Скрытые марковские модели на графах сборки

Попов Владимир Витальевич, гр. 17.Б04-мм

Санкт-Петербургский государственный университет
Математико-механический факультет
Кафедра статистического моделирования

Научный руководитель: к.ф.-м. н., д. Коробейников А.А.
Рецензент: к.ф.-м.н., д. Пржибельский А.Д.

Санкт-Петербург
2021г.

Введение

- Геном — носитель генетической информации.
- Ген — функциональный участок генома.

Одной из важных задач биоинформатики является предсказание генов — определение кодирующих участков (генов) в геноме.

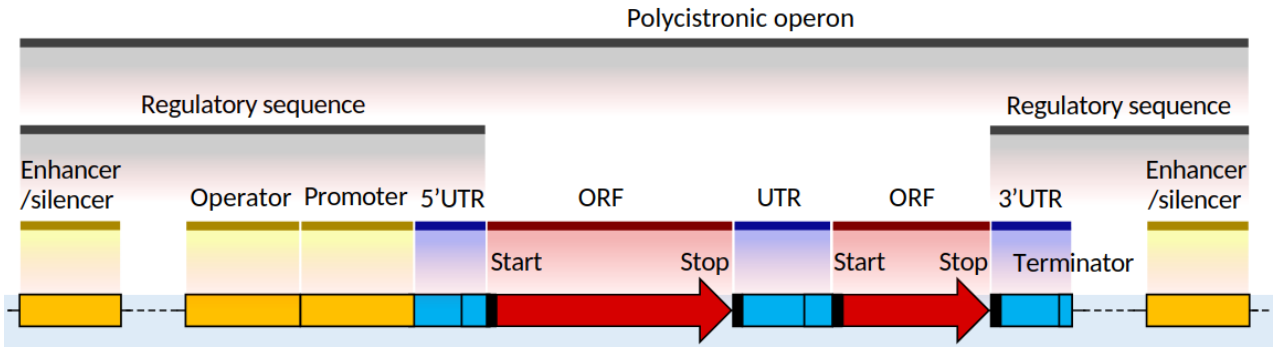


Рис.: Схема структуры гена

Ген имеет сложную структуру, особенности которой помогают при решении задачи предсказания.

Предсказание генов на строке

Существует множество алгоритмов для решения задачи предсказания генов, в частности:

- GLIMMER [A.L. Delcher, D. Harmon, 1999],
- MetaGene [H. Noguchi, J. Park, 2006],
- Prodigal [D. Hyatt, G. Chen 2010],
- FragGeneScan [M. Rho, H. Tang, 2010].

Предсказание генов на графе

Все существующие методы работают с геномом в виде строки.
Проблема: не существует метода для считывания целого генома.
Вместо этого считываются небольшие подстроки. Множество подстрок часто представляют в виде графа.

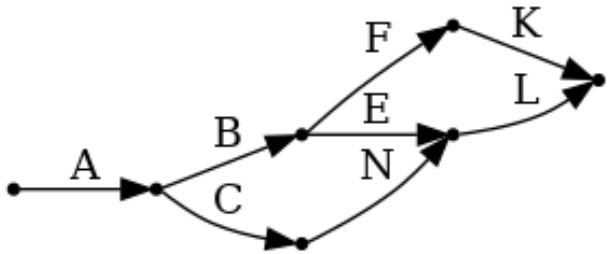


Рис.: Пример орграфа, A,B,. . . , L — строки над алфавитом A,C,G,T

Задача работы: разработать и реализовать эффективный алгоритм предсказания генов на ацикличных орграфах.
Предлагаемое решение: обобщение метода FragGeneScan.

Скрытая марковская модель

Работа метода FragGeneScan основана на скрытых марковских моделях.

Определение

X_t, Y_t — случайные процессы с конечными множествами значений X и Q. Скрытая марковская модель — пара (X_t, Y_t), такая, что:

- X_t — однородная цепь Маркова.
- $\mathbb{P}\big(Y_j \,\big|\, X_i, Y_i, 0 \le i \le j\big) = \mathbb{P}\big(Y_j \,\big|\, X_j\big)$, $\forall\, j \ge 1$.

