Система нахождения регионов на IG V-chains

Яковлев Павел

СПбАУ РАН

21 декабря 2013 года

BLAST

Принципы:

- Использование индекса k-меров (3-меры для аминокислот и 11-меры для нуклеотидов) для быстрого поиска в огромных базах
- Продление найденных k-меров в имеющейся базе

Проблемы в иммуноглобулинах:

- Высокая вариабельность (может быть более 1 мутации на каждый 11-мер)
- Огромное количество комбинаторных вариантов

Примеры

$$k = 4$$

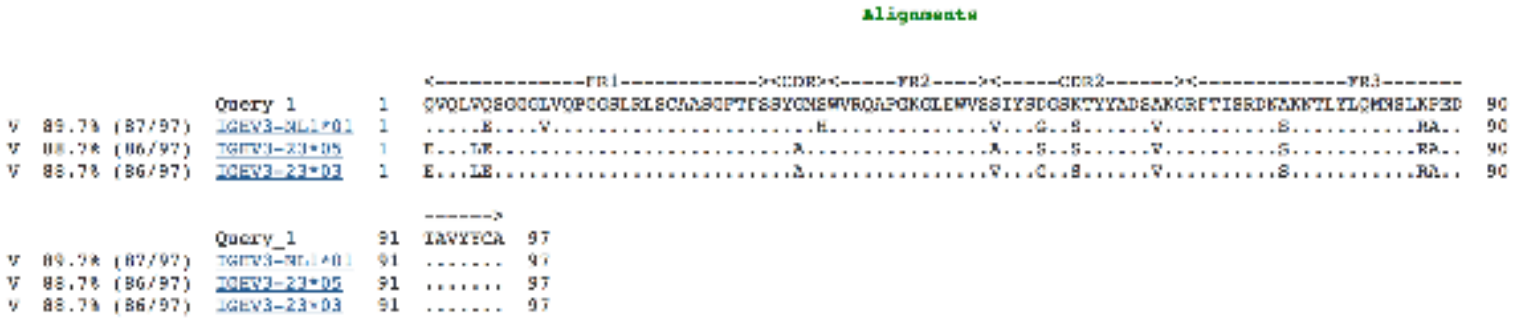
```
Запрос:
ACGCGT
ACGC
 CGCG
  GCGT
```

```
База:
 ACCCGTG
 ACGTCTG
 AAACCCC
 ATTCTTT
```

