

Летняя практика в компании BIOCAD

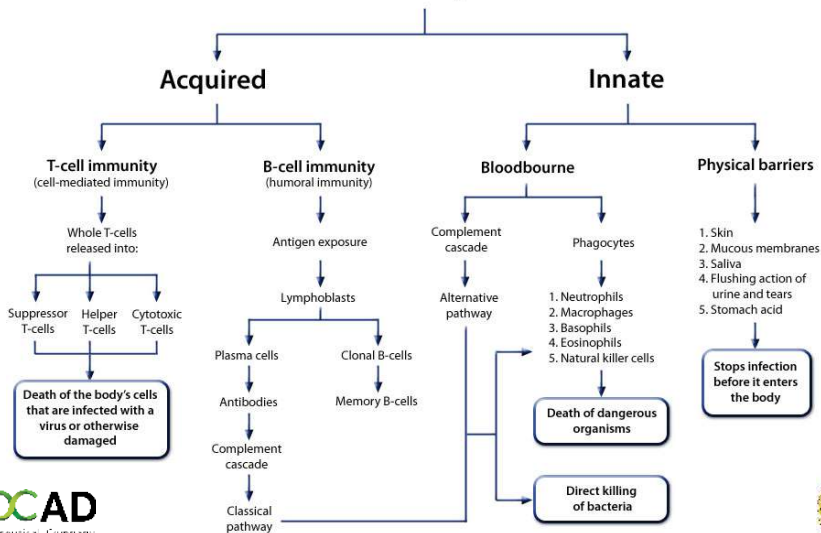
Система автоматической обработки иммуноглобулинов

Павел Яковлев

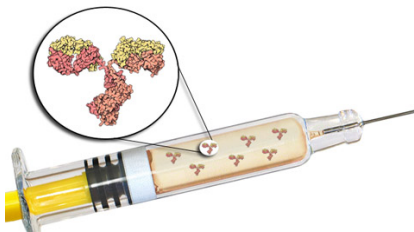
СПбАУ РАН, BIOCAD



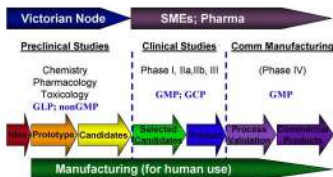
Immune system



Иммуноглобулины и лекарства



Идея: вместо химических средств широкого спектра получить высокоспецифичные антитела на конкретную мишень.



Проблемы:

- Человек не может синтезировать все необходимые антитела
- Антитела животных не годятся для человека в чистом виде
- При иммунизации многие животные могут погибнуть



IG pipeline: массовая обработка иммуноглобулинов

Цель: разработать высокопроизводительный pipeline, способный проводить все операции над иммуноглобулинами от обработки сырого выхода с секвенатора до предсказания свойств кандидатов в лекарства *in silico*.

Задачи:

- Разделить сырые риды на классы (например, типы цепей), исправить ошибки, кластеризовать полученные данные
- Определить классы ридов с ошибками в баркодах
- Провести детекцию генов-составляющих и грубую разметку на регионы
- Сохранить данные в контейнер, позволяющий быстро отвечать на типичные вопросы биологов, используя доступные аппаратные мощности
- Предусмотреть возможность сохранения и предсказания *любых* физических характеристик иммуноглобулинов
- Научиться отвечать на популярные однотипные вопросы биологов



Проблемы существующих решений

Начальные решения использовали сторонние продукты: **IgBLAST**, **IMGT/V-QUEST**, **MUSCLE** и др.

Недостатки:

- Наличие различных нотаций определения регионов в разных утилитах
- *Ограниченность применения (до региона FR3)*
- *Невозможность обработки сырых неисправленных данных*
- Неудобный формат вывода
- Плохо написанный медленный и нечитаемый код (IgBLAST)
- Возможность аннотирования последовательностей только типами регионов
- *Отсутствия удобного формата данных для хранения результатов и их последующего переиспользования*
- *Скорость работы (!!!)*



Шаг 1. Исправление и кластеризация



Путь от животного к достоверным последовательностям.

