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:

Tống Mạnh Đạt - 20173008

Giáo viên hướng dẫn: TS. Nguyễn Hồng Quang

Hà Nội, 02/2022

**Dự đoán các peptides chống tạo mạch**

## Sinh viên thực hiện :

Tống Mạnh Đạt, MSSV : 20173008

Lớp : KTMT 01, 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 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. 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 cặn 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>

## 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.

## Mô tả bài toán

### 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-antiangiogenic 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.

### 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-anigo | Non-anti-anigo | Anti-anigo | Non-anti-anigo |
| 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 )

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

Table 3 : Tập dữ liệu - 2

### Các độ đo đánh giá:

Table

Description automatically generated

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.

## Thực hiện hệ thống

### 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.

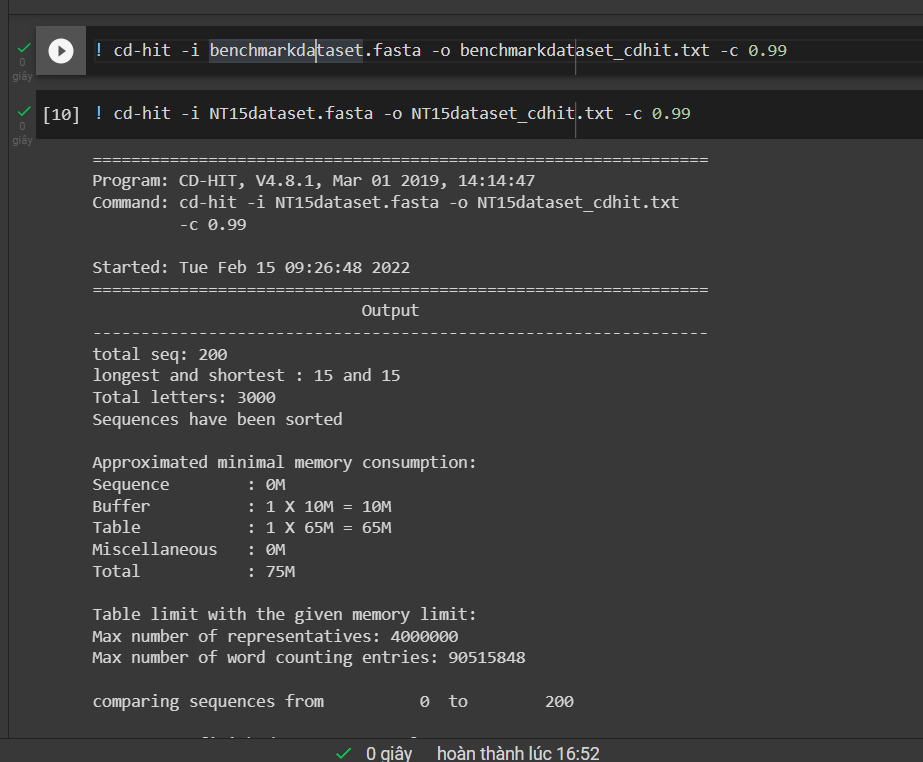


Image 1 : Tiền xử lý dữ liệu

### Xử lý dữ liệu đầu vào

Định dạng dữ liệu ban đầu:

>AA135

YTMNPRKLFDY

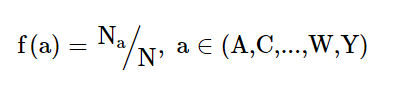
>neg1

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 :

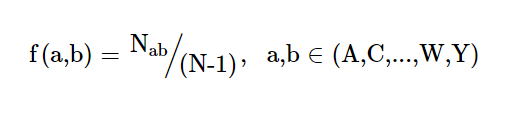


Với f(a) là đặc trưng AAC của aminoacid loại a trong trình tự

Na = tổng số lượng aminoacid loại a

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

Đặc trưng DPC - 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 20x20 = 400 đặc trưng



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 .