Ни один k-мер из запроса не присутствует в базе

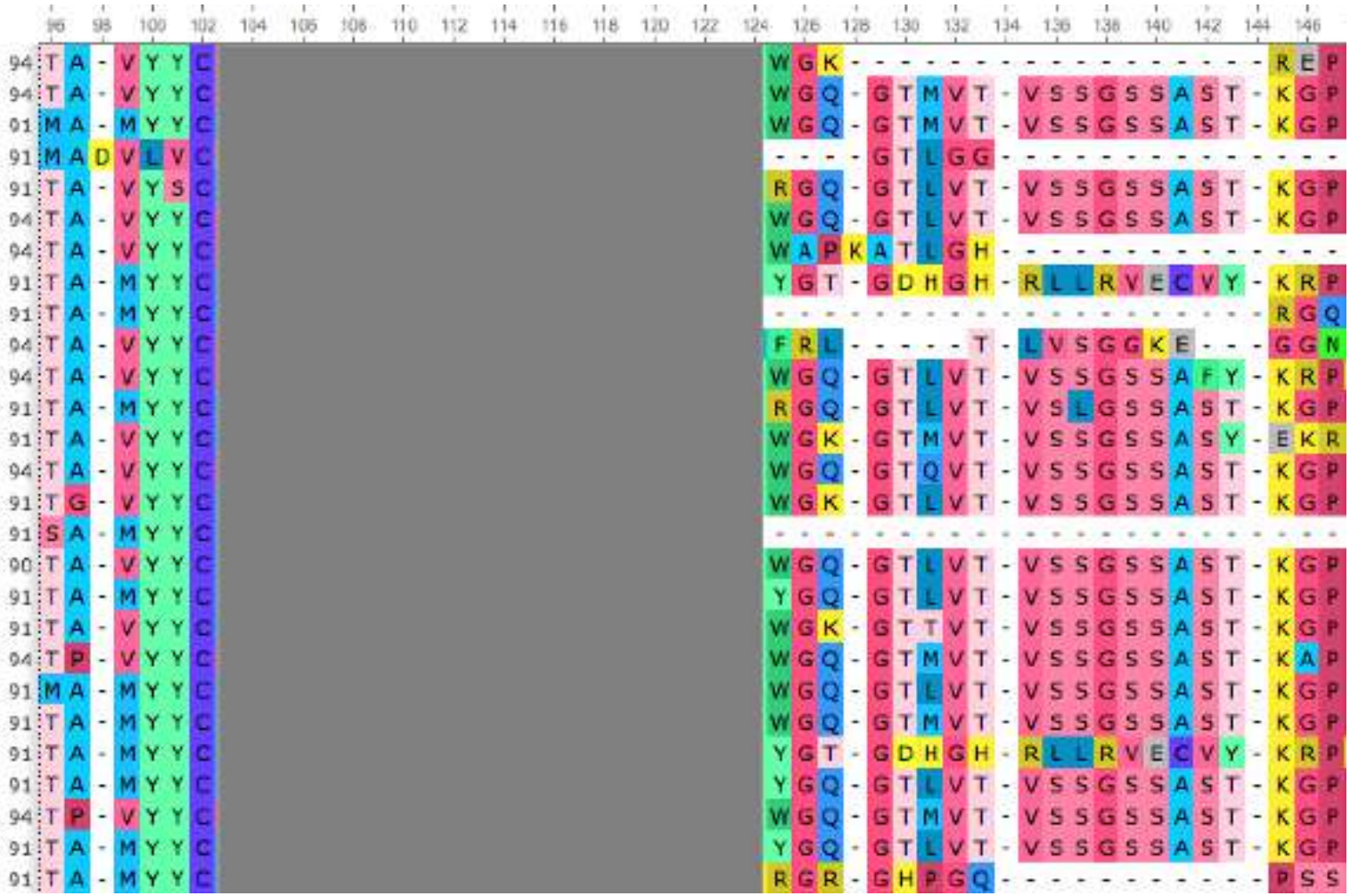
IgBLAST

- Для гермлайнов не используется эвристика с k-мерами, т.о. возможна работа только с малыми базами - ограничение V-геном (т.е. разметка не далее FR3)
- Для сторонних баз эвристика применяется, но нет фильтра на использование только иммуноглобулинов (в выравнивания могут случайно попасть иные белки)
- Используется множественное выранивание - долго и вызывает проблемы, связанные с накоплением ошибки (см. далее)



