Локализация дисульфидных связей в молекулах антител

Студент - Кощенко Екатерина Васильевна
Руководитель - Вяткина Кира Вадимовна
СПБАУ РФНб 2016г.

Введение.

Антитела - белки, вырабатываемые иммунной системой в ответ на появление в организме чужеродных субстанций.

Антитело: две идентичные тяжелые цепи + две идентичные легкие, соединенные друг с другом дисульфидными связями.

Исследование аминокислотного состава антител является одной из важных задач фармакологии.

Одна из ее составляющих: аннотация масс-спектра, снятого с антитела.

Цель.

Дан тандемный масс-спектр, снятый с антитела с известной аминокислотной последовательностью. Требуется его аннотировать, т.е. указать для каждого пика, для которого это возможно, какому фрагменту молекулы анализируемого антитела он соответствует. Следует иметь в виду, что такой фрагмент может представлять собой участки легкой и тяжелой цепей антитела, соединенные дисульфидной связью. Также возможны различные модификации внутри фрагмента.

Проблемы.

1. Возможны потери молекул H_2O и NH_3.

2. Возникновение внутримолекулярных дисульфидных связей.

3. Существование химер - фрагменты, представляющие собой участки легкой и тяжелой цепей антитела, соединенные дисульфидной связью. Такие фрагменты представляют для нас наибольший интерес, так как они помогают локализовать дисульфидные связи в молекуле.

Задачи.

1. Аннотировать пики, соответствующие
    a. простым фрагментам
    b. химерическим
2. Разработать формат вывода полученных результатов

Решение задачи.

1. **Аннотировать масс-спектр простыми фрагментами.**

Для этой задачи была написана функция peak_search(...), в основе которой лежит перебор.

Перебираются начала и конца строки, подставляются модификации. После этого проверяется, есть ли получившаяся масса (с заданной пользователем точностью) в масс-спектре.

Решение задачи.

2. Найти химеры.

> **⚠ ВАЖНО ⚠**
>
> Легкая цепь всегда оканчивается цистеином.
>
> Последний цистеин легкой цепи дает дисульфидную связь с тяжелой.

Перебираются суффиксы легкой цепи с модификациями, полученная масса фиксируется. Теперь передаем эту массу в старую функцию peak_search(), ищем соответствующий участок тяжелой цепи.

Реализация.

Почему именно перебор?

Перебор, время работы: $(|l\_chain| \cdot (<<(|h\_chain|^2)) \cdot P(x)$

Генерация по цепи всех возможных подстрок.

⚠ Проблема: 20 аминокислот, длина легкой цепи >200, тяжелой >400.

Реализация.

Как хранился результат?

Для обращения к результату по значению пика использовался unordered_map<long double, MyStruct>

Лучше было использовать свою хэш-таблицу.

Архитектура.

class Spectrum {};

Отвечает за работу с файлами. Считывает цепи антитела и масс-спектр из входных файлов, выводит результаты.

Что нас интересовало:

аннотированный масс-спектр (и обычные, и химеры)

