

TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

_____ * _____



Bài tập lớn Tin Sinh Học

Đề tài: Dự đoán các peptides chống tạo mạch

Sinh viên thực hiện:

Tổng Mạnh Đạt - 20173008

Giáo viên hướng dẫn: TS. Nguyễn Hồng
Quang

Dự đoán các peptides chống tạo mạch

Nhóm Sinh viên thực hiện :

Tổng Mạnh Đạt, MSSV : 20173008

Lớp : KTMT 06, K62

Đặng Quang Anh - 20172942

Lớp : KTMT07-K62

Giáo viên hướng dẫn:

TS. Nguyễn Hồng Quang,

Giảng viên Bộ môn Kỹ thuật máy tính,

Trưởng phòng thí nghiệm Tin học y sinh, Trung tâm nghiên cứu

Quốc tế về Trí tuệ Nhân tạo (BK.AI),

Viện Công nghệ thông tin và Truyền thông,

Trường Đại học Bách Khoa Hà Nội

Tóm tắt :

Tổng quan :

Peptit được coi là một liệu pháp điều trị quana trọng đang được thử nghiệm đối với các bệnh phụ thuộc vào quá trình tạo mạch bởi độc tính thấp cùng hiệu quả cao. Các peptides chống tạo mạch nhiều triển vọng trong các nghiên cứu tiền lâm sàng và lâm sàng đối với bệnh ung thư . Do đó, việc dự đoán được peptide chống tạo mạch là những ứng cử viên đầy hứa hẹn trong điều trị ung thư. Trong nghiên cứu này, đã thu thập các peptit chống tạo mạch từ các tài liệu và phân tích mức độ ưa thích dư lượng trong các peptit này. Các chất dư như Cys, Pro, Ser, Arg, Trp, Thr và Gly được ưu tiên trong khi Ala, Asp, Ile, Leu, Val và Phe không được ưu tiên trong các peptit này. Có sự ưu tiên về vị trí của Ser, Pro, Trp và Cys ở vùng tận cùng N và Cys, Gly và Arg ở vùng tận cùng C của các peptit chống tạo mạch.

phương pháp đề xuất : sử dụng phân loại rừng ngẫu nhiên (random forest) kết hợp với đặc trưng AAC của peptides .

Kết quả đạt được : dự đoán các peptit chống tạo mạch với độ chính xác là 86% trên một tập dữ liệu chuẩn khách quan. Ngoài ra, kết quả cho thấy các đặc điểm quan trọng sau đây của peptit chống tạo mạch: (i) liên kết disulfua hình thành cặp Cys đóng vai trò quan trọng trong việc ức chế tăng sinh mạch máu; (ii) Các nang nằm ở vùng tận cùng C có thể làm giảm hoạt động định dạng nội mô và ngăn chặn sự phát triển của khối u; và (iii) Các peptit giàu disulfua theo chu kỳ góp phần ức chế hình thành mạch và di chuyển tế bào, chọn lọc và ổn định.

Từ khóa : anti-angiogenic peptides, bioinformatics, randomforest,machine learning, classification, Pfeature, bioinformatics

Link dataset, code,báo cáo :

https://github.com/tmd22121999/Antimicrobial_Peptide

<https://drive.google.com/drive/folders/17ANGkcSxGVkQpTL8CyK9Ug5M1rJm4Ptf?usp=sharing>

1. Giới thiệu

Tổng quan về bài toán:

Dù công nghệ phát triển rất nhanh, nhưng Ung thư vẫn là một trong các bệnh nan y khó chữa và gây tử vong nhất nếu không được phát hiện ra sớm. Qua nhiều nghiên cứu, người ta thấy rằng hình thành mạch rất quan trọng đối với cơ chế bệnh sinh của các bệnh khác nhau ở người, đặc biệt là các khối u rắn. Việc dự đoán được các peptide chống tạo mạch giúp con người có bước tiến triển vọng trong việc điều trị ung thư. Do đó, dự đoán chính xác các peptit chống tạo mạch là cực kỳ quan trọng để hiểu được các đặc tính lý sinh và sinh hóa của chúng, làm cơ sở cho việc phát hiện ra các loại thuốc chống ung thư mới. Nghiên cứu này nhằm mục đích tìm hiểu mô hình tính toán hiệu quả để dự đoán và xác định đặc điểm của các peptit chống tạo mạch.

Nghiên cứu được phát triển bằng cách sử dụng công cụ phân loại rừng ngẫu nhiên kết hợp với đặc trưng AAC của peptides.

Các nghiên cứu đã có và định hướng nghiên cứu.

Phương pháp	Mô hình phân lớp	Đặc trưng của Sequence	Independent Test	Web Server
AntiAngioPred	SVM	AAC (20)	Có	Có
Blanco et al.'s method	glmnet	AAC, DPC, TC (200)	Không	Không
AntAngioCOOL	PART	PseAAC, k-mer composition, RAAC, PCP, AC (2,343)	Không	Không
TargetAntiAngio	RF	AAC, PseAAC, Am-PseAAC (48)	Có	Có
Nghiên cứu này	RF, Ada boost, extratrees	AAC, DPC	Không	Không

Table 1 : Các nghiên cứu về dự đoán chống tạo mạch đã có

Kết quả đạt được :

Xây dựng mô hình: Phần xây dựng mô hình sẽ sử dụng các bộ trình tự peptides từ dữ liệu có sẵn (dưới dạng các biến X) để xây dựng mô hình hồi quy dự đoán các peptides chống tạo mạch (biến Y).

Tìm hiểu, phân tích các đặc trưng của peptides quan trọng trong việc dự đoán peptides chống tạo mạch hay không

So sánh các mô hình: so sánh một số mô hình hồi quy dự đoán peptides chống tạo mạch bằng cách sử dụng thư viện lazyp dự đoán trong python.

Định hướng nghiên cứu so sánh các mô hình, thực nghiệm và tìm mô hình tốt nhất, tìm hiểu các peptide chống tạo mạch.

2. Mô tả bài toán

2.1. Chi tiết bài toán :

Bài toán:

Đầu vào : cho mỗi chuỗi peptides ngắn. ví dụ :
ADNWQSFDRWKDH.

Đầu ra : dự đoán có phải có chức năng chống tạo mạch
(anti-angiogenic peptides: Antiangio) hay không
(non-angiogenic peptides: Negative).

Định dạng dữ liệu:

>AA135

YTMNPRKLFDY

>neg1

ADNWQSFDRWKDH

Với “AA135”, “neg1” là tên các peptide và “YTMNPRKLFDY”,
“ADNWQSFDRWKDH” là trình tự của peptide tương ứng.

2.2. Tập dữ liệu sử dụng:

Tập dữ liệu gồm 2 file fasta được lấy từ nghiên cứu
TargetAntiAngio - Dự đoán và phân tích peptides chống tạo
mạch.

benchmarkdataset.fasta làm tập Train : 137 peptide
sequences Thuộc lớp peptide chống tạo mạch và 137 peptides
ngẫu nhiên sử dụng làm peptides không chống tạo mạch.

NT15dataset.fasta làm tập Test: it chứa 99 peptides chống
tạo mạch và 101 Không chống tạo mạch.

Dataset	Benchmarkdataset - tập train		NT15dataset - tập test	
	Anti-anig o	Non-anti-anig o	Anti-anig o	Non-anti-anig o
Tổng dữ liệu	137	137	99	101

Table 2 : Tập dữ liệu - 1

Cách chia tập dữ liệu khác : Gộp 2 tập benchmarkdataset và NT15 sau đó chia ngẫu nhiên thành 2 tập dữ liệu train và test với tỷ lệ (8 -2)

Table 3 : Tập dữ liệu - 2

Dataset	Tập train		Tập test	
	Anti-anig o	Non-anti-anig o	Anti-anig o	Non-anti-anigo
Tổng d ữ liệu	189	189	47	49

2.3. Các độ đo đánh giá:

$$Ac = \frac{TP + TN}{(TP + TN + FP + FN)}$$

$$Sn = \frac{TP}{(TP + FN)}$$

$$Sp = \frac{TN}{(TN + FP)}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Với Ac - Accuracy là độ chính xác cho các dự đoán của mô hình.