Проблема множественного выравнивания

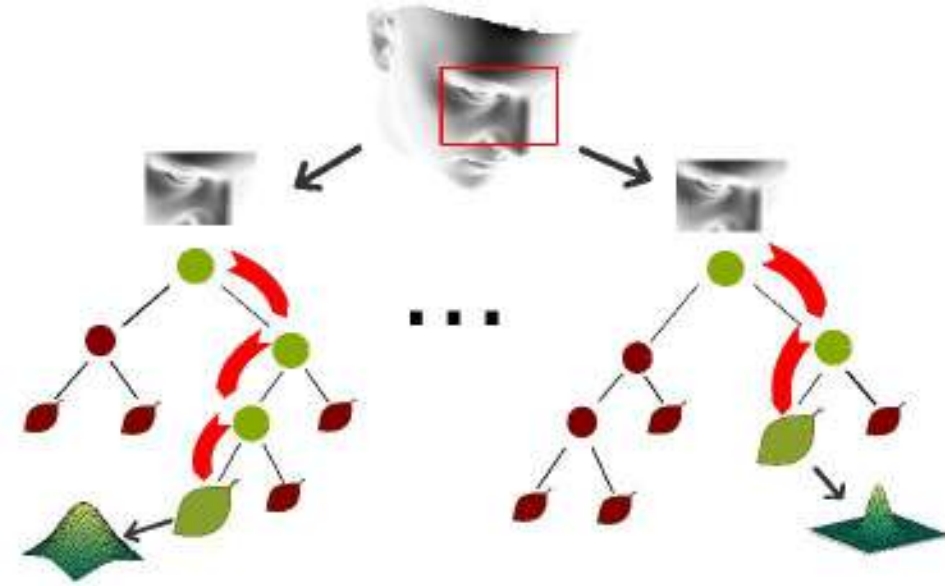
Еще один пример

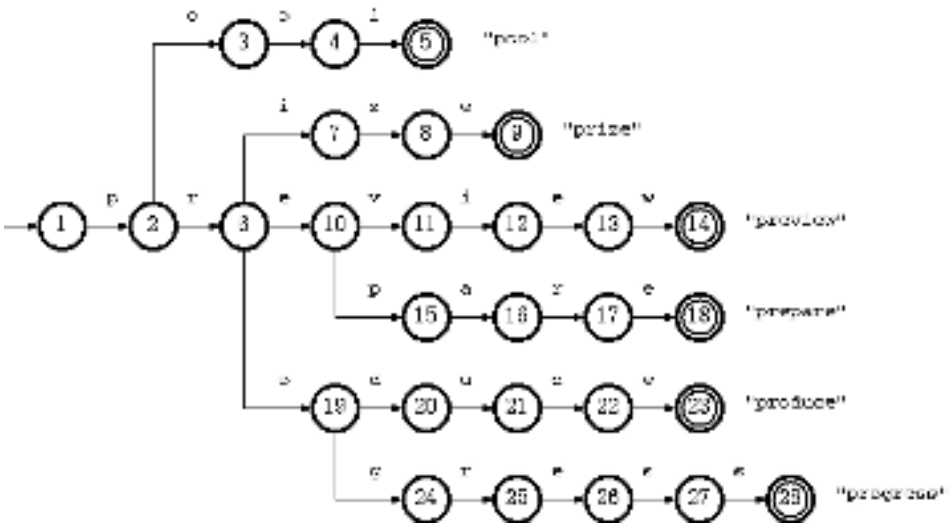


Альтернативные методы

В настоящий момент разработаны и тестируются два метода.

1. Метод случайных деревьев
2. Метод аннотирующего контейнера





Случайные деревья

- Принцип работы аналогичен человеку, выискивающему регионы глазами
- Метод машинного обучения
- Предсказание быстрее любого метода, основанного на выравниваниях, а обучение требуется очень редко
- Требует размеченную обучающую выборку только для тренировки модели
- Подготовка данных аналогична методам предсказания вторичных структур (GOR, PHD, JPred): последовательность разбивается на пересекающиеся участки нечетной длины, каждому участку ставится в соответствие регион, которому принадлежит средняя аминокислота

Проблемы:

- Чем больше окно, тем выше точность, но меньше скорость обучения
- Обучающая выборка должна быть близка к предсказываемым данным (в идеале - гермлайны соответствующего вида)
- Требуется вносить шум в обучающую выборку для избежания overfitting

Аннотирующий контейнер

- Позволяет хранить полные последовательности иммуноглобулинов в нуклеотидных или аминокислотных последовательностях с произвольными аннотациями для каждой буквы
- Хранит иммуноглобулины в виде дерева, что позволяет "склеивать"их префиксами (у иммуноглобулинов вариабельность растет к концу)
- Позволяет искать неточные входения паттернов (например, поиск CDR3 с одной заменой)
- Позволяет получать TOP N наилучих парных выравниваний последовательности-запроса на хранящиеся элементы
- По полученным парным выравниваниям позволяет аннотировать последовательность (в том числе последовательности, содержащие иммуноглобулины с "мусором"на краях и кусочки иммуноглобулинов)
- Высокая точность результатов за счет подхода похожего на BLAST, но лишенного его недостатков

Проблема аннотирующего контейнера

Данная проблема встречается только при использовании глобального выравнивания.

