Кластеризация на картах контактов ДНК с помощью максимизации модулярности

Валерия Ковалева

Научный руководитель: Сергей Нечаев

МФТИ

19 июня 2017

Цель исследования

Открытый вопрос

Существует ли иерархия в укладке ДНК, действительно ли сложенная молекула ДНК устроена фрактально.

Мотивация

Исследование структуры ДНК важно для медицины и понимания устройства живого организма.

Предложенный подход

- рассмотреть молекулу как взвешенный граф
- построить на нём системы кластеров разных масштабов
- исследовать устойчивость границ при изменении масштаба

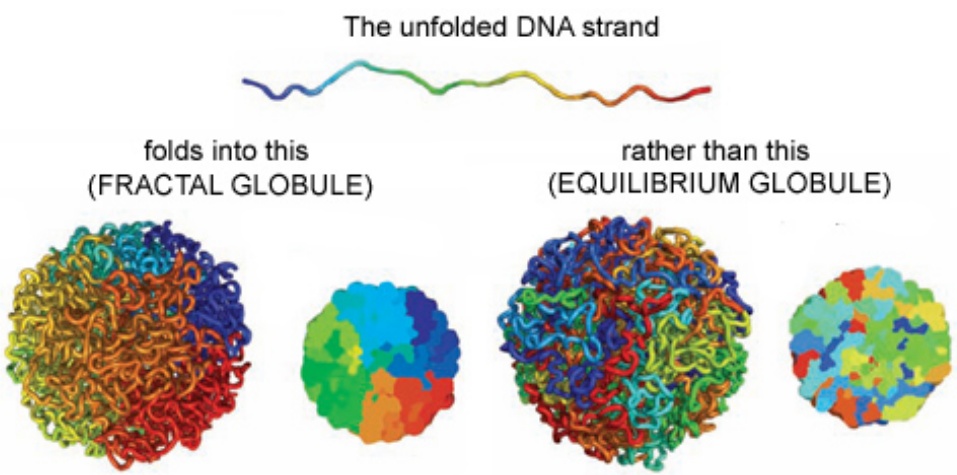
Карты контактов

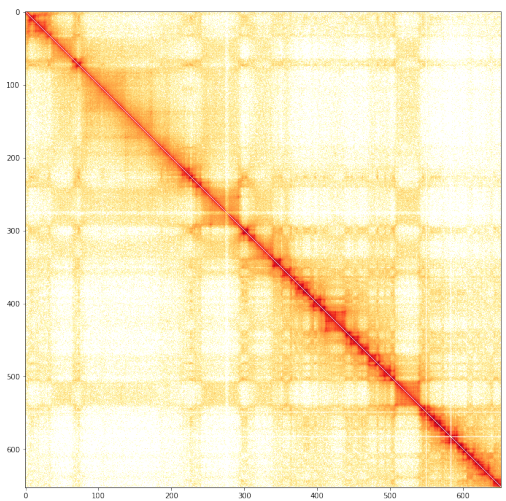
Определение

Карта контактов – симметричная бинарная матрица с нулевой диагональю, например, составленная так:

$$A_{ij} = \mathbb{1}\{r_{ij} < r_{\text{хар}}\},$$

где r_{ij} - расстояние между i и j мономером в пространстве, а $r_{\text{хар}}$ - характерное расстояние между мономерами.





