры РНК имеют ограничения, обусловленные в том числе недостаточным пониманием связи между сворачиванием РНК и ее последовательностью. Применение нейронных сетей может помочь выявить неизвестные правила фолдинга РНК, поскольку создаваемые при обучении нейросетей представления часто отражают скрытые структуры в данных.
Чтобы понять, как последовательность влияет на вторичную структуру РНК, мы сформулировали следующую постановку задачи.

Структура РНК задается порядком входящих в нее нуклеотидов. Если мы нарушим нуклеотидную последовательность, то мы можем ожидать и нарушение структуры. Нейросеть, обученная отличать последовательность РНК от последовательности, в которой порядок нуклеотидов изменен, может обнаружить важные для образования структуры позиции и/или нуклеотиды. Для решения этой задачи мы взяли датасет из последовательностей тРНК и подготовили его “дупликат”, в котором в каждой последовательности нуклеотиды в 5 случайных парах были поменяны местами. Затем мы обучили рекуррентную нейросеть определять, какая тРНК подается ей на вход (нормальная или измененная) и проанализировали, какие участки последовательности влияли на предсказание сильнее всего. Мы обнаружили, что нуклеотиды, вовлеченные во вторичную структуру тРНК, имеют большее значение для предсказания. Дальнейшая работа направлена на проверку того, что больше влияет на предсказание: вовлеченность в элементы вторичной структуры или консервативность нуклеотида.
Параллельно мы разработали иной подход. Нейросеть, предсказывающая пару для определенного нуклеотида в последовательности, при обучении может сформировать представления, отражающие важные для формирования вторичной структуры признаки. Для решения этой задачи мы расширили датасет, не ограничиваясь лишь тРНК, и сконструировали рекуррентную нейросеть. При анализе результатов мы выявили, что наша модель в части случаев делает очень точное предсказание. Также мы сравнили результат предсказания нашей модели с экспериментальными данными по доступности нуклеотидов в РНК. Сравнение выявило согласованность наших данных с экспериментальными: нуклеотиды, предсказанные как спаренные нашей моделью, имеют тенденцию быть недоступными в эксперименте. Дальнейшая работа по этой задаче направлена на подбор параметров модели, которые повысят долю случаев точного предсказания и обеспечат надежное предсказание пары G-U.

12.04.2020

Сплайсинг

Калмыкова Светлана Дмитриевна - Conserved panhandle RNA structures are associated with splicing markup and end-processing of vertebrate genes

Svetlana Kalmykova 1 *, Timofei Ivanov 1 , Stepan Denisov 1 , Roderic Guigo 2 , and

Dmitri D. Pervouchine 1,3

1 Skolkovo Institute for Science and Technology, Moscow 143025, Russia
2 Center for Genomic Regulation and UPF, Barcelona 08003, Spain
3 Moscow State University, Moscow 119991, Russia

Eukaryotic genes are expressed as single-stranded RNA molecules that fold into complicated secondary and tertiary structures. Remarkably, many of these structures contain long stretches of complementarity nucleotides with the free energy of interaction exceeding that of the folding of large protein domains. In this work, we present an extended catalog of the so-called panhandle RNA structures, e.g., pairs of conserved complementary regions (CCR) in the human protein-coding transcriptome that may interact over long distances. These structures are associated with RNA processing signals and footprints of RNA-binding proteins (RBP), depleted of population polymorphisms, show evidence of compensatory evolution, and are supported by RNA structure probing data. Global trends in the positioning of these structures suggest that, together with RBP binding sites, they constitute a global network of regulatory signals that define gene-level splicing markup, including RNA bridges and double-stranded regions that approximate distant exons or suppress cryptic splice sites. These regulatory RNA structures are highly dynamic and form co-transcriptionally depending on the elongation rate of RNA Pol II. An intriguing observation is that, in addition to previously known association with intron borders, panhandle

RNA structures also tend to loop out transcript ends, suggesting their involvement in 5’- and 3’-end processing. We hypothesize that, at least in some genes, RNA structure may serve as an intra-molecular crosslink that controls the dynamic competition between pre-mRNA splicing, cleavage, and polyadenylation.

Власенок Мария Александровна - De novo identification of alternative polyadenylation from RNA-seq data

Данчурова Анастасия Александровна - Авто- и кросс-регуляция белка SRSF7 посредством альтернативного сплайсинга, сопряженного с нонсенс-опосредованным распадом

Анастасия Александровна Данчурова

(Сколковский институт науки и технологии, Магистерская программа “Науки о жизни”, 2 курс.)
Логвина Наталья Александровна
Зацепин Тимофей Сергеевич
Первушин Дмитрий Давидович

В результате нонсенс-мутаций и аберрантного сплайсинга в образующихся транскриптах могут появляться преждевременные стоп-кодоны, что может привести к синтезу укороченных и дисфункциональных белков. Чтобы это предотвратить, в эукариотических клетках работает особая система контроля, известная как нонсенс-опосредованный распад (nonsense-mediated decay, NMD), которая способна распознавать и разрушать такие транскрипты. Однако, согласно последним исследованиям, NMD в сочетании с альтернативным сплайсингом (AS-NMD) может представлять собой не только систему защиты от ошибок сплайсинга, но и общий механизм регуляции экспрессии генов. В частности, многие РНК связывающие белки взаимодействуют с пре-мРНК и влияют на альтернативный сплайсинг, приводя к образованию транскрипта с преждевременным стоп-кодоном. Несколько новых подобных авторегуляторных механизмов были недавно предсказаны для человеческих РНК связывающих белков SRSF7, SFPQ, U2AF1 и RPS3. Кросс-регуляторные взаимодействия AS-NMD, в которой РНК связывающий белок регулирует сплайсинг и нонсенс-опосредованный распад пре-мРНК другого белка, изучен в гораздо меньшей степени. В этой работе мы используем транскриптомный анализ публичных данных по нокдаунам факторов NMD и большой панели РНК связывающих белков, данных eCLIP большой панели РНК связывающих белков и анализ коэкспрессии генов, чтобы предсказать предполагаемые кросс-регуляторные связи. Мы сосредоточились на регуляторной подсети фактора сплайсинга SRSF7, который играет важную роль в процессах регуляции апоптоза в некоторых типах рака. С этой целью мы разработали экспериментальную систему для биологической валидации регуляторной подсети SRSF7, которая включает в себя нокдаун и оверэкспрессию SRSF7 в клеточной линии HEK293 в комбинации с нокдауном факторов системы NMD. Эта работа демонстрирует, что только небольшая часть кросс-регуляторных взаимодействий AS-NMD была изучена до настоящего времени и что кросс-регуляция через AS-NMD может быть гораздо более распространенным механизмом, чем считалось ранее.

Микова Валерия Михайловна - A study of intronic polyadenylation and splicing in hepatocellular carcinoma

Recent studies have shown that intronic polyadenylation (IPA), i.e. premature termination of transcription, is frequently observed in diverse cancer types and can mimic the functional outcome of genetic alterations that lead to truncated proteins. In particular, these products may lack tumor suppressor functions, which they otherwise would have had in the case of translating full-length transcripts. In this project we used public cancer data sources to identify IPA events that are associated with hepatocellular carcinoma. Current approaches use the combination of 3-'seq and poly(A)-seq data to identify IPA.
We identified tumor-associated IPA events using RNA-seq data alone, namely by extracting short reads that contain poly(A)-stretches that partially align to the genome. For found IPA events we investigated association with read coverage changes and splicing of intron, where alternative polyadenylation takes place. To identify the mechanisms that drive alternative polyadenylation we investigated cis-matational events, i.e mutations that are potentially related to alternative polyadenylation in particular genes, and trans-mutational events, i.e mutations and expression changes in RBPs, that are potentially responsible for global changes in polyadenylation landscape.