- Все выравнивания строятся независимо
- По имеющимся выравниваниям строится консенснусная аннотация
- Ошибка не накапливается, как во множественном выравнивании, **но** все равно иногда может произойти

```
YYCAR VG.......WDQLDA WGQGTTVTVSSGSSASTKGPSVFPAG
44444 555555555666666 66666666666666666666666666
FR3-> <----CDR3-----> <----------FR4----------->
```

Почему так происходит?

BLOSUM62

- Универсальная матрица, стандарт для аминокислотных выравниваний
- Построена исходя из эволюционных процессов на базе группы гомологичных медленно меняющихся белков из различных баз данных
- Предназначена для сравнения гомологичных белков различных видов, а потому устанавливает максимальные веса для наиболее консервативных аминокислот с течением эволюции

Подходит ли такая матрица для гипервариабельных иммуноглобулинов?
Что произойдет, если заменить ее?

Эксперимент

Поднимем веса **W**, **G**, **Q** и опустим вес **D**:

```
YYCAR VG.......WDALDA WGQGTTVTVSSGSSASTKGPSVFPAG
44444 555555555555555 6666666666666666666666666
FR3-> <----CDR3-----> <----------FR4----------->
```

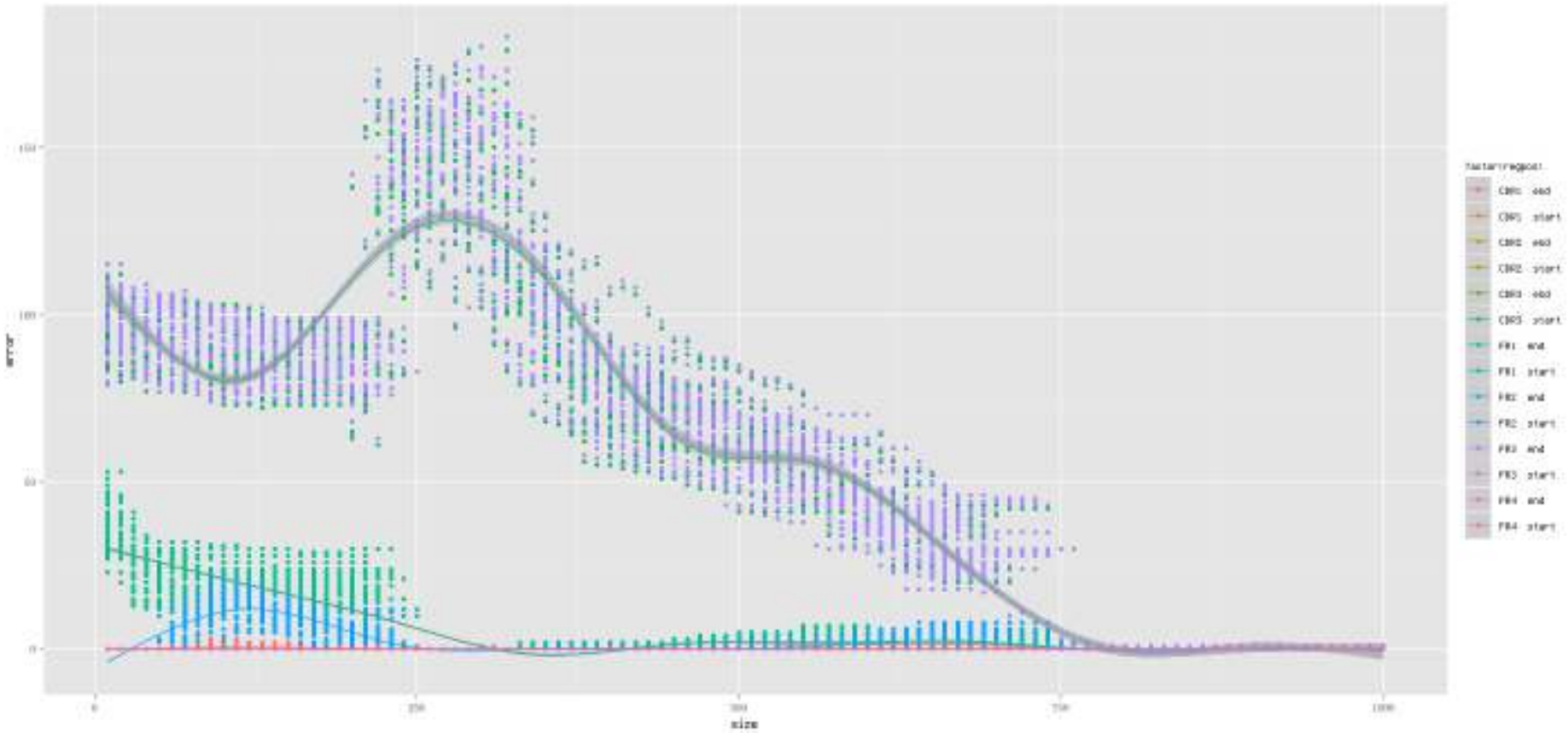
Гипотеза

- Иммуноглобулины - уникальные белки, эволюционные процессы в которых идут в миллионы раз быстрее, чем даже в прокариотах
- Использование классических матриц не оправдано в связи с различной биологической природой, а значит и различными статистическими моделями
- Гермлайновые последовательности весьма консервативны и позволяют составить альтернативную статистическую модель для иммуноглобулинов, учитывая как синонимичность замен, так и "устойчивость"аминокислот

