Исследование базы данных Immune Epitope Database на предмет необходимой информации для моделирования процессинга антигенов для HLA первого класса

Антон Смирнов

Feb 16, 2023

Оглавление

Введение

Immune Epitope Database представляет из себя крупнейшую базу данных и базу знаний о T- и B- клеточных эпитопах. Она курируется Национальным институтом аллергии и инфекционных болезней США. Количество информации в ней на 16 февраля 2023 года представлено в таблице 1.

Table 1: Метрики БД

Характеристика	Количество
Peptidic Epitopes	1,554,329
Non-Peptidic Epitopes	3,168
T Cell Assays	453,372
B Cell Assays	1,378,855
MHC Ligand Assays	4,697,592
Epitope Source Organisms	4,337
Restricting MHC Alleles	983
References	23,554

Дамп базы данных был скачан с сайта iedb.org и развернут на локальном сервере MySQL.

```
mysql --version
```

mysql Ver 15.1 Distrib 10.6.11-MariaDB, for debian-linux-gnu (x86_64) using EditLine wrapper

Скрипт восстановления БД.

mysql -u stotoshka -p
mysql> create database iedb;
mysql> use iedb;
mysql> source <path to sql_script>;

Необходимые библиотеки

```
library(RMySQL)
library(dplyr)
library(stringr)
library(knitr)
library(vroom)
```

Подключимся к базе данных

```
con = RMySQL::dbConnect(RMySQL::MySQL(),
                        dbname='iedb',
                        host='localhost',
                        port=3306,
                        user='stotoshka',
                        password='meowmeow')
```

Подготовка данных

Поиск расположения необходимых данных

Для начала посмотрим каких типов эксперименты хранятся в этой базе данных.

```
SELECT DISTINCT category FROM assay_type;
```

Table 2: 4 records

category
T Cell
MHC
B Cell
Naturally Processed

В данный момент на не интересуют эксперименты на B-лимфоцитах. Посмотрим, какие эксперименты проводят на T-клетках.

```
SELECT DISTINCT assay_type FROM assay_type WHERE category = 'T Cell';
```

assay_type
RNA/DNA detection
ICS
ELISPOT
cytometric bead array
ELISA
biological activity
in vivo assay

assay_type
x-ray crystallography
binding assay
CFSE
multimer/tetramer
bioassay
51 chromium
3H-thymidine
reporter gene assay
BrdU
surface plasmon resonance (SPR)
in vivo skin test
radio immuno assay (RIA)
intracellular staining
in vitro assay
High throughput multiplexed assay
any method

Эти эксперименты связаны с детекцией распознавания антигена Т-клеточным рецептором. В данный момент они нам скорее всего не подходят, потому что эти эксперименты не ставили задачу определения связывания эпитопа и HLA. Они могут внести сильное смещение в прогноз, так как содержат большое количество положительных примеров.

```
SELECT DISTINCT assay_type FROM assay_type WHERE category = 'MHC';
```

assay_type
purified MHC/competitive/fluorescence
lysate MHC/direct/radioactivity
cellular MHC/competitive/radioactivity
cellular MHC/T cell inhibition
purified MHC/direct/radioactivity
cellular MHC/direct/radioactivity
purified MHC/competitive/radioactivity
purified MHC/direct/fluorescence
x-ray crystallography
any method

assay_type
binding assay
cellular MHC/direct/fluorescence
lysate MHC/direct/fluorescence
cellular MHC/competitive/fluorescence
purified MHC/direct/phage display
cellular MHC
purified MHC
lysate MHC/competitive/radioactivity
High throughput multiplexed assay
lysate MHC

```
SELECT DISTINCT assay_type FROM assay_type WHERE category = 'Naturally Processed';
```

Table 5: 7 records

assay_type
secreted MHC/mass spectrometry
cellular MHC/mass spectrometry
coelution
Edman degradation
T cell recognition
High throughput multiplexed assay
mass spectrometry