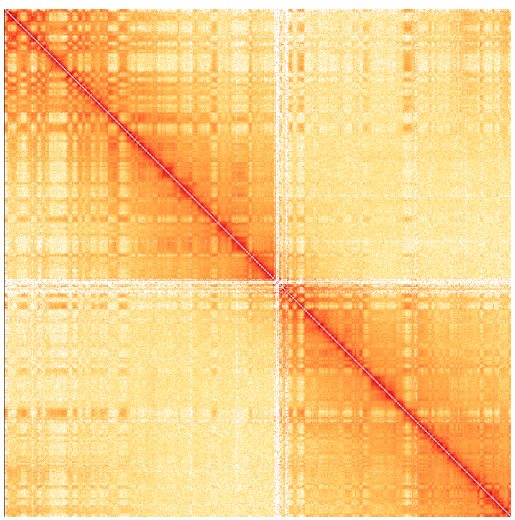
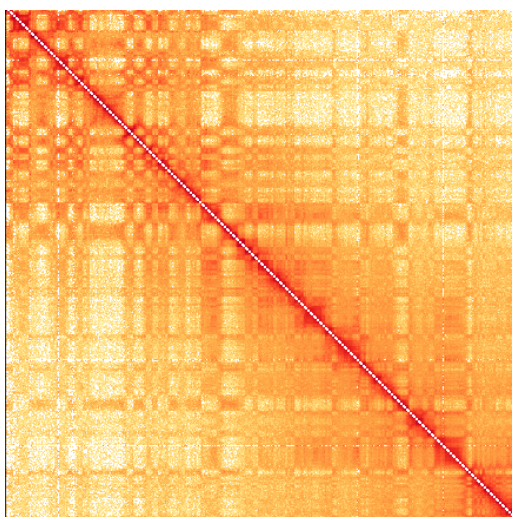
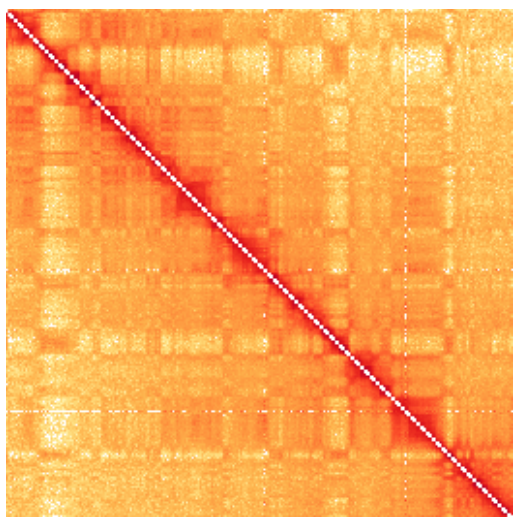
Топологические домены

Топологический домен

Участок цепи, мономеры которого взаимодействуют между собой много больше, чем со всеми остальными (“блоки” на диагонали).

Проблема

Нет общепринятного мнения, нет метрики качества и не с чем сравнить полученное разбиение.

Кластеризация

Идея

Мономеры внутри одного домена больше взаимодействуют между собой и будут лежать в одном кластере.

Модулярность

W - матрица весов $N \times N$, **C** - матрица принадлежности вершин кластерам $N \times k$.

$$Q[\mathbf{W}, \mathbf{C}] = \frac{1}{2w} \sum_{i,j} \left(w_{ij} - \frac{w_i w_j}{2w} \right) \sum_{\alpha} c_{i\alpha} c_{j\alpha}$$

Модулярность тем больше, чем больше отличается каждый кластер от подграфа графа Эрдеша-Реньи.

Модулярность

Матрица модулярности

$$\mathbf{B}: \; B_{ij} = w_{ij} - \frac{w_i w_j}{2w}$$

Задача линейного программирования

$$\frac{1}{2w} \sum_{i,j} B_{ij}(1 - x_{ij}) \to \max$$

$$\text{s.t. } x_{ik} \leq x_{ij} + x_{jk} \quad \forall i, j, k$$

$$x_{ij} \in \{0, 1\} \quad \forall i, j$$

- задача NP-сложная
- в силу линейности кластеров вдоль цепи и в предположении иерархичности можно решать жадным поиском

Предел разрешения

Существует предел разрешения $\sim \sqrt{w}$, не выделяются кластеры с суммой весов рёбер меньше $\sqrt{w/2}$, общее число кластеров $\sim \sqrt{w}$.

Уменьшение предела разрешения

$$\widetilde{\mathbf{W}} = \mathbf{W} + r\mathbf{I}$$

На вершины навешиваются петли веса r (“сопротивление”).

Новый предел разрешения

Общее число кластеров $\sim \sqrt{w + rN}$, предел разрешения по порядку равен $\sqrt{w/(1 + \frac{Nr}{w})}$.

