Анализ экспериментальных данных для определения участков ДНК, связывающих регуляторные белки

Воронцов И.Е.[1], Кулаковский И.В.[1,2], Медведева Ю.А.[1,3,5], Колпаков Ф.А.[4], Евшин И.С.[4], Макеев В.Ю.[1,2,5,6,*]

[1] Институт общей генетики РАН им. Н.И. Вавилова, ул. Губкина 3, Москва, 119991

[2] Институт молекулярной биологии РАН им. В.А. Энгельгардта, ул. Вавилова 32, Москва 119991

[3] ФИЦ биотехнологии РАН, Институт биоинженерии РАН, пр-т 60-летия Октября д. 7, корп. 1, Москва 117312

[4] Институт вычислительных технологий СО РАН, пр-т Академика Лаврентьева, 6, Новосибирск, 630090

[5] НИЦ «Курчатовский институт» – ГосНИИгенетика, 1-й Дорожный проезд, д. 1, Москва, 117545

[6] Московский физико-технический институт, Институтский переулок, д. 9., г. Долгопрудный, Московская область, 141701

email: vsevolod.makeev@vigg.ru

doi: 10.21519/0234-2758-2018-S-14-15

Аллельные варианты регуляторных регионов генома являются доминирующим фактором в формировании многих фенотипических признаков. Нуклеотидные замены в регуляторных районах генов могут изменять сродство конкретных участков ДНК к специфически связывающим их факторам транскрипции, что в свою очередь может влиять на экспрессию генов и, следовательно, фенотип. Связывание транскрипционных факторов зависит не только от последовательности нуклеотидов ДНК, но и от ее доступности для связывания белков. У эукариот именно доступность ДНК для связывания регуляторов является главным фактором, определяющим специфическую активность генов в конкретных типах клеток, которая в конечном счете определяет индивидуальные свойства клеточных типов. Учитывая, что в клетке имеются сотни типов транскрипционных факторов (порядка 250 для *E. coli* и порядка 1600 для *H. sapiens*), прямое экспериментальное определение всех участков ДНК, связывающих регуляторные белки конкретных типов, во всех случаях (например, для сотен клеточных типов человека) является нереалистичным.

В своей работе мы разрабатываем методы для вычислительного предсказания участков ДНК, связывающих конкретный регуляторный белок в конкретных условиях, а также для оценки изменений аффинности связывания при точечных заменах нуклеотидов в этих участках. Для решения этих задач анализируются факторы, влияющие на специфическое связывание факторов транскрипции, включая степень доступности хроматина, транскрипционную активность соседних генов, мотивы последовательности сайтов связывания транскрипционных факторов и их кофакторов. Цель состоит в том, чтобы разделить переменные, характеризующие состояние клетки и специфичность связывания фактора транскрипции. Мы обнаружили, что

сравнительная значимость доступности хроматина существенно зависит от клеточного типа и конкретного белка-регулятора. Это затрудняет прогнозирование профиля связывания фактора транскрипции для конкретного типа клеток. Тем не менее, использование современных методов машинного обучения позволяет получить осмысленные предсказания, по крайней мере для наиболее экспериментально изученных регуляторов. Полученная путем таких предсказаний информация о возможном вкладе конкретных нуклеотидных замен в связывание белка – регулятора транскрипции может быть использована для определения приоритетных вариантов, генетически ассоциированных с определенными признаками. Предсказание типа клеток, в которых активен сегмент ДНК, содержащий конкретную регуляторную замену, в свою очередь, может иметь важное значение для идентификации органов, которые в первую очередь поражаются при наследственных патологиях или при соматических мутациях.