Аннотирование и гуманизация последовательностей вариабельных доменов иммуноглобулинов

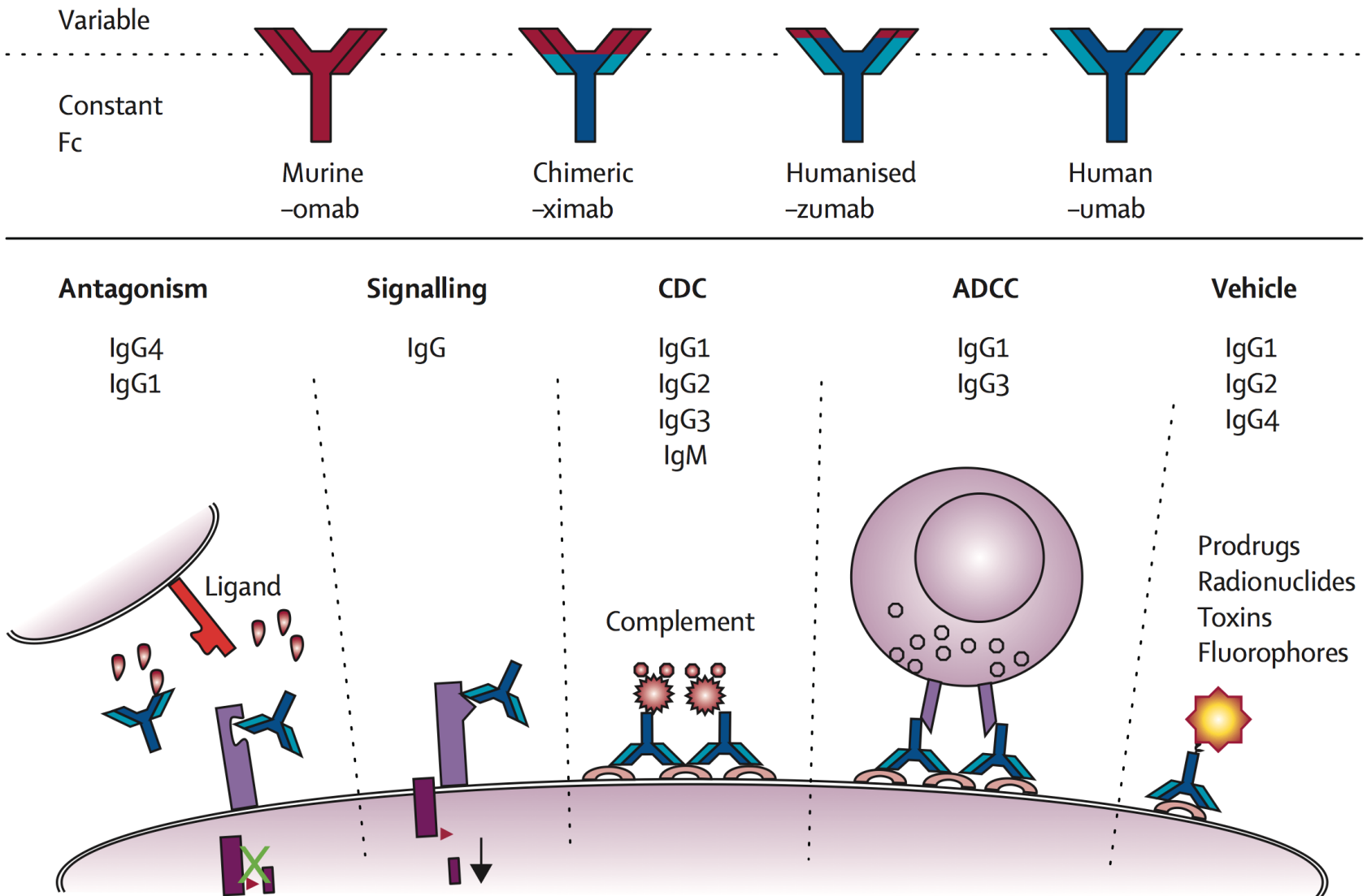
Павел Яковлев

СПбАУ РАН

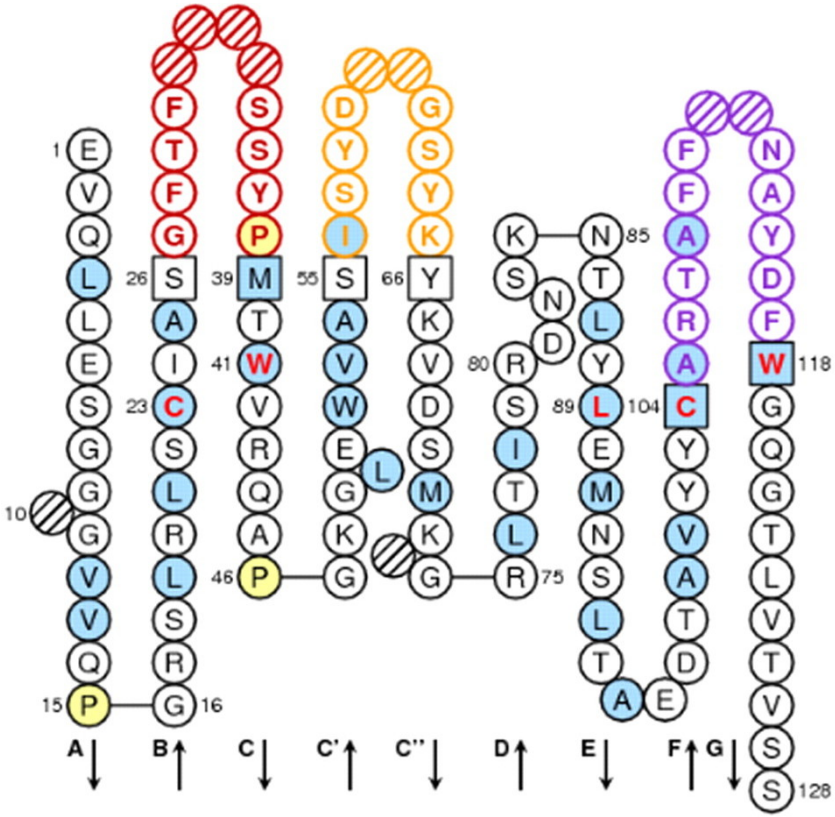
Руководитель: Карабельский А.В., к.б.н., BIOCAD
Рецензент: Порозов Ю.Б., к.м.н., НИУ ИТМО

5 июня 2014

Действие терапевтических иммуноглобулинов



Аннотирование антител



Гермлайны:

- Домен тяжелой цепи:
  51 VH, 75 DH, 6 JH
- Домен легкой цепи:
  40 VK, 5 JK
  31 VL, 4 JL

Структурные семейства:

7 семейств по тяжелой цепи и 16 (6 + 10) семейств по легкой цепи

Регионы:

4 каркасных региона – FR

3 петлевых региона – CDR

Числа для количества семейств и гермлайнов приведены для антител человека

Пример вывода

Регионы:

```
