HUST

TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI

HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

ONE LOVE. ONE FUTURE.

TRƯỜNG ĐẠI HỌC
BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Tin sinh học

Xây dựng mô hình dự đoán peptides chống tạo mạch

Giảng viên hướng dẫn: TS. Nguyễn Hồng Quang

Sinh viên thực hiện:
Tống Mạnh Đạt **20173008**

ONE LOVE. ONE FUTURE.

1.Giới thiệu

- 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.
- Peptit được coi là một liệu pháp điều trị quna 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.

1.Giới thiệu

- Các nghiên cứu đã có :

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

2.Mô tả bài toán

- 2.1.Chi tiết bài toán :

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

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.

Đầu ra : dự đoán có phải có chức năng chống tạo mạch (anti-angiogenic peptides: Antiangio - 1) hay không (non-antiangiogenic peptides: Negative - 0).

2.Mô tả bài toán

- 2.1.Tập dữ liệu :

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-anigo	Non-anti-anigo	Anti-anigo	Non-anti-anigo
Tổng dữ liệu	137	137	99	101

2.Mô tả bài toán

- 2.2.Tập dữ liệu :

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)

Dataset	Tập train		Tập test	
	Anti-anigo	Non-anti-anigo	Anti-anigo	Non-anti-anigo
Tổng dữ liệu	189	189	47	49

2.Mô tả bài toán

- 2.2.Các độ đo :

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

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

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

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

HUST

hust.edu.vn fb.com/dhbkhn

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.

4.Thực hiện hệ thống:

- 4.2.Xử lý đầu vào :
- Đặc trưng AAC:

$$f(a) = {N_a}/{N},\ a \in (A,C,...,W,Y)$$

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

Na = tổng số lượng aminoacid loại a

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

- Đặc trưng DPC

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

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

Nab = tổng số lượng depeptit loại ab

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

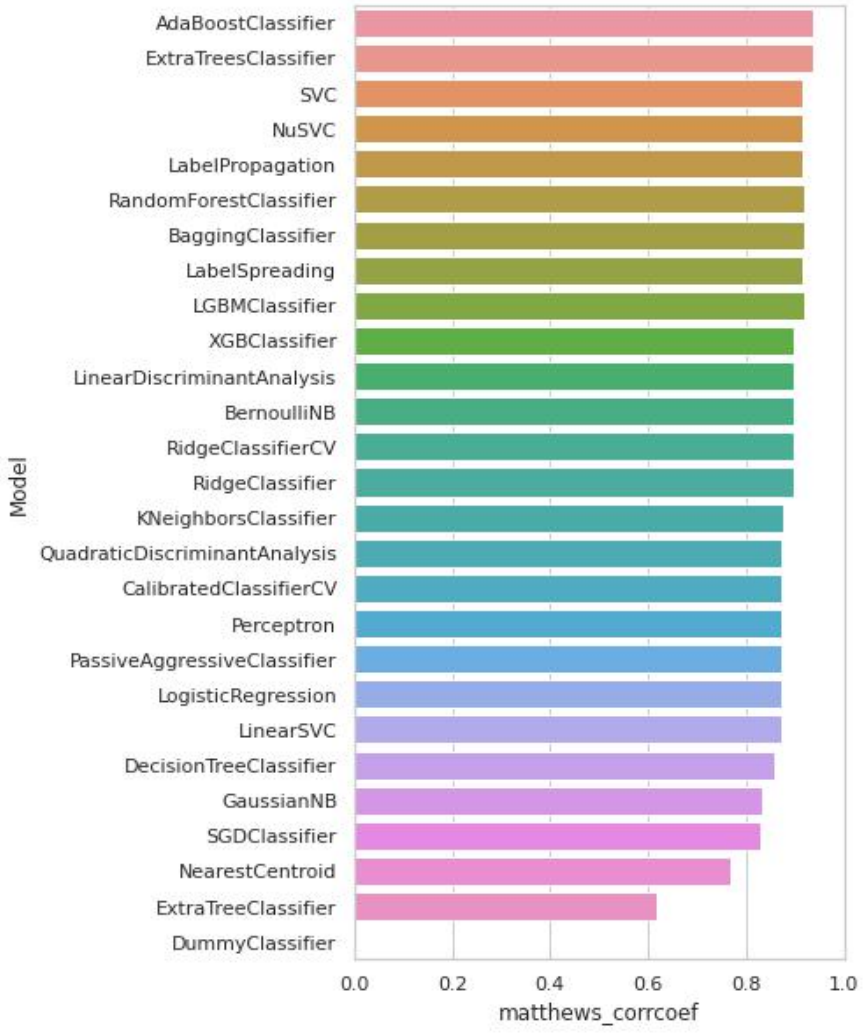
HUST

hust.edu.vn fb.com/dhbkhn

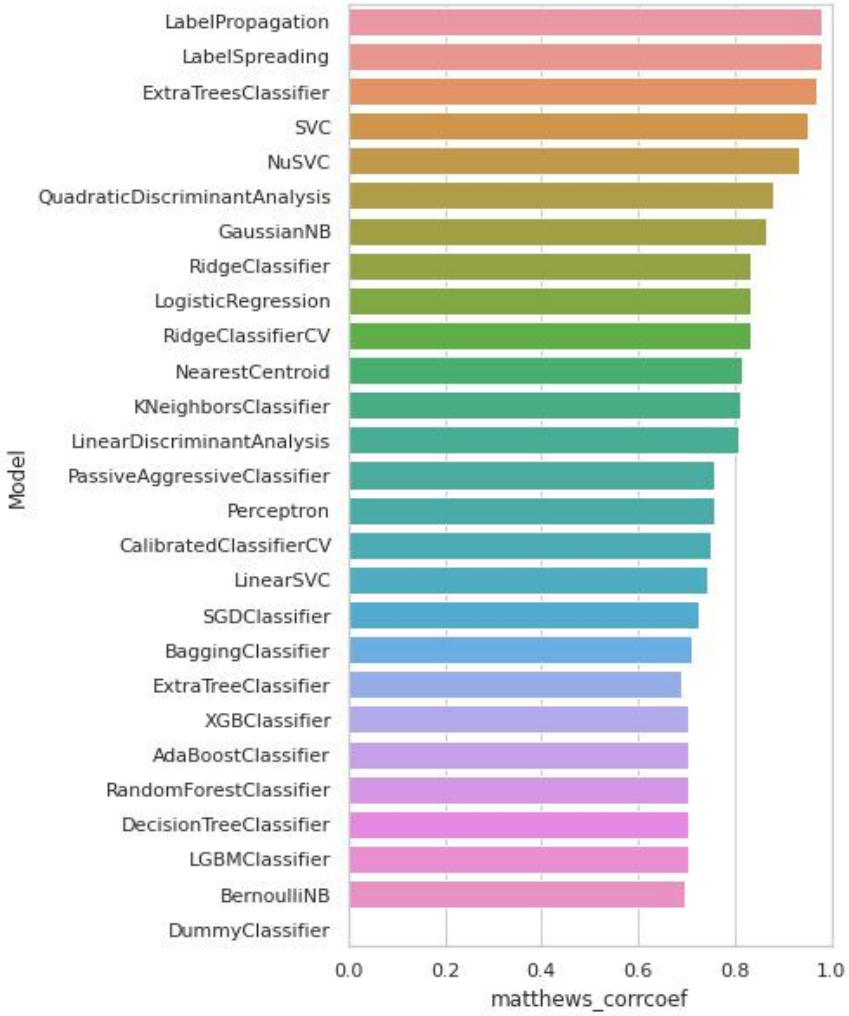
4.Thực hiện hệ thống:

- 4.3 Xây dựng mô hình và các kết quả :

Các kết quả :


AdaBoostClassifier
ExtraTreesClassifier
SVC
NuSVC
LabelPropagation
RandomForestClassifier
BaggingClassifier
LabelSpreading
LGBMClassifier
XGBClassifier
LinearDiscriminantAnalysis
BernoulliNB
RidgeClassifierCV
RidgeClassifier
KNeighborsClassifier
QuadraticDiscriminantAnalysis
CalibratedClassifierCV
Perceptron
PassiveAggressiveClassifier
LogisticRegression
LinearSVC
DecisionTreeClassifier
GaussianNB
SGDClassifier
NearestCentroid
ExtraTreeClassifier
DummyClassifier
Model
0.0
0.2
0.4
0.6
0.8
1.0
matthews_corrcoef

Các kết quả :


LabelPropagation
LabelSpreading
ExtraTreesClassifier
SVC
NuSVC
QuadraticDiscriminantAnalysis
GaussianNB
RidgeClassifier
LogisticRegression
RidgeClassifierCV
NearestCentroid
KNeighborsClassifier
LinearDiscriminantAnalysis
PassiveAggressiveClassifier
Perceptron
CalibratedClassifierCV
LinearSVC
SGDClassifier
BaggingClassifier
ExtraTreeClassifier
XGBClassifier
AdaBoostClassifier
RandomForestClassifier
DecisionTreeClassifier
LGBMClassifier
BernoulliNB
DummyClassifier
Model
0.0
0.2
0.4
0.6
0.8
1.0
matthews_corrcoef