Мазаев Лев Сергеевич - Regulatory role of upstream open reading frames in NMD-dependent degradation of human mRNAs

Раменский Василий Евгеньевич - Альтернативный сплайсинг и положительный отбор revisited

Василий Евгеньевич Раменский (Национальный медицинский исследовательский центр терапии и профилактической медицины), Андрей Александрович Миронов (Факультет биоинженерии и биоинформатики МГУ им. Ломоносова), Михаил Сергеевич Гельфанд

(Институт проблем передачи информации РАН)

В работе 2008 года на основе анализа межвидовых замен и полиморфизма нами было показано, что альтернативно сплайсируемые экзоны человека подвержены действию положительного отбора. Тем самым впервые был установлен тип участков генома, эволюционирующих под действием такого отбора. Эта работа была сделана незадолго до распространения методов высокопроизводительного секвенирования, и в силу небольшого объема данных по полиморфизму человека статистическая значимость полученных результатов была ограничена. Было бы интересно

повторить эту работу, используя современные данные об альтернативном сплайсинге и полиморфизме у человека, а также изучить особенности альтернативного сплайсинга генов в зависимости от степени их гапло-недостаточности, определенной с помощью новейших популяционных данных.

Ссылка: Ramensky, V.E., Nurtdinov, R.N., Neverov, A.D., Mironov, A.A., and Gelfand, M.S. (2008). Positive Selection in Alternatively Spliced Exons of Human Genes. Am J Hum Genet 83, 94–98.

Рак и другие болезни

Барановский Артем - Transcriptome analysis reveals high tumor heterogeneity with respect to re-activation of stemness and proliferation programs

Berlin Institute for Medical Systems Biology, Berlin, Germany
Первушин Дмитрий
Папаценко Дмитрий
Абстракт:
Significant genetic, epigenetic, hence functional alterations in signaling pathways and transcriptional regulatory programs together represent the hallmarks of most human cancers. These among all encompass the reactivation of “stemness”, registered by the expression of embryonic stem cell (ESC) marker molecules: Pou5f1, Sall4 and Sox2. Master regulators of pluripotency were reported to be highly expressed in particularly aggressive tumours, while they were absent in both, low-grade malignancies and normal tissues. Here, we analyzed a large panel of RNA-seq data from The Cancer Genome Atlas (TCGA) Consortium in order to specifically reveal the expression of (pluripotency) stemness-related and proliferation-related genes across the collection of different tumor types. Using a novel metric that captures the similarity in the expression profile of a tumor to that of the ESCs, we showed that the intensity of the stemness signature varies greatly between different tumor types, suggesting that the expression of stem cell markers is not a universal determinant of cancer. We observed a high degree of variation in the expression of pluripotency- and proliferation-related genes not only inter-, but also intratumorally, which was independent of tumor heterogeneity and had an inconsistent association with tumor aggressiveness, higher hazard ratios, and prognosis. Therefore, not all tumors are comparable in terms of reactivation of stemness and proliferation programs, contravening the current consensus.

Шпудейко Полина Сергеевна - Изучение репертуаров Т-клеточных рецепторов лимфоцитов, инфильтрирующих опухоль

Алексеева Евгения - Филогенетический анализ В-клеточных линий человека

Алексеева Евгения Искандеровна (1), Микелов Артем Ильич (1, 2), Шугай Михаил
Александрович (1, 2), Звягин Иван Владимирович (2), Базыкин Георгий Александрович (1)
1 Сколтех, Москва, Россия
2 Институт Биоорганической Химии им. Академиков М. М. Шемякина и Ю. А.
Овчинникова Российской Академии Наук, Москва, Россия

Для производства антител с высоким сродством к антигену расширение активированных
В клеточных клональных групп сопряжено с эволюционным процессом созревания
аффинности. Помимо этого с развитием иммунного ответа В клетки дифференцируются в
плазмобласты и плазматические клетки для активного производства антител, а также
переключают изотип антитела. Степень дифференцировки В клеток и представленность
изотипов в клональной группе отражает ее стадию в иммунном ответе, однако
эволюционные механизмы, стоящие за развитием клональной группы мало изучены. В
данной работе мы представляем филогенетический анализ В клеточных репертуаров,
полученных за три забора периферической крови из пятерых доноров в течение одного
года. В репертуарах выделены три клеточные фракции - В клеток памяти, плазмобласты и
плазматические клетки и установлены изотипы производимых иммуноглобулинов. Мы
наблюдаем, что действие отбора и численная динамика клональных групп зависит от их
клеточного состава и производимого изотипа. Клональные группы, прошедшие через В
клеточную дифференцировку и переключение изотипа, находятся под действием
положительного отбора и резко меняют свою численность в репертуаре, в то время как
клональные группы, преимущественно состоящие из В клеток памяти, не переключают
изотип и эволюционируют нейтрально. Данные результаты свидетельствуют о разных

эволюционных механизмах в клональных группах в режимах иммунной памяти или
активного иммунного ответа и требуют продолжения исследования.

Гурылева Мария Вячеславовна - IndieForest: машинное обучение на индикаторах и рангах для ранней диагностики онкогенных заболеваний

Московский государственный университет имени М.В.Ломоносова,
Факультет биоинженерии и биоинформатики, Москва, Россия
Пензар Дмитрий Дмитриевич, Фаворов Александр Владимирович,

Миронов Андрей Александрович

Онкологические заболевания характеризуются высокой скоростью прогрессирования. Поэтому особенно важна их ранняя диагностика. Методы, используемые для диагностики сейчас, сильно зависят от человеческого фактора, а также ограничены разрешающей способностью приборов. Анализ экспрессий генов может помочь детекции рака на ранних стадиях. Для такой задачи наиболее подходящим подходом видится использование алгоритмов машинного обучения.[1]
Один из таких алгоритмов - случайный лес (RF) - хорошо зарекомендовал себя в биоинформатических задачах.[2] Лес — это голосование большого числа решающих деревьев, каждое из которых строится на случайной подвыборке образцов и переменных. Решающее дерево состоит из набора элементарных решений, которые сравнивают значения переменных с порогами. Пороги подбираются при обучении и становятся частью классификатора. Из-за этого классифицируемые данные экспрессии надо нормализовать вместе с обучающей выборкой, что ограничивает применимость RF.
В то же время существует семейство непараметрических методов, основанных на попарных сравнениях экспрессий генов внутри одного образца, что делает их независящими от монотонной нормализации.[3]

Цель данного проекта — соединить два подхода и использовать для предсказания различных типов рака RF, обученный на результатах попарных сравнений экспрессий генов образца.
Данная идея была реализована на языке программирования R с использованием пакета randomFosest и набора пакетов для обработки данных tidyverse. Тестирование моделей проводилось на данных из баз данных TCGA (https://portal.gdc.cancer.gov/) и GEO (http s://www.ncbi.nlm.nih.gov/geo/). Для избежания переобучения отбирались наиболее вариабельные дифференциально экспрессирующиеся гены. Далее рассматривались три метода: классический - основанный на стандартных показателях экспрессии генов, ранговый - показатели экспрессии переводились в ранги, метод индикаторов - в качестве

признака использовалась величина I[A, B], равная 1 если A > B и -1 иначе. Сравнение результатов для бинарной классификации проводилось относительно алгоритма K-Top-Scoring-Pair (KTSP), реализованного в пакете switchBox.[3]
На бинарной классификации было показано, что уже на 100 генах и индикаторный, и ранговый методы превосходят алгоритм KTSP по таким показателям, как accuracy и ROCauc. Было показано, что данные методы применимы и для множественной классификации различных типов рака, в отличие от стандартного KTSP.
Источники и литература
1) Ming, C., Viassolo, V., Probst-Hensch, N. et al. Machine learning techniques for personalized breast cancer risk prediction: comparison with the BCRAT and BOADICEA models. Breast Cancer Res 21, 75 (2019). https://doi.org/10.1186/s13058-019-1158-4 1 Конференция «Ломоносов 2020» 2) Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/ A:1010933404324

3) Afsari, B., Fertig, E.J., Geman, D., Marchionni, L., 2015. switchBox: an R package for k–Top Scoring Pairs classifier development. Bioinformatics 31, 273–274. https://doi.org/ 10.1093/bioinformatics/btu622

Кононкова Анна Дмитриевна - IndieForest: машинное обучение на индикаторах и рангах для ранней диагностики онкогенных заболеваний

Анна Кононкова (Сколковский институт науки и технологий, аспирантура ИППИ РАН)

