Структурный и эволюционный анализ человеческих центромер

Кунявская Ольга Александровна

научный руководитель: Певзнер П.А.

НИУ ВШЭ — Санкт-Петербург

июнь 2021 г.

Введение

Центромера — важный участок хромосомы

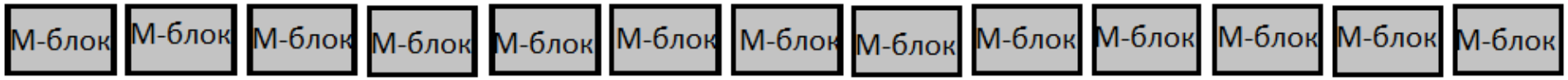
- Играет ключевую роль в процессе деления клетки
- Нарушения в центромере могут вызывать рак и бесплодие
- Расшифрована только в 2019 году
- Не изучена связь между генетической последовательностью центромер и болезнями
- Для дальнейших биомедицинских и эволюционных исследований необходима аннотация: разбиение на смысловые блоки.

Задача аннотации центромеры

Центромера – строка над алфавитом {A, G, C, T}

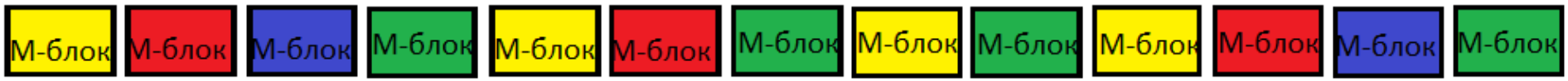
AGCTGCTT...

Мономерные блоки – похожесть на 65-100%, длина ≈ 171нт



Мономерные блоки можно разбить на кластеры

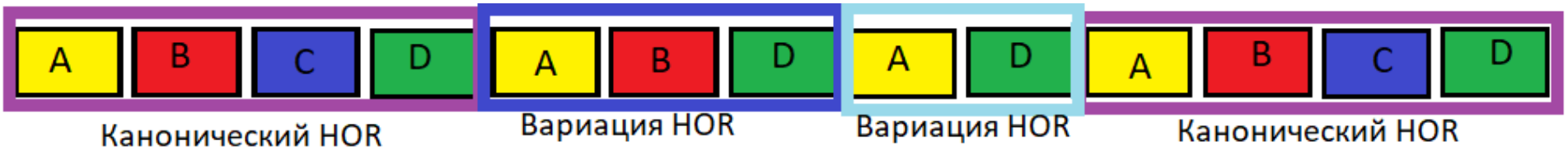
Усредненный блок в кластере – мономер



Моноцентромера — центромера над алфавитом мономеров

Мономеры образуют повторы — HOR

Ограничение: канонический HOR состоит из разных мономеров



Методы аннотации центромер

Существует только **полуручная аннотация:**

- Медленно: сейчас проаннотирован 1 человек, планируется – 100
- Не описана экспертная методика. В том числе нет математического определения мономеров и целевой функции, которая бы позволяла сравнивать две разные аннотации

Полностью автоматическая аннотация – большая задача, которая сейчас решается в рамках T2T консорциума.

Части:

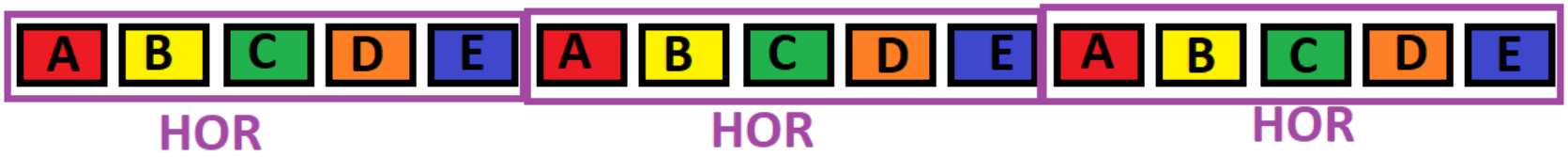
1. Выделение мономерных блоков
2. **Выделение мономеров (текущая работа)**
3. Декомпозиция на HOR

Выделение мономеров

Инструменты-аналоги: Alpha-CENTAURI, HORdetect

- Не работают с целыми центромерами
- Не учитывают контекст: некоторые мономеры ошибочно объеденены/разделены и из-за этого нельзя выделить канонический HOR

Желаемый вывод:



Пример вывода:

Цель и задачи

Цель:

Создать метод автоматического выделения мономеров, которые бы позволили декомпозировать центромеру на HOR.