Sn hay Sensitivity là độ nhạy. Sensitivity cao đồng nghĩa với việc True Positive Rate cao, tức tỉ lệ bỏ sót các điểm thực sự positive là thấp.

Sp - Specificity là độ đặc hiệu : tỉ lệ true Negative trên tổng số các trường hợp Negative thực sự. Specificity cao đồng nghĩa với việc True Negative Rate cao, tức tỉ lệ bỏ sót các điểm thực sự Negative là thấp.

Hệ số tương quan Matthews (MCC)

MCC được giới thiệu bởi Brian W. Matthews vào năm 1975, dùng để đánh giá phẩm chất của mô hình phân loại nhị phân. Mục tiêu của MCC là khắc phục vấn đề dữ liệu bị mất cân bằng.

3. Đề xuất mô hình :

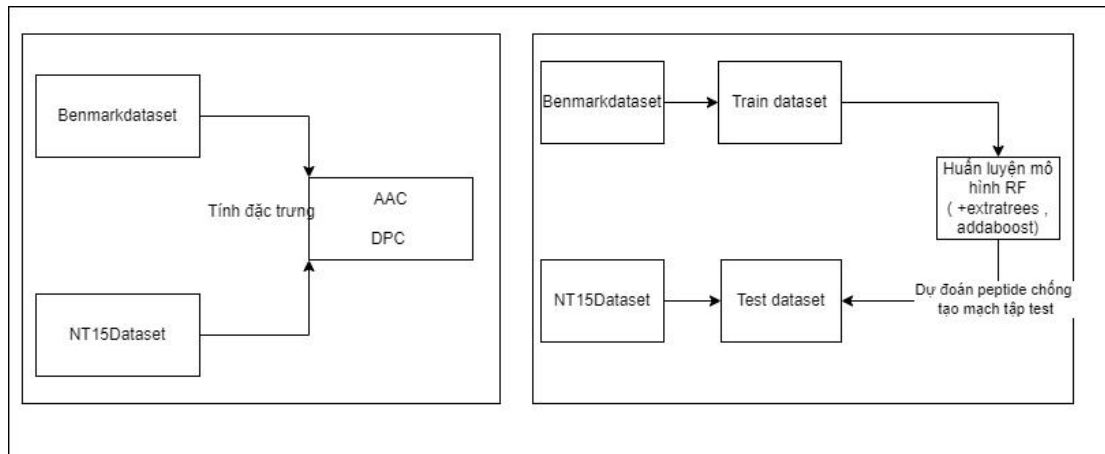


Image 1 : Mô hình đề xuất

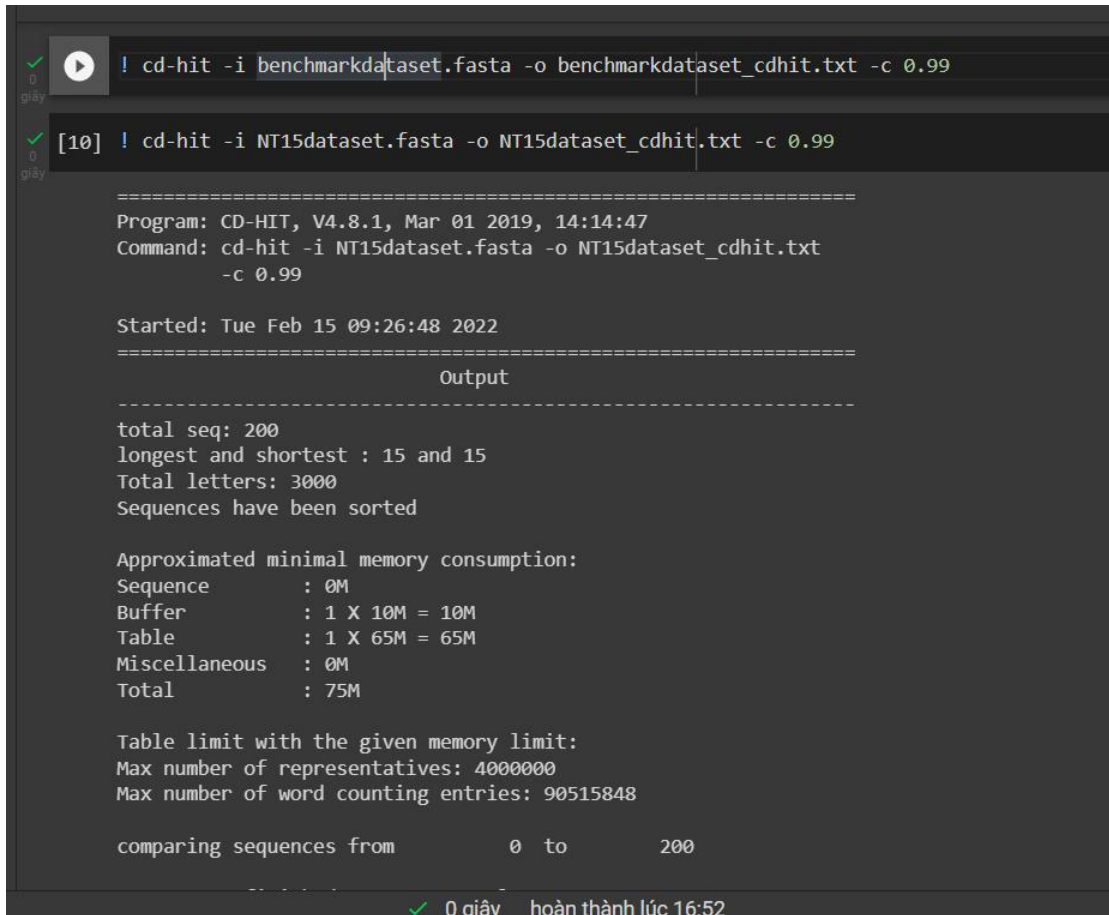
Với 2 tập dữ liệu benchmark và NT15, tính đặc trưng cho trình tự peptides bằng các đặc trưng AAC và DPC. Sử dụng tập benchmark làm tập train, huấn luyện mô hình random forest (huấn luyện thêm mô hình extratrees và adaboost để dự đoán peptides chống tạo mạch trên tập NT15 là tập test. Tính toán có độ đo để nhận xét, so sánh các mô hình). với benchmarkdataset.fasta có 137 peptide sequences Thuộc lớp peptide chống tạo mạch và 137 peptides ngẫu nhiên sử dụng làm peptides không chống tạo mạch. NT15dataset.fasta chứa 99 peptides chống tạo mạch và 101 Không chống tạo mạch

4. Thực hiện hệ thống

4.1. Tiền xử lý dữ liệu

Ban đầu tập benchmark dataset chứa 257 trình tự peptides thuộc lớp peptides chống tạo mạch. Tập dữ liệu được tiền xử lý bằng cách sử dụng CD-HIT lọc các peptide có trình tự giống

nhau trên 90%. Sau đó là các peptide chứa ký tự đặc biệt như U, X.

