Вычисление диаметра в задаче сортировки однохромосомных геномов транспозициями

Студент 5 курса СПбАУ Илья Минкин
Руководитель Макс Алексеев

Геномные перестройки

- В ходе эволюции геном изменяется
- Геном может изменяться по-разному
- Незначительные изменения — точечные мутации, вставки/удаления небольших фрагментов
- Крупные изменения — геномные перестройки, которые меняют расположение значительных участков хромосомы
- В результате геномных перестроек хромосомы некоторых организмов выглядят как разные перестановки одинаковых блоков

Геномные перестройки

- Существует несколько операций над геномом, которые относят к перестройкам
- Reversal — изменяется порядок следования некоторых «блоков», из которой составлена хромосома
- Transposition — меняются местами два соседних участка хромосомы
- Transereversal — комбинация reversal и transposition
- Задача — восстановить сценарий, по которому происходила эволюция организмов

Сортировка транспозициями

- Заданы два генома A и B, каждый из которых состоит из 1 хромосомы
- Предполагается, что хромосомы обоих геномов представляют собой наборы одних и тех же уникальных блоков, но расположенных в разном порядке
- Требуется найти кратчайшую последовательность транспозиций, превращающую один геном в другой

Сортировка транспозициями

- Заданы два генома A и B, каждый из которых состоит из 1 хромосомы
- Предполагается, что хромосомы обоих геномов представляют собой наборы одних и тех же уникальных блоков, но расположенных в разном порядке
- Требуется найти кратчайшую последовательность транспозиций, превращающую один геном в другой

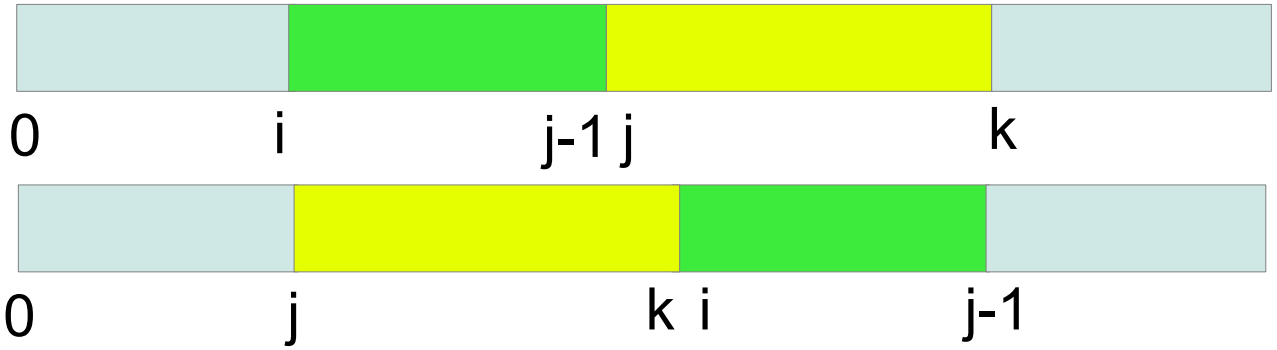
Формальная постановка задачи

- Пусть порядок генов на хромосоме представлен в виде числовой перестановки

  $\pi = \pi_1 \pi_2 \dots \pi_n, \quad \pi_i \in \{1,2, \dots n\}, \quad \pi \in S_n,$

  Где S_n это множество всех перестановок длины *n*

- Транспозиция $\rho_{ijk}, \quad 1 \leqslant i < j \leqslant k \leqslant n$ меняет местами 2 соседних фрагмента перестановки $\pi_i \pi_{i+1} \dots \pi_{j-1}$ и $\pi_j \pi_{j+1} \dots \pi_k$



- Транспозиционный граф это $G=(V,E), V=S_n$
  $E=\{\pi\sigma \mid \exists i<j\leqslant k, \pi\cdot\rho_{ijk}=\sigma\}$
- Транспозиционным расстоянием $d_T(\pi,\sigma)$ между парой перестановок π и σ называется расстояние между ними в графе *G*
- Можно показать, что
  $d_T(\pi,\sigma)=d_T(id,\pi\sigma^{-1})=d_T(id,\sigma\pi^{-1}),$
  где $id=1\,2\ldots n$ — тождественная перестановка
- Транспозиционным диаметром *TD(n)* называется диаметр графа *G*
  $TD(n)=max\{d_T(id,\pi) \mid \pi\in S_n\}.$

- Задача сортировки транспозициями — для данной числовой перестановки π найти кратчайшую последовательность транспозиций, которая переводит π в id
- Содержательный пример - Пусть требуется отсортировать перестановку «*2 1 5 4 3*»

2 1 5 4 3 → 2 1 4 3 5
2 1 4 3 5 → 2 1 3 4 5
2 1 3 4 5 → 1 2 3 4 5

История вопроса

- В 1996 году был опубликован 1.5 приближенный алгоритм для задачи сортировки транспозициями