Задачи:

1. Формализовать требования к мономерам
2. Выбрать метод разбиения на мономерные блоки
3. Разработать и реализовать метод кластеризации мономерных блоков в мономеры
4. Добавить учёт контекста мономеров
5. Сравнить с референсными мономерами

Требования к мономерам

Необходимо найти минимальное множество мономеров (строк) M, такое что:

1. длина каждого мономера $\sim$ 171 (150-190) [1]
2. 99% центромеры можно разбить на мономерные блоки так, чтобы расстояние до ближайшего мономера $\leq$ 5% [2]
3. $\exists$ строка S из мономеров (канонический HOR) такая что[3]:
   - в S входят все мономеры ровно по одному разу
   - тандемные повторы вида $SSS\ldots$ покрывают хотя бы 30% центромеры

---

[1] Miga, K. H. (2020). Centromere studies in the era of ‘telomere-to-telomere’genomics //Experimental cell research

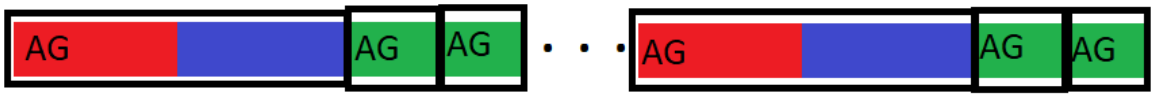
[2] Alexandrov I. et al.(2001) Alpha-satellite DNA of primates: old and new families //Chromosoma

[3] McNulty, S. M., Sullivan, B. A. (2018). Alpha satellite DNA biology: finding function in the recesses of the genome //Chromosome Research

Методы разбиения на мономерные блоки

key-string method

- Часто ошибочно разбивает мономерный блок на насколько частей или склеивает мономерные блоки вместе.



TandemRepeatFinder, RepeatMasker

- Решают общую задачу поиска неточных тандемных повторов (мономерные блоки похожи на 65-100%)
- Возможен разный сдвиг мономерных блоков

StringDecomposer, HMM based methods

- Необходимы готовые мономеры

Выделение мономеров

Состояние итерации: текущее приближение мономеров

Одна итерация:

1. Разбиение центромеры на блоки с помощью StringDecomposer и текущих мономеров
2. Кластеризация блоков на расстоянии >5% от текущих мономеров
3. Добавление консенсуса наибольшего кластера к текущим мономерам

Условие остановки: 99% центромеры разбито на мономерные блоки, каждый из которых находится на расстоянии $\leq 5\%$ от ближайшего мономера

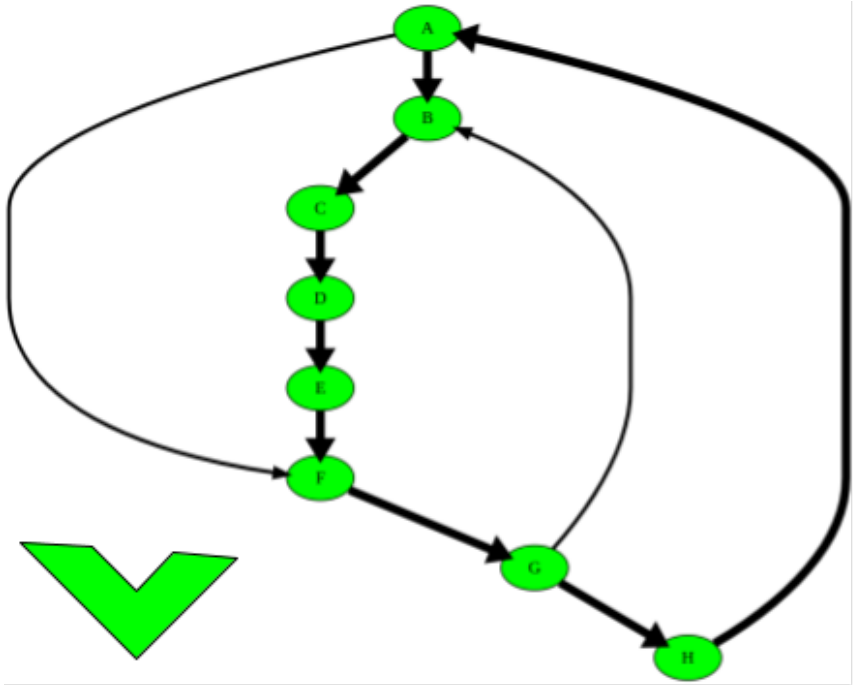
Исходное приближение: Один усредненный мономер (известен для приматов)

