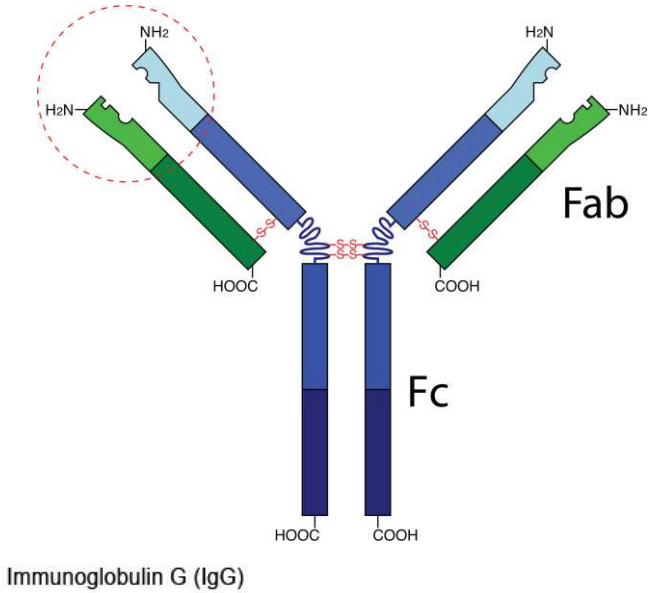
# ИЕРАРХИЧЕСКАЯ КЛАСТЕРИЗАЦИЯ ДАННЫХ NGS

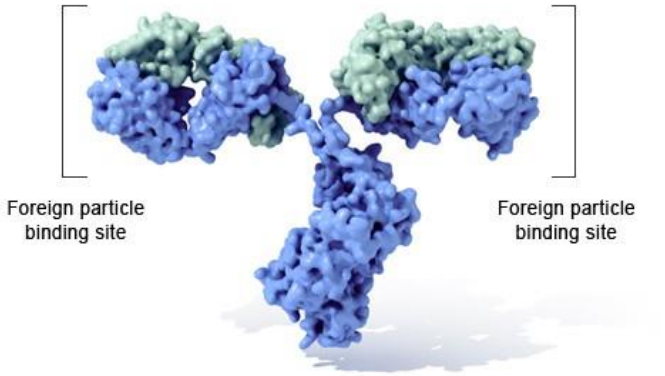
Ирина Смолина, Олег Яснев

Руководитель: Павел Яковлев

# Предметная область




Immunoglobulin G (IgG)




U.S. National Library of Medicine

- Иммуноглобулин – это полезная штука
- Фрагменты Fab иммуноглобулинов гипервариабельны
- Сложно отделять ошибки секвенирования от реальной вариабельности

# Постановка задачи

## Дано

- Много коротких и очень похожих ридов
- Секвенирование высоковариабельных участков технологией 454

## Надо

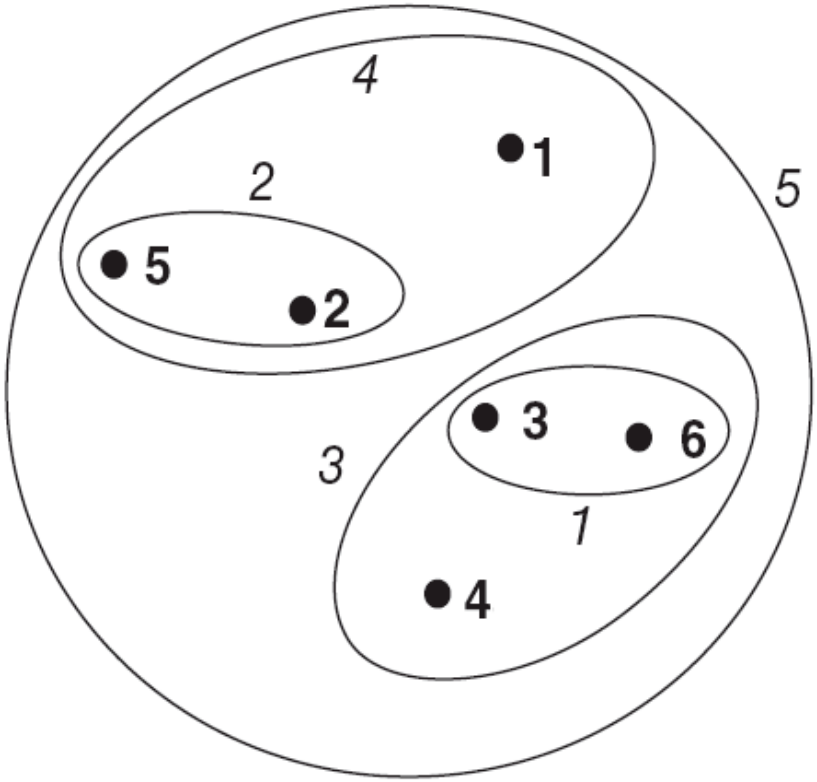
- Исправить ошибки
- Провести иерархическую кластеризацию

