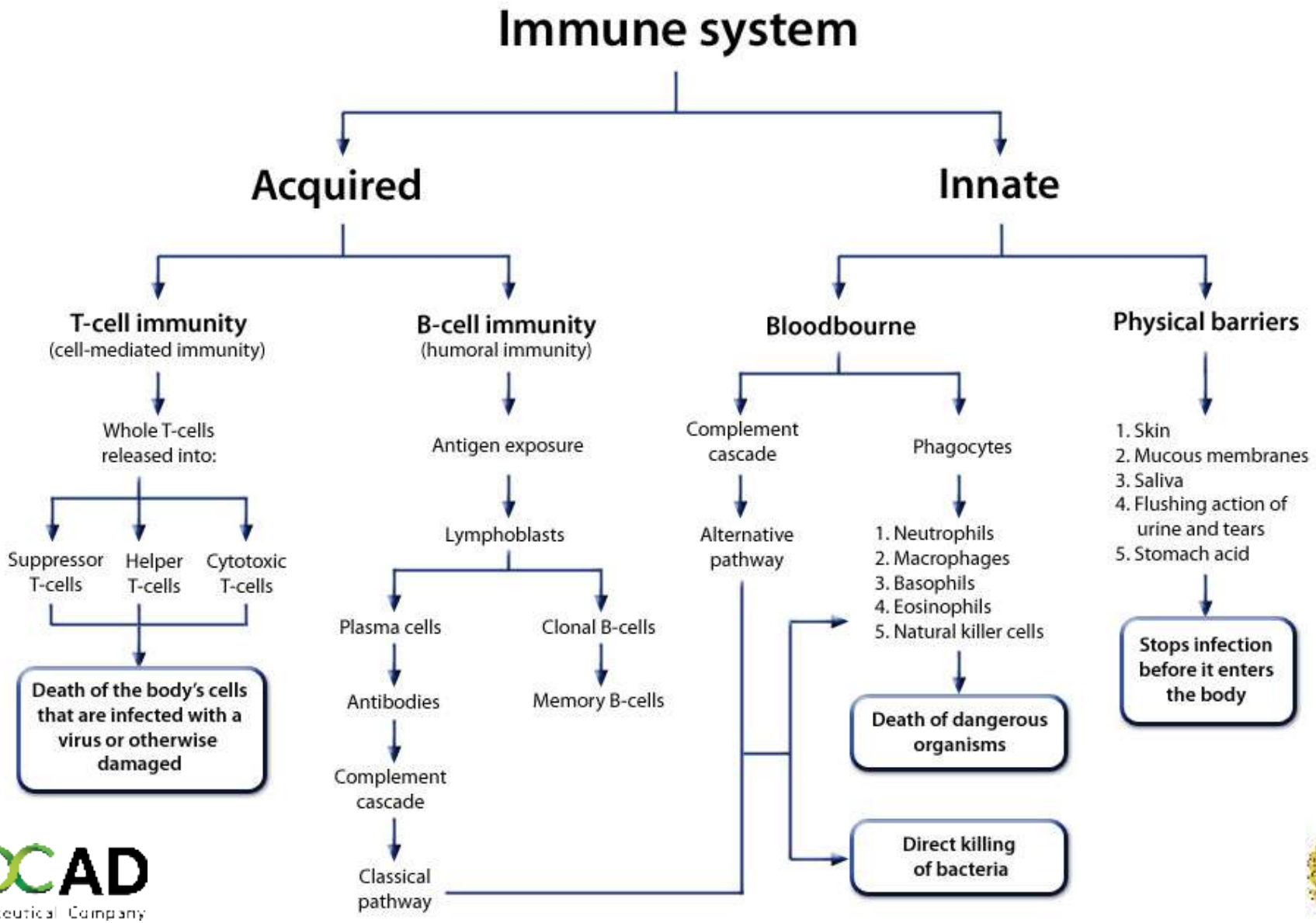
Биоинформатика в разработке лекарственных средств

