Аннотация.

Работа посвящена разработке метода, позволяющего найти точную оценку вероятности наблюдения заданного IUPAC мотива по случайным причинам в последовательности нуклеотидов заданной длины. В ходе работы разработаны методы, основанные на различных подходах и имеющие разные области применения. Для каждого метода написана компьютерная программа, отличающаяся от других по скорости счёта и области применения. Среди полученных программ имеются следующие программы:

- Программа для нахождения точной оценки вероятности наблюдения заданного IUPAC мотива по случайным причинам в последовательности нуклеотидов заданной длины в прямой или обратной цепях ДНК с низкой скоростью вычислений.
- Программа для нахождения точной оценки вероятности наблюдения заданного IUPAC мотива по случайным причинам в последовательности нуклеотидов заданной длины в прямой цепи ДНК с более высокой скоростью вычислений, в сравнении с первой программой.

С помощью первой программы была оценена доля и выявлены типы мотивов для которых показаны наибольшие отклонения оценки вероятности согласно аппроксимационной формуле.

Тема работы

Уточненная оценка вероятности наблюдения заданного IUPAC мотива по случайным причинам в последовательности нуклеотидов заданной длины

Состав коллектива

- Куликова Дарья Константиновна, Механико-математический факультет НГУ, кафедра Дискретной математики и информатики, 4-й курс, группа 19133
- Вишневский Олег Владимирович, к.б.н., н.с. ИЦиГ СО РАН, научный руководитель.

1. Постановка задачи

Целью данной работы являлась разработка метода и компьютерной программы для расчета точной оценки вероятности хотя бы однократного наблюдения заданного мотива в последовательности заданной длины.

Решаемые задачи:

1. Разработка простого метода для вычисления точной оценки вероятности.
2. Сравнение простого метода с аппроксимационной формулой.
3. Разработка оптимизированного метода оценки встречаемого мотива.
4. Сравнение результатов работы опимизированной программы с простым точным методом.

2. Современное состояние проблемы (на момент начала работы).

В настоящее время для оценки ожидаемой вероятности наблюдения рассматриваемого контекстного сигнала в одиночной последовательности по случайным причинам используется аппроксимационная формула. Она рассчитывается очень быстро, но, поскольку является аппроксимацией и не учитывает внутренней контекстной структуры мотива, является неточной.

3. Описание работы

Для подсчёта вероятности хотя бы однократного наблюдения заданного IUPAC мотива в последовательности заданной длины был разработан алгоритм и написана основанная на нём программа. В основе данного алгоритма лежит идея о том, что вероятность наблюдения мотива в последовательности заданной длины равна сумме вероятностей всех последовательностей заданной длины, в которых встречается данный мотив.

Этот алгоритм также подходит для вычисления вероятности по меньшей мере однократного наблюдения IUPAC мотива в прямой или обратной цепи нуклеотидной последовательности заданной длины. Одновременно с поиском вхождения мотива в последовательности, также происходит поиск вхождения комплементарного ему мотива, и если хотя бы один или же оба мотива встретились в некоторой позиции, мотив считается найденным.

Для более быстрого вычисления вероятности наблюдения мотивов в работе Боевой и соавторов был предложен подход, основанный на методе деревьев префиксов. Этот метод универсальный, но избыточный для решаемой задачи. В связи с этим был разработан новый алгоритм, идейно опирающийся на метод деревьев префиксов, предназначенный для расчета вероятности наблюдения IUPAC-мотива.

Его суть следующая:

Для мотива строятся дерево префиксов и его матрица смежности. Для этого используется алфавит, состоящий из букв, которые встречаются в заданном мотиве, либо возникают при построении дерева. Также, поскольку нас интересует хотя бы одно вхождение, вершины, состояния которых соответствуют заданному мотиву, будут конечными.

По этому дереву по ниже приведённым рекурсивным формулам рассчитывается вероятность.

$P_L(\boldsymbol{M})$= P(G_L(k)) = P(G_{L-1}(k)) + $\sum_{q'}$ P$\left(C_{L-1}(q)\right)$ · P((q, k)).
P(C_n(q)) = $\sum_{q'}$ P$\left(C_{n-1}(q')\right)$ · P((q, q′)),
P$\left(G_0(k)\right) = 0$;
P$\left(C_0(0)\right) = 1$, P$\left(C_0(q)\right) = 0$, $q \neq 0$.