Поиск эффективных методов лечения рака остается актуальной задачей. Несмотря на большое количество исследований в данной области, не все из них находят практическое применение. Изучение факторов транскрипции (ТФ), вовлеченных в регуляцию генов, ассоциированных с раком, может не только расширить представления о причинах и механизмах канцерогенеза на молекулярном уровне, но и дает надежду на успешное применение на практике. Существенная роль транскрипционных факторов в канцерогенных процессах была установлена более 15 лет назад (Bushweller, 2019). Активность транскрипционных факторов может меняться в различных типах рака как напрямую (в результате хромосомных транслокаций, делеций или вставок, точечных мутаций), так и опосредованно – в связи с мутациями в участках связывания ДНК. Важно отметить, что в последние годы появились методы, направленные на регуляцию работы транскрипционных факторов (Bhagwat et al,2015; Bushweller, 2019). Например, активно ведется разработка низкомолекулярных ингибиторов ТФ, а также молекул PROTACs (бифункциональная молекула, в которой присутствует 2 ковалентно связанных лиганда, один из которых взаимодействует с целевым белком (фактором транскрипции), а второй с лигазой Е3, что в конечном итоге

приводит к распаду целевого белка). Некоторые ТФ вовлечены в регуляцию генов с повышенной экспрессией в раковых образцах по сравнению с контрольными сразу в нескольких типах рака (например, ген FOXA1 (M.Fournier, 2016)), что делает их особенно привлекательными потенциальными мишенями для терапии.
Целью данной работы является определение ранее малоизученных ТФ, вовлеченных в регуляцию генов, которые играют непосредственную роль в формировании и росте злокачественной опухоли в нескольких типах рака.

Исследование целевых ТФ проводится для трех типов рака на основе данных из базы TCGA (The Cancer Genome Atlas): рак молочной железы (BRCA, 96 парных образцов), аденокарцинома простаты (PRAD, 51 парный образец) и аденокарцинома легкого (LUAD, 56). В данной работе в основу поиска ТФ ложится их перепредставленность для группы генов с более высокой экспрессией в раке по сравнению с нормой для всех трех типов рака. Для связи «ген – транскрипционный фактор» использована база данных GeneHancer.
Предварительные результаты по бикластеризации отношения экспрессии “рак/контроль” генов с более высокой экспрессией в раке для всех трех образцов в совокупности не позволяют выделить крупные кластеры ко-экспрессии, в которые выходили бы в равной мере образцы всех трех типов рака. А именно, образцы рака аденокарциномы простаты представлены в выделенном кластере в меньшинстве (PRAD - 12 образцов, BRCA и LUAD ~40). Бикластеризация, не ограниченная генами с более высокой экспрессией в раке, не позволяет получить кластеры, объединяющей три типа рака с приемлемым процентным соотношением образцов каждого типа рака.
Исследование перепредставленных ТФ для 149 генов с наиболее высокой экспрессией во всех трех типах рака (получены пересечением результатов дифференциальной экспрессии) показывает функциональное обогащение процессами, связанными с делением. 61 из этих генов и 4 ТФ образуют плотный кластер при анализе в веб-ресурсе STRING. Однако эти ТФ являются хорошо изученными, а факт обогащения GO-категориями деления и клеточного цикла – тривиальным.
В перспективе для решения задачи рассматривается как переход на другие типы рака, так и применение принципиально других подходов для поиска более активных в раке генов и объединяющих их факторов транскрипции.

Вышкворкина Юлия Михайловна - Вычислительный поиск генов-мишеней для перепрограммирования клеток с использованием генных регуляторных сетей
Целью данной работы является выявление факторов транскрипции, которые определяют решение о судьбе клетки, и, таким образом, могут использоваться для перепрограммирования клетки. Планируется анализ данных ChIP-seq и single cell ATAC-seq, чтобы найти ключевые

транскрипционные факторы в регуляторной сети дифференцировки клеток.

Курилович Анна – рак

Сафина Ксения - Describing the HIV epidemics in Russian population

Волобуева Мария Евгеньевна - Разработка метода классификации клеток крови на основании гистологических снимков
М.(МГУ ФББ 3 курс, Лаборатория Математических методов в биологии НИИ им. Белозерского)
Алексеевский А.
Шеваль Е.
Пензар Д.

В состав крови входят два компонента: форменные элементы (клетки) крови и плазма. К форменным элементам относятся эритроциты (красные клетки крови), тромбоциты и лейкоциты (белые клетки крови). Клетки крови обычно исследуют на мазках или пленках, которые получают, равномерно распределяя каплю тонким слоем по предметному стеклу. После этого мазок быстро высушивают на воздухе. В таких мазках клетки отчётливо видны и различаются между собой. Эта работа посвящена созданию методов автоматической классификации клеток крови на основании фотографий мазков крови.
Данные содержат фотографии мазков с лейкоцитами, которые были окрашены по методу Гимзе. Лейкоциты делятся на 5 классов: нейтрофилы, эозинофилы, базофилы, моноциты, лимфоциты. Помимо фотографий здоровых клеток имеются снимки с клетками больных лейкозом. Первичной задачей работы является написать автоматический алгоритм классификации здоровых клеток на 5 классов. Затем усложнять алгоритм, чтобы он эффективно распознавал различные заболевания по фотографиям клеток.

Черницов Александр Валерьевич - Применение методов машинного обучения для предсказания коронарной недостаточности

Эволюция белков

Драненко Наталия Олеговна - Реконструкция эволюции семейств транспортёров металлов CorA и ZntB

Н.О. Драненко, Skoltech

(Научный руководитель М.С. Гельфанд)

Постановка задачи:
Реконструкировать эволюцию белкового семейства CorA с подсемейством ZntB, идентифицировать позиции белка, определяющие функциональную специфичность и в коллаборации с лабораторией Альберта Гуськова (Университет Гронингена) провести экспериментальную проверку результатов.

Результаты:
Из базы данных EggNog выбран ортологический ряд семейства CorA, содержащий также представителей подсемейства ZntB. Всего 1994 бактериальных белка.
Для этих последовательностей построено филогенетическое дерево методом ML.
На филогенетическом дереве были размечены белки, для которых есть данные о специфичности, полученные в ходе экспериментов. Сопоставление структуры дерева с имеющимися данными экспериментов позволило сделать предположение том, что транспортёры цинка (ZntB) и транспортёры магния (CorA) образуют на дереве две монофилетические клады.
Для этих клад был проведён анализ позиций, определяющих специфичность, получен набор таких позиций.
Также было проведено предсказание РНК-переключателей с помощью Infernal и данных из rfam об известных структурах M-box РНК-переключателей, регулирующих экспрессию магниевых транспортёров. Такие переключатели были обнаружены как на предположительно магниевых

ветках, так и на предположительно цинковых, что не соотвествует исходной гипотезе о монофилетичности этих клад.
В данный момент производится поиск мотивов связывания известных цинковых репрессоров для определения принадлежности белков к цинковым транспортёрам.

Дальнейшие планы:
Разбить имеющиеся последовательности по специфичности белков, предсказать позиции, определяющие специфичность, и проверить полученные результаты экспериментально.

Коростелёв Юрий Дмитриевич - Корреляции эффекторного домена LacI с кофактором

Новикова Валерия Сергеевна - Вычислительный анализ компенсированных болезнетворных мутаций у человека
Новикова Валерия Сергеевна (Московский физико-технический институт), Раменский Василий Евгеньевич (ФГБУ «Национальный медицинский исследовательский центр терапии и профилактической медицины»

Минздрава России)

Компенсированной болезнетворной несинонимичной заменой называется болезнетворный для человека аллель, который наблюдается у ортологов близких видов. Значимость данного феномента определяется тем, что при интерпретации эффекта замены её наличие у относительно близких видов является аргументом в пользу безвредности данной замены. В ряде случаев (до 10%) это не так: замена, болезнетворная для человека, тем не менее фиксируется у ортологов. Можно предположить, что у этих видов в этих же белках происходит другая замена, компенсирующая эффект исходной. Подобный механизм также типичен и для патогенов, вырабатывающих устойчивость к лекарственным препаратам или иммунной системе человека. Целью данного проекта является поиск потенциальных компенсаторов болезнетворных мутаций человека и описание механизмов компенсации. Разработанные методы могут способствовать более детальному изучению молекулярной эволюции, благодаря чему можно понять, как возникают и развиваются функции того или иного белка.
Был разработан вычислительный метод анализа филогенетической информации для поиска замен, уникальных для группы видов; произведён анализ 335 болезнетворных и 4959 безвредных для человека замен, описанных в базе ClinVar и фиксировавшихся в последовательностях хотя бы одного из 26 секвенированных приматов. В общей сложности для 229