A terminal window showing the execution of the CD-HIT command. The command is: `cd-hit -i benchmarkdataset.fasta -o benchmarkdataset_cdhit.txt -c 0.99`. The output shows the program version (V4.8.1), the command used, the start time (Tue Feb 15 09:26:48 2022), and the output statistics. The output statistics include: total seq: 200, longest and shortest : 15 and 15, Total letters: 3000, Sequences have been sorted. The approximated minimal memory consumption is shown as: Sequence : 0M, Buffer : 1 X 10M = 10M, Table : 1 X 65M = 65M, Miscellaneous : 0M, Total : 75M. The table limit with the given memory limit is: Max number of representatives: 4000000, Max number of word counting entries: 90515848. The output also shows: comparing sequences from 0 to 200. The terminal status at the bottom indicates: 0 giây hoàn thành lúc 16:52.

```
! cd-hit -i benchmarkdataset.fasta -o benchmarkdataset_cdhit.txt -c 0.99

[10] ! cd-hit -i NT15dataset.fasta -o NT15dataset_cdhit.txt -c 0.99

=====
Program: CD-HIT, V4.8.1, Mar 01 2019, 14:14:47
Command: cd-hit -i NT15dataset.fasta -o NT15dataset_cdhit.txt
        -c 0.99

Started: Tue Feb 15 09:26:48 2022
=====
                        Output
-----

total seq: 200
longest and shortest : 15 and 15
Total letters: 3000
Sequences have been sorted

Approximated minimal memory consumption:
Sequence      : 0M
Buffer        : 1 X 10M = 10M
Table         : 1 X 65M = 65M
Miscellaneous  : 0M
Total         : 75M

Table limit with the given memory limit:
Max number of representatives: 4000000
Max number of word counting entries: 90515848

comparing sequences from      0 to      200

=====
0 giây hoàn thành lúc 16:52
```

Image 2 : Tiền xử lý dữ liệu

4.2. Xử lý dữ liệu đầu vào

Định dạng dữ liệu ban đầu:

>AA135

YTMNPRKLFDY

>neg1

ADNWQSFDRWKDH

Với “AA135”, “neg1” là tên các peptide và “YTMNPRKLFDY”,

“ADNWQSFDRWKDH” là trình tự của peptide tương ứng.

Trước khi đưa dữ liệu vào huấn luyện mô hình ta cần mã hóa dữ liệu này thành các đặc trưng của trình tự 1 peptides:

Đặc trưng AAC - Amino Acid : là đặc trưng đơn giản nhất. AAC có 20 đặc trưng ứng với mỗi aminoacid mỗi đặc trưng được tính toán bằng cách Lấy tổng số lượng aminoacid loại tương ứng chia cho tổng số lượng aminoacid toàn mạch :

$$f(a) = N_a / N, \quad a \in (A, C, \dots, W, Y)$$

Với $f(a)$ là đặc trưng AAC của aminoacid loại a trong trình tự

N_a = tổng số lượng aminoacid loại a

N = tổng số lượng aminoacid trên toàn bộ trình tự .

Đặc trưng DPC - Depeptit : Tương tự với AAC tuy nhiên mỗi đặc trưng không phải 1 aminoacid mà là 1 depeptit (2 aminoacid liền nhau có thứ tự) tức là DPC gồm $20 \times 20 = 400$ đặc trưng

$$f(a,b) = N_{ab} / (N-1), \quad a,b \in (A, C, \dots, W, Y)$$

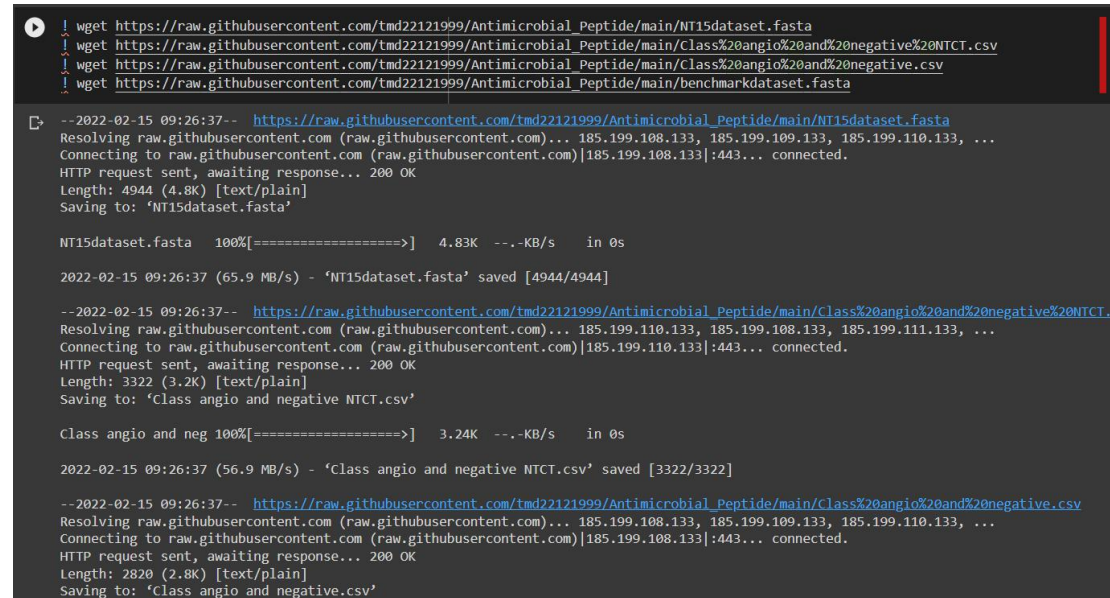
$f(a,b)$ là đặc trưng AAC của depeptit loại ab trong trình tự

N_{ab} = tổng số lượng depeptit loại ab

N = tổng số lượng aminoacid trên toàn bộ trình tự .

4.3. Xây dựng mô hình

Download các file dữ liệu :



```
! wget https://raw.githubusercontent.com/tmd22121999/Antimicrobial_Peptide/main/NT15dataset.fasta
! wget https://raw.githubusercontent.com/tmd22121999/Antimicrobial_Peptide/main/Class%20angio%20and%20negative%20NTCT.csv
! wget https://raw.githubusercontent.com/tmd22121999/Antimicrobial_Peptide/main/Class%20angio%20and%20negative%20NTCT.csv
! wget https://raw.githubusercontent.com/tmd22121999/Antimicrobial_Peptide/main/benchmarkdataset.fasta

--2022-02-15 09:26:37-- https://raw.githubusercontent.com/tmd22121999/Antimicrobial_Peptide/main/NT15dataset.fasta
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.108.133, 185.199.109.133, 185.199.110.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.108.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 4944 (4.8K) [text/plain]
Saving to: 'NT15dataset.fasta'

NT15dataset.fasta 100%[=====>] 4.83K --.-KB/s in 0s

2022-02-15 09:26:37 (65.9 MB/s) - 'NT15dataset.fasta' saved [4944/4944]

--2022-02-15 09:26:37-- https://raw.githubusercontent.com/tmd22121999/Antimicrobial_Peptide/main/Class%20angio%20and%20negative%20NTCT.csv
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.108.133, 185.199.109.133, 185.199.110.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.108.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 3322 (3.2K) [text/plain]
Saving to: 'Class angio and negative NTCT.csv'

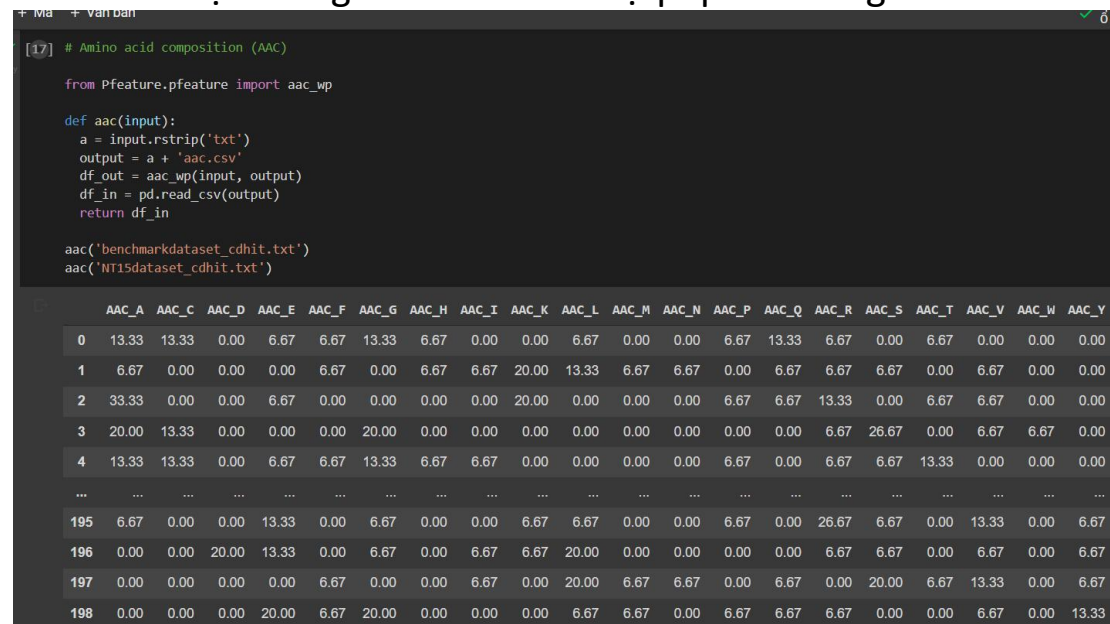
Class angio and neg 100%[=====>] 3.24K --.-KB/s in 0s

2022-02-15 09:26:37 (56.9 MB/s) - 'Class angio and negative NTCT.csv' saved [3322/3322]

--2022-02-15 09:26:37-- https://raw.githubusercontent.com/tmd22121999/Antimicrobial_Peptide/main/Class%20angio%20and%20negative%20NTCT.csv
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.108.133, 185.199.109.133, 185.199.110.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.108.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 2820 (2.8K) [text/plain]
Saving to: 'Class angio and negative NTCT.csv'
```

