

Секция «Математика и механика»

Вероятностно-статистические методы при решении одной генетической задачи

Рытова Анастасия Игоревна

Студент

МГУ - Московский государственный университет имени М.В. Ломоносова,

Механико-математический факультет, Москва, Россия

E-mail: nrche@mail.ru

С привлечением вероятностно-статистических методов проводится анализ генетических данных, полученных в ходе эксперимента с клетками А549 легочной карциномы человека. В эксперименте искусственно повышается выработка регуляторного белка Pterp1, который оказывает влияние на клеточный цикл в организме. Белок Pterp1 регулирует передачу информации от генов к РНК. Это происходит путём связывания белка со специфичными участками ДНК, имеющими отношение к регулируемому гену. В результате повышения уровня белка Pterp1 с клетками происходит важное изменение: до эксперимента они росли кластерами, а после — образование кластеров перестаёт наблюдаться, и клетки начинают развиваться независимо. Предполагается, что повышенный уровень белка Pterp1 приводит к активации особой группы генов, которые отвечают за миграцию и инвазию клеток, т.е. за способность клеток злокачественной опухоли отделяться от нее и проникать в окружающие ткани. Ставится задача обнаружения этой группы генов.

Экспериментальные данные. Последовательность ДНК — это слово порядка 3 млрд. букв алфавита $\{A, C, T, G\}$. Занумеровав все буквы слова по порядку натуральных числами, каждой непрерывной последовательности натуральных чисел, кодирующей определённый участок ДНК, поставим в соответствие координаты ее начала и конца. Технически можно обнаружить положение на ДНК тех генов, с которыми связался белок до эксперимента (список А), и тех генов, с которыми он связался во время эксперимента (список В). То есть известно два списка специфичных участков ДНК (вместе с соответствующими им генами) с координатами начала и конца для каждого. Также для каждого гена определяется параметр, называемый в литературе fold enrichment (FE), который характеризует силу связывания белка с данным геном. Предполагается, что гены, отвечающие за миграцию и инвазию клеток, находятся как среди уникальных генов каждого списка, так и среди одинаковых генов, но с существенным различием параметра FE.

Алгоритмы поиска и статистический анализ. Для двух независимых множеств A и B , элементами которых являются конечные последовательности натуральных чисел x , соответствующие специфичным участкам ДНК, с которым связался белок, построен алгоритм нахождения последовательностей из подмножеств $A \Delta B$ и $A \cap B$. Значение параметра FE обозначим $r_A(x)$ при $x \in A$ или $r_B(x)$ при $x \in B$. Для $x \in A \cap B$ с помощью парного критерия Стьюдента было проведено сравнение $r_A(x)$ и $r_B(x)$ и установлено статистически значимое снижение FE после эксперимента для всех изучаемых генов. Этот алгоритм обобщен для нахождения совпадающих участков генов из списков А и В при различных мерах совпадения. После реализации алгоритма и проведения статистического анализа была определена группа генов, которая может быть связана

с изменениями в клетке, а также получены некоторые сведения о распределении силы связывания белка с рассмотренными генами.

Слова благодарности

Автор благодарит ведущего научного сотрудника РКНПК Д. Н. Пенькова за постановку задачи и предоставление данных и научного руководителя доцента кафедры теории вероятностей Е. Б. Яровую за помощь и постоянное внимание к работе.