Система автоматической обработки иммуноглобулинов

Павел Яковлев

СПбАУ РАН, BIOCAD

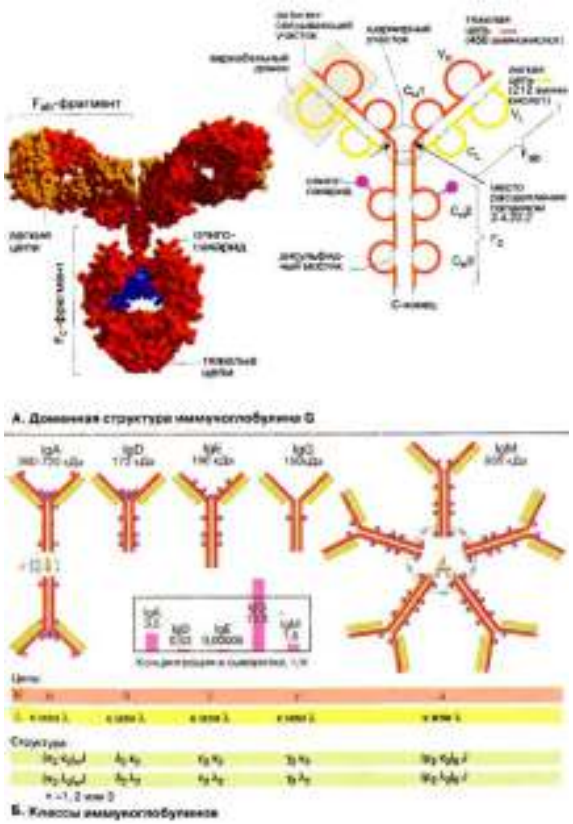
Иммунология в одном слайде



BIOCAD
Biopharmaceutical Company

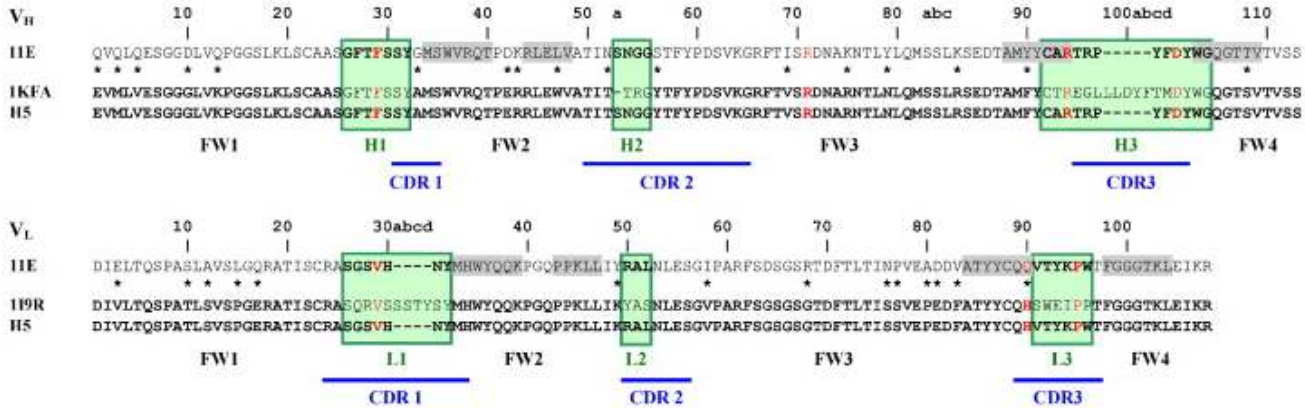
Иммунология во втором слайде

Взгляд биолога:



Взгляд информатика:

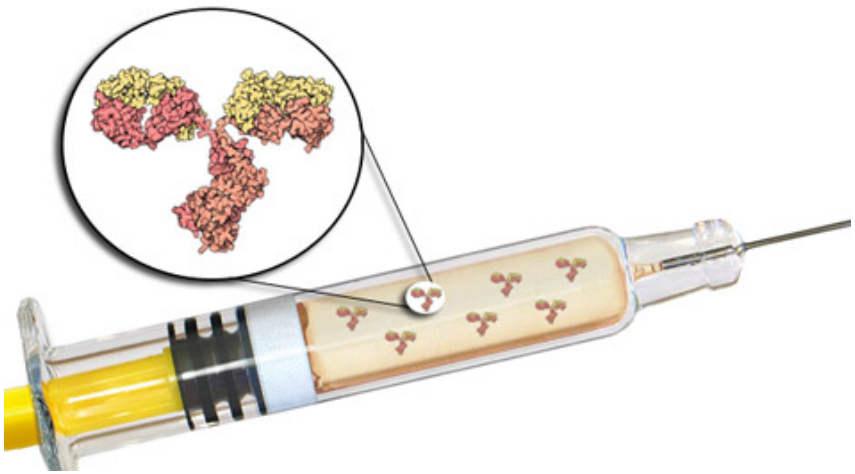
Интересный участок: вариабельные домены.