Image 3 : Tải về các file dữ liệu

Hàm tính đặc trưng AAC của trình tự peptides và ghi ra file csv :



```
[17] # Amino acid composition (AAC)

from Pfeature.pfeature import aac_wp

def aac(input):
    a = input.rstrip('txt')
    output = a + 'aac.csv'
    df_out = aac_wp(input, output)
    df_in = pd.read_csv(output)
    return df_in

aac('benchmarkdataset_cdhit.txt')
aac('NT15dataset_cdhit.txt')
```

	AAC_A	AAC_C	AAC_D	AAC_E	AAC_F	AAC_G	AAC_H	AAC_I	AAC_K	AAC_L	AAC_M	AAC_N	AAC_P	AAC_Q	AAC_R	AAC_S	AAC_T	AAC_V	AAC_W	AAC_Y
0	13.33	13.33	0.00	6.67	6.67	13.33	6.67	0.00	0.00	6.67	0.00	0.00	6.67	13.33	6.67	0.00	6.67	0.00	0.00	0.00
1	6.67	0.00	0.00	0.00	6.67	0.00	6.67	6.67	20.00	13.33	6.67	6.67	0.00	6.67	6.67	6.67	0.00	6.67	0.00	0.00
2	33.33	0.00	0.00	6.67	0.00	0.00	0.00	0.00	20.00	0.00	0.00	0.00	6.67	6.67	13.33	0.00	6.67	6.67	0.00	0.00
3	20.00	13.33	0.00	0.00	0.00	20.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	6.67	26.67	0.00	6.67	6.67	0.00
4	13.33	13.33	0.00	6.67	13.33	6.67	6.67	6.67	0.00	0.00	0.00	0.00	6.67	0.00	6.67	6.67	13.33	0.00	0.00	0.00
...
195	6.67	0.00	0.00	13.33	0.00	6.67	0.00	0.00	6.67	6.67	0.00	0.00	6.67	0.00	26.67	6.67	0.00	13.33	0.00	6.67
196	0.00	0.00	20.00	13.33	0.00	6.67	0.00	6.67	6.67	20.00	0.00	0.00	0.00	6.67	6.67	0.00	6.67	0.00	6.67	0.00
197	0.00	0.00	0.00	0.00	6.67	0.00	0.00	6.67	0.00	20.00	6.67	6.67	0.00	6.67	0.00	20.00	6.67	13.33	0.00	6.67
198	0.00	0.00	0.00	20.00	6.67	20.00	0.00	0.00	0.00	6.67	6.67	0.00	6.67	6.67	6.67	0.00	0.00	6.67	0.00	13.33

Image 4 Hàm tính đặc trưng AAC của trình tự peptides và ghi ra file csv

Loại bỏ đặc trưng có phương sai thấp:

```

# Feature selection (Variance threshold)
from sklearn.feature_selection import VarianceThreshold

fs = VarianceThreshold(threshold=(.8 * (1 - .8)))
fs.fit_transform(X)
#X2.shape
X2 = X.loc[:, fs.get_support()]
X2

```

Image 5 : Loại bỏ đặc trưng có phương sai thấp

Map dữ liệu đầu ra với 0 tượng trưng cho nhãn lớp peptides không chống tạo mạch còn 1 là peptide chống tạo mạch :

```

# Encoding the Y class label
y = y.map({"Antiangio": 1, "Negative": 0})
y

```

0	1
1	1
2	1
3	1
4	1
..	
195	0
196	0
197	0
198	0
199	0

Name: Class, Length: 468, dtype: int64

Image 6 : map dữ liệu đầu ra

Huấn luyện mô hình và kiểm tra kết quả dự đoán trên tập test :

```
✓ [104] # Build random forest model
0 giây
from sklearn.ensemble import RandomForestClassifier

rf = RandomForestClassifier(n_estimators=500)

rf.fit(X_train, y_train)

↳ RandomForestClassifier(n_estimators=500)

✓ [30] X_train.shape
0 giây
(268, 20)

▼ Apply the model to make predictions

✓ [105] y_train_pred = rf.predict(X_train)
y_test_pred = rf.predict(X_test)
```

Image 7 : Huấn luyện mô hình

4.4. So sánh các mô hình

Chúng ta sẽ sử dụng thư viện lazypredict so sánh nhanh một số thuật toán ML xây dựng mô hình dự đoán peptide chống tạo mạch

Cài đặt thư viện :



Image 8 : Cài đặt thư viện lazypredict

Sau đó ta xử lý đầu vào và huấn luyện nhanh 29 mô hình với tập train và tập test được lấy ra từ tập dữ liệu kết hợp của benchmark và NT15 với tỷ lệ 8-2 :

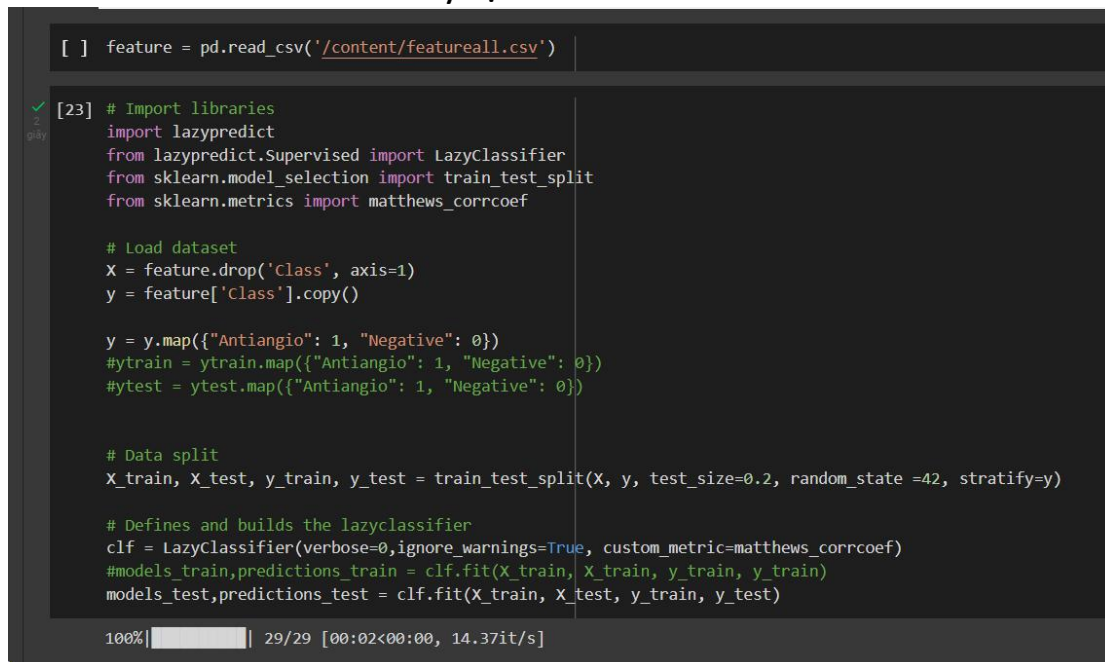


Image 9 :xử lý đầu vào, chia tập dữ liệu và huấn luyện nhanh 29 mô hình

Sau đây là top những mô hình có độ chính xác cao nhất khi dự đoán peptide chống tạo mạch trong 29 mô hình trên :

