Методы распознавания кодирующих последовательностей белков на геномных графах сборки

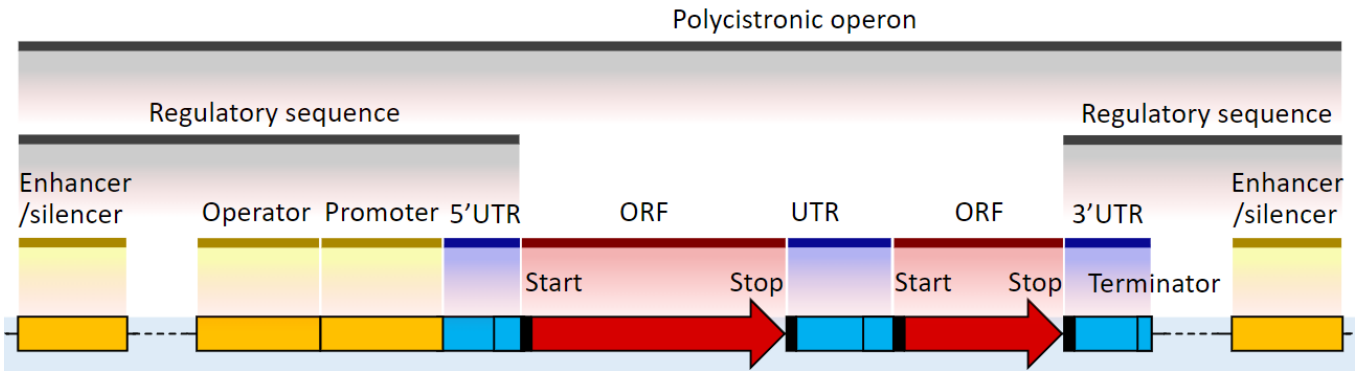
Логинов Андрей Сергеевич, группа 22.Б03-мм

Санкт-Петербургский государственный университет
Математическое моделирование, программирование и искусственный интеллект
Научный руководитель: к. ф.-м. н., доцент Шпилёв П. В.
Рецензент: к. ф.-м. н. Коробейников А. И.

7 июня 2024г.

Введение

- Геном — совокупность наследственной информации об организме. Хранится в виде ДНК.
- Кодирующая последовательность — участок ДНК, который кодирует белок.
- Распознавание кодирующих последовательностей или предсказание генов — одна из ключевых задач в биоинформатике.



У генов нет формального определения, но есть сложная структура, которая помогает в распознавании.

Существующие методы

Существуют различные методы предсказания генов:

- Glimmer (Delcher и др. 1999);
- GeneMark (Lukashin и Borodovsky 1998);
- Prodigal (Hyatt и др. 2010);
- FragGeneScan (Rho, Tang и Ye 2010).

Большая часть предназначена для работы с геномом в виде строки над алфавитом $\{A, T, G, C\}$.

Графы сборки

На практике геном чаще считывается в виде множества подстрок, которые представлены в виде графа:

- Рёбра графа — подстроки;
- Вершины — перекрытие ребра-предка и ребра-потомка.

С ростом числа развилок число путей через граф растёт экспоненциально.

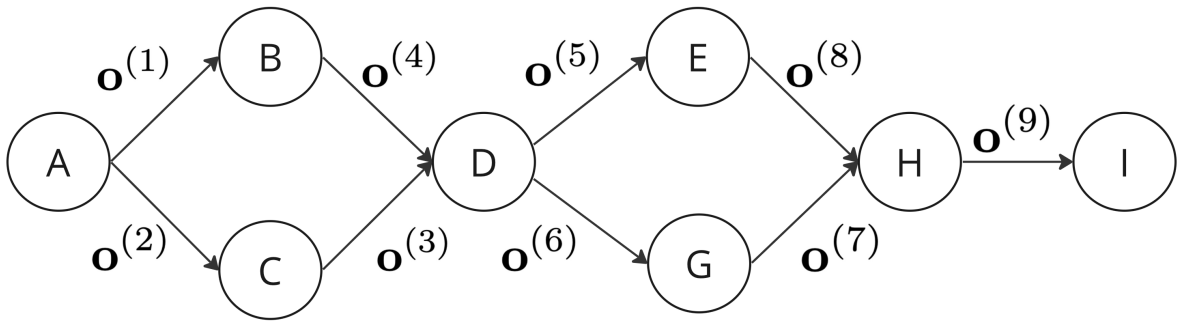


Рис.: Пример графа

Проблема: необходима разработка метода распознавания кодирующих последовательностей на графах сборки.
Предлагаемое решение: адаптация метода FragGeneScan.

Скрытые Марковские Модели

Метод FragGeneScan основан на скрытых марковских моделях.

Определение

X_t, Y_t — случайные процессы с дискретным временем и конечными множествами значений $\mathcal{X} = (x_1, \ldots, x_n)$, $\mathcal{Y} = (y_1, \ldots, y_m)$.
Пара (X_t, Y_t) называется скрытой марковской моделью (СММ), если

- *X_t — однородная марковская цепь*
- $\mathbb{P}(Y_i | X_1, \ldots, X_i, Y_1, \ldots, Y_{i-1}) = \mathbb{P}(Y_i | X_i),\ \forall i \geq 1.$

Состояния процесса X_t будем называть скрытым, а процесса Y_t — наблюдаемыми.

Часто при использовании СММ задача сводится к восстановлению последовательности скрытых состояний по имеющимся наблюдениям.