масс-спектр с распечатанными фрагментами аннотации

для каждого цистеина химеры с его участием

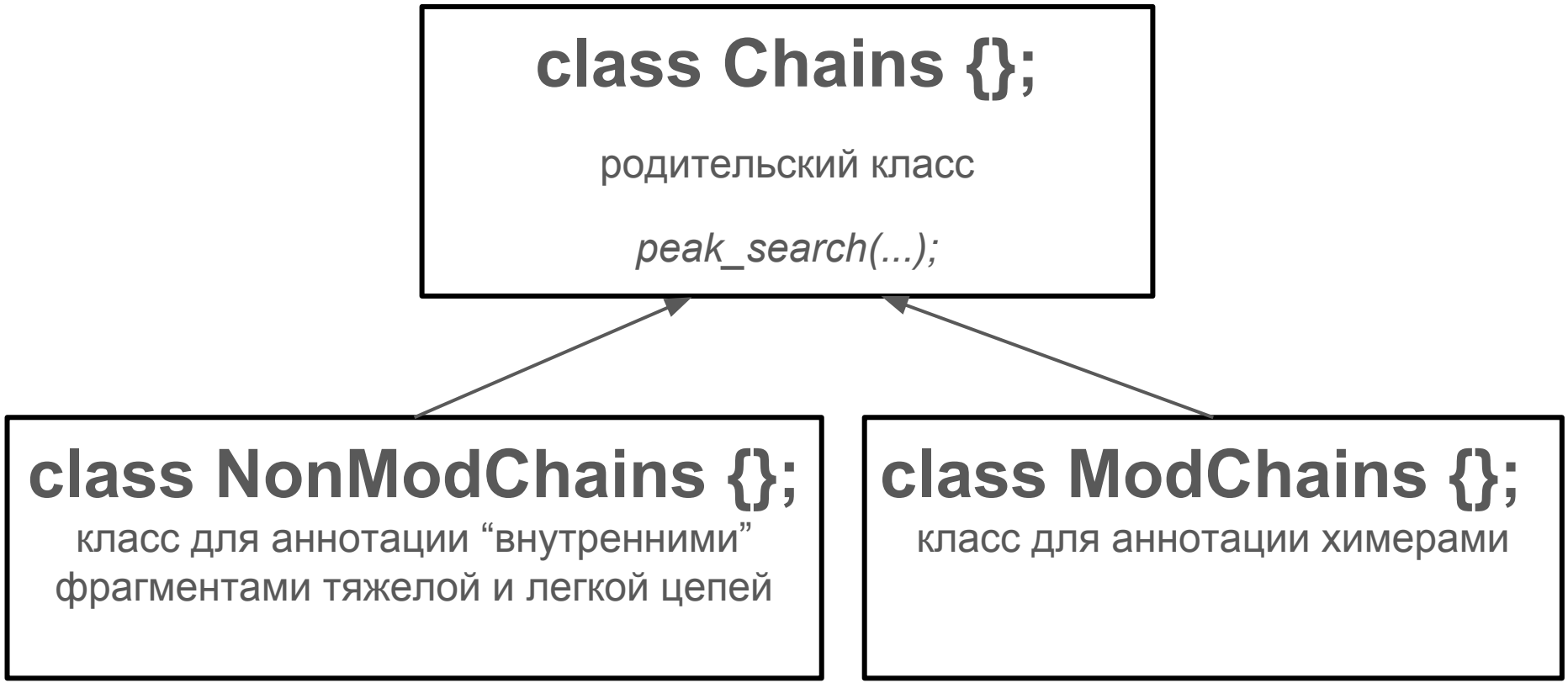
другие занятные результаты.

Архитектура.

class Antibody {};

Тут лежат результаты обработки масс-спектра. Аннотация простыми фрагментами обеих цепей, аннотация химерами, количества пиков, которые удалось аннотировать каким-то способом.

Архитектура.



Результат.

1. Аннотируются почти все пики (исключения - маленькие массы). Использованная точность - 0.00001ptm.
2. Время работы ~1-2 минуты.
3. Найдено много химер (813 / 833).

Есть ли аналоги?

Существующие программные инструменты для аннотации масс-спектров не позволяют обработать химерические фрагменты.

Было предложено несколько методов локазизации дисульфидных связей в антителах (например, Shen et al.,2010 и Zhang et al., 2016). Однако мы работаем с масс-спектрометрическими данными, полученными по специальной технологии, разработанной Tsybin et al. в Ecole Polytechnique Federale de Lausanne (EPFL), Швейцария. Это данные очень высокого качества, и для их обработки необходимо предложить новые методы.

Что не получилось.

Придумать алгоритм, который будет работать быстрее перебора. Хотя есть ли смысл?

Использовать что-то, что дает доступ к любому элементу быстрее, чем unordered_map.

Визуализация.

Скрины.