Типы экспериментов ‘MHC’ и ‘Naturally Procecessed’, так как методы, содержащиеся в них, направлены на определение связывания эпитопа и MHC.

```
SELECT DISTINCT at2.category,at2.assay_type FROM mhc_elution me
INNER JOIN assay_type at2 ON as_type_id  = at2.assay_type_id;
```

category	assay_type
Naturally Processed	cellular MHC/mass spectrometry
Naturally Processed	coelution
Naturally Processed	secreted MHC/mass spectrometry

category	assay_type
Naturally Processed	T cell recognition
Naturally Processed	Edman degradation
Naturally Processed	mass spectrometry

```
SELECT DISTINCT at2.category,at2.assay_type FROM mhc_bind mb
INNER JOIN assay_type at2 ON as_type_id  = at2.assay_type_id
```

category	assay_type
MHC	purified MHC/competitive/radioactivity
MHC	lysate MHC/direct/fluorescence
MHC	cellular MHC/competitive/fluorescence
MHC	x-ray crystallography
MHC	lysate MHC/direct/radioactivity
MHC	cellular MHC/direct/fluorescence
MHC	purified MHC/direct/phage display
MHC	cellular MHC/competitive/radioactivity
MHC	cellular MHC/T cell inhibition
MHC	purified MHC/direct/fluorescence
MHC	purified MHC/competitive/fluorescence
MHC	cellular MHC/direct/radioactivity
MHC	purified MHC/direct/radioactivity
MHC	cellular MHC
MHC	purified MHC
MHC	binding assay
MHC	lysate MHC
MHC	High throughput multiplexed assay

Данные типы находятся в таблицах ***mhc_bind*** и ***mhc_elution***.

Подготовка mhc_bind

Необходимо, чтобы присутствовала информация об аллеле HLA, эпитопе, источнике и результате эксперимента.

```
CREATE TABLE filtered_bind
AS (SELECT reference_id, curated_epitope_id, as_char_value,
as_location, category, assay_type,units,
as_num_value,as_inequality,as_comments, mhc_allele_restriction_id
   FROM mhc_bind
   INNER JOIN assay_type ON as_type_id = assay_type_id
   WHERE reference_id IS NOT NULL AND
       curated_epitope_id IS NOT NULL AND
       as_char_value IS NOT NULL AND
       mhc_allele_restriction_id IS NOT NULL);
```

Создадим индекс для быстродействия.[1]

```
CREATE INDEX filtered_bind_index ON filtered_bind (curated_epitope_id,mhc_allele_restriction_id);
```

Количество строк

```
SELECT COUNT(*) FROM filtered_bind;
```

Table 8: 1 records

COUNT(*)
810474

Подготовка mhc_elution

Отфильтруем строки с пустыми полями

```
CREATE TABLE filtered_elution
AS (SELECT reference_id, curated_epitope_id, as_char_value, as_location,
category, assay_type, as_num_value, as_inequality, units,
as_num_subjects, as_num_responded, as_response_frequency, as_comments,
as_immunization_comments,h_sex, h_age, mhc_allele_restriction_id,
h_organism_id, ant_type, ant_ref_name, ant_object_id, apc_cell_type,
apc_tissue_type, apc_origin
```

[1] MySQL не поддерживает индексацию представлений, поэтому я вынужден создавать таблицы.

```
    FROM mhc_elution
    INNER JOIN assay_type ON as_type_id = assay_type_id
    WHERE reference_id IS NOT NULL AND
        curated_epitope_id IS NOT NULL AND
        as_char_value IS NOT NULL AND
        mhc_allele_restriction_id IS NOT NULL);
```

```
CREATE INDEX filtered_elution_index ON filtered_elution (curated_epitope_id,mhc_allele_restriction_id);
```

Количество строк

```
SELECT COUNT(*) FROM filtered_elution;
```

Table 9: 1 records

COUNT(*)
3887118

Подготовка mhc_restrinction и epitopes

Также необходимо к имеющимся таблицам присоединить информацию о MHC аллелях и эпитопах. Делаю раздельно в связи с долгим временем выполнения, если делать вместе.

Необходимо, что последовательность была линейная, без модификаций и была в наличии референсная последовательность.

```
CREATE TABLE filtered_epitope_cur
AS (SELECT curated_epitope_id, e_name,
source_antigen_accession, description ,e_region_domain_flag,
e_ev, linear_peptide_seq, e_ref_start,
e_ref_end, `database`, name, sequence, organism_name
    FROM curated_epitope
    INNER JOIN epitope_object eo ON eo.object_id = e_object_id
    INNER JOIN epitope e ON e.epitope_id = eo.epitope_id
    INNER JOIN source s ON s.accession = source_antigen_accession
    WHERE linear_peptide_seq IS NOT NULL AND
        linear_peptide_modification IS NULL AND
```

```
        sequence IS NOT NULL);
```

```
CREATE INDEX filtered_epitope_cur_index ON filtered_epitope_cur (curated_epitope_id);
```

```
SELECT COUNT(curated_epitope_id) FROM filtered_epitope_cur;
```