### Xây dựng mô hình

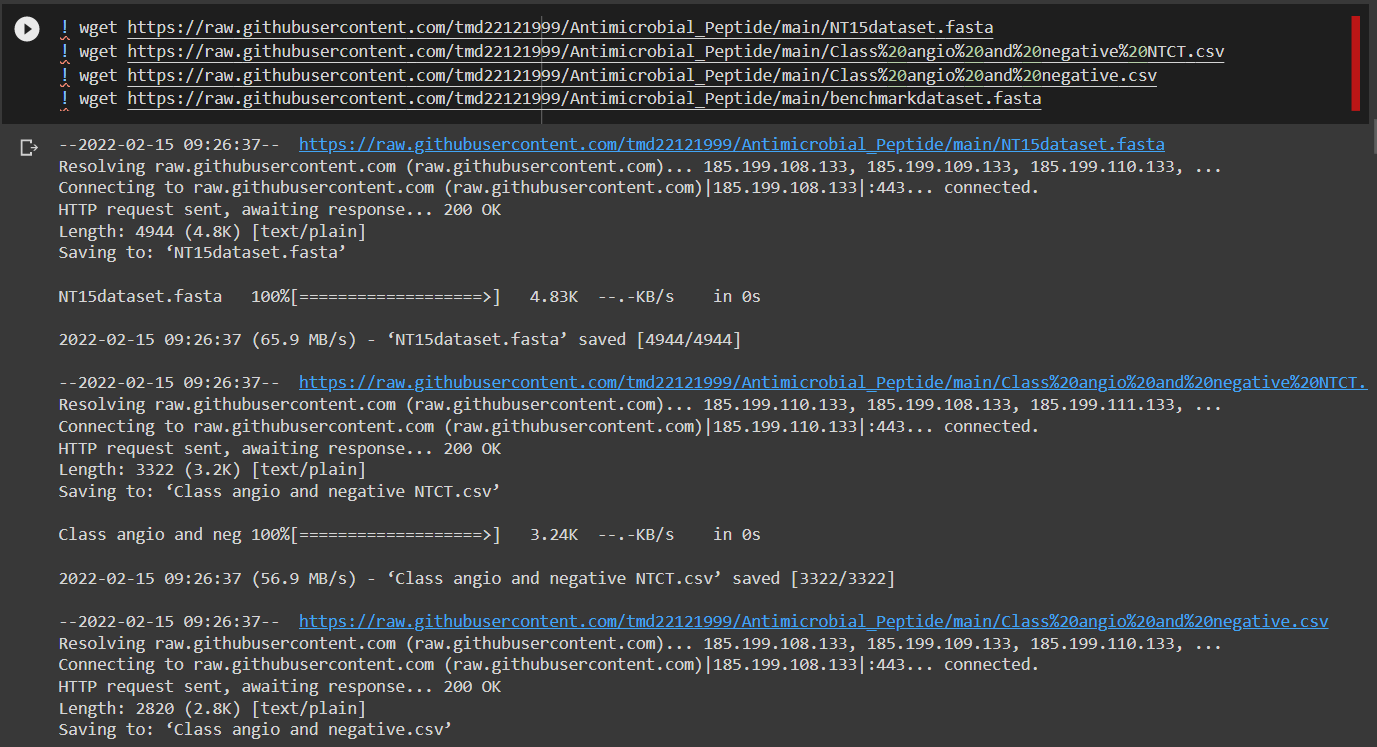
Download các file dữ liệu : 