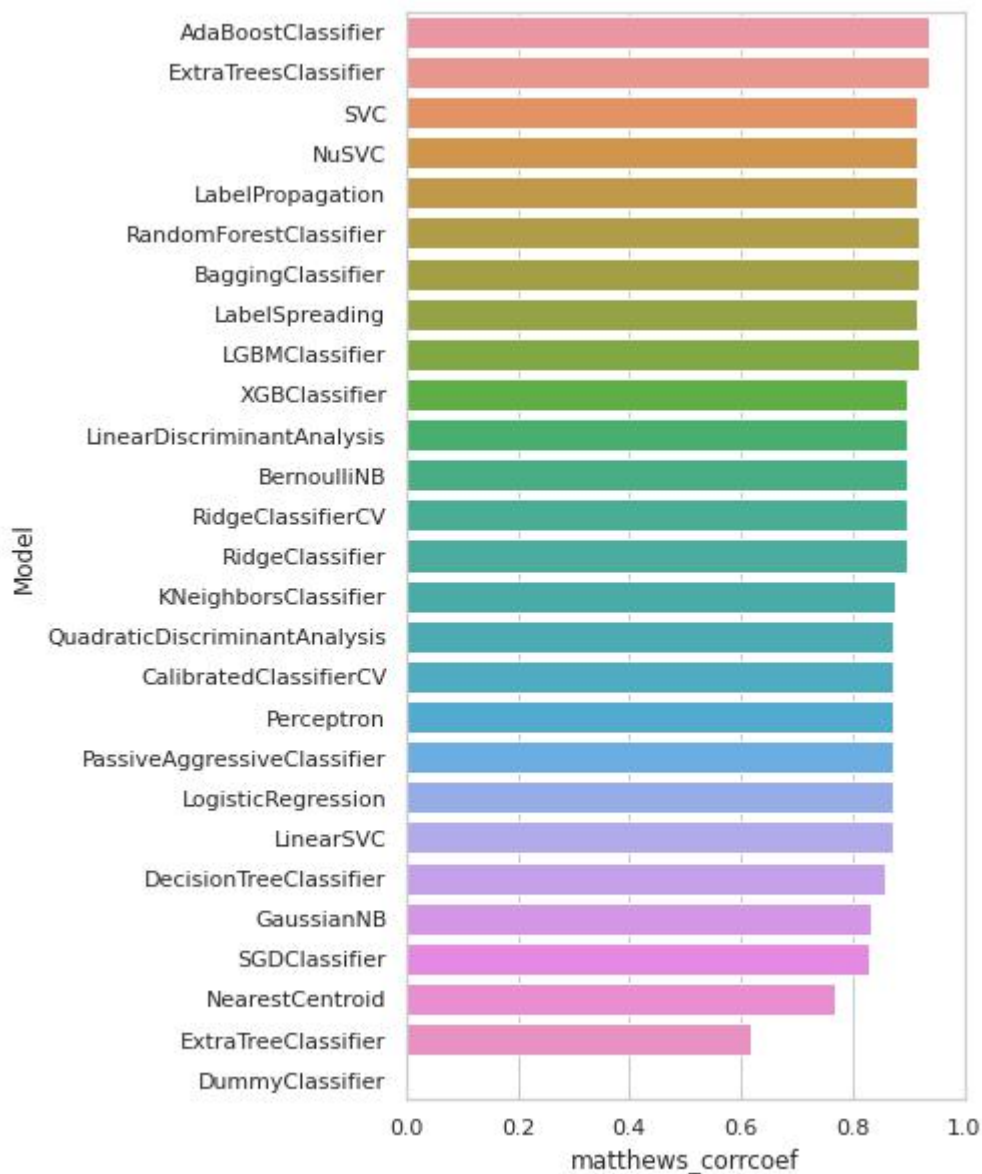


Image 10 :top những mô hình có độ chính xác cao nhất -1

Từ biểu đồ trên ta thấy một số giải thuật như phân lớp adaboost, random forest, extratrees có độ chính xác cao.

Tuy nhiên khi sử dụng tập benchmark làm tập train và NT15 làm tập test thì giải thuật phân lớp adaboost cũng như random forest giảm độ chính xác đáng kể, còn phân lớp extratrees vẫn khá cao:

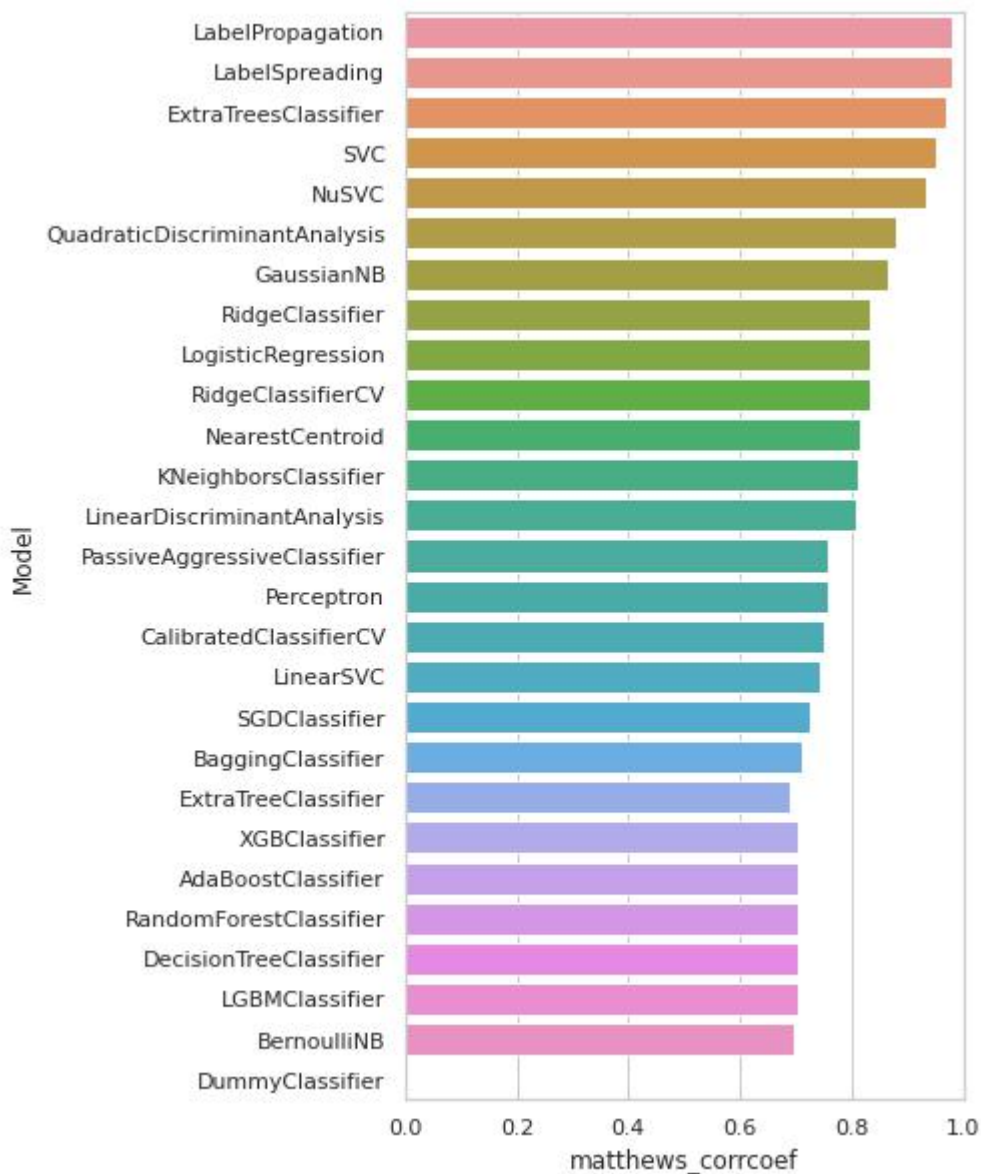


Image 11 : top những mô hình có độ chính xác cao nhất - 2

So sánh các độ đo đánh giá của 3 mô hình sử dụng các thuật toán phân lớp adaboost, random forest, extratrees với tập dữ liệu benchmark làm tập train và NT15 làm tập test:

Tính toán các độ đo đánh giá của 3 mô hình sử dụng các thuật toán phân lớp adaboost, random forest, extratrees

```
[144] TP = result3[0][0]
      TN = result3[1][1]
      FP = result3[0][1]
      FN = result3[1][0]
      AC.append((TP+TN)/(TP+TN+FP+FN))
      Sn.append(TP/(TP+FN))
      Sp.append(TN/(TN+FP))
      MCC.append((TP*TN-FP*FN)/(math.sqrt((TP+FP)*(TP+FN)*(TN+FP)*(TN+FN))))

[145] comparedata = {"method" : ["RandomForestClassifier", "ExtraTreesClassifier", "AdaBoostClassifier"], "AC":AC, "Sn":Sn, "Sp":Sp, "MCC":MCC}

[146] Comparedf = pd.DataFrame(comparedata)

# Print the output.
print(Comparedf)
```

Image 12 : Tính toán các độ đo đánh giá

Ta được bảng sau :

	method	AC	Sn	Sp	MCC
0	RandomForestClassifier	0.860	0.854369	0.865979	0.720060
1	ExtraTreesClassifier	0.925	0.898148	0.956522	0.851973
2	AdaBoostClassifier	0.775	0.780000	0.770000	0.550028

Image 13 : So sánh các độ đo đánh giá của 3 mô hình

Từ bảng này ta thấy kết quả từ phân lớp extratrees vượt trội hơn 2 phương pháp còn lại và với phân lớp extratrees tỷ lệ tìm ra đúng peptide không chống tạo mạch chính xác cao hơn đáng kể so với tỷ lệ tìm ra đúng peptide chống tạo mạch chính xác. Trong khi 2 phương pháp còn lại 2 tỷ lệ này khá cân bằng.

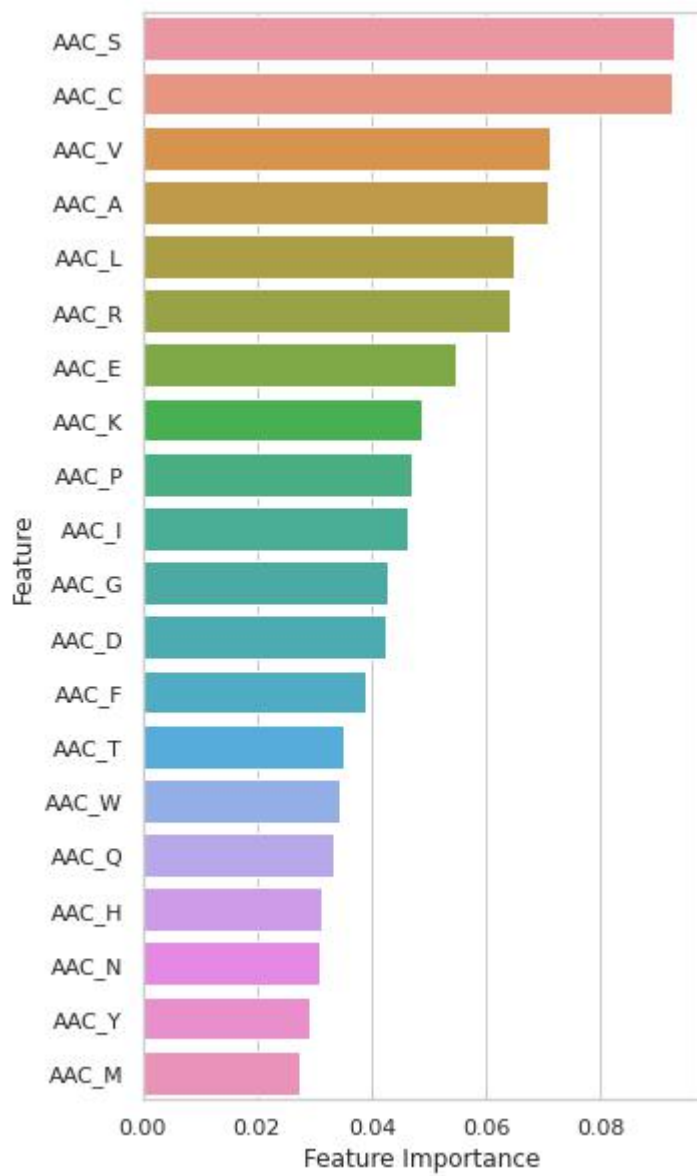


Image 14 : top những đặc trưng AAC quan trọng

5. Thử nghiệm và kết quả

Môi trường thử nghiệm :

-model name : 2-core Intel(R) Xeon(R) CPU @ 2.20GHz

-Ram : 12Gb

-Ổ cứng : 107Gb

-GPU 0: Tesla K80 (UUID:

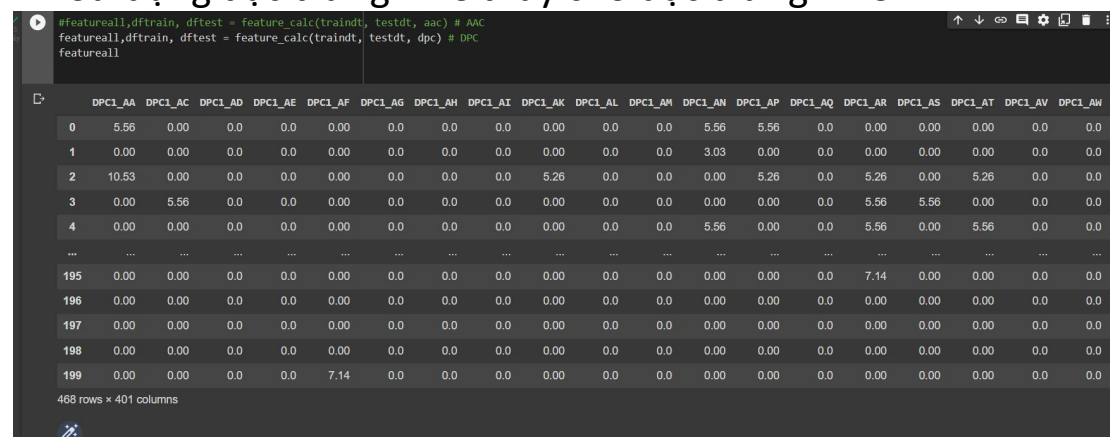
GPU-184900f1-5c99-060a-fd45-fbbe5fd940ec)

Các phần mềm, bộ công cụ toolkits sử dụng:

-Google colab.

Các thử nghiệm khác:

Sử dụng đặc trưng DPC thay cho đặc trưng AAC



The screenshot shows a Jupyter Notebook interface. At the top, there is a code cell with the following Python code:

```
#featureall,dfttrain, dfttest = feature_calc(traindt, testdt, aac) # AAC
featureall,dfttrain, dfttest = feature_calc(traindt, testdt, dpc) # DPC
featureall
```

Below the code cell is a large data table with 401 columns and 468 rows. The columns are labeled with DPC1_ followed by a two-letter code (e.g., DPC1_AA, DPC1_AC, ..., DPC1_ZZ). The rows contain numerical values, mostly 0.00, with some non-zero values like 5.56, 10.53, 3.03, 5.26, 7.14, and 7.14. The table is truncated on the right side, indicated by an ellipsis.