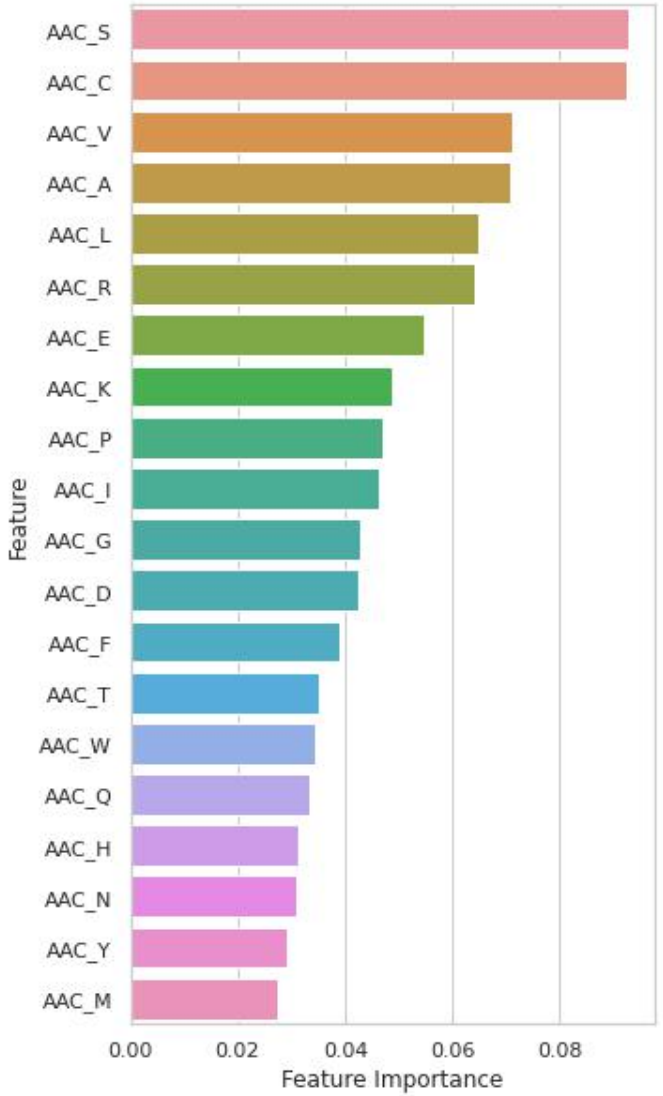
Các kết quả :

- So sánh các độ đo của 3 phương pháp :

	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

Các kết quả :

- Những đặc trưng AAC quan trọng trong việc dự đoán peptide chống tạo mạch



HUST

hust.edu.vn fb.com/dhbkhn

Các thử nghiệm khác :

Các kết quả :

- 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) với :

- Đặc trưng DPC

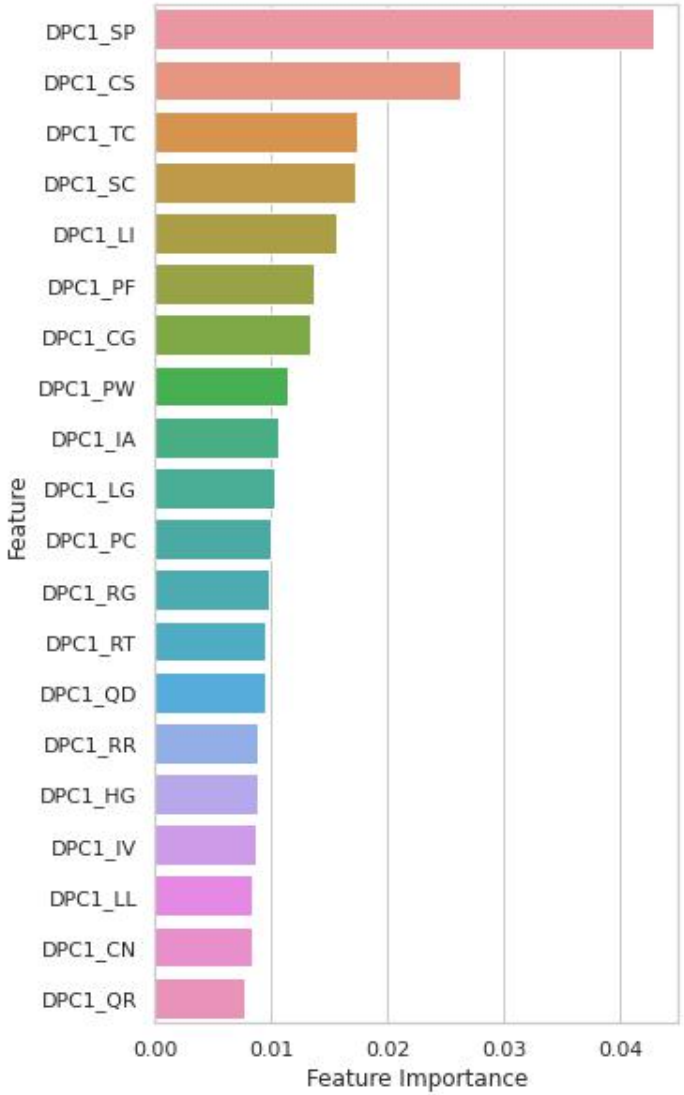
	method	AC	Sn	Sp	MCC
0	RandomForestClassifier	0.755319	0.772727	0.740000	0.511682
1	ExtraTreesClassifier	0.840426	0.847826	0.833333	0.681005
2	AdaBoostClassifier	0.734043	0.761905	0.711538	0.470757

- Đặc trưng AAC

	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

Các kết quả :

- Những đặc trưng DPC quan trọng trong việc dự đoán peptide chống tạo mạch



HUST

THANK YOU !