Оценка оптимальности результатов работы жадного иерархического алгоритма для решения задачи о наименьшей общей надстроке

Швецова А. А.

Научный руководитель: к.ф.-м.н. Д. Н. Москвин

Научный консультант: д.ф.-м.н. А. С. Куликов

Санкт-Петербургская школа физико-математических и компьютерных наук

НИУ ВШЭ – Санкт-Петербург

2020

Задача о надстроке

Условие

Дано n строк, ни одна из которых не является подстрокой другой. Нужно найти строку минимальной длины, включающую все данные строки в качестве подстрок.

Замечание

Если строки не являются подстроками друг друга, то задача эквивалентна нахождению оптимальной перестановки строк.

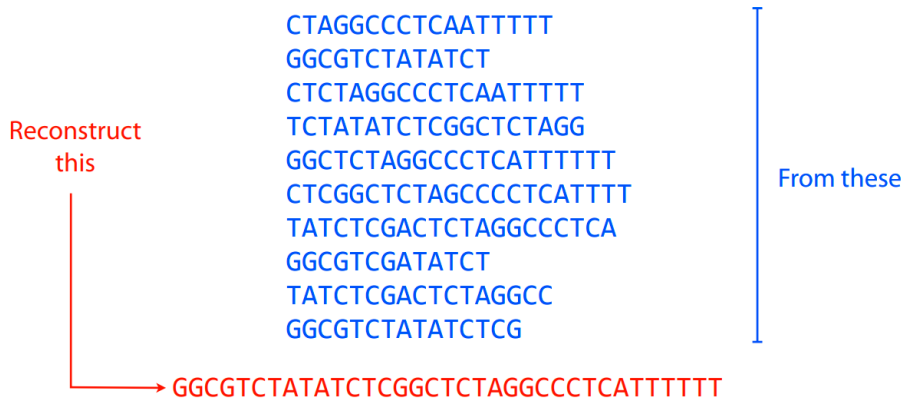
Пример надстроки

Входные данные: aaa, cae, aec, eee

Пример ответа: aaa $\rightarrow$ aec $\rightarrow$ cae $\rightarrow$ eee $=$ aaaecaeee

Мотивация

- Задача используется при секвенировании ДНК [1][2]
- Важны и точность алгоритма, и скорость работы
- Задача о надстроке NP-полная, поэтому часто используют приближенные решения.



[1] Jiang, T., & Li, M. (1996). DNA sequencing and string learning. Mathematical Systems Theory, 29(4), 387–405. doi:10.1007/bf01192694

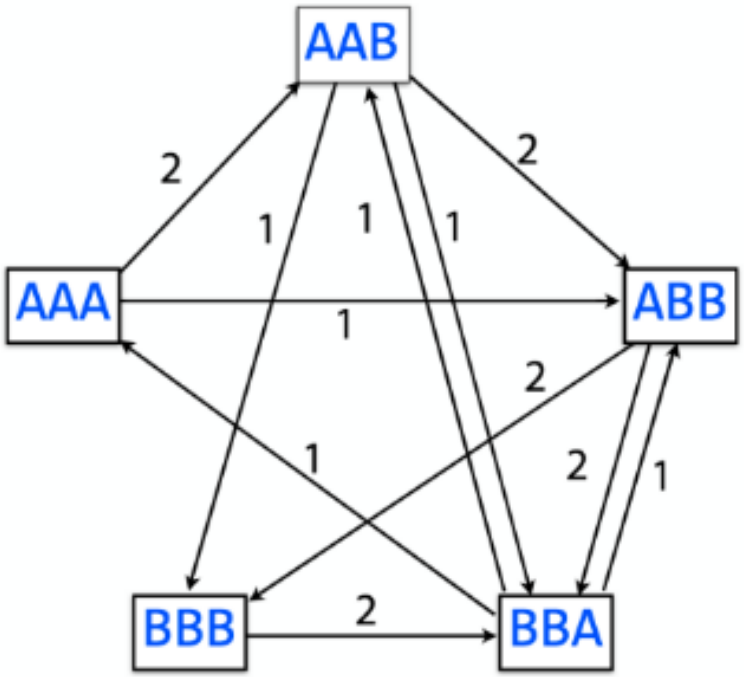
[2] Kececioglu, J. D., & Myers, E. W. (1995). Combinatorial algorithms for DNA sequence assembly. Algorithmica, 13(1-2), 7–51. DOI:10.1007/bf01188580