annotated.txt

1198	21197.46995	i(20-212)-H2O	b201-H2O-NH3+(S-S)	ly1+S-S+hi(4-203)-H2O+(S-S)
1199	DELTA MASS=	-0.00798196	0.10367804	0.10936804
1200				
1201	21459.61864	i(3-198)-H2O	i(6-208)-NH3	ly1+S-S+hi(8-209)-NH3+(S-S)
1202	DELTA MASS=	0.15113174	0.12645174	0.13949174
1203				
1204	21485.49491	i(4-199)+(S-S)	i(12-213)-H2O	ly1+S-S+hi(9-210)-NH3+2(S-S)
1205	DELTA MASS=	-0.05047644	-0.05981644	0.03649356
1206				
1207	21517.55628	i(11-207)-NH3	i(17-217)-H2O+(S-S)	ly1+S-S+hi(2-204)-H2O-NH3+(S-S)
1208	DELTA MASS=	-0.08072682	-0.06104682	0.08760318
1209				
1210	21549.75722	i(7-204)-NH3	i(3-206)+(S-S)	ly1+S-S+hi(16-216)-2H2O
1211	DELTA MASS=	0.16676492	0.19298492	0.20420492
1212				
1213	21571.60119	i(2-198)-NH3+(S-S)	i(5-208)	ly1+S-S+hi(1-203)
1214	DELTA MASS=	0.08050684	0.04136684	0.05574684
1215				
1216	21629.61444	i(7-205)-2H2O	i(5-209)-H2O-NH3+2(S-S)	ly1+S-S+hi(2-205)-H2O+2(S-S)
1217	DELTA MASS=	-0.04439412	0.07895588	0.07812588
1218				

Скрины.

pictured.txt

```
6376  PEAK:  13881.83986
6377
6378  LIGHT FOUND:                    i(54-180)-2H2O+(S-S)
6379  LYSGVPSRFSGSRSGTDFTLTISSLQPEDFATYYCQQHYTTPPTFGQGTKVEIKRTVAAPSVFIFPPSDEQLKSGTASVVCLLNNFYPREAKVQWKVDNALQSGNSQESVTEQDSKDSTYSLSSTLT
6380  DELTA MASS = 0.10049131
6381
6382  HEAVY FOUND:                     i(5-133)-2H2O+(S-S)
6383  VESGGGLVQPGGSLRLSCAASGFNIKDTYIHWVRQAPGKGLEWVARIYPTNGYTRYADSVKGRFTISADTSKNTAYLQMNSLRAEDTAVYYCSRWGGDGFYAMDYWGQGTLVTVSSASTKGPSVFPLAP
6384  DELTA MASS = 0.08174131
6385
6386  MODIFIED FOUND:             ly1+S-S+hi(27-152)+(S-S)
6387  C<--
6388  -->FNIKDTYIHWVRQAPGKGLEWVARIYPTNGYTRYADSVKGRFTISADTSKNTAYLQMNSLRAEDTAVYYCSRWGGDGFYAMDYWGQGTLVTVSSASTKGPSVFPLAPSSKSTSGGTAALGCLVKDY
6389  DELTA MASS = 0.12305131
6390  --------------------
```

Скрины.

modified.txt

```
1
2  CYSTEIN 5
3  ------------------------------------------
4  6935.550284:              ly1+S-S+hi(207-269)+(S-S)
5  9048.616151:            ly1+S-S+hi(173-256)+2(S-S)
6  11189.57138:            ly1+S-S+hi(124-230)+2(S-S)
7  11217.53063:             ly1+S-S+hi(155-259)+(S-S)
8  11317.57978:                   ly1+S-S+hi(129-236)
9  11646.79263:             ly1+S-S+hi(121-232)+(S-S)
10 11760.79371:             ly1+S-S+hi(125-237)+(S-S)
11 11841.90328:            ly1+S-S+hi(127-240)+2(S-S)
12 11841.90362:                   ly1+S-S+hi(129-242)
13 11842.90019:            ly1+S-S+hi(122-235)+2(S-S)
14 12111.01774:             ly1+S-S+hi(123-238)+(S-S)
15 12281.13924:             ly1+S-S+hi(111-229)+(S-S)
16 12467.25384:             ly1+S-S+hi(133-251)+(S-S)
17 12726.40722:                   ly1+S-S+hi(127-248)
```

Скрины.

cys_process.txt

```
3537  NUMBER OF MODIFIED SEGMENTS THAT COVER CYSTEINS WITH THEIR LEFTEST/RIGHTEST END:
3538  cystein 2:  223
3539  cystein 3:  184
3540  cystein 4:  20
3541  cystein 5:  3
3542  cystein 6:  1
3543  cystein 7:  4
3544  cystein 8:  9
3545  cystein 9:  23
3546  cystein 10:  3
```

Спасибо за внимание!

Репозиторий с проектом: KatyaKos/Proj