Image 15 : Sử dụng đặc trưng DPC

Kết quả khi sử dụng tập train và tập test được lấy ra từ tập dữ liệu kết hợp của benchmark và NT15 với tỷ lệ 8-2 (lấy ngẫu nhiên) :

	method	AC	Sn	Sp	MCC
0	RandomForestClassifier	0.755319	0.772727	0.740000	0.511682
1	ExtraTreesClassifier	0.840426	0.847826	0.833333	0.681005
2	AdaBoostClassifier	0.734043	0.761905	0.711538	0.470757

Image 16 : So sánh độ đo đánh giá 3 mô hình khi sử dụng đặc trưng DPC -1

So với khi train mô hình khi sử dụng đặc trưng AAC với tập dữ liệu lấy ngẫu nhiên theo cách trên :

	method	AC	Sn	Sp	MCC
0	RandomForestClassifier	0.797872	0.804348	0.791667	0.595880
1	ExtraTreesClassifier	0.872340	0.843137	0.906977	0.747392
2	AdaBoostClassifier	0.691489	0.687500	0.695652	0.383065

Image 17 : So sánh độ đo đánh giá 3 mô hình khi sử dụng đặc trưng DPC -2

Top những đặc trưng quan trọng DPC trong việc dự đoán peptides chống tạo mạch

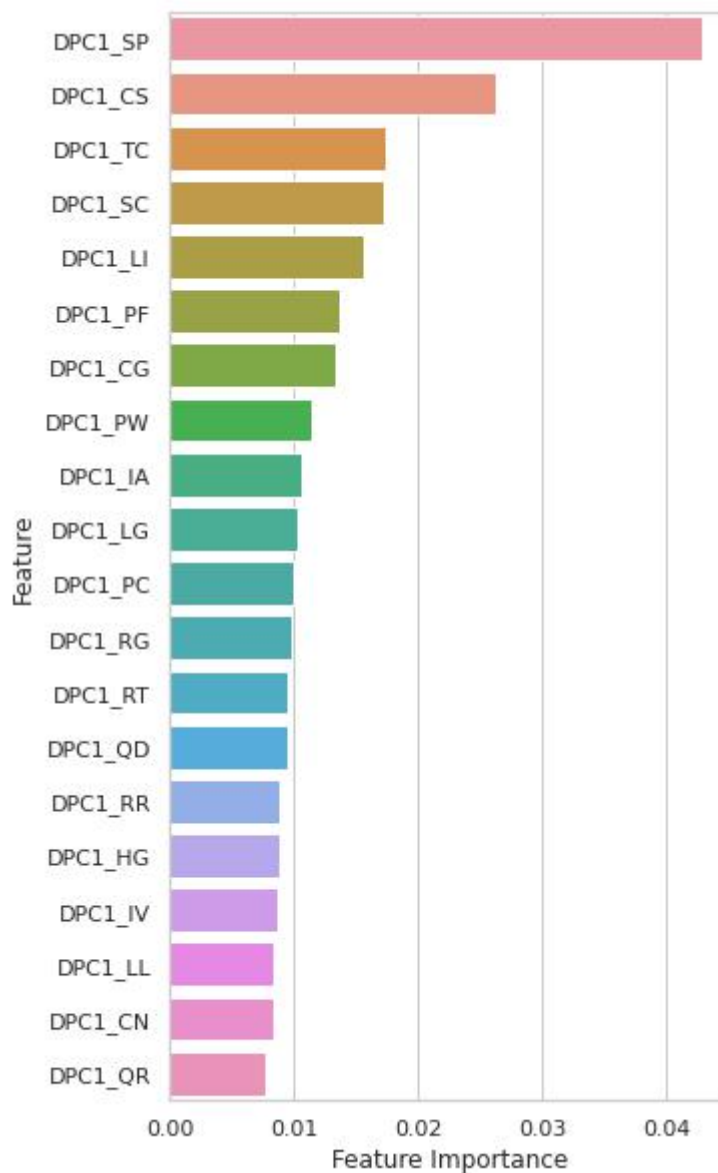


Image 18 : Top những đặc trưng DPC quan trọng

6. Thảo luận

Với tập dữ liệu benchmark và NT15, Qua kết quả thu được ta thấy khi sử dụng thuật toán phân lớp extratrees xây dựng mô hình dự đoán peptides chống tạo mạch có độ chính xác tốt hơn so với 2 phương pháp randomtrees và adaboost cũng như các độ đo khác. Ngoài ra việc sử dụng đặc trưng AAC tuy đơn giản nhưng lại rất hiệu quả để xây dựng mô hình dự đoán. Trong đó đặc trưng AAC_S có tầm quan trọng nhất trong việc dự đoán peptide chống tạo mạch. Với đặc trưng DPC ta thấy aminoacid S đứng liền trước P, trong trình tự peptide quan trọng trong việc dự đoán.

Khó khăn : tin sinh học còn khá mới lạ nên còn ít tài liệu để tham khảo. Kiến thức còn kém để làm tốt công việc.

7. Tổng kết và phương hướng phát triển

Các mô hình đã thực hiện cho kết quả với độ chính xác tốt với tập dữ liệu chuẩn khách quan khi sử dụng các đặc trưng AAC và DPC để mã hóa trình tự peptide đầu vào.

Hướng phát triển :

Xây dựng webserver để người dùng sử dụng để dự đoán peptide chống tạo mạch.

Tìm hiểu kỹ hơn về các peptide chống tạo mạch để có thể hỗ trợ tốt trong các nghiên cứu tiền lâm sàng và lâm sàng đối với bệnh ung thư. Để có thể bác sĩ sau này có thể dựa trên nghiên cứu chế tạo thuốc chữa ung thư

8. Tài liệu tham khảo

1. Hanahan D., Weinberg R.A. Hallmarks of cancer: The next generation. *Cell*. 2011;144:646–674. doi: 10.1016/j.cell.2011.02.013.
2. Siegel R.L., Miller K.D., Jemal A. Cancer statistics, 2018. *CA Cancer J. Clin.* 2018;68:7–30. doi: 10.3322/caac.21442.
3. Zhang H., Chen J. Current status and future directions of cancer immunotherapy. *J. Cancer*. 2018;9:1773–1781. doi: 10.7150/jca.24577.
4. Zugazagoitia J., Guedes C., Ponce S., Ferrer I., Molina-Pinelo S., Paz-Ares L. Current challenges in cancer treatment. *Clin. Ther.* 2016;38:1551–1566. doi: 10.1016/j.clinthera.2016.03.026.
5. Laengsri V, Nantasenamat C, Schaduengrat N, Nuchnoi P, Prachayasittikul V, Shoombuatong W. TargetAntiAngio: A Sequence-Based Tool for the Prediction and Analysis of Anti-Angiogenic Peptides. *International Journal of Molecular Sciences*. 2019; 20(12):2950.
6. Kubota Y. Tumor angiogenesis and anti-angiogenic therapy. *Keio J. Med.* 2012;61:47–56. doi: 10.2302/kjm.61.47.
7. Sund M., Zeisberg M., Kalluri R. Endogenous stimulators and inhibitors of angiogenesis in gastrointestinal cancers: Basic science to clinical application. *Gastroenterology*. 2005;129:2076–2091. doi: 10.1053/j.gastro.2005.06.023.
8. Lenz H.-J. Antiangiogenic agents in cancer therapy. *Oncology*. 2005;19:17–25.
9. Senger D.R., Claffey K.P., Benes J.E., Perruzzi C.A., Sergiou A.P., Detmar M. Angiogenesis promoted by vascular endothelial growth factor: Regulation through $\alpha 1\beta 1$ and $\alpha 2\beta 1$

integrins. *Proc. Natl. Acad. Sci. USA*. 1997;94:13612–13617.
doi: 10.1073/pnas.94.25.13612.