Table 10: 1 records

COUNT(curated_epitope_id)
3143463

Нам нужны человеческие (NCBI Taxonomy ID 9606) MHC 1 класса.

```
CREATE TABLE filtered_mhc
AS (SELECT mhc_allele_restriction_id,
restriction_level,displayed_restriction, organism_ncbi_tax_id,class, chain_i_name
   FROM mhc_allele_restriction
   WHERE restriction_level = 'complete molecule' AND
        organism_ncbi_tax_id = 9606 AND
        class = 'I');
```

```
CREATE INDEX filtered_mhc_index ON filtered_mhc (mhc_allele_restriction_id);
```

```
SELECT COUNT(*) FROM filtered_mhc;
```

Table 11: 1 records

COUNT(*)
12133

Объедним вместе

```
CREATE TABLE epi_mhc_bind
AS (SELECT fb.curated_epitope_id, fb.mhc_allele_restriction_id, reference_id, as_char_value, as_location,
category, assay_type,units,as_num_value,as_inequality,as_comments,
e_name, source_antigen_accession, description ,e_region_domain_flag, e_ev, linear_peptide_seq,
e_ref_start,e_ref_end, `database`, name, sequence, organism_name,
restriction_level,displayed_restriction, organism_ncbi_tax_id,class, chain_i_name
   FROM filtered_bind fb
   INNER JOIN filtered_mhc fm ON fm.mhc_allele_restriction_id = fb.mhc_allele_restriction_id
   INNER JOIN filtered_epitope_cur fe ON fe.curated_epitope_id = fb.curated_epitope_id);
```

```
CREATE INDEX epi_mhc_bind_index ON epi_mhc_bind (curated_epitope_id,mhc_allele_restriction_id);
```

```
SELECT COUNT(*) FROM epi_mhc_bind;
```

Table 12: 1 records

COUNT(*)
168071

```
CREATE TABLE epi_mhc_elution
AS (SELECT reference_id, fel.curated_epitope_id, as_char_value, as_location,
category, assay_type, as_num_value, as_inequality, units,
as_num_subjects, as_num_responded, as_response_frequency, as_comments,
as_immunization_comments,h_sex, h_age, fel.mhc_allele_restriction_id,
h_organism_id, ant_type, ant_ref_name, ant_object_id, apc_cell_type,
apc_tissue_type, apc_origin,
e_name, source_antigen_accession, description ,e_region_domain_flag, e_ev, linear_peptide_seq, e_ref_start,e_ref_end,
restriction_level,displayed_restriction, organism_ncbi_tax_id,class, chain_i_name
   FROM filtered_elution fel
   INNER JOIN filtered_mhc fm ON fm.mhc_allele_restriction_id = fel.mhc_allele_restriction_id
   INNER JOIN filtered_epitope_cur fe ON fe.curated_epitope_id = fel.curated_epitope_id);
```

```
CREATE INDEX epi_mhc_elution_index ON epi_mhc_elution (curated_epitope_id,mhc_allele_restriction_id);
```

```
SELECT COUNT(*) FROM epi_mhc_elution;
```

Table 13: 1 records

COUNT(*)
1099078

Сохраним данные

```
bind.assays = RMySQL::dbReadTable(con, "epi_mhc_bind")
elution.assays = RMySQL::dbReadTable(con, "epi_mhc_elution")
write.table(bind.assays, "../data/source/iedb_bind_assays.tsv", sep = "\t",row.names = F,fileEncoding = "UTF-8")
write.table(elution.assays, "../data/source/iedb_elution_assays.tsv", sep = "\t",row.names = F,fileEncoding = "UTF-8")
```

Анализ bind assays

Rows: 168071 Columns: 28

-- Column specification --

Delimiter: "\t"

chr (21): as_char_value, as_location, category, assay_type, units, as_inequa...

dbl (7): curated_epitope_id, mhc_allele_restriction_id, reference_id, as_nu...

i Use `spec()` to retrieve the full column specification for this data.

i Specify the column types or set `show_col_types = FALSE` to quiet this message.

'data.frame':	168071 obs. of 28 variables:

 $ curated_epitope_id   : num 8188433 8188434 8188435 8188436 8188437 ...