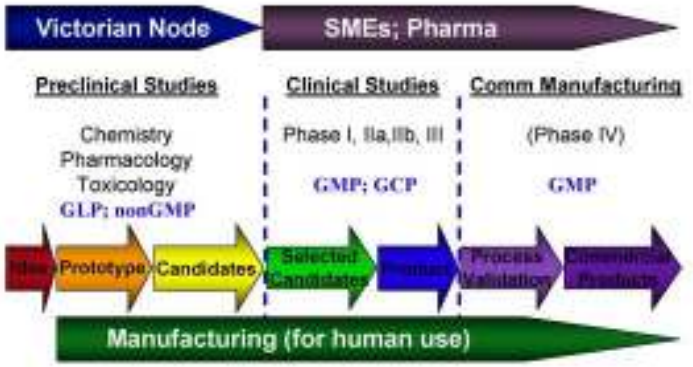
- 2 аминокислотных последовательности $\sim 140aa$
- разделены на 7 регионов
- 3 региона (CDRs) - hotspots, отвечают за специфичность
- 4 региона (FRs) - каркасные, менее вариабельные

BIOCAD
Biopharmaceutical Company

Иммуноглобулины и лекарства



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



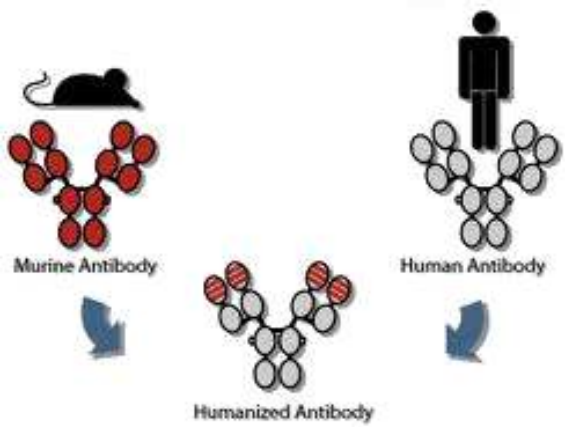
Проблемы:

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

BIOCAD
Biopharmaceutical Company

Первые работы: гуманизация

Взгляд биолога:



Взгляд информатика:

```
Fd138-80:
    VL: DIVMTQSHKFMSTSVGDRVSITCKASQDVGSAVVWHQQKSGQSPKLLIYWASTRHTGVPDRFTGSGSGTDFTLTITNVQSEDLADYFC...
        FR1                         CDR1    FR2                 CD2 FR3
        DIVMTQSHKFMSTSVGDRVSITCKAS QDVGSA VVWHQQKSGQSPKLLIY WAS TRHTGVPDRFTGSGSGTD
        FR1                         CDR1    FR2                 CD2 FR3
        DIQMTQSPSFLSASVGDRVTITCKAS QDVGSA LVWYQQKPGKAPKLLIY WAS TLHTGVPSRFSGSGSGTD   VL-con1
        FR1                         CDR1    FR2                 CD2 FR3
    VH: QVQLQQSDAELVKPGASVKISCKVSGYTFTDHTIHWMKQRPEQGLEWFGYIYPRDGHTRYSEKFKGKATLTADKSASTAYMHLNSLTS...
        FR1                         CDR1      FR2               CDR2      FR3
        QVQLQQSDAELVKPGASVKISCKVS GYTFTDHT IHWMKQRPEQGLEWFGY IYPRDGHT RYSEKFKGKATL
        FR1                         CDR1      FR2               CDR2      FR3
        QVQLVQSGAEVKKPGASVKISCKVS GYTFTDHT MHWVKQAPGQGLEWMGY IYPRDGHT RYSEKFQGRVTL   VH-con1
        FR1                         CDR1      FR2               CDR2      FR3
```

Задачи:

- Найти регионы на последовательностях
- Найти человеческие гомологи последовательностей
- Выровнять последовательности и произвести необходимые замены по одному из алгоритмов

BIOCAD
Biopharmaceutical Company

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)*
- *Невозможноть обработки сырых неисправленных данных*
- Неудобный формат вывода
- Плохо написанный медленный и нечитаемый код (IgBLAST)
- Возможность аннотирования последовательностей только типами регионов
- *Остутствия удобного формата данных для хранения результатов и их последующего переиспользования*
- *Скорость работы **(!!!)***

BIOCAD
Biopharmaceutical Company

IG pipline: массовая обработка иммуноглобулинов

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

Задачи:

- Разделить сырые риды на классы (например, типы цепей), исправить ошибки, скластеризовать 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 предказывания *любых* физических характеристик иммуноглобулинов
- Научиться отвечать на популярные однотипные вопросы биологов

BIOCAD
Biopharmaceutical Company

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 стурктуры для быстрого поиска и анализа представленности