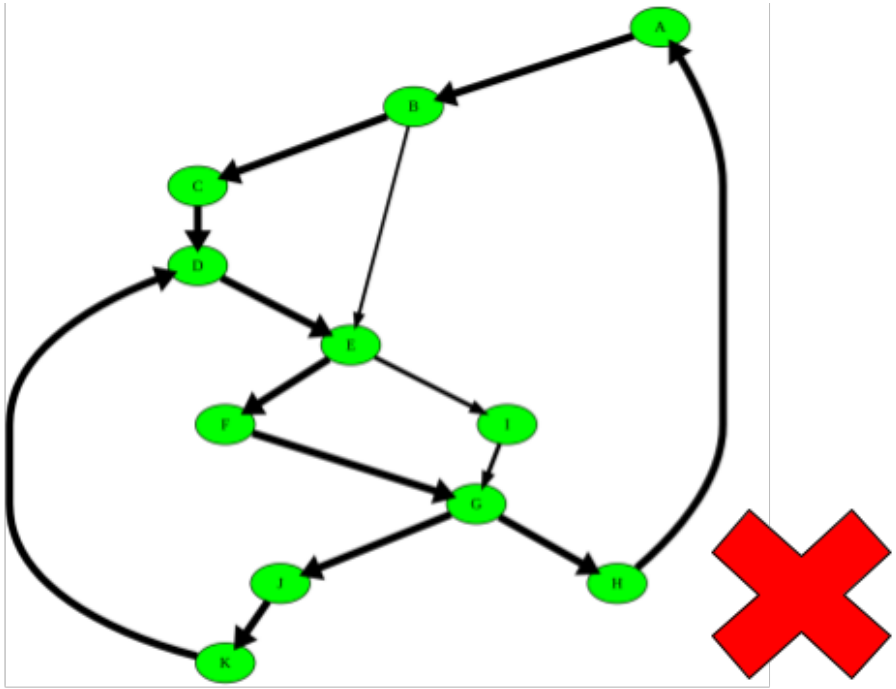
Требования к мономерам

Необходимо найти минимальное множество мономеров (строк) M, такое что:

- **+** длина каждого мономера $\sim$ 171 (150-190)
- **+** 99% центромеры можно разбить на мономерные блоки так, что бы расстояние до ближайшего мономера $\leq$ 5%
- **-** $\exists$ строка S из мономеров (канонический HOR) такая что:
  - в S входят все мономеры ровно по одному разу
  - тандемные повторы вида $SSS\ldots$ покрывают хотя бы 30% центромеры

Учёт позиций мономеров: мономерный граф



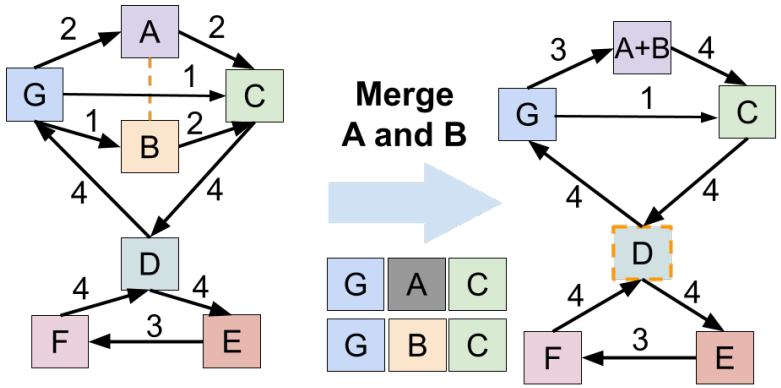


Вершины графа: мономеры
Вершины соединены **ребром**, если один мономер идет за другим в моноцентромере
Вес ребра: как часто один мономер идёт за другим
Переформулировка требования: Мономерный граф для данного множества мономеров должен содержать гамильтонов цикл.

