Предсказание связывания пептидов с TAP транспортером.

Антон Смирнов

Apr 12, 2023

Оглавление

Введение

TAP1/2 - важное звено в процессинге антигенов для MHC I. Он транспортирует пептиды из цитоплазмы в полость эндоплазматической сети. Экспериментальных данных по связыванию пептидов с этим транспортером крайне мало. Методика определения аффинности своеобразная. Определяют IC_{50} в эксперименте с микросомами относительно некоторого меченого ^{125}I пептида. Чаще всего в публикациях используется RRYNASTEL, по методике описанной в работе Ван Эндерта(Endert et al. 1994).

Материалы

Было найдено два датасета.

1. База данных MHC BN(Lata, Bhasin, and Raghava 2009)
2. Данные, использованные при обучении TAPREG(Diez-Rivero et al. 2010)
    - Два экспериментальных датасета(Daniel et al. 1998; Toseland et al. 2005)
    - Сгенерированные с помощью “Gibbs sampler with an exhaustive method and maximum blosum 62 relatedness scores of 25, 30, 35, and 37.”(Neuwald, Liu, and Lawrence 1995)

Фильтрация данных

```
library(dplyr)
library(vroom)
library(stringr)
library(readxl)
library(caret)
library(ggplot2)
#MHC BN
mhcbn = vroom("data/source/TAP/MHCBN_2023-02-16.tsv", delim = "\t") %>%
  select(-1) %>%
  filter(`MHC Allele` == "TAP") %>%
  filter(`Host Organism` == "HUMAN") %>%
  filter(!is.na(Comment)) %>%
  mutate(dup_check = paste(`Peptide Sequence`, Comment)) %>%
```

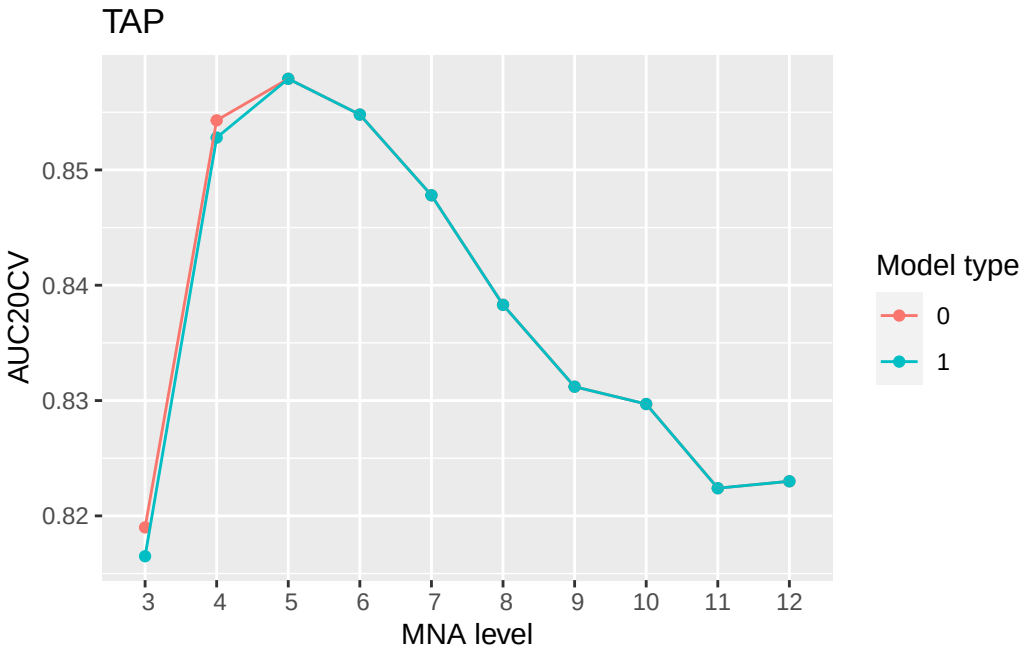
```
  filter(!duplicated(dup_check)) %>%
  filter(grepl("Relative", Comment)) %>%
  filter(!grepl("approx", Comment)) %>%
  filter(grepl("nM|uM",Comment,ignore.case = T)) %>%
  mutate(unit = str_extract(Comment,"\\(([^)]+)\\)",group = T),
         value = if_else(grepl("u",unit),
                         sapply(str_split(Comment,"="),function(x){as.double(x[2])}) * 1000,
                         sapply(str_split(Comment,"="),function(x){as.double(x[2])}))) %>%
  filter(!is.na(value)) %>%
  mutate(log_IC50_rel = log10(value)) %>%
  select(all_of(c("Peptide Sequence", "log_IC50_rel"))) %>%
  rename("PEPTIDE" = "Peptide Sequence")
#activity = if_else(pIC50 < (9 - log10(800)), 0, 1)
table(mhcbn$activity)
#TAPREG supplementary
# 1 - 10.4049/jimmunol.161.2.617 + 10.1186/1745-7580-1-4
# 2 - 5 613-peptide dataset using the purge utility of the Gibbs Sampler (10.1002/pro.556004082)
# with an exhaustive method and maximum blosum 62 relatedness scores of 25, 30, 35, and 37.
# 6 - 723 unique 9-mer CD8 T cell epitopes obtained from the IMMUNEEPITOPE and EPIMHC databa
# 7- tapreg parameters

#let tapreg log(IC50_relative) as is pIC50_relative in mhc bn
first = read_excel("data/source/TAP/TAPREG/prot_22535_sm_supptable1.xls") %>%
  rename("log_IC50_rel" = "log(IC50_relative)")

second = read_excel("data/source/TAP/TAPREG/prot_22535_sm_supptable2.xls") %>%
  rename("log_IC50_rel" = "log(IC50_relative)")

third = read_excel("data/source/TAP/TAPREG/prot_22535_sm_supptable3.xls") %>%
  rename("log_IC50_rel" = "log(IC50_relative)")
```

```
fourth = read_excel("data/source/TAP/TAPREG/prot_22535_sm_supptable4.xls") %>%
  rename("log_IC50_rel" = "log(IC50_relative)")

fifth = read_excel("data/source/TAP/TAPREG/prot_22535_sm_supptable5.xls") %>%
  rename("log_IC50_rel" = "log(IC50_relative)")

total = bind_rows(mhcbn, first,second,third,fourth,fifth) %>%
  group_by(PEPTIDE) %>%
  summarise(median_log10_IC50_rel = median(log_IC50_rel)) %>%
  mutate(activity = if_else(median_log10_IC50_rel <= (log10(800)), 0, 1))
```