Параметры распределения СММ:

- P — матрица переходных вероятностей X_t.
- p — вектор начального распределения X_t.
- B — матрица вероятностей, где $b_i(q_j) = b_{ij} = \mathbb{P}\big(q_j \,|\, x_i\big)$.

Обозначим множество параметров $\lambda = (P, p, B)$.

CMM FragGeneScan

Рассмотрим модель FGS. Дана строка S над алфавитом Q — геном.
$Q = \{A, C, G, T\}, X = \{Start, Coding, Stop, Noncoding\}$.
Блок кодирующих состояний — это специальная CMM, которая позволяет промоделировать структуру гена.

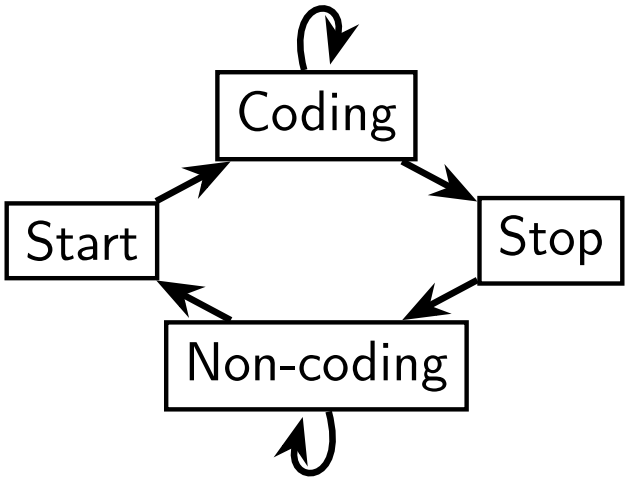


Рис.: Модель FragGeneScan

Возникает задача — найти строку состояний из X соответствующую символам строки S.

Алгоритм Витерби

Дано: СММ H с множествами значений X, Q и параметрами λ, а также последовательность $S = s_1, \ldots, s_L$ и путь $x = x_1, \ldots, x_L$.
Определим функцию правдоподобия:

$$\mathbb{P}(x, S|\lambda) = p_{x_1} b_{x_1}(s_1) p_{x_1 x_2} b_{x_2}(s_2) \times \cdots \times p_{x_{L-1} x_L} b_{x_L}(s_L).$$

Задача поиска наиболее подходящего списка состояний
Дано: $S = s_1, \ldots, s_L$, $s_i \in Q$.
Найти: $x^* = \operatorname{argmax}_{x:|x|=L} \mathbb{P}(x, S|\lambda)$.
Задачу решает алгоритм Витерби.

Решение задачи предсказания генов

Дана модель FGS H с известными параметрами, геном — строка $S = s_1 \dots s_L$ над алфавитом $Q = \{ A, C, G, T \}$.

- Алгоритм Витерби — получаем скрытый путь $x^* = (x_1 \dots x_L)$.
- Кодирующая последовательность — это подстрока S, скрытый путь для которой начинается со старта, заканчивается стопом, а все внутренние состояния — кодирующие.
- Для их нахождения используется алгоритм поиска кодирующих последовательностей.
  Обозначим: N — не кодирующие состояния, C — кодирующие состояния, O – старт, E — стоп.

Пример

$S = s_1 \dots s_{17}$, $x^* = (N, N, O, C, C, C, C, E, N, N, N, O, C, C, C, E, N)$.
Кодирующие последовательности: $s_3 \dots s_8$ и $s_{12} \dots s_{16}$.

Предсказание генов на графах

Перейдём к решению задачи предсказания генов на орграфах.
Идея обобщения алгоритма Витерби:

- Нужно обработать каждую строку
- В вершине сохраняем вероятности перехода от предков
- Марковость — состояния скрытого пути вычисляются последовательно

Обход графа производится в порядке топологической сортировки.