C-оптимальный алгоритм

C-оптимальный алгоритм

Назовём алгоритм c-оптимальным, если длина генерируемой им надстроки превышает оптимальную не больше чем в c раз.

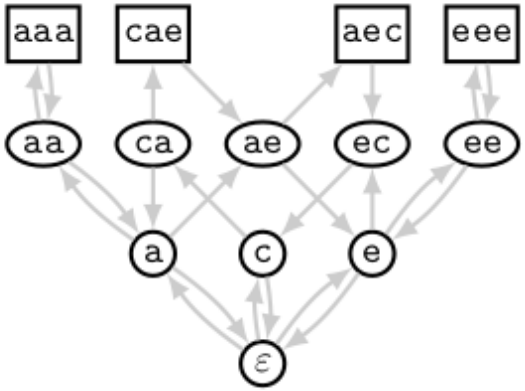
Обзор работ

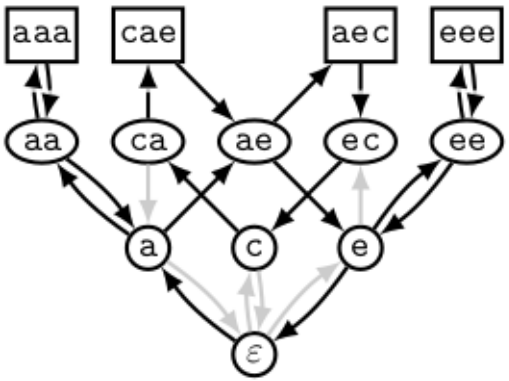


- Наиболее популярный и универсальный подход: Асимметричная задача коммивояжера (ATSP)
- Используется в качестве сведения с 1990 г.
- Лучший результат: полиномиальный алгоритм с точностью $2\frac{11}{23}$. [3]
- Основные проблемы: отсутствие чёткой структуры, неприспособленность под задачу.

[3] Mucha, M. (2013). Lyndon Words and Short Superstrings. Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, 958–972. DOI:10.1137/1.9781611973105.69

Иерархический граф [4]