EVQLLESGGGVVQPGRSLRLSCIAS GFTFSSYP MTWVRQAPGKGLEWVAS ISYDGSYK
<---------FR1-----------> <-CDR1-> <------FR2------> <-CDR2->
YKVDSMKGRLTISRDNSKNTLYLEMNSLTAEDTAVYYC ARTAFFNAYDF WGQGTLVTVSS
<----------------FR3-----------------> <--CDR3---> <---FR4--->
```

Гермлайны:

№	V-ген	D-ген	J-ген	Схожесть
1	IGHV3-10	IGHD4-4*{1}	IGHJ3	0,83
2	IGHV3-10	IGHD4-4*{1}	IGHJ4	0,83
3	IGHV3-10	IGHD3-2*{2}	IGHJ3	0,80

Семейство: VH3

Замены:

```
EVQLLESGGGVVQPGRSLRLSCIAS GFTFSSYP MTWVRQAPGKGLEWVAS ISYDGSYK
    V                     V
    *                     K
YKVDSMKGRLTISRDNSKNTLYLEMNSLTAEDTAVYYC ARTAFFNAYDF WGQGTLVTVSS
                       Q     R
```

Цели и задачи работы

Цель: создание эффективного метода аннотирования иммуноглобулинов, а также получения вариантов замен для их гуманизации.