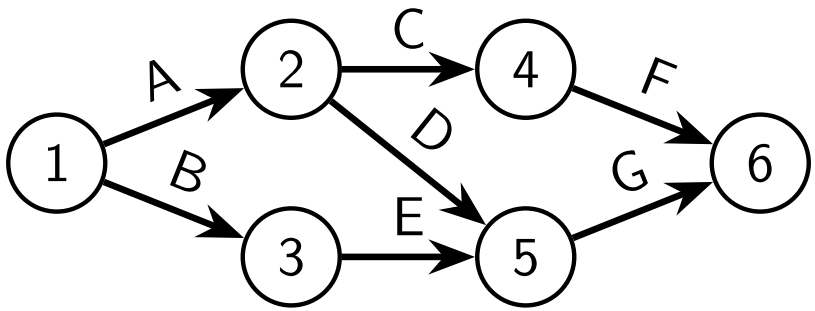


Рис.: Пример порядка обхода графа

Решение задачи предсказания генов на графе

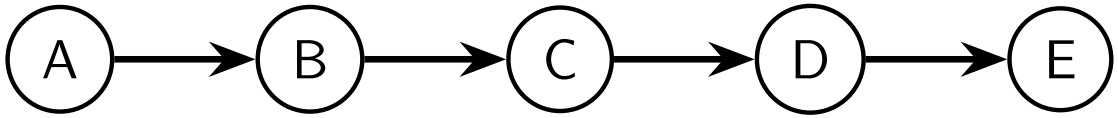
Схема алгоритма решения задачи предсказания генов на графах:

1. Обобщённый алгоритм Витерби для орграфов.
2. Алгоритм поиска кодирующих последовательностей в выравнивании.

Результат — множество кодирующих последовательностей.
Этот алгоритм был реализован на языке программирования C++.

Численные эксперименты

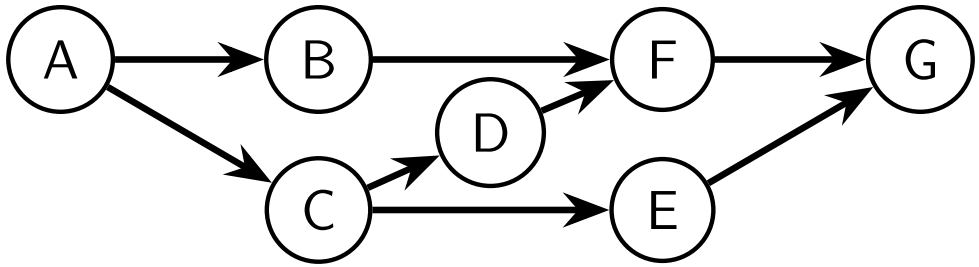
- Для проверки был взят геном $Escherichia\ coli$, разбит на 5 частей, которые были расположены на ребрах графа:



- В результате получено 4565 кодирующих последовательностей.
- Точно такие же последовательности были получены и в результате работы FragGeneScan на строке.

Численные эксперименты

Рассмотрим более сложный граф:



На пути $ABFG$ расположим первые 4200 символов генома $E.coli$, на остальных — случайные отрезки генома той же длины.

Истинные кодирующие последовательности на пути $ABFG$ найдены верно, а на общих участках путей получены одинаковые последовательности.

Численные эксперименты

Определение

Пусть дана СММ с параметрами λ. Значимость последовательности $S = s_0 \dots s_N$ со скрытым путём $x = x_0 \dots x_N$ равна $\ln \dfrac{\mathrm{P}(x, S|\lambda)}{\mathrm{P}(S|R)}$, где $\mathrm{P}(S|R)$ — вероятность того, что строка S была получена случайно.

Кодирующие последовательности имеют разную информативность. Расположим на ребрах использованного ранее графа:

1. случайные подстроки из генома,
2. случайно сгенерированные последовательности,

и сравним распределения полученных значимостей с истинными. Возьмём геномы *Staphylococcus aureus* и *Rhodobacter sphaeroides*.