Порог в 800 нМ был взять по данным из аналогичных сервисов: там прогнозируют результат по ранговой шкале, где 0 баллов для IC50>1000 нМ, 10 баллов IC50<0,03 нМ(Daniel et al., n.d.). Пептиды с баллами 0-2 считаются неактивными(Zhang et al. 2006). В итоге обучающая выборка состояла из 683 пептидов: 388 неактивных и 295 активных. Модель строилась с помощью программы PASS на MNA дескрипторах.

Результаты

Оптимальный уровень для моделирования - 5.



Была проведена 5-кратная кросс-валидация.

```
import pandas as pd
import os
from sklearn import metrics
from glob import glob
import numpy as np
folds = glob(os.path.join("/media/stotoshka/STotoshka/cross_val", "*.CSV"))
#print(folds)
```

```
union = pd.DataFrame()
#print("Parse results")
for f in folds:
    tbl = pd.read_csv(f, sep=";", header=4, decimal=",")
    union = pd.concat([union, tbl])
#print(f"Folds union {union.shape}")

union = union.drop(columns=["Substructure Descriptors", "New Descriptors", "Possible Activities at Pa
union = union.rename(columns={union.columns[0]: "activity"})
activities = union.columns[1:]
prediction = union.copy(deep = True)
result = pd.DataFrame(columns=["AUROC", "Average precision", "Precision", "Accuracy", "BA", "Recal
#print("a\troc_auc\tpr_auc\tprecision\taccuracy\tba\trecall\tf1\tsensitivity\tspecifity")
#print("a\ttp\ttn\tfp\tfn")
pred = np.where(prediction.loc[prediction["1"].notnull(), "1"] <= 0, 0, 1)
true = prediction.loc[prediction["1"].notnull(), "activity"]
try:
    roc_auc = round(metrics.roc_auc_score(true, pred),4)
    pr_auc = round(metrics.average_precision_score(true, pred),4)
    precision = round(metrics.precision_score(true,pred),4)
    accuracy = round(metrics.accuracy_score(true, pred),4)
    ba = round(metrics.balanced_accuracy_score(true, pred),4)
    recall = round(metrics.recall_score(true, pred),4)
    f1 = round(metrics.f1_score(true, pred),4)
    tn, fp, fn, tp = metrics.confusion_matrix(true, pred).ravel()
    sensitivity = round(tp / (tp + fn),4)
    specifity = round(tn / (tn + fp),4)
        #print(f"{a}\t{roc_auc}\t{pr_auc}\t{precision}\t{accuracy}\t{ba}\t{recall}\t{f1}\t{sensitivity}\t{specifity}"
        #print(f"{a}\t{tp}\t{tn}\t{fp}\t{fn}")
    result.loc[0] = [roc_auc, pr_auc,precision,accuracy,ba,recall,f1,sensitivity,specifity]
```

```
except ValueError as ve:
    print(f"{a} {ve}")
```

```
library(knitr)
library(reticulate)
kable(py$result)
```