  [Bafna, Pevzner, 1996]

- Они также посчитали *TD(n)* для $n \leq 10$ и предположили, что *TD(n)* = $\lceil (n+1)/2 \rceil$ для всех *n*

- Были доказаны оценки $\left\lceil \frac{n+1}{2} \right\rceil \leq TD(n) \leq \left\lfloor \frac{2n-2}{3} \right\rfloor$

- Высказано предположение, что перестановка

  rev = n n - 1 n - 2 … 2 1

  всегда будет принадлежать к числу самых удаленных

- В 2001 году было опровергнуто предположение о точном значении *TD(n)* и самой удаленной перестановке *rev*

  [H. Eriksson et al., 2006]
- *TD(13) = 8,* что на 1 больше, чем $\lceil (13+1)/2 \rceil = 7$
- Перестановка «*4 3 2 1 5 13 12 11 10 9 8 7 6*» требует 8 транспозиций для сортировки и является более удаленной, чем *rev*
- Они также доказали новые оценки и вычислили точные значения *TD(n), n ≤ 15*

$$\left\lceil \frac{n+1}{2} \right\rceil \leq TD(n) \leq \left\lfloor \frac{2n-2}{3} \right\rfloor$$

- В 2006 году был опубликован 1.375 приближенный алгоритм с временем работы $O(n^2)$

  [I. Elias and T. Hartman., 2006]
- Они также улучшили нижнюю оценку показав, что при $n \leq 14$

$$TD(n) \geq \left\lfloor \frac{n+1}{2} \right\rfloor + 1$$

- Наконец, в 2011 году было доказано, что сортировка транспозициями это *NP*-полная задача

  [L. Bulteau, G. Fertin, I. Rusu., 2011]

Точные значения транспозиционного диаметра

[H. Eriksson et al., 2006], http://oeis.org/A065603

n	*TD(n)*	*n*	*TD(n)*
1	0	9	5
2	1	10	6
3	2	11	6
4	3	12	7
5	3	13	8
6	4	14	8
7	4	15	9
8	5	16	9 или 10?

Цель работы

- Попытаться вычислить точные значения *TD(n)* для как можно большего числа *n*
- Попробовать построить общий вид наиболее удаленных перестановок (относительно транспозиционного расстояния)

С чего можно начать

- Пусть перестановка длины *n,* требующая максимально возможное число транспозиций для сортировки, называется $d_{max}(n)$
- Можно найти все такие перестановки для $n \leq 10$ при помощи обычного обхода в ширину
- Ничего нового мы не вычислим, но можно будет посмотреть, как выглядят $d_{max}(n)$

Результаты

n	Диаметр	Количество перестановок	Доля наиболее удаленных перестановок
1	1	1	100%
2	1	1	50%
3	2	1	16.7%
4	3	1	4.2%
5	3	31	25.8%
6	4	45	6.3%
7	4	1513	30.0%
8	5	2836	7.0%
9	5	114327	31.5%
10	6	255053	7.0%
11	6	12537954	31.4%

Анализ

- Результаты совпадают с ранее опубликованными, см. http://oeis.org/A164366
- Чтобы проанализировать структуру самых удаленных перестановок для разных *n,* можно воспользоваться диаграммами циклов, которые ввели в своей работе Bafna и Pevzner

Cycle graph

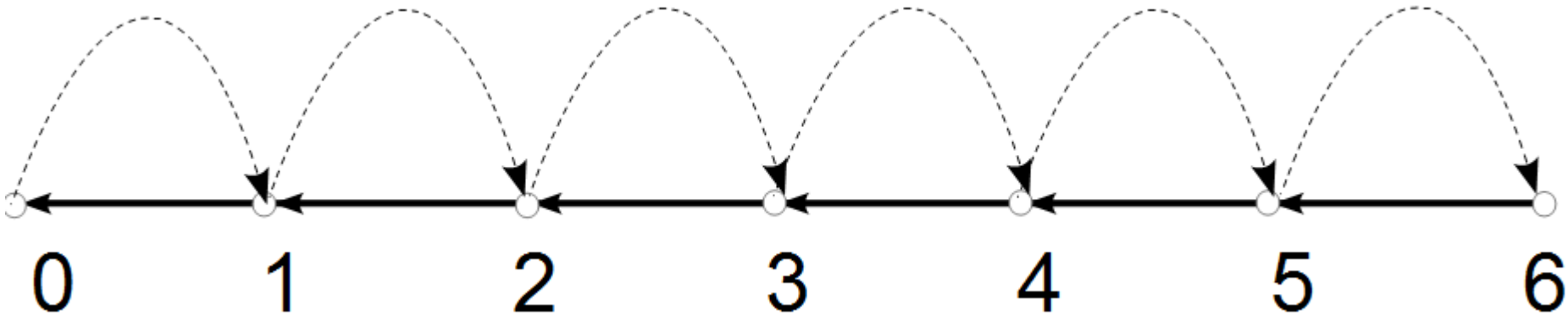
- К перестановке длины *n* приписывается *0* в начало, в конец дописывается *n + 1*
- Каждое число в перестановке это вершина графа
- Два типа ребер:

1. «Черные» - ориентированное ребро ведет из π_i в π_{i-1}
2. «Серые» - ориентированное ребро ведет из π_i в $\pi_i + 1$

- Такой граф однозначным образом разбивается на циклы, в котором чередуются ребра разных цветов

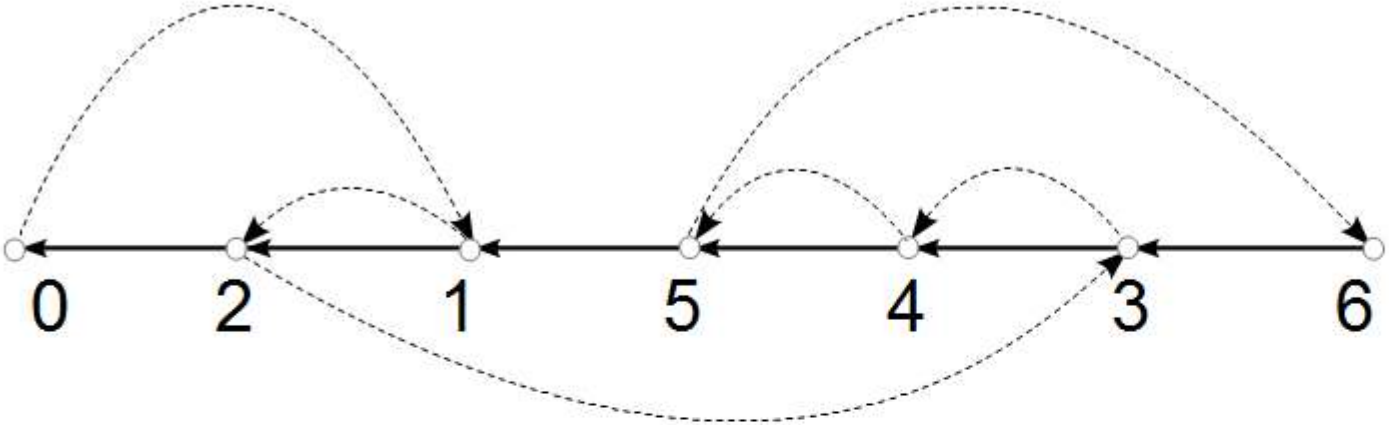
Содержательный пример №1

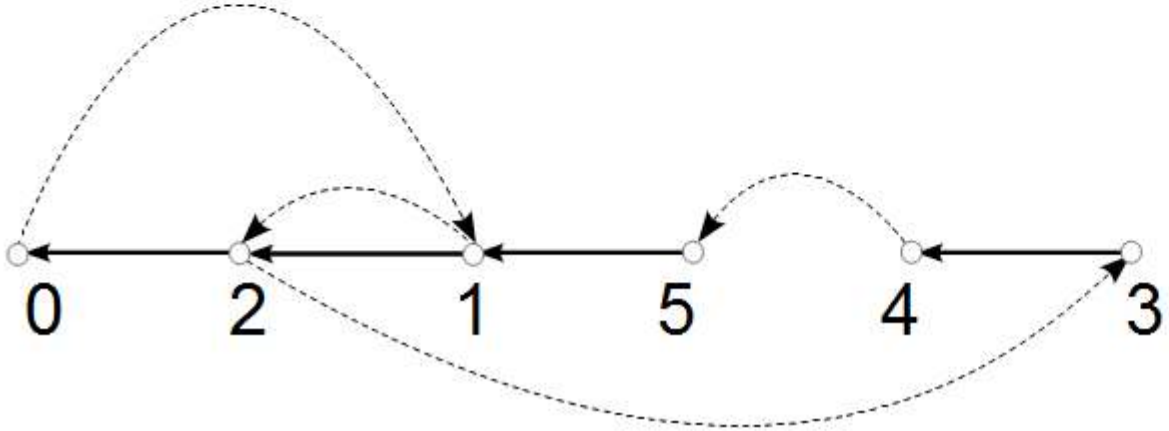
$\pi = id$

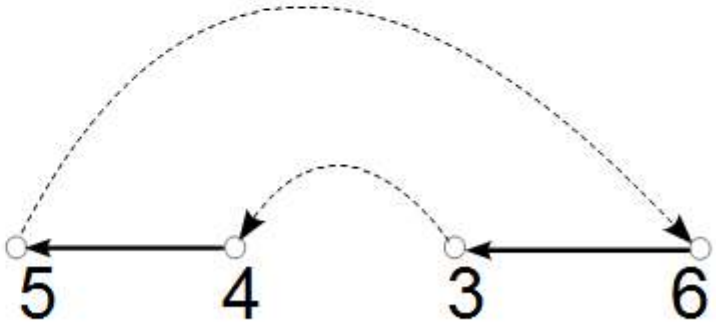


Содержательный пример №2

π = *«2 1 5 4 3»*







Что с этим делать

- Количество циклов на cycle graph перестановки влияет на ее транспозиционное расстояние
- Самых удаленных перестановок для $n \leq 11$ очень много
- Попробуем собрать некоторую «статистику» для самых удаленных перестановок, а именно, из циклов какого размера они состоят

Паттерны

- Пусть диаграмма циклов перестановки π разбивается на *k* циклов, а количество ребер в *i-м* цикле это *sc(i)*
- Запишем числа *sc(i)* в порядке неубывания, получившаяся строка это ***паттерн***
- Пример — перестановка *«2 1 5 4 3»* разбивается на циклы *«0 1 2 3 4 5 1 2 0»* и *«3 4 5 6 3»*
- Тогда паттерн перестановки *«2 1 5 4 3»* это *«4 8»*
- Какие бывают паттерны и сколько им соответствует наиболее удаленных перестановок для разных *n?*

$n = 5, 6, 7$

n	Паттерн	Количество наиболее удаленных перестановок, подходящих под паттерн	Количество всех перестановок, подходящих под паттерн
5	2 10	6	48
	6 6	1	12
	4 8	24	24
6	14	45	180
7	2 14	360	1440
	6 10	136	608
	4 12	720	720
	8 8	276	276
	4 4 4 4	21	21

$n = 8, 9$

n	Паттерн	Количество наиболее удаленных перестановок, подходящих под паттерн	Количество всех перестановок, подходящих под паттерн
8	4 4 10	63	1728
	18	2773	8064
9	2 4 4 10	630	17280
	2 18	27730	80640
	4 4 6 6	175	4440
	6 14	10470	31680
	10 10	4402	13248
	4 16	40320	40320
	8 12	27360	27360
	4 4 4 8	3240	3240

$n = 11$

n	Паттерн	Количество наиболее удаленных перестановок, подходящих под паттерн	Количество всех перестановок, подходящих под паттерн