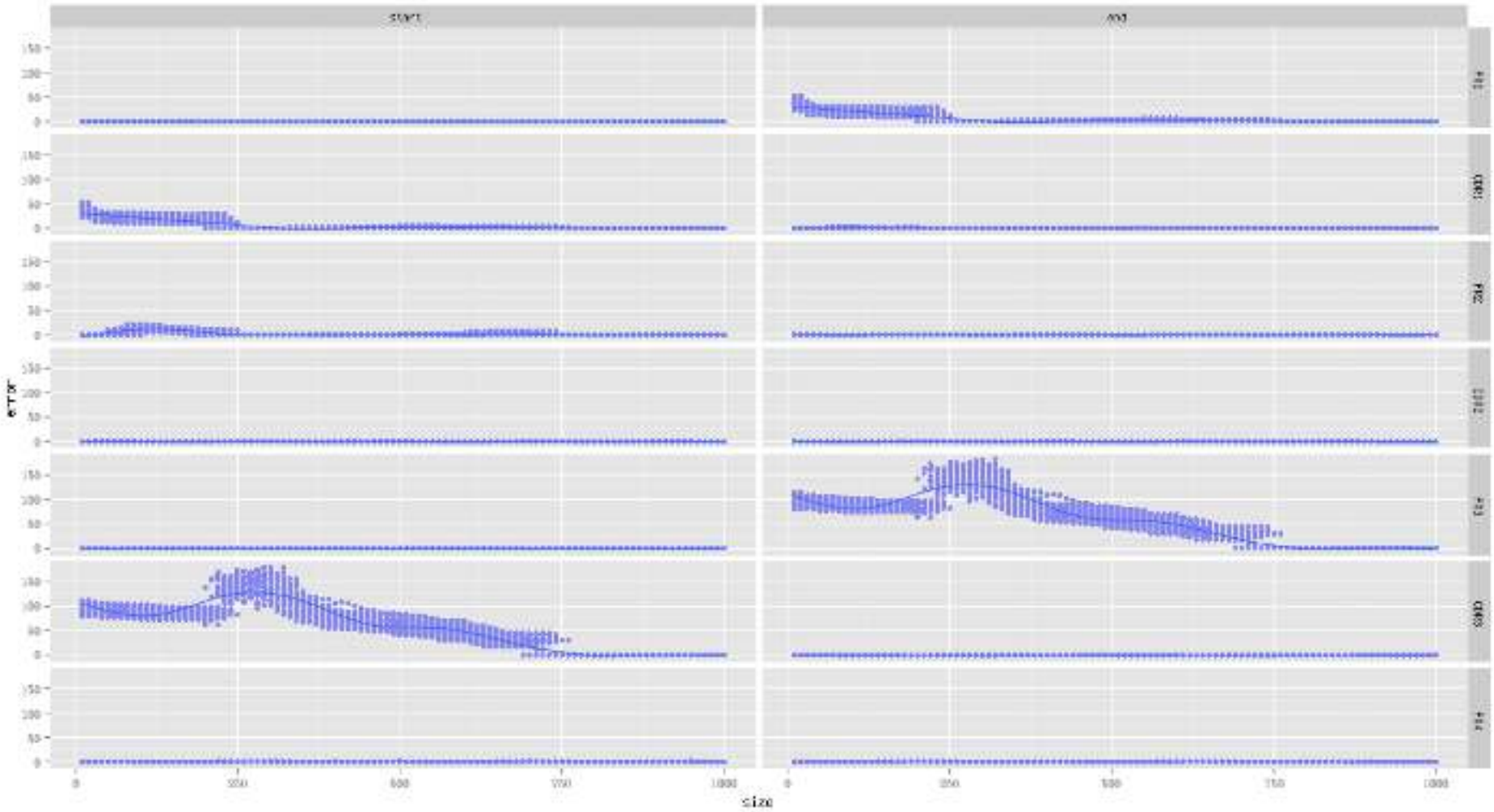
В настоящий момент производится расчет альтернативной матрицы выравнивания на основе гермлайнов человека, мыши, крысы, кролика и макаки.

Альтернатива: использование semiglobal alignment для аннотирования.

Результаты I



Результаты II



Результаты III

Гермлайны

- Случайная выборка объема 10000 из размеченных гермлайнов тяжелых цепей (всего около 22000)
- Разбиение 30:70 - референсные и тестовые последовательности
- 30 прогонов теста со сравнением с истинной разметкой

REGIONS:	errors	signed / unsigned	signed / unsigned
FR1 start:	0	0.000 / 0.000	0.000 / 0.000
FR1 end:	0	0.000 / 0.000	0.000 / 0.000
CDR1 start:	0	0.000 / 0.000	0.000 / 0.000
CDR1 end:	0	0.000 / 0.000	0.000 / 0.000
FR2 start:	0	0.000 / 0.000	0.000 / 0.000
FR2 end:	0	0.000 / 0.000	0.000 / 0.000
CDR2 start:	0	0.000 / 0.000	0.000 / 0.000
CDR2 end:	0	0.000 / 0.000	0.000 / 0.000
FR3 start:	0	0.000 / 0.000	0.000 / 0.000
FR3 end:	0	0.000 / 0.000	0.000 / 0.000
CDR3 start:	7	-0.002 / 0.002	-1.857 / 1.857
CDR3 end:	0	0.000 / 0.000	0.000 / 0.000
FR4 start:	0	0.000 / 0.000	0.000 / 0.000
FR4 end:	0	0.000 / 0.000	0.000 / 0.000

Результаты IV

Реальная ситуация:

$IGHV3-22\_IGHD3-4*2\_IGHJ1$	...	65	96	97	114	115	125
$IGHV3-22\_IGHD3-4*2\_IGHJ1$	...	65	96	99	114	115	125
$IGHV5-1\_IGHD2-4*1\_IGHJ6$	...	67	98	99	115	116	126
$IGHV5-1\_IGHD2-4*1\_IGHJ6$	...	67	98	100	115	116	126

- Отсутствие аннотаций обуславливается наличием гэпов во всех ТОП-выровнявшихся референсов в данной позиции.
- Возможно исправление ситуации с помощью фильтрации, присоединяющего непроаннотированные позиции на краях к ближайшему CDR-региону.
- Точность теста в таком случае достигает 100%.

Результаты V

IgBASE + IgBLAST

- 302 пары сиквенсов (VH + VL) из базы IgBASE
- Все сиквенсы размечены до FR3 региона включительно с помощью IgBLAST
- В качестве референсной базы для аннотирующего контейнера используется случайная смесь последовательностей тяжелых и легких гермлайнов человека
- Наблюдается системная разница сдвига начала FR1 региона: IgBLAST использует локальное выравнивание, которому свойственно "отрезать"края, при этом достоверно известно, что в IgBASE лежат полные последовательности вариабельных доменов
- Редкие ошибки на краях CDR2 региона имеют случайный характер

REGIONS:	errors	signed / unsigned	signed / unsigned
FR1 start:	16	0.088 / 0.088	1.625 / 1.625
FR2 end:	8	0.007 / 0.027	0.250 / 1.000
CDR2 start:	8	0.007 / 0.027	0.250 / 1.000
CDR2 end:	1	-0.003 / 0.003	-1.000 / 1.000
FR3 start:	1	-0.003 / 0.003	-1.000 / 1.000

Результаты VI

ROSIE

- 336 пар сиквенсов из публичных успешно завершенных запросов к сервису ROSIE
- Границы регионов найдены методами машинного обучения и уточнены рассчетом фолдинга с помощью пайплайна Rosetta Antibody
- Аннотация проведена в номенклатуре Chothia, имеющей системное отличие от Kabat
- Большинство ошибок имеют случайный характер и часто связаны с отсутствием аннотации, что убирается соответствующим фильтром
- Наличие системной ошибки сдвига на одну позицию для конца FR4 региона обуславливается тем, что Rosetta ищет только границы регионов, аннотирую FR4 регионом сиквенс до его завершения, даже если в конце присутствует часть следующего, консервативного, домена

Результаты VII

Heavy

REGIONS:	errors	signed / unsigned	signed / unsigned
FR1 start:	2	-0.009 / 0.009	-1.500 / 1.500
FR1 end:	1	0.003 / 0.003	1.000 / 1.000
CDR1 start:	1	0.003 / 0.003	1.000 / 1.000
CDR2 end:	2	0.006 / 0.006	1.000 / 1.000
FR3 end:	24	-0.003 / 0.098	-0.042 / 1.375
CDR3 start:	29	-0.006 / 0.107	-0.069 / 1.241
FR4 end:	39	0.116 / 0.116	1.000 / 1.000

Light

REGIONS:	errors	signed / unsigned	signed / unsigned
FR1 start:	12	-0.054 / 0.054	-1.500 / 1.500
FR2 end:	10	0.006 / 0.030	0.200 / 1.000
CDR2 start:	10	0.006 / 0.030	0.200 / 1.000
CDR3 start:	1	-0.003 / 0.003	-1.000 / 1.000
FR4 end:	142	0.747 / 0.747	1.768 / 1.768

Выводы

- Разработанная алгоритмическая модель превосходит существующие известные аналоги по точности и объему обрабатываемых данных, а также сравнима с ними по скорости работы
- Из биологических и статистических данных следует возможность создания новой альтернативной таблицы аминокислотных замен, характерной для иммуноглобулиновых последовательностей
- Предложенный метод может обеспечить стабильную работу имеющихся методов для определения регионов