Задачи:

- обеспечить хранение больших референсных баз (нуклеотидных и аминокислотных);
- определять ближайшие гены гермлайны и семейство антитела;
- осуществлять поиск регионов для последовательности, содержащей вариабельный домен или его часть;
- предоставлять варианты для точечной гуманизации антител.

Существующие решения

Получение вариантов замен для гуманизации:

- решения отсутствуют в открытом доступе.

Поиск гермлайнов и ближайших гомологов:

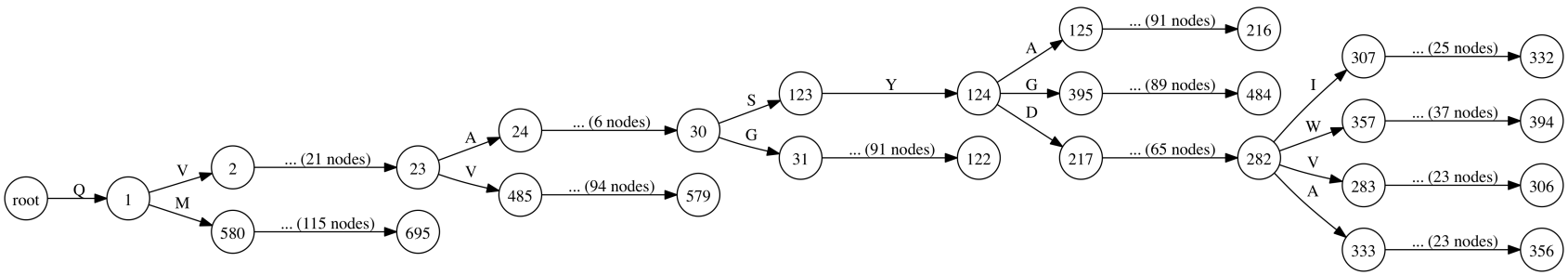
- выравнивание на референсную базу (*V-BASE2*, *IMGT/V-QUEST*, *IgBLAST*).

Определение регионов:

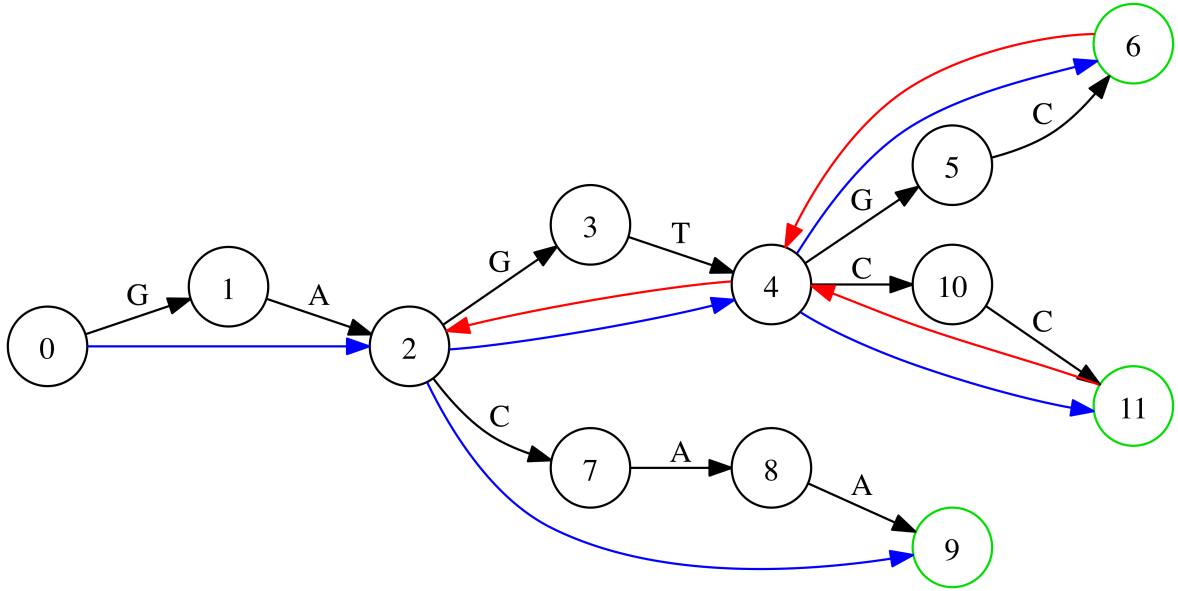
- Два основных метода:
  - выравнивание на референсы (*IMGT/V-QUEST*, *IgBLAST*).
  - поиск паттернов (*Rosetta Antibody/ROSIE*, *proABC*)