BIOCAD
Biopharmaceutical Company

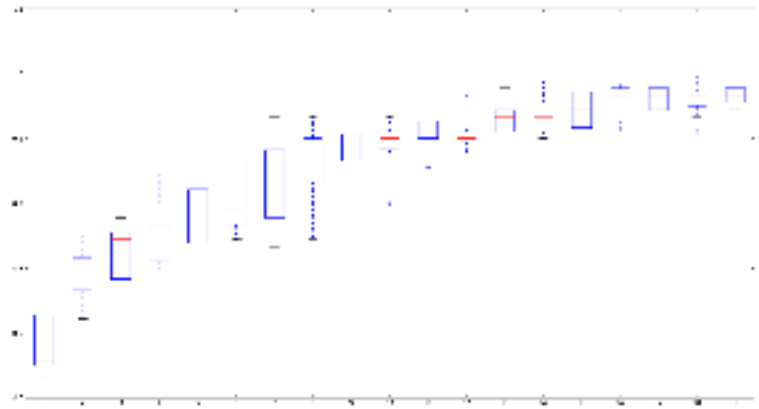
Шаг 2. Первичный анализ

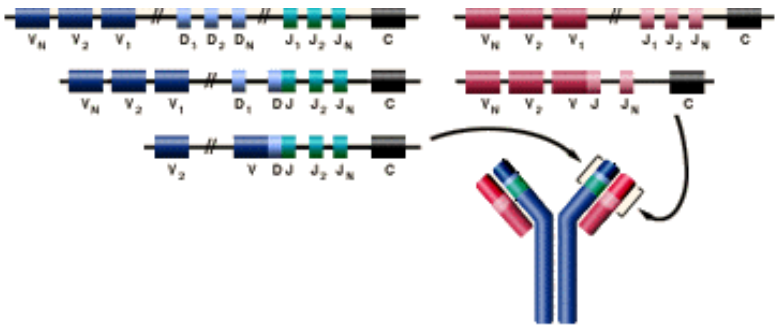
Три основных задачи:

- 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, HMM

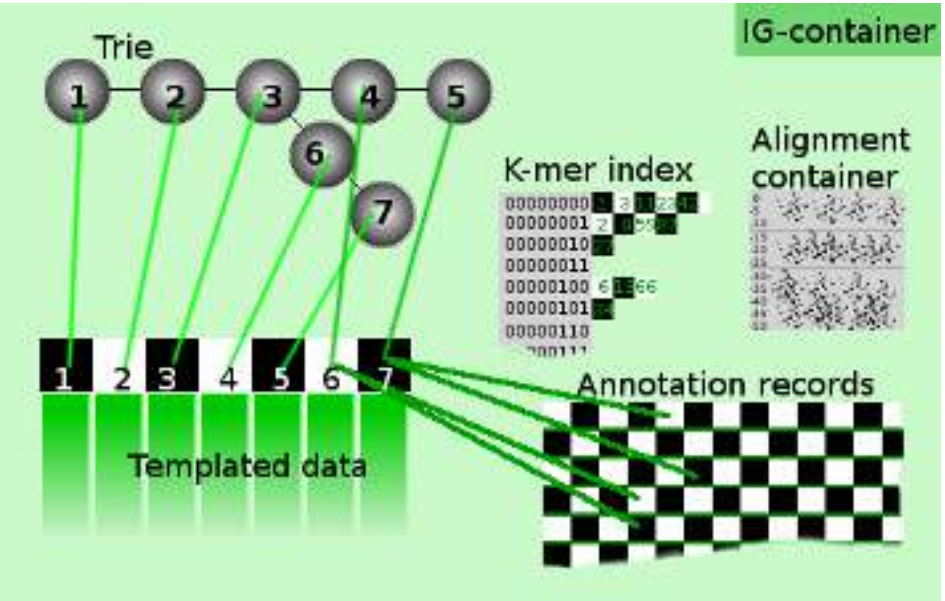




Seq_10000_1	24	312	301	343	345	389
Seq_10001_1	23	311	300	342	344	388
Seq_10002_1	26	322	98	129	325	398
Seq_10003_1	24	312	301	343	345	389
Seq_10004_1	24	312	301	343	345	387
Seq_10005_2	25	291	280	322	324	368
Seq_10006_1	31	290	279	321	323	367
Seq_10007_1	24	312	301	343	315	388
Seq_10008_1	23	310	299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

BIOCAD
Biopharmaceutical Company

Шаг 3. Хранение и последующая обрабокта запросов



База данных для анализа:

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

Laptop tests (35к последовательностей)[1]:

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

BIOCAD
Biopharmaceutical Company

[1] Core i7 3-gen, 8G RAM, GNU/Linux

Дальнейшие перпективы

Что дальше?

- Подготовка баз для оценки вариабельности позиций в кандидатах, специфичных к конкретному антигену
- Оценка характерных переходов аминокислот при гуманизации
- Создание баз для оценки иммуногенности кандидатов
- Аннотирование кандидатов по различным сайтам (deamidation, isomerization, cleavage, oxidation, glycosylation...)

А потом...

- Оценка термодинамических свойств
- Предсказание структуры Fab и полноразмерных антител
- Предсказание аминокислотной последовательноти по рассчитанной структуре - дизайн белков

BIOCAD
Biopharmaceutical Company

The end

Спасибо за внимание!

Вопросы?

yakovlev@biocad.ru

BIOCAD
Biopharmaceutical Company