Где P(C_n(q)) – вероятность генерации последовательностей длины n оканчивающейся в вершине q, не являющейся конечной.
P(C_n(q)) = $\sum_{q'}$ P$\left(C_{n-1}(q')\right)$ · P((q, q′)), сумма пор всем возможным q', где q' – такая вершина, что в вершину q из неё можно попасть за одно ребро, (q,q') – ребро из q в q', P((q, q′)) - вероятность буквы, соответствующей ребру (q,q').
Где P(G_n(q)) – вероятность генерации последовательностей длины n оканчивающейся в конечной вершине q.

Эту формулы можно интерпретировать следующим образом.
Вероятность встречи мотива в последовательности длины L равна вероятности встречи мотива в последовательности длины L-1 плюс вероятность последовательности, оканчивающейся в вершине, из которой можно попасть в конечную по одному ребру, умноженная на вероятность буквы, соответствующей этому ребру.
Вероятность последовательности длины L закончиться в вершине q при проходе по дереву равна вероятности последовательности длины L-1, заканчивающейся в вершине q', из которой можно попасть в вершину q по одному ребру, умноженная на вероятность буквы, соответствующей этому ребру.

4. Полученные результаты.

В ходе работы был разработан алгоритм и написана программа, вычисляющая точную теоретическую оценку вероятности по меньшей мере однократного наблюдения IUPAC мотива в нуклеотидной последовательности. Программа может быть применена только к прямой цепи или к двум сразу. Однако данная программа является весьма ресурсоёмкой по времени. С её помощью определялась корректность работы

более сложных алгоритмов, а также была оценена доля и выявлены типы мотивов для которых показаны наибольшие отклонения оценки вероятности согласно аппроксимационной формуле.

На основе алгоритма основанного на методе деревьев был разработан алгоритм для оценки вероятности наблюдения IUPAC мотивов по случайным причинам в прямой цепи и написана программа, основанная на этом методе. Программа имеет более высокую скорость в сравнении с первой.

Также были предложены модификации алгоритма, для оценки вероятности встречи мотива в любой из двух цепей.

Для оценки отличий оценок вероятности переборного метода и аналитической формулы использовалась величина отклонения $\Delta = (P-W)/P = 1-W/P$, где P – результаты, полученные переборным методом, а W – результаты, полученные с помощью аппроксимационной формулы.

Получено, что число мотивов имеющих отклонение растёт с ростом длины последовательности.

Таблица 2.2.1 Процент мотивов, имеющих отклонение меньше 5%. Приведены данные для мотивов длины 5 на последовательности длины 7, 8, 9, 10.

Длина последовательности	**7**	**8**	**9**	**10**
Процент мотивов имеющих погрешность меньше 5%	80,23	77,07	73,63	71,64

По результатам вычислений программы, основанной на переборном методе, запущенной для всех возможных мотивов длины 4 на последовательности длины 11, получены следующие результаты:

- Наибольшую перепредставленность по точной формуле относительно аппроксимационной имеют мотивы с сильным самоперекрытием и вырожденными буквами.
- Наименьшее отклонение имеют мотивы, не имеющие жёстких ограничений на самоперекрытие.
- Наиболее недопредставленными по точной формуле относительно упрощенной являются мотивы, запрещающие самоперекрытие.

У сильно перепредставленыых и недопредставленных мотивов наблюдается высокий уровень погрешности, из чего следует необходимость использования более точных методов, нежели аппроксимационная формула. В то же время, аппроксимационная формула может быть применена к мотивам, не имеющим строгих запретов на самоперекрытие.

Эффект от использования кластера в достижении цели работы

Кластер ИВЦ НГУ позволял производить запуски программ с высоким временем исполнения, предоставляя для этого ресурсы и увеличивая скорость их выполнения.

Публикации

Куликова Д.К. Уточненная оценка вероятности наблюдения заданного IUPAC мотива по случайным причинам в последовательности нуклеотидов заданной длины//Биология: Материалы 61-й Междунар. науч. студ. конф. 17–26 апреля 2023 г. / Новосиб. гос. ун-т. — Новосибирск : ИПЦ НГУ, 2023 (Принято в печать).