Численные эксперименты

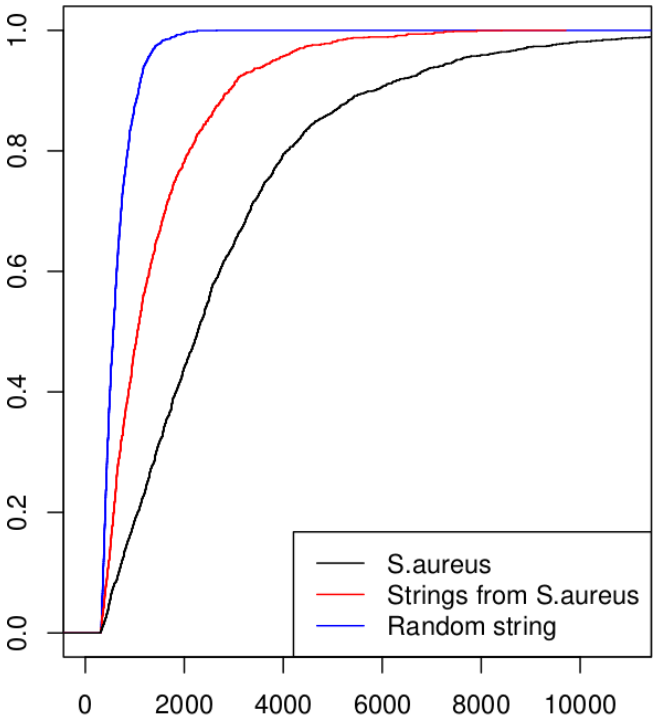


Рис.: $S.aureus$

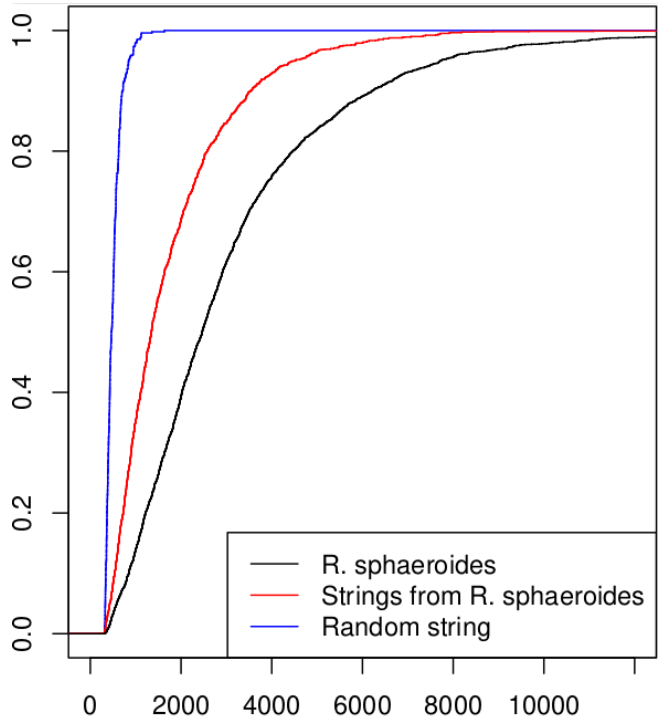


Рис.: $R.sphaeroides$

Рис.: Распределение значимостей последовательностей на случайных строках (синие), на случайных подстроках генома (красные) и истинных (черные).

Распределения смещены $\Rightarrow$ можно использовать значимость для классификации.

Выводы

- Алгоритм проверен на простом графе из пяти вершин, результаты полностью совпадают с работой FragGeneScan на строке.
- При усложении структуры графа корректно обрабатываются все пути на графе, и успешно находятся кодирующие последовательности на нескольких ребрах.
- Алгоритм протестирован на различных геномах. Отмечено, что отношение правдоподобия может использоваться для отличения истинных кодирующих последовательностей от случайных.

Заключение

В работе была рассмотрена задача предсказания генов на графах.

1. Изучена скрытая марковская модель FragGeneScan и способ решения задачи предсказания генов на строке.
2. Алгоритм метода обобщен на случай генома, представленного в виде ориентированного ациклического графа.
3. Алгоритм реализован на языке программирования C++ и протестирован как на случайных строках, так и на реальных геномах.

В дальнейшем алгоритм может быть обобщён на графы с циклами и применен для решения задачи предсказания генов на графах сборки.