Ключевое наблюдение

Соматические мутации в процессе созревания антитела происходят по всей длине вариабельного домена.
Наибольшее количество мутаций среди FR регионов происходит в FR3.
Еще больше вариантов имеет CDR3 регион за счет дополнительного очага мутаций - V(D)J-рекомбинации.
Вывод: вариабельность доменов нарастает к концу, тогда как их префиксы часто совпадают.



Выравнивание



1. создается стэк матриц выравнивания;
2. (опционально) выбираются ветви для обхода;
3. дерево обходится в глубину;
4. конечный автомат, обрабатываются разные типы вершин дерева:
   - на каждой развилке наращивается матрица на пройденную подстроку (push);
   - на каждом листе выдается ответ;
   - при «прыжке» обратно по дереву убирается хвост матрицы, соотвествующий ветви, которая была полностью пройдена (pop).

Аннотирование и получение вариантов замен

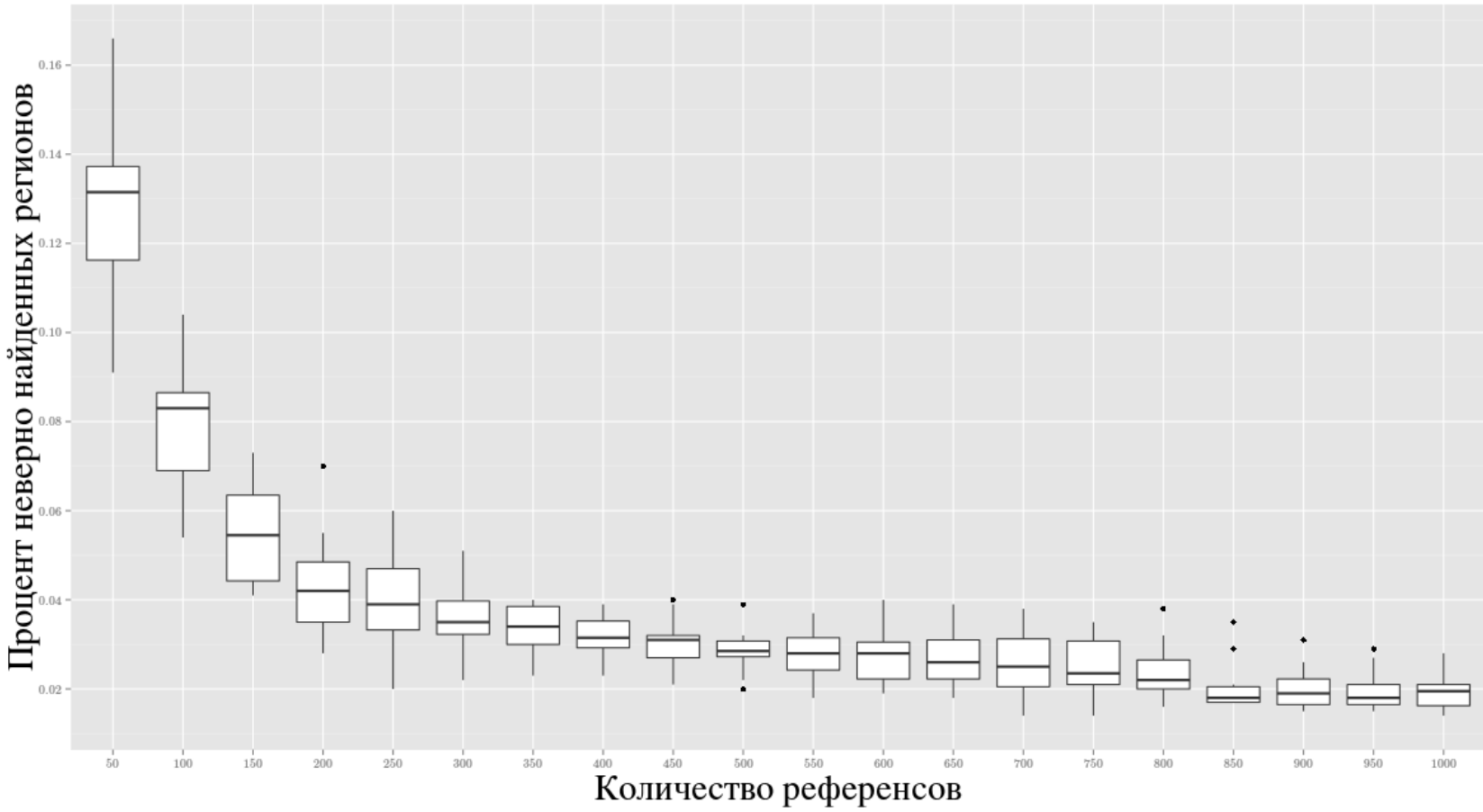
Аннотирование:

- по лучшим N выравниваниям;
- по выравниваниям со схожостью выше P.

Гуманизация:

- полные референсы, выбор вариантов только внутри FR регионов;
- только FR регионы, самостоятельная референсная база для каждого FR региона.

Зависимость точности аннотирования от количества используемых референсов



Сравнения I

- Поиск ближайших гермлайнов и гомологов дал одинаковые результаты с IgBLAST на всех имеющихся данных.
- IgBLAST (метод выравнивания) и ROSIE (метод поиска паттернов).
- Данные человеческих и гуманизированных последовательностей.

Частота различных предсказаний:

$$\mathrm{DDR} = \frac{D}{T},$$

Средние относительные сдвиги:

$$AS_s = \frac{\sum(B_1 - B_2)}{\mathrm{D}},\ AS_u = \frac{\sum|B_1 - B_2|}{\mathrm{D}}$$

Сравнения II

Таблица: Различия в предсказаниях с IgBLAST

регион	DDR (%)	AS_s	AS_u
FR1 начало	8.88	1.19	1.19
FR1 конец	1.32	-2.50	2.50
FR2 начало	0.00	0.00	0.00
FR2 конец	0.33	2.00	2.00
FR3 начало	0.33	-4.00	4.00
FR3 конец	20.39	-1.34	1.34

При использовании локального выравнивания и только V-генов в качестве референсной базы результаты полностью совпали.

Сравнения III

Таблица: Различия в предсказаниях с ROSIE

регион	DDR (%)	AS_s	AS_u
FR1 начало	6.86	-1.30	1.30
FR1 конец	1.27	0.55	1.00
FR2 начало	$<$0.01	1.24	1.75
FR2 конец	$<$0.01	0.22	1.00
FR3 начало	0.00	0.00	0.00
FR3 конец	12.85	-0.47	1.49
FR4 начало	0.11	-0.64	1.13
FR4 конец	11.61	1.00	1.00

Мишени в разработке

Мишени	Терапевтическая область	Стадия разработки	Старт доклинических испытаний
IL-17	Аутоиммунные	Доклинические испытания	Инициированы
EGFR	Онкология	Доклинические испытания	Инициированы
c-Met	Онкология	Оптимизация	Q1 2015
HER3	Онкология	Оптимизация	Q1 2015
IL-6R	Аутоиммунные	Оптимизация	Q1 2015
IL-12/23	Аутоиммунные	Скрининг кандидатов	Q3 2015
Ang2	Онкология	Скрининг кандидатов	Q3 2015

Благодарности

BIOCAD:

- Карабельский Александр
- Улитин Андрей
- Неманкин Тимофей
- Свечников Иван

НИУ ИТМО:

- Порозов Юрий
- Князев Сергей

СПбАУ:

- Лебедев Сергей
- Колмогоров Михаил
- Жирков Игорь
- Малыгина Татьяна

СПбГУ:

- Лебедева Екатерина
- Логачев Кирилл

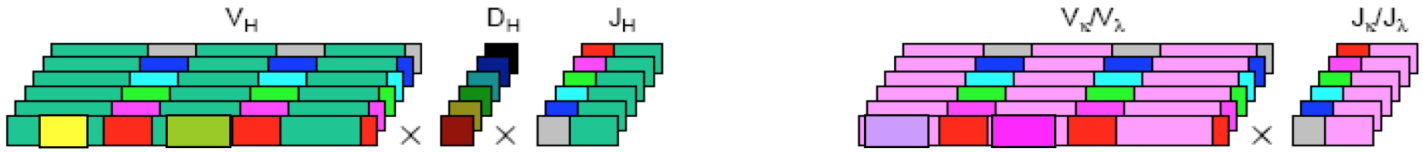
Спасибо за внимание!

