Секция «Биоинформатика»

Оценка специфичности РНК в контексте взаимодействий РНК-ДНК-хроматиновых взаимодействий.

Научный руководитель – Миронов Андрей Александрович

Питиков Егор Николаевич

Студент (специалист)

Московский государственный университет имени М.В.Ломоносова, Факультет биоинженерии и биоинформатики, Москва, Россия

E-mail: pitikovegor@gmail.com

Изучение строения и функций различных некодирующих хроматин-ассоциированных РНК является важной задачей молекулярной биологии[1-2], однако уровень шума в all-to-all экспериментах для обнаружения таких контактов катастрофически высок[3]. Данная работа предлагает подход, основанный на методиках машинного обучения для решения этой проблемы на экспериментах Red-C для клеток линии K562.

В работе предпринята попытка предсказать специфичность ДНК-РНК контакта, опираясь на его к-мерный состав для экспериментов Red-C на клетках линии K562 (использовались к-меры от 1 до 6).

Обучающие выборки выбирались двумя методиками. В первом случае как специфичные контакты выбирались контакты РНК с высоким хроматиновым потенциалом, как неспецифика - трансхромосомные контакты матричных РНК, MALAT1 и NEAT1 были исключены из анализа. Выборка была валидирована по количеству контактов разных хромосом, количеству специфичных и неспецифичных контактов. Обучение проводилось методом k-fold кросс-валидации, количество эпох определялось точностью предсказаний на валидирующей выборке - остановка после 5 раундов, не повышающих точность предсказаний.

Во втором случае были составлены “кластеры” контактов - группы контактов со схожей позицией ДНК и РНК части. Контакт входил в кластер при условии наличия хотя бы одного контакта из данного кластера с ДНК и РНК частями на расстоянии менее килобазы от него, РНК MALAT1 и NEAT1 были исключены из анализа. Процессы валидации выборки и обучения модели были схожи.

Были обучены модели на основе бэггинга, бустинга и случайного леса. Модели на основе бэггинга показали максимальную точность и этот подход стал основным в дальнейшей работе.

При использовании хроматинового потенциала как критерия специфичности днРНК и белок-кодирующие РНК получили схожие доли специфичных контактов (0,27% и 0,3% специфичных контактов соответственно).

Подход,основанный на числе близких контактов давал значительно больший процент специфичных контактов днРНК по сравнению с матричными (7,55% и 0,28% соответственно), что позволяет считать его более точным.

Был проведен анализ значимости признаков для моделей. Обнаружено, что большая часть из них - ГЦ-богатые (GC-context у контактов, отбранных “кластерной” моделью составил 55,7%, для “потенциальной” модели составил 44,8%). Часто среди значимых к-мер у моделей обоего типа встречались паттерны типа GG(N) 2-4 GG.

С помощью программ Queseq и DFilter проведен пикколинг. днРНК в среднем получали более значимые пики, но количество пиков у белок-кодирующих РНК было выше, что можно объяснить соотношением данных типов РНК в клетке.

Источники и литература

1) Li, X. et. al, Chromatin-associated RNAs as facilitators of functional genomic interactions. Nat Rev Genet 20, 503–519 (2019)
2) Holoch, D. et. al, RNA-mediated epigenetic regulation of gene expression. Nat Rev Genet 16, 71–84 (2015).
3) Sergey V Razin et. al, Studying RNA–DNA interactome by Red- C identifies noncoding RNAs associated with various chromatin types and reveals transcription dynamics, Nucleic Acids Research, Volume 48, Issue 12, 09 July 2020, Pages 6699–6714