Статистика границ

Идея

Изменяя r в большом диапазоне, получим доменную иерархию - системы кластеров различных масштабов - и исследуем её устойчивость.

Статистика границ

Число раз, когда она являлась границей между кластерами, при проходе по диапазону.

Предложение (Тамм, Москалец)

Статистика границ сортируется. Чем устойчивее структура, тем быстрее зависимость будет убывать и тем более выпуклой она будет.

Эксперименты

Синтетические данные

- сумма большого числа шумовых матриц
- сумма диагональных матриц
- глобулы двух заранее известных типов

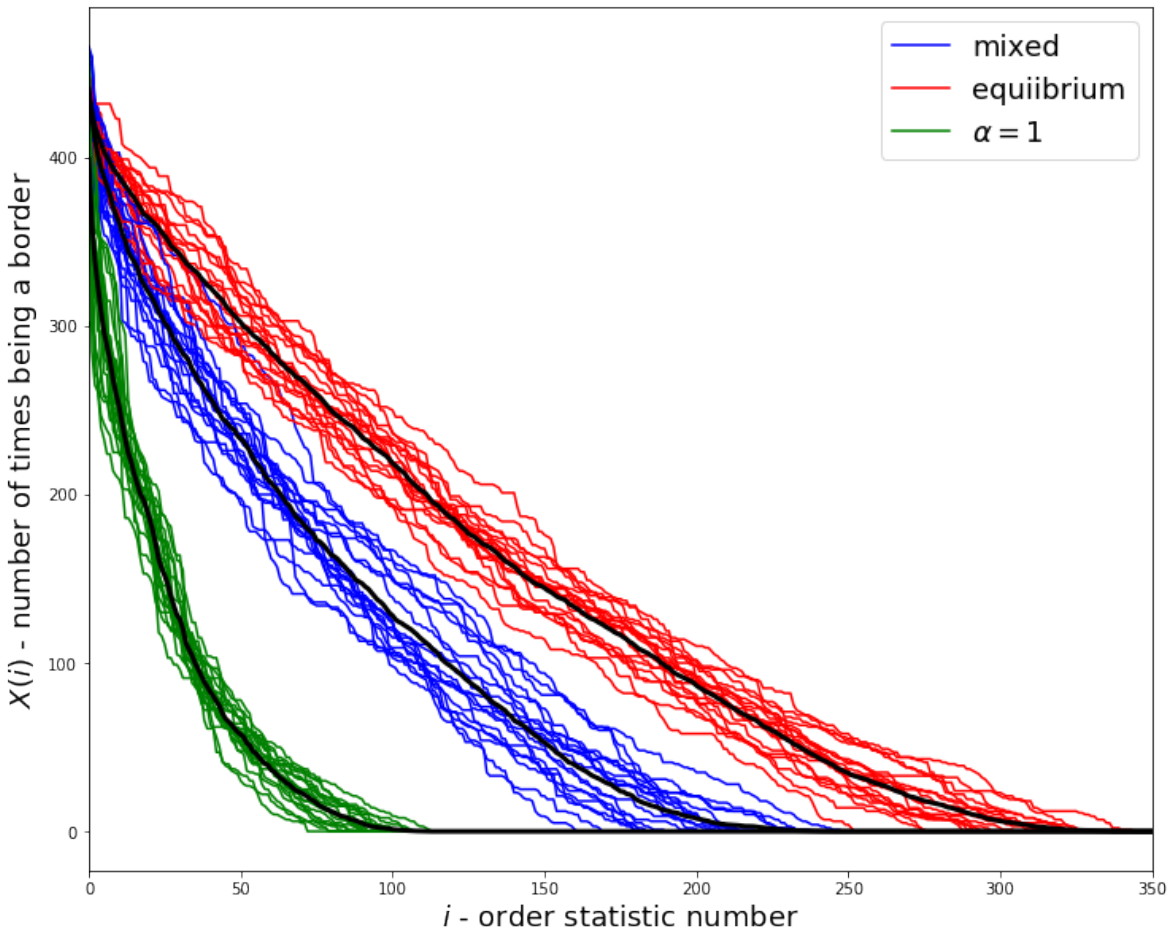
Экспериментальные данные

Hi-C карты предоставила лаборатория Л. Мирного в MIT.