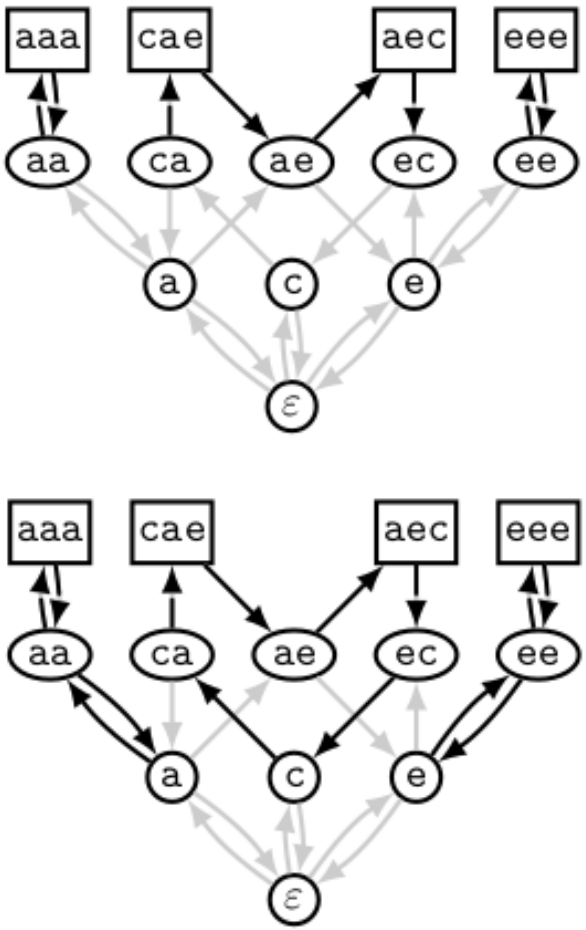
aaa → aec → cae → eee

aaaecaeee

- Вершины графа – всевозможные подстроки строк входного набора, в т.ч. пустая строка ε.
- Вершины сгруппированы в слои по длине.
- Рёбра графа бывают двух видов:
  - $S \to Sb$ – рёбра вверх.
  - $aS \to S$ – рёбра вниз.
- Решения – эйлеровы циклы проходящие через входные строки и ε

[4] Golovnev, A., & Kulikov, A. & Logunov, A., & Mihajlin, I. (2019). Collapsing superstring conjecture. DOI: 10.4230/LIPIcs.APPROX/RANDOM.2019.26

Greedy Hierarchical Algorithm (GHA)

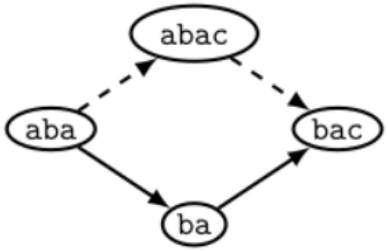


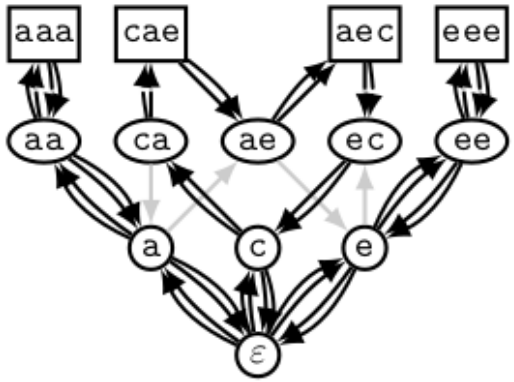
- Полиномиальный алгоритм для решения задачи о надстроке
- Оптимальность работы в общем случае грубо оценена в 3.5.
- 2-оптимален для строк длины 3 и 4
- Оптимален для строк длины 2 и k-спектров строк

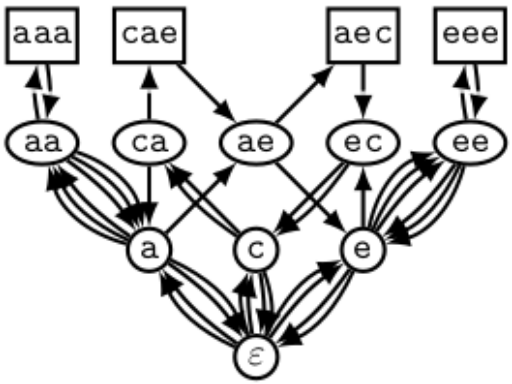
Алгоритм:

- Послойно поддерживать баланс входящих-исходящих рёбер в вершинах, сохраняя связность, где необходимо.

Collapsing Algorithm (CA)







Нормализация: переставим между собой операции добавления и удаления символа.
Алгоритм:

- Удвоить произвольный цикл-решение.
- Послойно применить к нему операцию нормализации, сохраняя связность, где необходимо.

CA & GHA

Авторами работы было доказано следующее утверждение:

Если результаты работы $CA(D \sqcup D) = GHA(G)$ вне зависимости от начального решения D, то GHA 2-оптимален.