# Этапы решения задачи

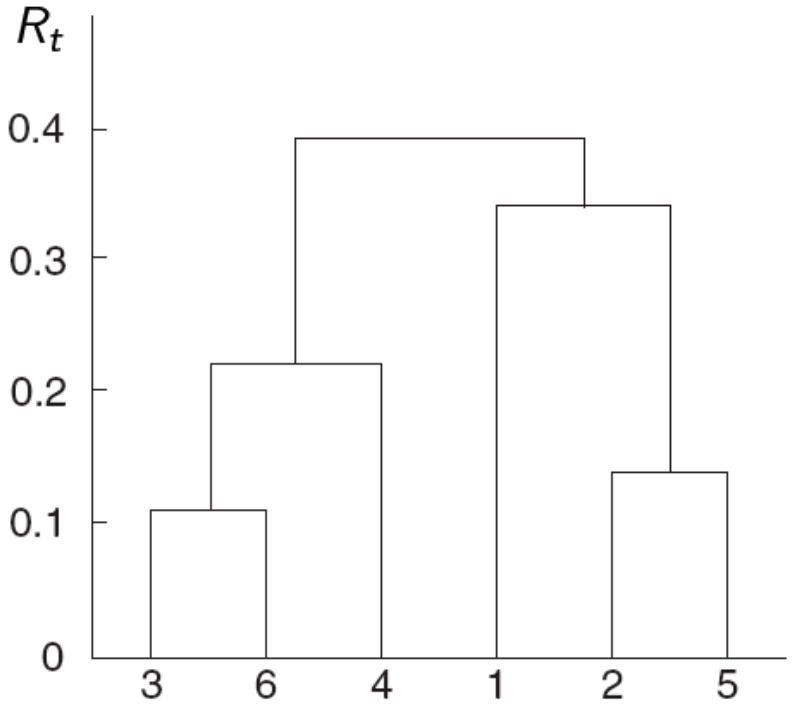
1. Получить все попарные выравнивания ридов.
2. Провести иерархическую кластеризацию и построить филогенетическое дерево ридов.
3. Провести схлопывание похожих ридов, используя функцию правдоподобия.
4. Перестроить полученное дерево.

# Иерархическая кластеризация

## Диаграмма вложения




## Дендрограмма



# Важные наблюдения

- Большинство ошибок 454 содержится в гомополимерах
- Нужно считать расстояние между ридами с учетом ошибок в гомополимерах
- Clustal Omega:
  - умеет делать попарные выравнивания
  - умеет делать иерархическую кластеризацию
  - не умеет работать с ошибками в гомополимерах

# Варианты решения задачи

- Вариант 1
  - научить Clustal Omega работать с ошибками в гомополимерах
  - написать функцию выравнивания
  - написать функцию правдоподобия
- Вариант 2
  - использовать Clustal Omega как есть
  - написать функцию правдоподобия
  - написать схлопывание листьев и перестроение дерева

# Текущий прогресс

- Изучили теорию и предметную область
- Изучаем Clustal Omega
- Работаем над нужными алгоритмами
- Анализируем варианты решения задачи