болезнетворных и 1896 нейтральных из таких замен может быть однозначно определено время возникновения.
Следующим этапом работы будет анализ найденных пар "исходная замена - компенсатор" с точки зрения типов замен, локализации в структуре белка и влияния на стабильность белка.
Проект поддерживается грантом РФФИ № 20-54-12008

Зайченока Мария - Анализ и предсказание эффекта коротких инделов в белках

Лёвина Татьяна Борисовна - Редактирование мРНК и частота вызванных им несинонимичных замен в стуктурных и неструктурных частях белков у мягкотелых головоногих моллюсков

Татьяна Б. Лёвина [1] , Михаил А. Молдован [2, 3] , Михаил С. Гельфанд [2, 3, 4]
1 Факультет Биологии и Биотехнологии Высшей Школы Экономики

2 Сколковский Институт Наук и Технологий
3 Институт Проблем Передачи Информации им. Харкевича РАН
4 Факультет Компьютерных Наук Высшей Школы Экономики

Tatlyovina@mail.ru

1. Введение.
Редактирование РНК - это посттранскрипционный процесс, позволяющий осуществлять реализацию нескольких возможных протеомов в рамках одного генома. Одним из способов редактирования РНК является дезаминирование аденозина белками семейства ADAR (adenosine deaminase that acts on RNA). Дезаминирование приводит к замене аденозина инозином, который в процессе трансляции распознаётся как гуанин. Таким образом, редактирование РНК может влиять на аминокислотную последовательность белка.
Мишенями ADAR-белков для большинства организмов, у которых было обнаружено редактирование, чаще всего служат участки РНК, расположенные внутри интронов, тандемных повторов или 3', 5' некодирующих областей [1].
В случае мягкотелых моллюсков редактирование чаще приводит к заменам в аминокислотном составе белков, особенно в нервной ткани [2].
Было показано, что сайты редактирования эволюционируют иначе, чем нередактируемые аденины: редактируемые аденины с большей вероятностью заменяются в ходе эволюции на гуанин [3].

2. Постановка задачи
Мы решили посчитать частоты несинонимичных замен, вызванных редактируемыми аденинами, в структурированных и неструктурированных участках ADAR-белков и сравнить их.

Список литературы:
1. Nishikura K (2010) Functions and regulation of RNA editing by ADAR deaminases. Annu Rev Biochem 79:321–349
2. L. Bazak et al. A-to-I RNA editing occurs at over a hundred million genomic sites, located in a majority of human genes. Genome Res., 24 (2014), pp. 365-376
3. H. Ota ADAR1 Forms a Complex with Dicer to Promote MicroRNA Processing and RNAInduced Gene Silencing; Cell Volume 153, Issue 3, 25 April 2013, Pages 575-589

Ivankov Dmitry Nikolaevich - Prediction of the impact of mutation on the protein stability using free energy function conservation

Пак Марина Алексеевна - Study of influence of homology modeling on the prediction of protein stability change upon mutation

Воробьев Илья Сергеевич - Алгоритм поиска генотипов образующих гиперкуб в многомерном пространстве

Воробьев Илья Сергеевич, Сколтех.
Эпистаз – зависимость эффекта мутации от генетического контекста – является одним из главных факторов, препятствующих предсказанию фенотипа по генотипу. Поэтому представляется важным его изучение, как экспериментальными, так и вычислительными методами. Одним из перспективных направлений является экспериментальное определение фенотипа для множества генотипов. Для определения эпистаза в таких экспериментах необходимо, чтобы фенотип был определен для генотипов, образующих гиперкуб в многомерном пространстве генотипов. В больших экспериментах нахождение всех гиперкубов представляет собой ресурсоемкую задачу, даже с учетом существования эффективного алгоритма.
Текущая версия программы HypercubeME использует строковое представление генотипов. В то же время известно, что операции над целыми числами часто осуществляются намного быстрее. Нам удалось воспользоваться преимуществами библиотеки numpy в языке программирования Python для убыстрения алгоритма. Ключевым «трюком» оказалось переопределение произведения матриц, где операция умножения заменена на операцию сравнения: если два элемента одинаковых, то она выдает 0, в противном случае – 1. В итоге после такого «перемножения» результирующая матрица содержит расстояния между генотипами. Нам удалось убыстрить скорость работы алгоритма примерно в 3 раза.

18.04.2020

Факторы транскрипции

Кравченко Павел Андреевич - Объединение позиционно-весовых матриц в решающие деревья для распознавания сайтов связывания факторов транскрипции

Белоусова Евгения Александровна - Консервативность неконсенсусных позиций в сайтах связывания факторов транскрипции

Белоусова Е. А. Факультет Биоинженерии и Биоинформатики, МГУ им. М. В. Ломоносова, Москва, Россия
Для глобальных регуляторов существуют большие регулоги – группы регулонов из близких видов. В некоторых регулогах наблюдаются следующие феномены:
1. В ортологичных сайтах сохраняется позиция с неконсенсусным основанием.

2. В этой позиции сохраняется конкретное неконсенсусное основание.
Мы рассматривали регулятор CcpA, который обеспечивает транскрипционный ответ на появление быстроусвояемых углеводородов в среде. Чтобы оценить, насколько значима консервативность неконсенсусных позиций, эти позиции можно сравнить с третьими позициями четырехвырожденных кодонов генов, транскрипция которых регулируется данным фактором, что и было сделано коллегами в 2005 году [1]. Оказалось, что изучаемые неконсенсусные основания, действительно, более консервативны, чем третьи позиции. Этому может существовать несколько объяснений. Во-первых, возможно перекрывание сайтов связывания регуляторов транскрипции, и тогда консервативные «неконсенсусы» относятся к еще неизвестным регуляторным элементам. Во-вторых, поскольку замена неконсенсусного основания на консенсусное меняет сродство регулятора к сайту, а значит, и уровень транскрипции гена, если неконсенсусное основание однажды начало обеспечивать нужный уровень регуляции, то оно уже не может «переместиться» на другую позицию.
Таким образом, в данной работе мы хотим найти хорошо выраженные консервативные неконсенсусные основания в ортологичных сайтах, сравнить их консервативность с нейтрально (или почти нейтрально) эволюционирующими элементами и объяснить это явление.
Сравнение консервативности неконсенсусных оснований всех сайтов с синонимичными позициями генов
Для каждого консервативного неконсенсусного основания в сайте и для синонимичных оснований в гене мы считали так называемую долю «незамен» - долю этого основания в колонке выравнивания. Далее мы построили распределения доли «незамен» синонимичных букв во всех генах (рис. 1); и распределение доли «незамен» неконсенсусных букв во всех сайтах (рис. 2).

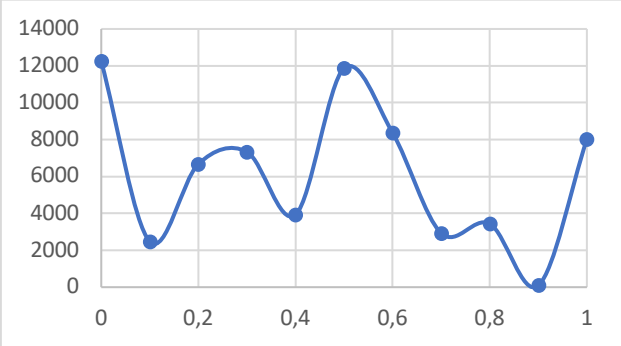


Рисунок 1. Распределения доли «незамен» синонимичных оснований во всех генах.

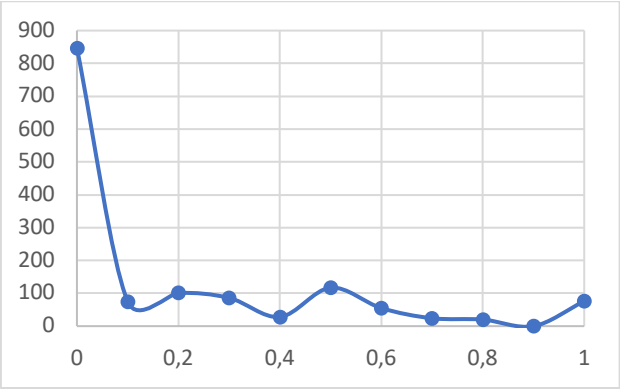


Рисунок 2. Распределения доли «незамен» неконсенсусных оснований во всех сайтах.