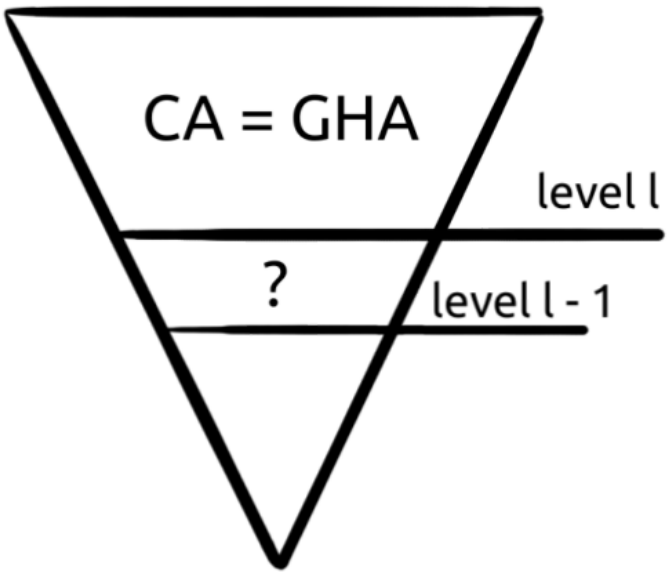
Цели и задачи

Цель

Предложить новую гипотезу о работе алгоритмов GHA и CA, из которой следовала бы эквивалентность их результатов.

1. Итеративно сравнить работу алгоритмов GHA и CA, выделить проблемные вершины.
2. Предложить свойство графа, доказывающее эквивалентность исполнения в проблемных вершинах.
3. Проверить, что гипотеза выполняется на случайных данных.

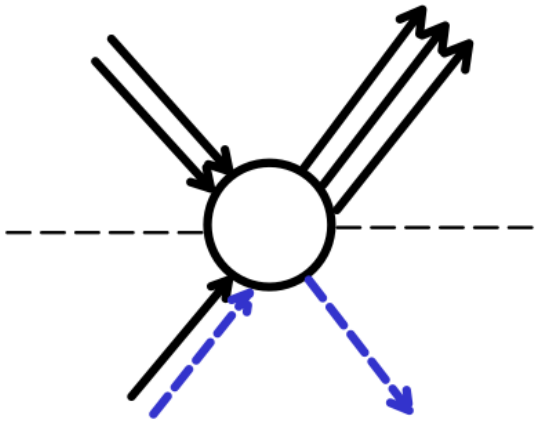
Базовые соображения

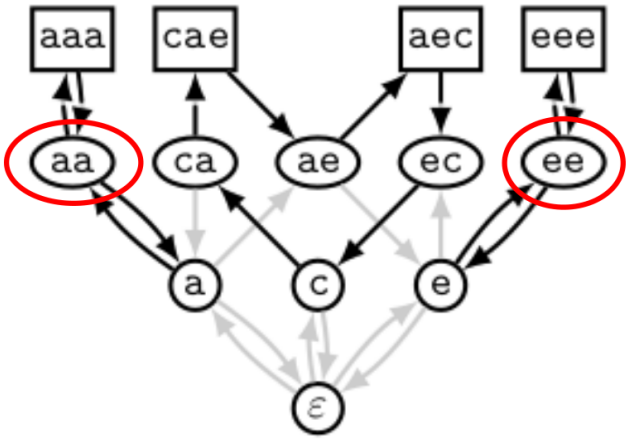


В основу доказательства входит следующая индукция по слоям:

- Пусть были обработаны слои от верхнего до $l + 1$, в результате чего GHA и CA совпали от слоя l и выше.
- Нужно доказать, что после обработки следующего слоя они совпадут до слоя $l - 1$.
- Докажем, что все нижние рёбра вершин уровня l совпали.