Запасные слайды

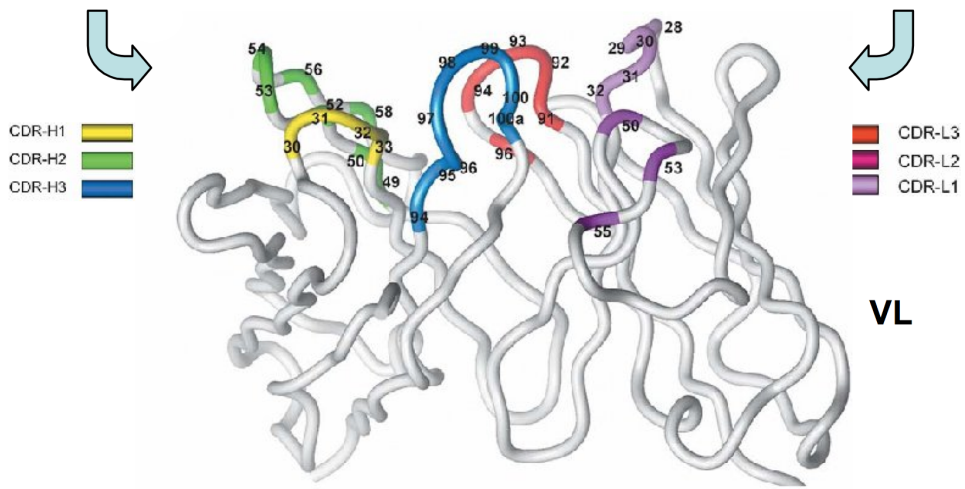
Таблица: Различия в предсказаниях между IgBLAST и ROSIE

регион	DDR (%)	AS_s	AS_u
FR1 начало	6.80	-1.26	1.26
FR1 конец	0.03	-1.00	1.00
FR2 начало	0.03	-1.00	1.00
FR2 конец	23.7	1.97	1.97
FR3 начало	0.80	1.00	1.00
FR3 конец	99.7	2.01	2.01

Образование вариабельных доменов антител



Рекомбинация гермлайн генов и соматические гипермутации



VL

Ig Fv фрагмент

Метод поиска паттернов

```
QVQLVQSGAEVKKPGASVKVSCKASGYTFT  G--YYMH  WVRQAPGQGLEWMG  WINP--NSGGTNYAQKFQG  RVTMTRDTSISTAYMELSRLRSDDTAVYYCAR
QVQLVQSGAEVKKPGASVKVSCKASGYTFT  S--YAMH  WVRQAPGQRLEWMG  WINA--GNGNTKYSQKFQG  RVTITRDTSASTAYMELSSLRSEDTAVYYCAR
QVQLVQSGAEVKKPGASVKVSCKASGYTFT  S--YDIN  WVRQATGQGLEWMG  WMNP--NSGNTGYAQKFQG  RVTMTRNTSISTAYMELSSLRSEDTAVYYCAR
```

- Позволяет находить только границы CDR.
- Определяет регионы только в конкретной номенклатуре.
- Не работает на данных с ошибками.
- Не работает на некоторых организмах с очень высокой вариабельностью гермлайнов (например, на кроликах).

Метод выравнивания на референсы

```
...LAISGLQSEDEADYHCMGSGIAVF...
...LTIKNIQEEDESDYYC-GSGIVVF...
...---FR3---------> <--CDR3...
```

- Требует огромную базу.
- Использует одно лучшее выравнивание для получения регионов.
- Работает только с V-геном (разметка не далее FR3 региона) или выравнивает гены независимо.