Надеялись увидеть, что, если неконсенсусные основания действительно консервативнее нейтрально эволюционирующих элементов, то второе распределение будет сдвинуто вправо относительно первого. Но здесь об этом вряд ли можно говорить: распределения имеют одинаковую форму. Выравнивания сайтов были не очень большие, и форма распределений обусловлена несколькими комбинациями деления целых чисел друг на друга.

Оценка консервативности веса в ортологичных сайтах

Одна из гипотез, объясняющих возможную консервативность «неконсенсусов», состоит в том, что сайтам перед ортологичными генами нужно сохранять постоянным сродство к транскрипционному фактору, то есть сохранять вес. Чтобы проверить, консервативен ли вес ортологичных сайтов, мы брали выравнивание сайтов и считали стандартное отклонение распределения весов. Затем 100 раз случайным образом перемешивали буквы внутри колонок выравнивания и каждый раз так же считали стандартное отклонение получившихся весов. Таким образом мы получали одно истинное стандартное отклонение весов и сто случайных. Мы упорядочивали по возрастанию 101 стандартное отклонение и получали место, которое заняла в этом ряду реальная величина, ранг. Эту процедуру мы повторили для 56 выравниваний сайтов толщиной >= 4 и получили 56 рангов, из которых также построили распределение (рисунок 3).

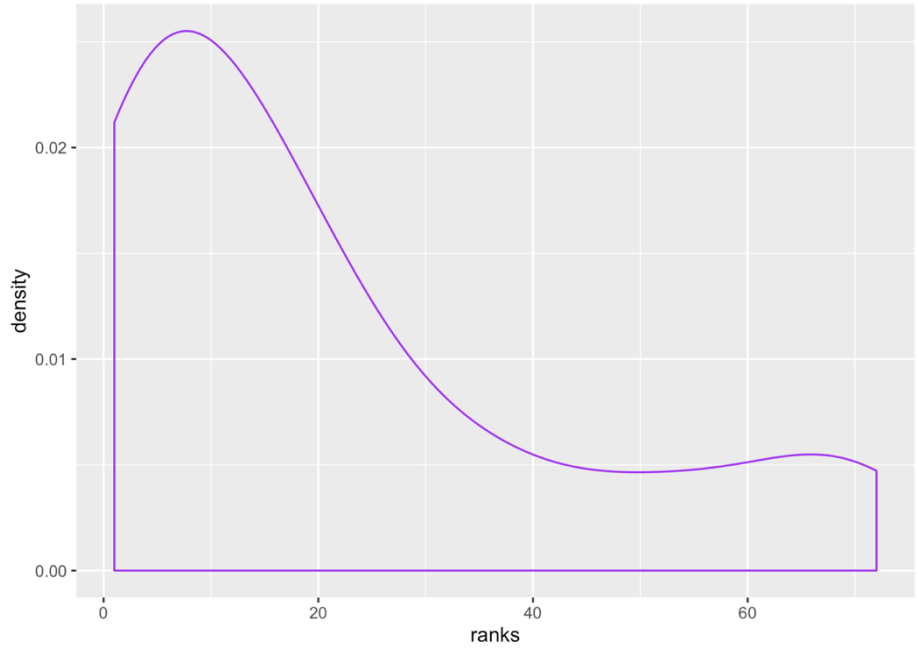


Рисунок 3. Распределение рангов: мест, которые занимали истинные стандартные отклонения весов сайтов среди ста стандартных отклонений, полученных случайным перемешиванием.

Это распределение показывает, что вес более консервативен в выравниваниях ортологичных сайтов, чем в выравниваниях, полученных случайным перемешиванием по столбцам. Однако известное нам распределение весов уже реального, потому что сайты, найденные в геноме – больше выбранного порога. То есть дисперсия веса при перемешивании по столбцам будет расти еще и по этой причине, что тоже вносит вклад в наблюдаемый эффект.

Планы

Получить большие выравнивания для сайтов и соответствующих генов, восстановить филогению видов и посчитать число реальных замен консенсусных нуклеотидов на неконсенсусные и наоборот в сайтах и число замен соответствующих синонимичных нуклеотидов в генах, сделать выводы о консервативности неконсенсусных нуклеотидов.

Оценить рост дисперсии веса сайтов при перемешивании нуклеотидов по столбцам при изменении порога веса сайтов. Оценить вклад этого эффекта в смещение приведенного выше распределения рангов.

Данная работа производится совместно с М. С. Гельфандом. Исследование поддержано грантами РНФ 18-14-00358 и РФФИ 18-34-01006.

Ссылки

Kotelnikova, E., Makeev, V., Gelfand, M.: Evolution of transcription factor DNA binding sites. Gene 347(2):255-63 (2005)

RegPrecise, http://regprecise.lbl.gov/RegPrecise/index.jsp

WebLogo, http://weblogo.berkeley.edu/logo.cgi

Агаева Зарифа Фарман кызы - Реконструкция регулонов метаболизма железа и марганца у α-протеобактерий

Суворова Инна Андреевна - Исследование структуры и расположения сайтов связывания факторов транскрипции

Тутукина Мария Николаевна - Регуляция метаболизма гексуронатов у кишечной палочки: роль UxuR и ExuR

Ракитин Денис - Экспериментальный и сравнительно-геномный анализ эволюции генетических регуляторных систем бактерий

Гатиятов Юрий - Поиск и анализ консервативных островков в межгенных областях хламидий

Шевкопляс Алексей Евгеньевич - Изучение консервативных регуляторных элементов в геномах Enterobacteriales при помощи нейросетей
Факультет биологии и биотехнологии ВШЭ
Червонцева Зоя Сергеевна

Регуляторные элементы генома, такие как сайты связывания транскрипционных факторов, рибопереключатели и термосенсоры играют важную роль в приспособлении бактерий к меняющимся условиям среды. Как правило, такие элементы расположены в межгенных областях и высоко консервативны у близкородственных видов, - а некоторые из элементов сохраняются и на больших филогенетических расстояниях. На этом наблюдении основан метод филогенетического футпринтинга, когда регуляторные элементы ищутся непосредственно в местах высокой консервативности. Однако для высоко мутабельных участков генома, к которым относятся межгенные области, часто бывает сложно получить надежное выравнивание для далеких видов. Поэтому в этой работе мы попытаемся обучить нейросеть детектировать консервативные участки в одной отдельно взятой последовательности - в надежде, что в процессе обучения нейросеть выучит основные регуляторные элементы.
На вход сети будет подан набор данных, полученный следующим образом. Будут выбраны некодирующие участки геномов представителей родов семейства Enterobacteriales, в которых количество видов больше или равно 5. Полученные участки близких видов будут выровнены между собой с помощью программы NPGE. По результату выравниваний нуклеотиды будут разделены на консервативные и неконсервативные. После этого вокруг каждого нуклеотида будет выбрано окно в последовательности, по которому нейросети и будет предложено

предсказать, является ли центральный нуклеотид консервативным.

Предварительные результаты коллег показывают, что в белок-кодирующих областях такое предсказание возможно, и мы надеемся перенести результат на межгенные области.