11	2 6 8 8	132	831600
	2 4 8 10	39600	1957824
	2 4 4 14	71940	1378080
	2 22	2948964	7257600
	6 6 6 6	21	38720
	4 6 6 8	11832	506880
	4 4 6 10	42612	598752
	6 18	1096368	2741760
	10 14	850860	2108160
	4 20	3628800	3628800
	8 16	2298240	2298240
	4 4 4 12	245520	245520
	12 12	1022400	1022400
	4 4 8 8	279180	279180
	4 4 4 4 4 4	1485	1485

n = 13

n	Паттерн	Количество наиболее удаленных перестановок, подходящих под паттерн	Количество всех перестановок, подходящих под паттерн
13	10 18	238	258435072
	14 14	189	118644480

- При *n = 13* красивая закономерность уже не работает
- Можно скорректировать гипотезу и предположить, что для «определяющих паттернов» сохраняется транспозиционное расстояние $\lceil (n+1)/2 \rceil$

Результаты

- Сделано предположение, что существует класс перестановок, для которых сохраняется транспозиционное расстояние $\lceil (n+1)/2 \rceil$ для всех нечетных *n,* и этот класс можно легко описать
- Можно добавить новые значения в OEIS

Ссылки на источники

1. L. Bulteau, G. Fertin, I. Rusu. Sorting by Transpositions Is Difficult. ICALP (1) 2011: 654-665.
2. V. Bafna and P. A. Pevzner. Sorting by transpositions. SIAM J. Discrete Math., 11(2):224–240, 1998.
3. I. Elias and T. Hartman. A 1.375-approximation algorithm for sorting by transpositions,IEEE/ACM Trans. Comput. Biol. Bioinform., 3 (2006), pp. 369–379.
4. L. Lu, Y. Yang. A lower bound on the transposition diameter. SIAM J. Discrete Math., 24(4):1242–1249, 2010.
5. H. Eriksson, K. Eriksson, J. Karlander, L. Svensson, and J. Wastlund. Sorting a bridge hand. Discrete Math., 241(1-3):289–300, 2001.
6. Последовательность OEIS A164366. http://oeis.org/A164366
7. R. de A. Hausen et al. On the Toric Graph as a Tool to Handle the Problem of Sorting by Transpositions. BSB '08 Proceedings of the 3rd Brazilian symposium on Bioinformatics: Advances in Bioinformatics and Computational Biology, 79 – 91, 2008