 $ mhc_allele_restriction_id: num 143 143 143 143 143 143 143 143 251 251 ...

 $ reference_id   : num 1038825 1038825 1038825 1038825 1038825 ...

 $ as_char_value   : chr "Positive" "Positive" "Positive-Low" "Positive-Low" ...

 $ as_location   : chr "Table 3" "Table 3" "Table 3" "Table 3" ...

 $ category   : chr "MHC" "MHC" "MHC" "MHC" ...

 $ assay_type   : chr "cellular MHC/direct/fluorescence" "cellular MHC/direct/fluorescence" "cellular MHC/direct/fluores

 $ units   : chr NA NA NA NA ...

 $ as_num_value   : num NA NA NA NA NA NA NA NA 59.8 48.2 ...

 $ as_inequality   : chr NA NA NA NA ...

 $ as_comments   : chr NA NA NA NA ...

 $ e_name   : chr "Epitope 1" "Epitope 2" "Epitope 3" "Epitope 4" ...

 $ source_antigen_accession : chr "BAS53332.1" "BAS53332.1" "BAS53332.1" "BAS53332.1" ...

 $ description   : chr "RVTGGVFLV" "GLLGFAAPF" "HLPDRVHFA" "LLDDEAGPL" ...

 $ e_region_domain_flag   : chr "Exact Epitope" "Exact Epitope" "Exact Epitope" "Exact Epitope" ...

 $ e_ev   : chr NA NA NA NA ...

 $ linear_peptide_seq   : chr "RVTGGVFLV" "GLLGFAAPF" "HLPDRVHFA" "LLDDEAGPL" ...

 $ e_ref_start   : num NA NA NA NA NA NA NA NA NA NA ...

 $ e_ref_end   : num NA NA NA NA NA NA NA NA NA NA ...

```
$ database              : chr  "GenPept" "GenPept" "GenPept" "GenPept" ...
$ name                  : chr  "P protein [HBV genotype B]" "P protein [HBV genotype B]" "P protein [HBV genotype B]" "P protein
$ sequence              : chr  "MPLSYQHFRKLLLLDDEAGPLEEELPRLADEGLNRRVAEDLNLGNLNVSIPWTHKVGNFTGLYSS
$ organism_name         : chr  "Hepatitis B virus genotype B (HBV genotype B)" "Hepatitis B virus genotype B (HBV genotype
$ restriction_level     : chr  "complete molecule" "complete molecule" "complete molecule" "complete molecule" ...
$ displayed_restriction : chr  "HLA-A*02:01" "HLA-A*02:01" "HLA-A*02:01" "HLA-A*02:01" ...
$ organism_ncbi_tax_id  : num  9606 9606 9606 9606 9606 ...
$ class                 : chr  "I" "I" "I" "I" ...
$ chain_i_name          : chr  "HLA-A*02:01" "HLA-A*02:01" "HLA-A*02:01" "HLA-A*02:01" ...
```

Все ли записи имеют ссылку на источник?

```
[1] TRUE
```

Все ли эксперименты имеют запись о результате?

```
[1] TRUE
```

Распределение результатов экпериментов.

Var1	Freq
Negative	85974
Positive	10887
Positive-High	25621
Positive-Intermediate	20650
Positive-Low	24939

Сделаем два уровня значений результатов: положительный и отрицательный.

Value	Freq
Negative	85974
Positive	82097

Распределение по типам экспериментов в процентах.

Var1	Freq
MHC	100

Var1	Freq
binding assay	0.12
cellular MHC	0.09
cellular MHC/competitive/fluorescence	3.17
cellular MHC/competitive/radioactivity	0.02
cellular MHC/direct/fluorescence	2.83
cellular MHC/direct/radioactivity	0.01
cellular MHC/T cell inhibition	0.12
lysate MHC/direct/fluorescence	0.00
lysate MHC/direct/radioactivity	0.08
purified MHC	0.58
purified MHC/competitive/fluorescence	14.03
purified MHC/competitive/radioactivity	38.08
purified MHC/direct/fluorescence	36.64
purified MHC/direct/radioactivity	3.95
x-ray crystallography	0.28

Как много записей имеют замерянное значение?

[1] 86.75

Сколько из них имеют точное значение и каков исход?

[1] 39090

Var1	Freq
Negative	15.95
Positive	84.05