Геномы, пан-геномы и метагеномы

Бочкарева Ольга - Bacterial paralogs evolve under negative selection acting against unwanted intragenomicrecombination

Перевощикова Кристина Юрьевна - Из плазмиды в хромосому, реконструкция эволюционных событий в геномах Vibrio

Сефербекова Заира Назимовна - Сравнительная геномика Shigella и других патогенных E.coli

Факультет биоинженерии и биоинформатики

sef.zaira@gmail.com
О. О. Бочкарёва, М. С. Гельфанд

Бактерии Shigella являются возбудителями шигеллёза — тяжёлой формы бактериальной дизентерии. Штаммы Shigella spp. подразделяются на четыре вида — S. dysenteriae, S. flexneri, S. boydii и S. Sonnei, но фактически являются
парафилитической группой внутри E.coli [2, 3].
Геномы Shigella spp. характеризуются высокой динамичностью, которая позволяет быстрее адаптироваться к внутриклеточному образу жизни [2, 3]. Динамичность обусловлена накоплением большого числа инсерционных последовательностей (IS элементов) [1], перемещение которых может приводить
к геномным перестройкам и, как следствие, к активации и инактивации генов [2,
3]. Изучение геномных перестроек Shigella spp. интересно как в контексте эволюционной микробиологии, так и с практической точки зрения, поскольку
это позволит улучшить понимание эволюции и патогенных свойств этих бактерий.
В данной работе было решено проверить, различаются ли частоты геномных перестроек в штаммах разных видов Shigella и объясняются ли эти различия содержанием IS элементов. Чтобы проверить, различаются ли по тем
же параметрам патогенные и непатогенные штаммы E. coli, аналогичный анализ

был выполнен и для остальных штаммов E. coli.
Для анализа мы использовали 380 геномов E. coli и 34 генома Shigella spp. Было найдено 238 универсальных однокопийных ортологичных рядов, и нуклеотидные последовательности каждого ряда были выравнены алгоритмом
Mafft. По конкатенату полученных выравниваний с помощью RaxML было построено дерево, на котором с использованием пакета GGRaSP на языке R было

выделено 6 кластеров. Полученные кластеры согласуются с известными филогруппами E.coli, расположение Shigella на дереве также согласуется с литературными данными.
С помощью сервиса ISsaga мы произвели поиск IS элементов в хромосомах всех исследуемых штаммов. Число IS элементов в геномах Shigella spp. оказалось значимо больше числа IS элементов в геномах других патогенных и
непатогенных штаммов E. coli. Распределение найденных IS элементов по семействам оказалось неравномерным. В частности, геномы S. flexneri, S. boydii
и S. Sonnei содержат значимо больше IS элементов из семейств IS1, IS3, IS4, IS91
по сравнению с другими штаммами E. coli из той же филогруппы.
Чтобы сравнить частоты перестроек на разных ветвях, мы реконструировали историю перестроек для 34 штаммов Shigella и 17 штаммов E.
coli с помощью MGRA. Всего было найдено 114 событий, 105 из которых соответствуют ветвям Shigella. Наибольшее число инверсий было реконструировано для ветви, отделяющей два штамма S. dysenteriae. Возможно,
накопление инверсий связано с переходом к внутриклеточному образу жизни,
независимо возникшим несколько раз, и резким накоплением мобильных элементов. Также наши результаты указывают на то, что перестройки являются
основным механизмом эволюции Shigella spp.
В дальнейшем также планируется произвести поиск событий гомологичной рекомбинации и оценить её частоту между разными штаммами E.coli и Shigella
spp. для проверки гипотезы о падении интенсивности гомологичной рекомбинации с увеличением числа геномных перестроек.

Источники:
1) Kunst F. et al. The virulence plasmid pWR100 and the repertoire of proteins secreted
by the type III secretion apparatus of Shigella flexneri // Molecular Microbiology

38(4), 2003, pp. 760–771.

2) Pupo G. M. et al. Multiple independent origins of Shigella clones of Escherichia coli
and convergent evolution of many of their characteristics // Proceedings of the National Academy of Sciences 97(19), 2000, pp. 10567–10572.
3) The H. C. et al. The genomic signatures of Shigella evolution, adaptation and geographical spread // Nature Reviews Microbiology 14(4), 2016, pp. 235–250.

Ходжаева Евгения Сергеевна - Организация метаболических оперонов в геномах бактерий из разных филумов

Ходжаева Евгения Сергеевна. Биологический факультет МГУ им. Ломоносова. Червонцева Зоя Сергеевна

Приспособление к разным экологическим нишам у бактерий часто связано с регуляцией их метаболических путей. Для того, чтобы она была эффективной, гены одного пути бывают собраны в опероны, а опероны – в локусы. Цель этого исследования – определить структуру организации генов 24-х основных метаболических путей (таких, как пути синтеза аминокислот, некоторых витаминов и азотистых оснований) бактерий из пяти филумов: Firmicutes, Cyanobacteria, Alpha-, Beta- и Gammaproteobacteria. Мы определяем
как локус любую последовательность генов, расположенных друг от друга на расстоянии не более 200 нуклеотидов. Для распределения генов по локусам мы
используем их координаты в геномах, взятые из базы данных PATRIC. По результатам этой работы мы планируем также определить возможные варианты
структуры локусов у предковых форм бактерий.

Рыбина Анна - Эволюция локуса катаболизма сульфоглюкозы и лактозы

Рыбина Анна Александровна (Сколтех). Казнадзей Анна Денисовна, Тутукина Мария Николаевна

В геноме Escherichia coli есть кассета из десяти генов, часть которых кодирует ферменты деградации сульфоквиновозы (yih-кассета) [1]. С помощью методов сравнительной геномики в нашей лаборатории было выдвинуто предположение, что эта кассета участвует и в утилизации лактозы, так как ее состав у E. coli сходен с составом кассеты бактерий класса Bacilli, отвечающей за катаболизм этого дисахарида [2]. С помощью ОТ-ПЦР в реальном времени было показано, что экспрессия четырех генов yih-кассеты, кодирующих альдолазу, изомеразу (yihTS), киназу (yihV) и фактор транскрипции (yihW), значительно возрастает во время роста

культуры на лактозе. С помощью сравнительной геномики в межгенных областях кассеты было выявлено несколько мест потенциального связывания глобального регулятора углеводного метаболизма cAMP-CRP, которое затем было подтверждено экспериментально. Однако до настоящего времени было неясно, насколько эффективно связывается с межгенными областями кассеты, влияет ли лактоза на связывание YihW и как именно, перекрывается ли YihW с CRP и каков мотив его узнавания [2,3]. Также yih-кассета ранее не подвергалась филогенетическому анализу.
Цель настоящей работы - изучить эволюцию и регуляцию yih-кассеты, в том числе, с точки зрения возможных мультифункциональных характеристик соответствующих белков.
Филогенетический анализ осуществляли с помощью поиска гомологичных белковых последовательностей (BLAST, HMMER), множественного выравнивания (MUSCLE) и построения филогенетических деревьев методом максимального правдоподобия (PhyML). Мы выяснили, что yih-кассета, в основном, присутствует в геномах семейства

Enterobacteriaceae и представлена в виде “короткой” (yihTUVW) или “длинных” форм (в основном, ompLyihOPQRSTUVW). При этом появление короткой и длинной формы скоррелировано, а разделение исходной предковой кассеты на два типа произошло один раз. Комбинации, включающие в себя как минимум три из четырех гомологов основных исследуемых генов (yihS, yihT, yihU, yihW), была обнаружена у некоторых представителей типа Actinobacteria (yihTUVW у Streptomyces sp. SCSIO 03032,
yihSTVW у Pseudarthrobacter phenanthrenivorans Sphe3 str. Sphe3) и у одного представителя типа Chloroflexi (yihTVW у Anaerolinea thermophila UNI-1 str. UNI-1). На филогенетических деревьях, построенных по нуклеотидным последовательностям соответствующих участков генома, данные организмы располагались в одной кладе с представителями типа Proteobacteria, что, может свидетельствовать о событиях горизонтального переноса.
С помощью электрофореза с задержкой в геле (EMSA) мы показали, что YihW эффективно связывается с регуляторной областью собственного гена и межгенной областью yihV/yihU уже в 8-кратном молярном избытке. Интересно, что при взаимодействии CRP и YihW с регуляторной областью yihW наблюдалась конкуренция с преимущественным связыванием CRP, тогда как с межгенной областью yihV/yihU белки связывались кооперативно. Эффективность связывания YihW с межгенными участками не менялась в присутствия глюкозы и галактозы, но существенно падала в присутствии лактозы. Это подтверждает гипотезу о том, что лактоза влияет на регуляцию экспрессии генов yih-кассеты и, возможно, является одним из эффекторов YihW.