Этапы обработки:

- 1 Разделение ридов по типам (MID)
- 2 Иерархическая кластеризация
- 3 Объединение кластеров нижних уровней - исправление ридов одних последовательностей
- 4 Выдача иерархической структуры для быстрого поиска и анализа представленности

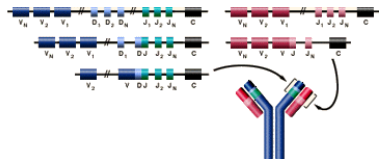
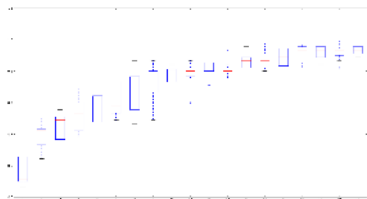
Шаг 2. Первичный анализ

Три основных задачи:

- Разделить последовательности на классы
- Определить гены-гермлайны для каждой цепочки, найти границы генов
- Осуществить нечеткую детекцию регионов

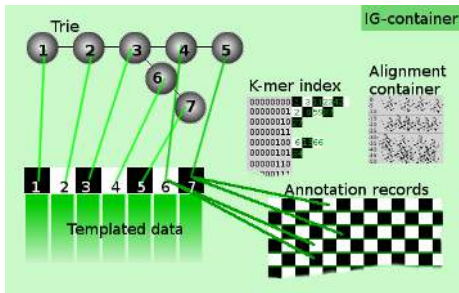
Подходы:

- BLAST-подобное выравнивание на основе индекса k-меров
- Машинное обучение: SVM, Random Trees



Seq_10000_1	24	312	301	343	345	389
Seq_10001_1	23	311	300	342	344	388
Seq_10002_1	26	322	98	129	325	398
Seq_10003_1	24	312	301	343	345	389
Seq_10004_1	24	312	301	343	345	387
Seq_10005_2	25	291	280	322	324	368
Seq_10006_1	31	290	279	321	323	367
Seq_10007_1	24	312	301	343	315	388
Seq_10008_1	23	310	299	341	313	386

Шаг 3. Хранение и последующая обработка запросов



База данных для анализа:

- Проаннотировать последовательность по регионам
- Найти иммуногенные участки
- Найти все последовательности в выборке, образовавшиеся из одного гермлайна
- и многое другое

Латпор tests (35к последовательностей)¹:

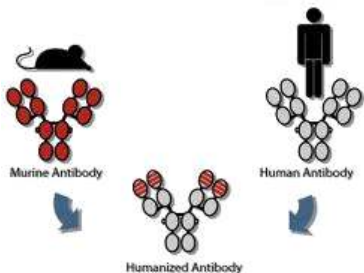
- Загрузка: 7 сек
- Поиск: 0.4 сек
- Выравнивание/аннотирование: 32-40 сек

BIOCAD

Биофармацевтическая Компания

¹Core i7 3-gen, 8G RAM, GNU/Linux





```
Fd138-00:
VL: DIVNTQSRKPHSTSVQDRVSI TCKASGQVQSGAVVWVHQKQKQSPRELLIYVASTRTDVPQRFKTSQGGSDFTLTIITVQSEDLADYPC...
FR1: CDR1 FR2 CDR2 FR3
DIVNTQSRKPHSTSVQDRVSI TCKAS GQVQSGA IYVWQKQKQKQSPRELLIYVASTRTDVPQRFKTSQGGSDFTLTIITVQSEDLADYPC...
FR1: CDR1 FR2 CDR2 FR3
DIQMTQSPFSLASVQDRVITCKAS GQVQSGA IYVWYQKQKQKQSPRELLIYVASTRTDVPQRFKTSQGGSDFTLTIITVQSEDLADYPC... VL-cos1
FR1: CDR1 FR2 CDR2 FR3
VH: QVQLQQSQAELVWPGASVKISCKVSGVYFIDMTLTIRWYQKQKQKQSPRELLIYVASTRTDVPQRFKTSQGGSDFTLTIITVQSEDLADYPC...
FR1: CDR1 FR2 CDR2 FR3
QVQLQQSQAELVWPGASVKISCKVSGVYFIDMTLTIRWYQKQKQKQSPRELLIYVASTRTDVPQRFKTSQGGSDFTLTIITVQSEDLADYPC...
FR1: CDR1 FR2 CDR2 FR3
QVQLQQSQAELVWPGASVKISCKVSGVYFIDMTLTIRWYQKQKQKQSPRELLIYVASTRTDVPQRFKTSQGGSDFTLTIITVQSEDLADYPC... VH-cos1
FR1: CDR1 FR2 CDR2 FR3
```

Спросите меня после презентации про:

- гуманизацию
- автоматизацию разбора результатов Sanger
- статистическую обработку скрининга
- кодон-оптимизацию
- подбор параметров сред культивирования
- мой NDA



Спасибо за внимание!

Вопросы?

yakovlev@biocad.ru

