[Рабочий] Еще пара подходов к поиску регионов на вариабельных доменах иммуноглобулинов

December 5, 2013

Abstract

Motivation:

Идея:

Потребность анализировать огромное количество последовательностей иммуноглобулинов, приходящих со скрининга, NGS и прочего. Требуется находить регионы для выполнения оценки разнообразия, определения типов цепей, кластеризации специфичных антител и выполнение точечной гуманизации.

Results:

Идея:

У нас получилось 2 инструмента, которыми можно удобно пользоваться. Один чувствителен к входным данным, другой потребовал альтернативной матрицы выравнивания. Однако оба демонстрируют довольно высокую точность и активно используются в нашем продакшене.

Availability:

Скачать все это можно абсолютно бесплатно здесь: [создать публичный репозиторий и оформить в него ig-snooper, ig-alicont, ig-container и ig-annotator]

Contact:

yakovlev@biocad.ru

1 Introduction

Идея (нужна быстрая массовая обработка):

High-throughput требует новых алгоритмов для быстрой обработки огромного количества данных. Точечный метод, распространенных ранее, когда каждый глобулинчик рассматривали самостоятельно более неприменим. Это позволяет во много раз сократить период разработки новых лекарств.

Идея (обзор существующих решений):

Современные инструменты для поиска регионов несовершенны, т.к. из-за большого числа вариантов фигово ищут регионы. Большинство из них (IgBLAST, IMGT/V-QUEST) находят их только на V-гене в силу своих ограничений. Более того, ни один из подобных инструментов (igblast, v-quest, proabc и другие) не может быть использован для построения пайплайна автоматической обработки, в котором этап поиска регионов обязательно должен быть.

Идея (мы предлагаем то, что будет лучше всех):

Мы хотим представить два новых метода, которые могут быть использованы не только для единичного поиска регионов, но и для обработки больших данных в автоматическом режиме. Оба эти инструмента возвращают результаты в машино-читаемом формате, а потому могут быть легко интегрированны в любую систему автоматической обработки.

2 Approaches

2.1 General

Идея (зачем нам два метода):

При разработке методов нам стало интересно попробовать сразу два подхода. При глобальной обработке иммуноглобулинов точность определения регионов не так важна, тогда как при исследовании малой выборки кандидатов, прошедших многие этапы отсева в мокрой лабе, хотелось бы иметь инструмент, который почти никогда не ошибается. Именно так и появились два тула, один из которых испольузет машинное обучение на подобии proabc (но решает некоторые дополнительные задачи, ему недоступные), а другой пользуется классическим методом выравнивания с применением некоторой эвристики, что позволяет избежать проблем, с которыми столкнулись igblast и v-quest.

2.2 Random Forest

Идея (биологическая связь между вторичной структурой и регионами, похожесть задач, использование принципов существующих решений):

Регионы задаются пространственной структурой иммуноглобулина. Сам иммуноглобулин - это *beta*-складка, часть которой лежит на поверхности, а часть убрана. Есть изоморфизм между вторичной структурой и положением регионов. Известные тулы типа PHD или JPred построены на машинном обучении, опираясь на соображение, что окружение аминокислоты, как и она сама задает вторичную структуру. У них получились неплохие результаты для произвольных белков, и мы решили вдохновиться этим для иммуноглобулинов.

Идея (почему random forest):

Подход с SVM или нейронной сетью недостаточно тру, так как сама природа регион-характерных последовательностей плохо ложится на разделяемое гиперплоскостями feature-space. Random Forest здесь более подходит идеалогически, поскольку рассматривает паттерны с вариациями в разных позициях иммуноглобулина так же, как это делает биолог, глазами ищущий знакомые паттерны.

Идея (ближайший аналог - proabc):

К той же мысли, видимо, пришли создатели proABC, однако их подход решает одну конкретную задачу, и не способен, например, определить регионы на куске сиквенса, более того, он использует дополнительные выравнивания и парный анализ цепей, тогда как мы хотим сделать возможным гнать поток из всякого трэша, в котором все равно искать регионы.

2.3 Annotation Container

Идея (бор):

Для алгоритмов поиска множества паттернов типа Ахо-Корасик используется префиксное дерево или бор, которое позволяет уменьшить количество данных, склеивая строки префиксами, где это возможно.