Джамалова Дильфуза Фазлиддин кизи - Re-classification of bacterial strains and species via pan-genome analysis

Николаева Дарья Дмитриевна - Особенности структуры пангенома у бактерий-специалистов и бактерий-генералистов

Руководитель: Гарушянц Софья Константиновна

Пангеном - это совокупность белок-кодирующих генов, присутствующих в наборе геномов одного вида или рода бактерий. Традиционно в структуре пангенома выделяют “универсальный геном” - гены, которые встречаются почти во всех рассматриваемых штаммах, и “периферию” - гены, встречающиеся у небольшого количества штаммов. Соотношение размера периферии к размеру универсального генома отличается у разных бактерий, и какие именно факторы определяют это соотношение, до сих пор остается непонятным. Мы предположили, что одним из определяющих факторов может являться количество экологических ниш, в которых встречается данный вид. Так, виды бактерий, способные существовать в разнообразных физико-химических условиях (виды-генералисты), должны, с одной стороны, иметь гены, необходимые для приспособления к конкретному местообитанию, а, с другой стороны, могут взаимодействовать с большим количеством бактерий других видов, от которых могут получать более разнообразные гены в результате горизонтального переноса. В то же время существуют виды-специалисты, которые привязаны к единственной экологической нише, хорошо приспособлены к ней и поэтому, вероятно, генетически более однородны. Получить информацию о структуре сообществ можно из метагеномных данных, когда совместно секвенируются все нуклеотидные последовательности, выделенные из данного местообитания. Такой

подход позволяет определить как качественный, так и количественный состав организмов разных экосистем.
Идея работы заключается в том, чтобы установить, существует ли связь между соотношением элементов структуры пангенома (универсального генома и периферии) и количеством местообитаний, в которых данный вид бактерий встречается. Ранее мы уже попытались ответить на этот вопрос, используя данные о бактериях-генералистах и специалистах [1] и предварительно получили положительный ответ, однако он требует подтверждения с использованием более масштабных данных. На этот раз используются данные Earth Microbiome Project (EMP) [2], на основе которых будут выбраны виды бактерий-генералистов и специалистов для построения и анализа пангеномов. Отдельное внимание уделено поиску подходящей для определения генералистов и специалистов классификации местообитаний, к которым принадлежат образцы EMP.

[1] Sriswasdi S., Yang C., Iwasaki W. Generalist species drive microbial

dispersion and evolution //Nature communications. – 2017. – Т. 8. – No. 1. – С. 1162.
[2] Thompson L. R. et al. A communal catalogue reveals Earth’s multiscale microbial diversity //Nature. – 2017. – Т. 551. – №. 7681.
Шелякин Павел Владимирович - Бактериальный микробиом и (1) загрязнение почвы керосином, (2) загярзнение почвы сернокислыми стоками с отвалов угольных шахт, (3) болезни кораллов

Сарана Юлия - Микробиомы тлей и соплей

Лебедев Юрий - Влияние жизнедеятельности дождевых червей на почвенный микробиом

1 Сколковский институт науки и технологий, Москва, Россия
2 Институт проблем экологии и эволюции им. А.Н. Северцова РАН, Москва, Россия

*lebedev_ym@yahoo.com

Введение. С почвами связан жизненный цикл почти 90% всех живых организмов, обитающих в наземных экосистемах. Обитатели почв создают сложные трофические цепи, напрямую влияя на круговорот различных биогенных элементов, на выполнение экосистемами их функций. В значительной степени это влияние оказывают почвообитающие микроорганизмы [1]. Микробные сообщества играют ключевую роль в поддержании ряда экосистемных функций (минерализация азота, разложение детрита, регуляция климата и др.). Видовое разнообразие бактерий положительно связано с функционированием экосистем [2].
Дождевых червей относят к так называемым «экосистемным инженерам» - к последним причисляют группы животных, наиболее значительно трансформирующих ландшафты. Дождевые черви составляют основную часть животной биомассы в большинстве наземных экосистем. Некоторые виды червей способны «пропустить» через свой пищеварительный тракт за сутки количество почвы, превышающее их собственный вес в несколько десятков раз. Однако доля органического вещества, усваиваемого при этом, относительно невысока (около 8%) [3]. Структура почвенного микробиома меняется в результате жизнедеятельности дождевых червей. Гликопротеины, выделяющиеся как

пищеварительной системой, так и кутикулой дождевых червей, могут инициировать повышение активности микроорганизмов. Пропуская почву через пищеварительный тракт, черви могут переваривать часть микробиоты, как следствие уменьшая микробную биомассу. В то же время пищеварительная система дождевых червей зачастую не имеет всех необходимых энзимов для переваривания части органического вещества.

Группы бактерий, способные помогать в переваривании, отбираются и стимулируются в кишечнике червя, в результате чего в почву может позднее выделяться повышенное количество некоторых таксонов бактерий. Также существуют и микроорганизмы-симбионты дождевых червей, постоянно обитающие в пищеварительном тракте. Несмотря на то, что результаты некоторых исследований, рассматривавших вопрос взаимодействия дождевых червей и микробиоты оказываются противоречивыми, общепризнанным остается тот факт, что в зоне жизнедеятельности дождевых червей наблюдается повышенная активность микроорганизмов, определяющая функционирование экосистем [4].

Дождевые черви разделяются на 3 морфо-экологические группы – поверхностнообитающие, почвенно-подстилочные и норники [5]. Известно, что влияние разных групп на микробиом отличается. Однако и внутри данных групп взаимодействие с микробиомом субстрата может отличаться – в то время, как в части работ, объектами которых были почвенно-подстилочные дождевые черви (род Eisenia), было обнаружено положительное влияние последних на бактериальную биомассу, в иных работах в иных условиях наблюдался обратный эффект [4].

Таким образом, влияние дождевых червей на почвенный микробиом определяется в значительной мере условиями, в которых происходит это взаимодействие. Цель данной работы – определить влияние различных условий окружающей среды на взаимодействие дождевых червей с микробиомом почвы. Также будет рассмотрено изменение структуры бактериальных сообществ при продвижении почвы по пищеварительному тракту червя.

Экспериментальная часть. Для лабораторного эксперимента будут использованы 2 вида дождевых червей – E.fetida и Dendrobaena veneta. Эксперимент будет проходить в мезокосмах по 1 червю на мезокосм, 3 повторности для каждого вида червей. В качестве условий были выбраны следующие виды воздействия: повышенное содержание органического вещества, повышенное содержание фосфора, повышенное содержание азота, загрязнение почвы кадмием, пониженная влажность, повышенная влажность. Для каждого воздействия будут использованы по 3 мезокосма для червей каждого вида, в качестве контроля будут использованы 3 мезокосма без животных. Также в качестве контроля будут использованы 9 мезокосмов (3 E.fetida, 3 D.veneta, 3 без животных), где почва не будет подвергаться какому-либо специальному воздействию. Таким образом, всего в лабораторном эксперименте будет использовано 63 мезокосма. Для определения собственного микробиома дождевого червя до эксперимента будет взято по 3 червя каждого вида и сутки выдержаны на фильтровальной бумаге для удаления копролитов. Микробиом каждого червя будет отбираться в 4 различных отделах пищеварительного тракта.

Список литературы

[1] R. Bardgett and W. van der Putten, "Belowground biodiversity and ecosystem
functioning", Nature, vol. 515, no. 7528, pp. 505-511, 2014. Available: 10.1038/nature13855.
[2] M. Delgado-Baquerizo et al., "Microbial diversity drives multifunctionality in
terrestrial ecosystems", Nature Communications, vol. 7, no. 1, 2016. Available:
10.1038/ncomms10541.
[3] M. Blouin et al., "A review of earthworm impact on soil function and ecosystem
services", European Journal of Soil Science, vol. 64, no. 2, pp. 161-182, 2013.
Available: 10.1111/ejss.12025.

[4] R. Medina-Sauza et al., "Earthworms Building Up Soil Microbiota, a Review", Frontiers in Environmental Science, vol. 7, 2019. Available: 10.3389/fenvs.2019.00081.
[5] Т.С. Перель, Распространение и закономерности распределения дождевых червей фауны СССР. =. Москва: Наука, 1979.