10. Johnson K.E., Wilgus T.A. Vascular endothelial growth factor and angiogenesis in the regulation of cutaneous wound repair. *Adv. Wound Care*. 2014;3:647–661.
doi: 10.1089/wound.2013.0517.

11. Shih T., Lindley C. Bevacizumab: An angiogenesis inhibitor for the treatment of solid malignancies. *Clin. Ther.* 2006;28:1779–1802.
doi: 10.1016/j.clinthera.2006.11.015.

12. Su Y., Yang W.-B., Li S., Ye Z.-J., Shi H.-Z., Zhou Q. Effect of angiogenesis inhibitor bevacizumab on survival in patients with cancer: A meta-analysis of the published literature. *PLoS ONE*. 2012;7:e35629. doi: 10.1371/journal.pone.0035629.

13. Kim A., Balis F.M., Widemann B.C. Sorafenib and sunitinib. *Oncologist*. 2009;14:800–805.
doi: 10.1634/theoncologist.2009-0088.

14. Grandinetti C.A., Goldspiel B.R. Sorafenib and sunitinib: Novel targeted therapies for renal cell cancer. *Pharmacother. J. Hum. Pharmacol. Drug Ther.* 2007;27:1125–1144.
doi: 10.1592/phco.27.8.1125.

15. Rosca E.V., Koskimaki J.E., Rivera C.G., Pandey N.B., Tamiz A.P., Popel A.S. Anti-angiogenic peptides for cancer therapeutics. *Curr. Pharm. Biotechnol.* 2011;12:1101–1116.
doi: 10.2174/138920111796117300.

16. Lee E., Lee S.J., Koskimaki J.E., Han Z., Pandey N.B., Popel A.S. Inhibition of breast cancer growth and metastasis by a biomimetic peptide. *Sci. Rep.* 2014;4:7139.
doi: 10.1038/srep07139.

17. Foy K.C., Liu Z., Phillips G., Miller M., Kaumaya P.T. Combination treatment with HER-2 and VEGF peptide mimics induces potent anti-tumor and anti-angiogenic responses in

vitro and in vivo. *J. Biol. Chem.* 2011;286:13626–13637.
doi: 10.1074/jbc.M110.216820.

18. Wong W. Combining anti-inflammatory and anti-angiogenic therapy. *Sci. Signal.* 2013;6:ec224.
doi: 10.1126/scisignal.2004747.

19. Chan L.Y., Craik D.J., Daly N.L. Dual-targeting anti-angiogenic cyclic peptides as potential drug leads for cancer therapy. *Sci. Rep.* 2016;6:35347.
doi: 10.1038/srep35347.

20. Chlenski A., Guerrero L.J., Peddinti R., Spitz J.A., Leonhardt P.T., Yang Q., Tian Y., Salwen H.R., Cohn S.L. Anti-angiogenic SPARC peptides inhibit progression of neuroblastoma tumors. *Mol. Cancer.* 2010;9:138.
doi: 10.1186/1476-4598-9-138.

21. Park S.W., Cho C.S., Jun H.O., Ryu N.H., Kim J.H., Yu Y.S., Kim J.S., Kim J.H. Anti-angiogenic effect of luteolin on retinal neovascularization via blockade of reactive oxygen species production. *Investig. Ophthalmol. Vis. Sci.* 2012;53:7718–7726.
doi: 10.1167/iovs.11-8790.

22. Kong J.S., Yoo S.A., Kim J.W., Yang S.P., Chae C.B., Tarallo V., Falco S.D., Ryu S.H., Cho C.S., Kim W.U. Anti-neuropilin-1 peptide inhibition of synoviocyte survival, angiogenesis, and experimental arthritis. *Arthritis Rheum. Off. J. Am. Coll. Rheumatol.* 2010;62:179–190. doi: 10.1002/art.27243.

23. Mahlapuu M., Håkansson J., Ringstad L., Björn C. Antimicrobial peptides: An emerging category of therapeutic agents. *Front. Cell. Infect. Microbiol.* 2016;6:194.
doi: 10.3389/fcimb.2016.00194.

24. Recio C., Maione F., Iqbal A.J., Mascolo N., De Feo V. The potential therapeutic application of peptides and peptidomimetics in cardiovascular disease. *Front. Pharmacol.* 2017;7:526. doi: 10.3389/fphar.2016.00526.

25. Lau J.L., Dunn M.K. Therapeutic peptides: Historical perspectives, current development trends, and future directions. *Bioorganic Med. Chem.* 2018;26:2700–2707. doi: 10.1016/j.bmc.2017.06.052.
26. Sulochana K., Ge R. Developing antiangiogenic peptide drugs for angiogenesis-related diseases. *Curr. Pharm. Des.* 2007;13:2074–2086. doi: 10.2174/138161207781039715.
27. Prediction of Anticancer Peptides Using a Low-Dimensional Feature Model
Qingwen Li, Wenyang Zhou, Donghua Wang, Sui Wang, and Qingyuan Li,* *Front Bioeng Biotechnol.* 2020; 8: 892. Published online 2020 Aug 12. doi: 10.3389/fbioe.2020.00892
28. Anti-angiogenic peptides for cancer therapeutics
Elena V. Rosca,§ Jacob E. Koskimaki,§ Corban G. Rivera, Niranjana B. Pandey, Amir P. Tamiz, and Aleksander S. Popel* *Curr Pharm Biotechnol.* Author manuscript; available in PMC 2011 Aug 1. *Curr Pharm Biotechnol.* 2011 Aug 1; 12(8): 1101–1116.

Mục lục

Catalog

Sinh viên thực hiện :	2
Giáo viên hướng dẫn:	2
Tóm tắt :	3
1. Giới thiệu.....	5
2. Mô tả bài toán.....	7
2.1. Chi tiết bài toán :	7
2.2. Tập dữ liệu sử dụng:	7
2.3. Các độ đo đánh giá:	8
3. Đề xuất mô hình :	11
4. Thực hiện hệ thống.....	11
4.1. Tiền xử lý dữ liệu.....	11
4.2. Xử lý dữ liệu đầu vào.....	12
4.3. Xây dựng mô hình.....	14
4.4. So sánh các mô hình.....	16
5. Thử nghiệm và kết quả.....	22
6. Thảo luận.....	24
7. Tổng kết và phương hướng phát triển.....	25
8. Tài liệu tham khảo.....	26
Mục lục.....	30

Table 1 : Các nghiên cứu về dự đoán chống tạo mạch đã có.....	6
Table 2 : Tập dữ liệu - 1.....	8
Table 3 : Tập dữ liệu - 2.....	8

Image 1 : Mô hình đề xuất	11
Image 2 : Tiền xử lý dữ liệu.....	12
Image 3 : Tải về các file dữ liệu.....	14
Image 4 : Hàm tính đặc trưng AAC của trình tự peptides và ghi ra file csv.....	14
Image 5 : Loại bỏ đặc trưng có phương sai thấp.....	15
Image 6 : map dữ liệu đầu ra.....	15
Image 7 : Huấn luyện mô hình.....	16
Image 8 : Cài đặt thư viện lazypredict.....	17
Image 9 : xử lý đầu vào, chia tập dữ liệu và huấn luyện nhanh 29 mô hình.....	17
Image 10 : top những mô hình có độ chính xác cao nhất -1	18
Image 11 : top những mô hình có độ chính xác cao nhất - 2.....	19
Image 12 : Tính toán các độ đo đánh giá.....	20
Image 13 : So sánh các độ đo đánh giá của 3 mô hình.....	20
Image 14 : top những đặc trưng AAC quan trọng.....	21
Image 15 : Sử dụng đặc trưng DPC	22
Image 16 : So sánh độ đo đánh giá 3 mô hình khi sử dụng đặc trưng DPC -1.....	22
Image 17 : So sánh độ đo đánh giá 3 mô hình khi sử dụng đặc trưng DPC -2.....	23
Image 18 : Top những đặc trưng DPC quan trọng	23