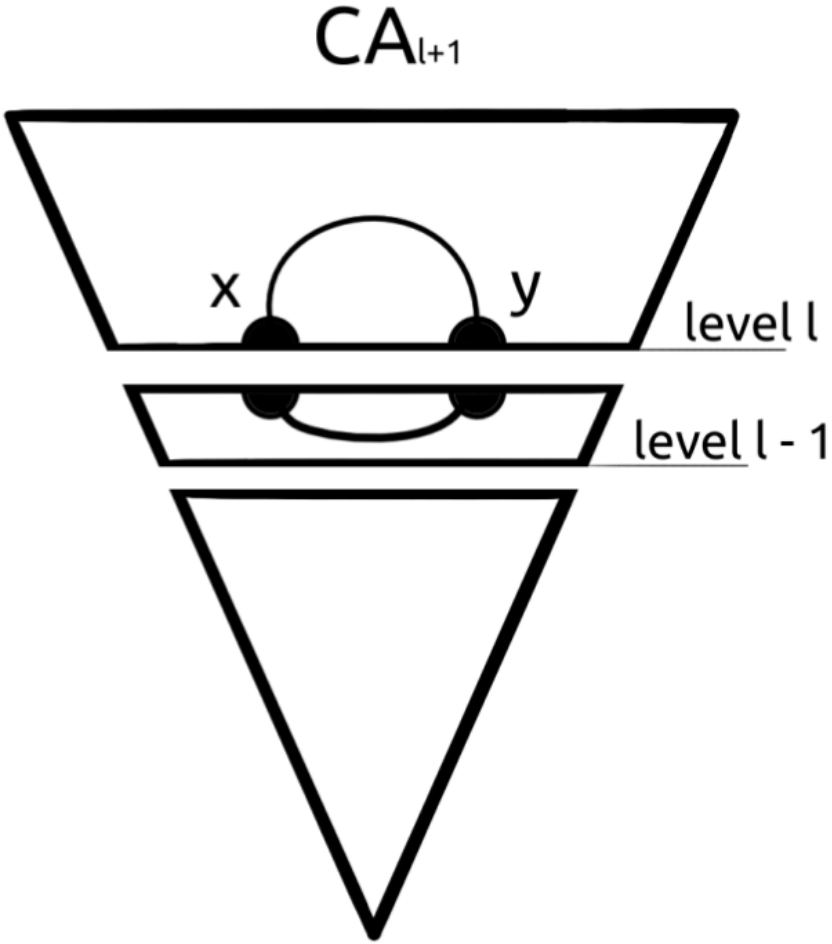
Итеративное сравнение GHA и CA





- Для большинства вершин: GHA просто поддержит баланс, CA нормализует всё, что нормализуется.
- Для "особых" вершин, сохраняющих связность: почему у разных алгоритмов они совпадают?

Достаточное условие



Разрежем граф $CA_{l+1}(D \sqcup D)$ по слоям l и $l-1$. Получим 3 его части: верхнюю, среднюю и нижнюю.

Достаточное условие

Если две вершины x и y слоя l связны в верхней части графа, то в средней части графа они также должны быть связны.

Если это условие выполняется, то множества особых вершин на уровне l у алгоритмов равны и тогда GHA = CA.

Проверка гипотезы

- Используя существующие алгоритмы генерации датасетов для задачи о надстроке [5], было сгенерировано 3 000 000 наборов входных данных.
- Для каждого набора было сгенерировано несколько перестановок, построены их обходы в графе и запущен CA.
- Послойно проверено выполнение гипотезы в процессе исполнения.

[5] Romero, Heidi Brizuela, Carlos & Tchernykh, Andrei. (2004). An Experimental Comparison of Approximation Algorithms for the Shortest Common Superstring Problem. Proceedings of the Fifth Mexican International Conference in Computer Science, ENC 2004. 27-34. 10.1109/ENC.2004.1342585

Результаты

- Найдены и описаны проблемные места в сравнении алгоритмов GHA и CA.
- Предъявлена новая, более простая гипотеза, из которой следует 2-оптимальность алгоритма GHA.
- Гипотеза проверена на миллионах наборов входных данных и скорее всего является верной (код представлен здесь: `github.com/annikura/CA-hypothesis-checker`).
- Ведётся активная работа в направлении формального доказательства гипотезы.