CMM FragGeneScan

Модель:

- Последовательность наблюдений — строка над $\{A, T, G, C\}$;
- Скрытые состояния — $\{Start, Stop, Coding, Non-coding\}$.
  Ген — подстрока, соответствующая скрытому состоянию $Coding$.

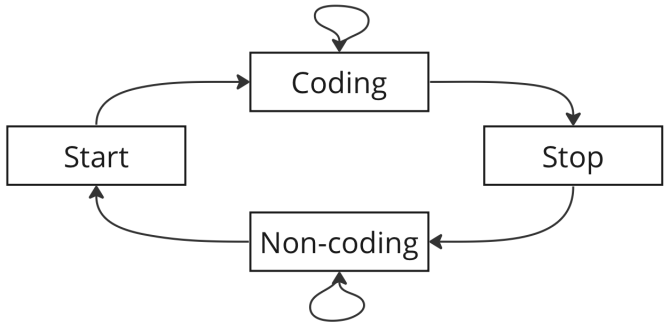


Рис.: Скрытые состояния CMM FragGeneScan

- Модель кодирующего региона задается CMM 2 порядка специального вида

Для нахождения последовательности скрытых состояний используется алгоритм Витерби (Viterbi 1967).
Марковость модель и алгоритм Витерби допускают обобщение на графы.

Предсказание генов на графах сборки

Алгоритм предсказания генов на строках метода FragGeneScan был обобщён для графов сборки.

Рис.: Алгоритм предсказания генов на строках метода FragGeneScan

Рис.: Предложенный алгоритм предсказания генов на графах сборки

Обобщение алгоритма Витерби

В алгоритм Витерби были внесены изменения:

- Жадная стратегия выбора предков для рёбер с несколькими предшествующими.
  Позволяет избежать полного перебора путей через граф.
- Запоминание выбранных предков.
  Необходимо для осуществления обратного хода.
- Использование алгоритма Тарьяна (Tarjan 1972) для определения порядка обхода рёбер.

Программная реализация

В качестве основы был взят исходный код FragGeneScan:

- Проведена переработка, исправлены неточности, которые могли приводить к произвольному результату (в том числе и на строках);
- Добавлены структуры данных и функции для обработки графов;
- Реализован обобщенный алгоритм Витерби;
- Реализовано восстановление последовательности и предсказание генов на графе.

Численные эксперименты. Критерии качества

- Сравнение с исходным методом проводилось на графах с известным расположением кодирующих последовательностей.
- Использованы метрики *точность* и *полнота*:

$$precision = \frac{TP}{TP + FP}$$
$$recall = \frac{TP}{TP + FN},$$

где TP — число верно найденных полных генов,
FP — число ошибочно предсказанных генов,
FN — число пропущенных генов.

Численные эксперименты

Для проверки был использован граф сборки генома *E. coli* MG1655.K-12:

- Были рассмотрены различные подграфы: тривиальный, без циклов и с циклами;
- Предложенный метод показал точность, такую же как исходный, но большую полноту.

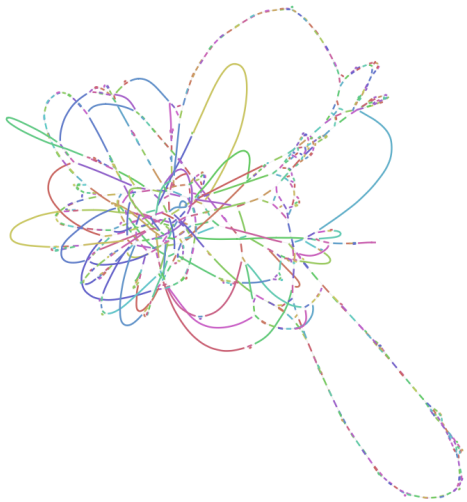


Рис.: Схема графа сборки генома *E.coli*.

Наибольший интерес представляют результаты работы на полном графе.

Полный граф

Доработанный метод FragGeneScan для графов показывает большую полноту, но меньшую точность:

Метод	FragGeneScan	FragGeneScan для графов
Всего генов	4657	
Предсказано генов	4586	5095
Верно предсказано	3421	3539
Точность	0.746	0.695
Полнота	0.736	0.762

Анализ ошибок

Большая часть верно предсказанных доработанным методом генов на графе лежит внутри 1 ребра:

	Гены на 1 ребре	Гены на 2 и более рёбрах
Число истинных	4521	126
Всего предсказано	4751	344
Верно предсказано	3525	14
Точность	0.742	0.041

- Ошибки в предсказании генов, лежащих на нескольких рёбрах, связаны с невозможностью повторного прохода через циклы в текущей реализации.
- Имеет место унаследованная от исходного метода проблема корректного распознавания начала гена.

Выводы

- На ориентированных ацикличных графах удалось достичь улучшения полноты и точности предсказания генов;
- На полном графе удалось добиться увеличения полноты, но снизилась точность;
- Возникающие ошибки связаны с некорректной обработкой циклов, а также унаследованы от исходного метода.

Предложенная доработка FragGeneScan для графов справляется с задачей предсказания генов лучше исходного метода.

Заключение

В работе была рассмотрена задача детектирования кодирующих последовательностей на графах сборки:

- Был изучен аппарат скрытых марквоских моделей и основанный на них метод FragGeneScan;
- Был предложен способ обобщения алгоритма Витерби и метода FragGeneScan для графов сборки;
- Разработанный метод показывает себя лучше исходного на простых графах и не хуже исходного на полном графе сборки;
- Реализация корректной обработки циклов позволит улучшить качество решения задачи предсказания генов на полном графе сборки.