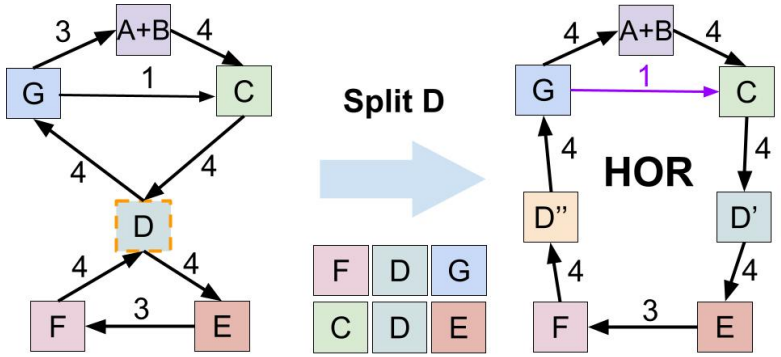
Учёт позиций мономеров

Преобразуем множество мономеров, так, чтобы в графе был гамильтонов цикл.

1 **Объединяем** мономеры, если:
- похожи консенсусы
- похожи контексты



2 **Разделяем** мономер, если:
- через него проходят ровно два независимых пути



Сравнение с известными мономерами

ID – номер хромосомы
A – множество мономеров, полученных в данной работе
M – множество мономеров, полученных экспертами[4]
MAX dist – максимальное редакционное расстояние между соответствующими мономерами

ID	|A|	|M|	MAX dist
1	6	6	5
2	4	4	0
3	17	17	1
4	19	19	3
5	6	6	4
6	18	18	0
7	6	6	0
8	11	11	0

ID	|A|	|M|	MAX dist
9	**8**	**7**	**12**
10	8	8	0
11	5	5	2
12	8	8	0
13	11	11	3
14	8	8	1
15	11	11	0
16	10	10	1

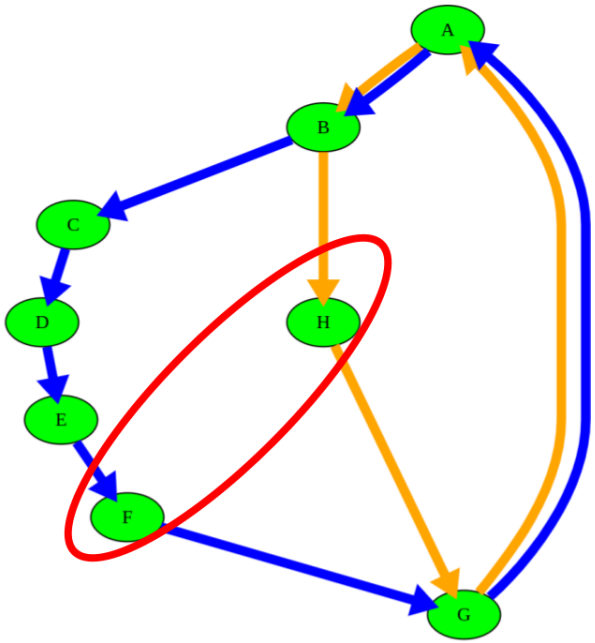
ID	|A|	|M|	MAX dist
17	30	30	0
18	12	12	2
19	**2**	**6**	**3**
20	16	16	5
21	11	11	4
22	8	8	1
X	12	12	1

Расхождение в несколько нуклеотидов допустимо: некоторые позиции вариабельные.

[4] внутрение данные международного T2T консорциум

Несостыковки с известными мономерами

Центромера 9



Надо объединить два мономера, но они сильно отличаются

Центромера 19 Нет согласованности между экспертами:

- Uralsky et al, 2019 – 6 мономеров.
- Rice, 2019 – 2 мономера.

Результаты

- На основе 15 статей сформулированы требования, которым должны удовлетворять мономеры, согласованные с биологами.
- В качестве инструмента для разбиения на мономерные блоки я решила использовать *StringDecomposer*, который требует на вход искомое множество мономеров.
- Разработан метод кластеризации мономеров, который итеративно уточняет разбиение на мономерные блоки и множество мономеров.
- Добавлена дополнительная обработка множества мономеров, которая исправляет ошибки кластеризации на основе контекста
- В 21 из 23 центромер выделенные мономеры совпали с выделенными экспертами.

По результатам работа принята на конференцию ISMB/ECCB 2021