- карты хромосом человека и мыши из разных тканей
- в разном разрешении, начиная от 1 кб
- индивидуальные и по популяции

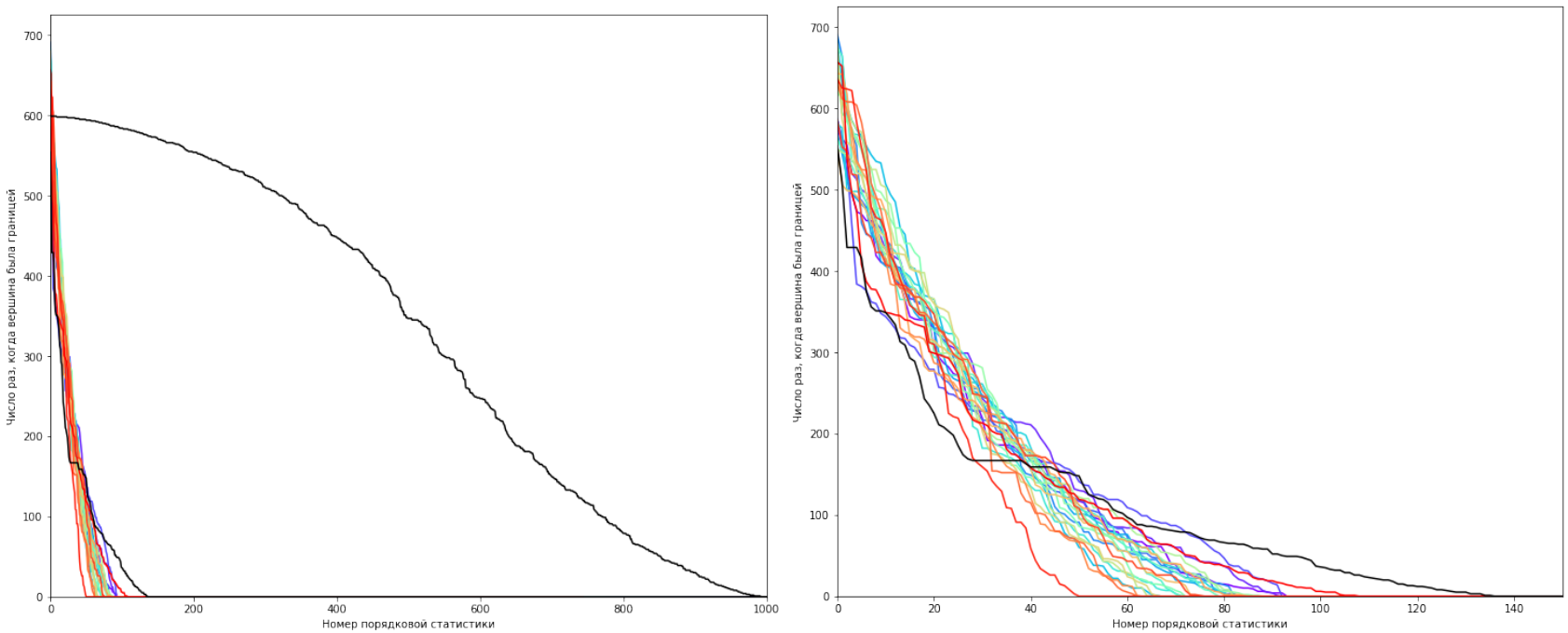
Эксперименты

Синим обозначены смешанные глобулы, красным – равновесные. Зелёным - матрицы с диагональным убыванием.



Эксперименты

Черным обозначены зависимости для случайных матриц, различными цветами спектра - разные хромосомы из одного набора.



Заключение

Статистика частот границ

- статистика быстрее убывает на данных, где заведомо известна фрактальность,
- на картах хромосом зависимость имеет маленькую дисперсию вне зависимости от размера карты
- пока неясно, как это охарактеризовать количественно