Эволюция и геномика эукариот

Селифанова Мария Витальевна - Длинные идентичные межвидовые элементы в растительных геномах и их роль в универсальной экстремальной консервативности у эукариот
Руководители: Дмитрий Александрович Коркин, Михаил Сергеевич Гельфанд
Соавторы: студенты Лаборатории экстремальной геномной консервативности ШМТБ
2019
Целью этого проекта является изучение участков экстремальной консервативности в геномах однодольных и двудольных растений. Эти участки представляют собой последовательности ДНК с абсолютной или почти 100% идентичностью в трёх и более геномах. Они были были получены с помощью нового вычислительного метода, разработанного в лаборатории Дмитрия Коркина и названы - «длинными идентичными межвидовыми элементами» (Long Identical Multispecies Elements, LIMEs). Проект включает в себя функциональную аннотацию растительных LIME-ов, изучение особенностей их расположения в геноме, а также создание общей карты растительных LIME’ов для геномов Arabidopsis thaliana и Physcomitrella patens.
На данный момент был разработан комплекс программ для полногеномного поиска, кластеризации и функциональной аннотации консервативных геномных элементов в растениях на основе BLAST

против баз данных геномных элементов и по координатам при помощи REST API Ensemble. В результате работы было выявлено, что LIME-ы формируют кластеры, которые, по видимому, являются функциональными единицами и присутствуют в геноме Arabidopsis thaliana в большом количестве копий. Более подробное изучение каждой функциональной группы консервативных элементов показало, что кластеры соответствуют наиболее консервативным участкам некоторых тРНК и малых ядерных

РНК, а также то, что многие из них них пересекаются одновременно с экзонами белков и с участками РНК генов.

Мыларщиков Дмитрий Евгеньевич - Поиск ортологичных некодирующих РНК с помощью синтеничного подхода
Факультет биоинженерии и биоинформатики МГУ им. М.В.Ломоносова
Руководитель: д.б.н., к.ф.-м.н., проф. Андрей Александрович Миронов

В ряде задач возникает необходимость поиска ортологов ранее не обнаруженных некодирующих РНК. Ввиду того, что гены некодирующих РНК менее консервативны, чем гены белков, необходимо использовать дополнительные подходы для повышения чувствительности. Так, для полногеномных скринингов используются множественные выравнивания геномов и обнаружение консервативных вторичных структур[1]. Для заданного набора транскриптов двух геномов используются подход наилучшего взаимного попадания (best reciprocal hit)[2], наличие транскриптов в синтеничных областях[3] или выравнивание с учётом вторичной структуры[4]. У всех этих подходов есть две проблемы: ортологичные транскрипты не ограничены набором уже аннотированных генов, а вторичная структура РНК не всегда является функциональной, потому необязательно будет консервативной[5].
Мы предлагаем метод поиска ортологов ранее не аннотированных некодирующих РНК с использованием синтении. Мы определяем синтеничные участки в двух геномах как ограниченные одним набором якорных генов- ортологов (в частности, белок-кодирующих ортологичных генов из базы OrthoDB[6]) и ведём поиск ортолога для некодирующей РНК в пределах синтеничного участка с помощью BLAST[7]. Чтобы оценить неслучайность находки, мы составляем распределение значений функции сходства последовательностей при выравнивании случайной подвыборки данных РНК и случайной подвыборки межгенных участков целевого генома. В соответствии с этим распределением мы определяем, какие участки выравнивания (hit scoring
pairs, HSPs) синтеничных участков значимо отличаются от выравнивания несинтеничных. Далее, мы выбираем только значимые HSP и из них собираем итоговое выравнивание данного транскрипта с синтеничным участком с помощью метода динамического программирования.
Метод поиска ортологов реализован в виде пакета для языка Python 3 и

доступен по ссылке: https://github.com/dmitrymyl/ortho2align/
В данный момент закончена разработка пакета и ведётся работа по оценке параметров алгоритма для поиска ортологов некодирующих РНК в пределах Млекопитающих. Параметрами являются расстояния от данных генов РНК до якорных генов в исходном геноме и расстояние между якорными генами в целевом геноме для объединения их в синтеничные участки.
Планируется применить алгоритм для поиска ортологов полувыделяемых РНК[8], среди которых было собрано несколько неаннотированных ранее. Консервативность таких РНК может указать на их функциональную роль в клетке.
Литература:
1. Washietl, S., Hofacker, I., Lukasser, M. et al. Mapping of conserved RNA secondary
structures predicts thousands of functional noncoding RNAs in the human genome. Nat Biotechnol 23, 1383–1390 (2005);
2. Necsulea, A., Soumillon, M., Warnefors, M. et al. The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature 505, 635–640 (2014);
3. Chen, J., Shishkin, A.A., Zhu, X. et al. Evolutionary analysis across mammals reveals
distinct classes of long non-coding RNAs. Genome Biol 17, 19 (2016);
4. Will, S., Siebauer, M.F., Heyne, S. et al. LocARNAscan: Incorporating thermodynamic stability in sequence and structure-based RNA homology search. Algorithms Mol Biol 8, 14 (2013);
5. Igor Ulitsky, David P. Bartel, lincRNAs: Genomics, Evolution, and Mechanisms,
Cell, Volume 154, Issue 1, 2013, Pages 26-46;
6. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and
viral genomes for evolutionary and functional annotations of orthologs Kriventseva
EK et al, NAR, Nov 2018;

7. Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, David J. Lipman,
Basic local alignment search tool, Journal of Molecular Biology, Volume 215, Issue
3, 1990, Pages 403-410;
8. Unusual semi-extractability as a hallmark of nuclear body-associated architectural
noncoding RNAs, Takeshi Chujo et al., The EMBO Journal (2017) 36: 1447–1462.

Гайдукова Софья Александровна - Эволюция сдвигов рамки считывания в транскриптомах инфузорий

Попов Алексей Алексеевич - Поиск следов положительного отбора в S. commune с помощью глубокого обучения
Попов Алексей Алексеевич, Селифанова Мария Витальевна, Столярова Анастасия Валерьевна, Базыкин Егор Александрович

В генетике selective sweep - это явление понижения нуклеотидного разнообразия вокруг мутации, которая находится под действием положительного отбора. Изучение данного процесса позволяет более точно определять участки, и даже гены, на которые ведётся отбор в популяции. Дополнительная польза изучения данного процесса заключается в возможности детектировать то, как давно ведётся отбор на данный участок, вкупе со многими другими параметрами, которые могут быть интересны популяционной генетике. На данный момент существует классификация “свипов” на сильные/слабые и проходящие/закрепившиеся.
Основной целью данного исследования является поиск “свипов” в популяциях Schizophyllum commune с помощью методов глубокого обученния. Schizophyllum commune - широко известный в узких кругах модельный объект популяционной генетики, который характеризуется самым высоким уровнем генетической “разнородности”. На первых этапах работы поиск “свипов” производится с помощью нейросети S/HIC. При этом обучение производится на данных симуляций со сходными с модельным объектом параметрами, призведенных с помощью инструмента Slim. В дальнейшем также планируется модификация S/HIC и варьирование параметров симуляции тренировочных данных для увеличения точности предсказания и борьбы с переобучением. Полученные данные мы хотим сравнить с результатами других алгоритмов поиска следов положительного отбора, использующих графы рекомбинации.

Безменова Александра - Зависимость скорости гомологичной рекомбинации и мутагенеза от уровня гетерозиготности хромосомы в базидиомицете commune Schizophyllum

Столярова Анастасия - Оценка числа мишеней положительного отбора по мутационным спектрам

Набиева Елена - Поиск изменений копийности по экзомных данным в «кариотипически нормальных» образцах

Кузнецов Иван Алексеевич - Ограничения аддитивной модели для роста человека
Кузнецов Иван Алексеевич (Сколковский Институт Науки и Технологий, Москва, Россия), Славский Сергей А., Шашкова Татьяна И., Базыкин

Георгий А., Аксенович Татьяна И., Кондрашов Фёдор А., Аульченко Юрий С.

Классический подход к анализу полигенных количественных признаков предполагает использование нормального приближения и аддитивности эффектов. На протяжении более ста лет рост человека служил модельным признаком для такого рода анализов. В нашей работе мы показываем, что общепринятый подход к анализу роста становится не применимым на больших выборках. В частности, мы демонстрируем существование слабых, но достоверных неаддитивных взаимодействий генетических факторов и факторов окружающей среды. Соответствие классической модели и современных данных может быть достигнуто за счёт усложнения используемой модели. С другой стороны, наблюдаемое несоответствие может быть исчерпано путём введения лог-нормального приближения для распределения роста человека.

ВЫСШАЯ ШКОЛА ЭКОНОМИКИ

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Межвузовская студенческая научная школа-конференция

Информационные технологии и системы. Биоинформатика.

11-18 апреля 2020 года