	AUROC	Average precision	Precision	Accuracy	BA	Recall	F1	Sens	Spec
0	0.7689	0.6548	0.7294	0.7716	0.7689	0.7492	0.7391	0.7492	0.7887

Без жесткого порога AUROC и AUC-PR будут несколько выше.

```
pred = np.where(prediction.loc[prediction["1"].notnull(), "1"] <= 0, 0, prediction.loc[prediction["1"].notnull
true = prediction.loc[prediction["1"].notnull(), "activity"]
curve = metrics.roc_curve(true,pred)
#plt.plot(curve)
print(f"AUC ROC = {metrics.roc_auc_score(true,pred)}")
```

AUC ROC = 0.7916477372007689

```
print(f"AUC PRC = {metrics.average_precision_score(true,pred)}")
```

AUC PRC = 0.7194385123388266

Выводы

Модели получились удолетворительного качества. Необходимо сравнить с аналогами. Существует недостаток экспериментальных данных. В дальнейшем возможно необходимо задавать малый вес выходу модели для принятия решения

Список литературы

Daniel, Soizic, Vladimir Brusic, Sophie Caillat-Zucman, Leonard Harrison, Daniela Riganelli, Fabio Gallazzi, Jürgen Hammer, and Peter M van. n.d. “Relationship Between Peptide Selectivities of Human Transporters Associated with Antigen Processing and HLA Class I Molecules,” 9.

Daniel, Soizic, Vladimir Brusic, Sophie Caillat-Zucman, Nicolai Petrovsky, Leonard Harrison, Daniela Riganelli, Francesco Sinigaglia, Fabio Gallazzi, Jürgen Hammer, and Peter M. van Endert. 1998. “Relationship Between Peptide Selectivities of Human Transporters Associated with Antigen Processing and HLA Class I Molecules.” *The Journal of Immunology* 161 (2): 617–24. https://doi.org/10.4049/jimmunol.161.2.617.

Diez-Rivero, Carmen M., Bernardo Chenlo, Pilar Zuluaga, and Pedro A. Reche. 2010. “Quantitative Modeling of Peptide Binding to TAP Using Support Vector Machine.” *Proteins: Structure, Function, and Bioinformatics* 78 (1): 63–72. https://doi.org/10.1002/prot.22535.

Endert, Peter M. van, Robert Tampé, Thomas H. Meyer, Roland Tisch, Jean-François Bach, and Hugh O. McDevitt. 1994. “A Sequential Model for Peptide Binding and Transport by the Transporters Associated with Antigen Processing.” *Immunity* 1 (6): 491–500. https://doi.org/10.1016/1074-7613(94)90091-4.

Lata, Sneh, Manoj Bhasin, and Gajendra PS Raghava. 2009. “MHCBN 4.0: A Database of MHC/TAP Binding Peptides and T-Cell Epitopes.” *BMC Research Notes* 2 (1): 61. https://doi.org/10.1186/1756-0500-2-61.

Neuwald, Andrew F., Jun S. Liu, and Charles E. Lawrence. 1995. “Gibbs Motif Sampling: Detection of Bacterial Outer Membrane Protein Repeats.” *Protein Science* 4 (8): 1618–32. https://doi.org/10.1002/pro.5560040820.

Toseland, Christopher P, Debra J Clayton, Helen McSparron, Shelley L Hemsley, Martin J Blythe,

Kelly Paine, Irini A Doytchinova, Pingping Guan, Channa K Hattotuwagama, and Darren R Flower. 2005. *Immunome Research* 1 (1): 4. https://doi.org/10.1186/1745-7580-1-4.

Zhang, Guang, Nikolai Petrovsky, Chee Kwoh, J Thomas August, and Vladimir Brusic. 2006. "PREDTAP: A System for Prediction of Peptide Binding to the Human Transporter Associated with Antigen Processing." *Immunome Research* 2 (1): 3. https://doi.org/10.1186/1745-7580-2-3.