Идея (биологическое обоснование):

За счет V(D)J рекомбинации и особенностей иммуноглобулинов вообще их вариабельность растет к концу, в начале же могут быть весьма схожие последовательности. Нередка ситуация, когда даже антитела на разные мишени имеют одинаковые не то что FR, но и первые 1-2 CDR региона. В таком случае использвоание бора для хранения иммуноглобулинов кажется отличной идеей.

Идея (разделенеие сиквенсов и аннотаций):

Хранение сиквенсов в виде бора, что позволяет проводить все алгоритмы на сильно меньших данных и, таким образом, сократить время их работы. Аннотации хранятся отдельно и обращение к ним идет при необходимости.

3 Methods

3.1 Datasets

Идея (дополнительные данные, которые нам пришлось сгенерировать):

Для обоих методов требуется некоторое количество данных, которые уже размечены, чтобы обучаться на них или использовать их в качестве референсного набора. Для этого мы подготовили размеченные человеческие гермлайны, взяв их из V-Base и просчитав все возможные конкатенации строк. Оба наших подхода работают и для нуклеотидов, и для аминокислот, так что данные были подготовлены и такие, и такие.

3.2 Random Forest

Идея (подготовка данных):

Скользящее окно, выходящее за рамки иммуноглобулина, определение центральной аминокислоты, сглаживающее окно.

Идея (weka):

Используется Weka с ее random forest. Реализация на python - подготовка данных, их передача и формирование результатов.

3.3 Annotation Container

Идея (реализация):

Попытка сделать очень эффективную реализацию на языке Scala.

Идея (построение дерева):

Строим бор из сиквенсов. Разделяем реализацию бора и данные в вершинах с возможностью быстрой подмены последних.

Идея (сохранение индекса k-меров):

Параллельно с добавлением в бор строим таблицу k-меров, в которой каждый k-mer указывает на список узлов, в которых он заканчивается. Наличие ANY-символа, меняющегося на любой.

Идея (добавление аннотаций):

Для каждого сиквенса можем добавить аннотации и константно переходить из аннотации в узел, а также из узла в список аннотаций.

Идея (поиск):

Алгоритм поиска на k-мерах с возможностью наличия ANY-символа.

Идея (выравнивание: общая суть):

Создание контейнера, позволяющего на некоторую постоянную строку выравнивать строку, состоящую из набора неперекрывающихся подстрок. Организация по принципу стэка для добавления новых строк.

Идея (выравнивание: алгоритмы):

Реализация 6 наиболее распространенных и нужных алгоритмов выравнивая: глобальный, локальный семиглобальный и их аналоги с аффинными гэпами. Пример необходимости semiglobal - поиск аннотаций для кусочка иммуноглобулина.

Идея (выравнивание: DFS-поиск выравниваний):

Обход дерева по DFS, выравнивае каждого ребра ровно один раз, константная память на протяжении всей работы алгоритма, fixed-size очередь с приоритетами или список для выдачи результата как top N или со степенью гомологичности более X.

Идея (аннотирование):

Подсчет консенсуса аннотаций по нескольким топовым выравниваниям.

3.4 Alternate matrix

Идея (блосум - не торт):

Блосум - эволюционная матрица, построенная на данных о консервативных белках у различных видов, разделяемых продолжительным эволюционным расстоянием. IG эволюционируют в миллионы раз быстрее, требуется альтернативная матрица.

Идея (составление матрицы):

Составлена на данных из человека (и кого-нибудь еще) по алгоритму, аналогичному тому, по которому составляли blosum.

4 Results

?

5 Discussion

Идея (улучшение случайных деревьев):

Требуется использовать обучающую выборку из близкородственного организма. Стоит повносить шум в обучающую выборку в некоторых хотспотах для исключения overfitting тренировочных данных.

Идея (улучшение поиска за счет изменение индекса k-меров):

Добавить вырожденные нуклеотиды по аналогии с ANY заменяющиеся только на некоторые нуклеотиды, а не на все.

Идея (улучшение выравнивания):

Сделать alicont масштабируемым по обеим координатам по требованию для исключения постоянных перевыделений памяти на его создание для каждого нового запроса.

Идея (хранения аннотаций):

Для всех алгоритмов аннотации не нужны, они треубется только на самом последнем этапе для выдачи информации. Хотелось бы хранить их на диске и подгружать только по мере надобности.

6 Conclusion

?

7 Acknowledgement

Спасибо БИОКАДу за наше счастливое детство!