Image 2 : Tải về các file dữ liệu

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

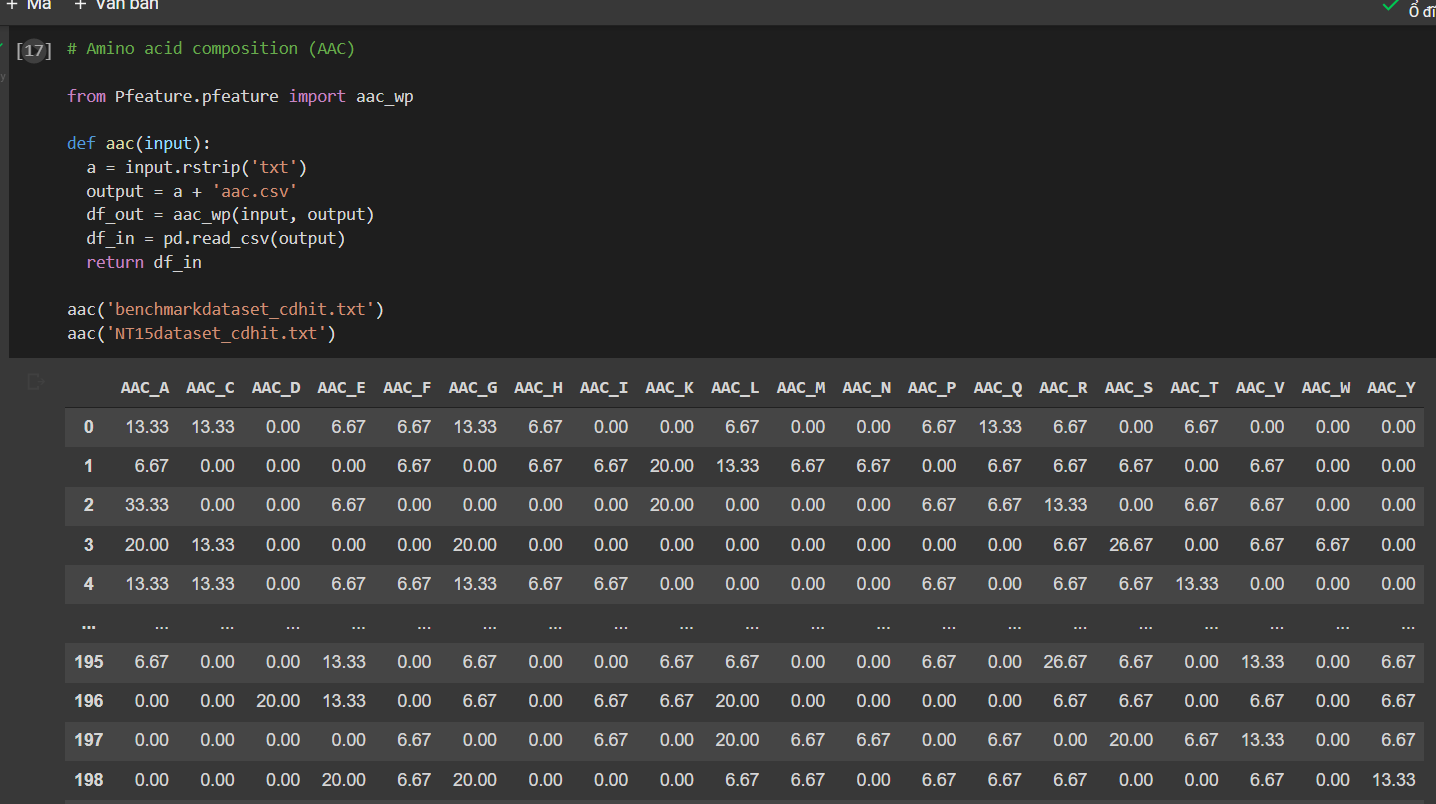


Image 3 Hàm tính đặc trưng AAC của trình tự peptides và ghi ra file csv

Loại bỏ đặc trưng có phương sai thấp:

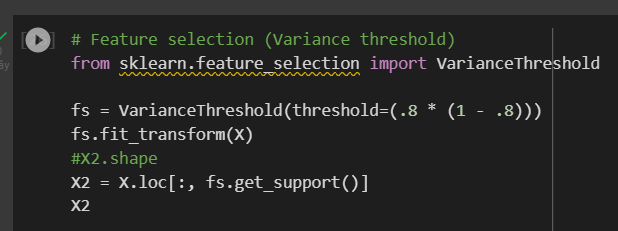


Image 4 : Loại bỏ đặc trưng có phương sai thấp

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

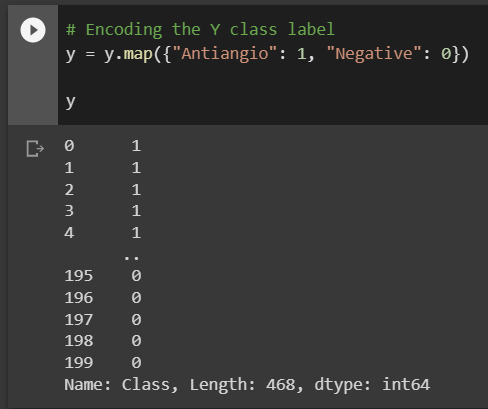


Image 5 : map dữ liệu đầu ra

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

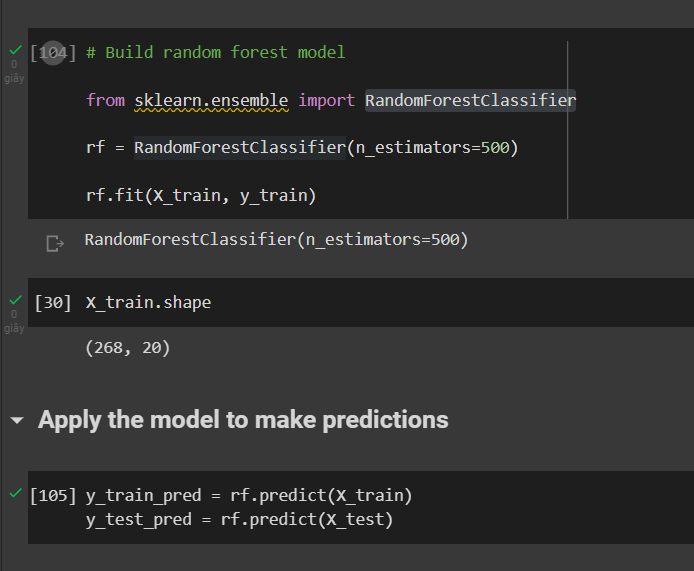


Image 6 : Huấn luyện mô hình

### 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 :

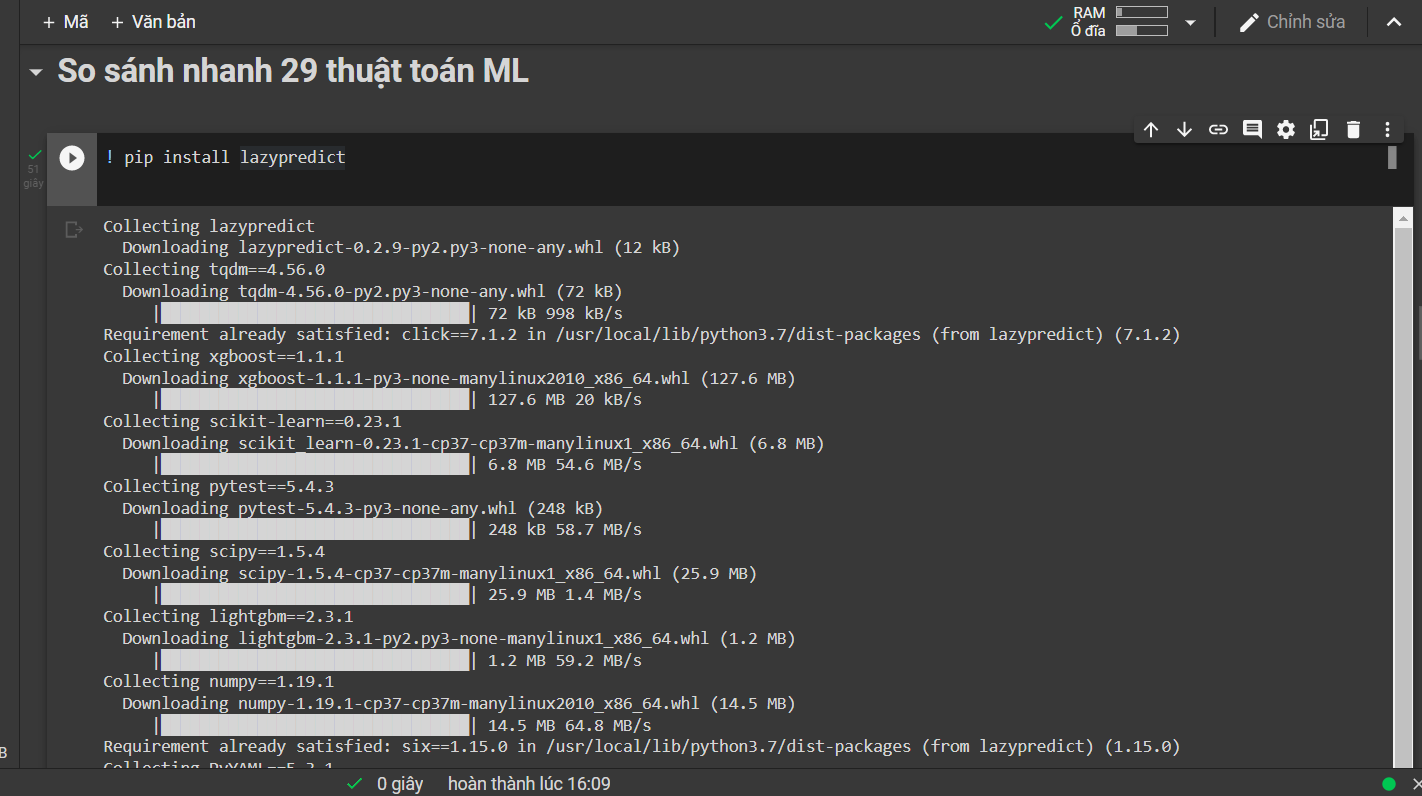


Image 7 : Cài đặt thư viện lazypredict

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 :

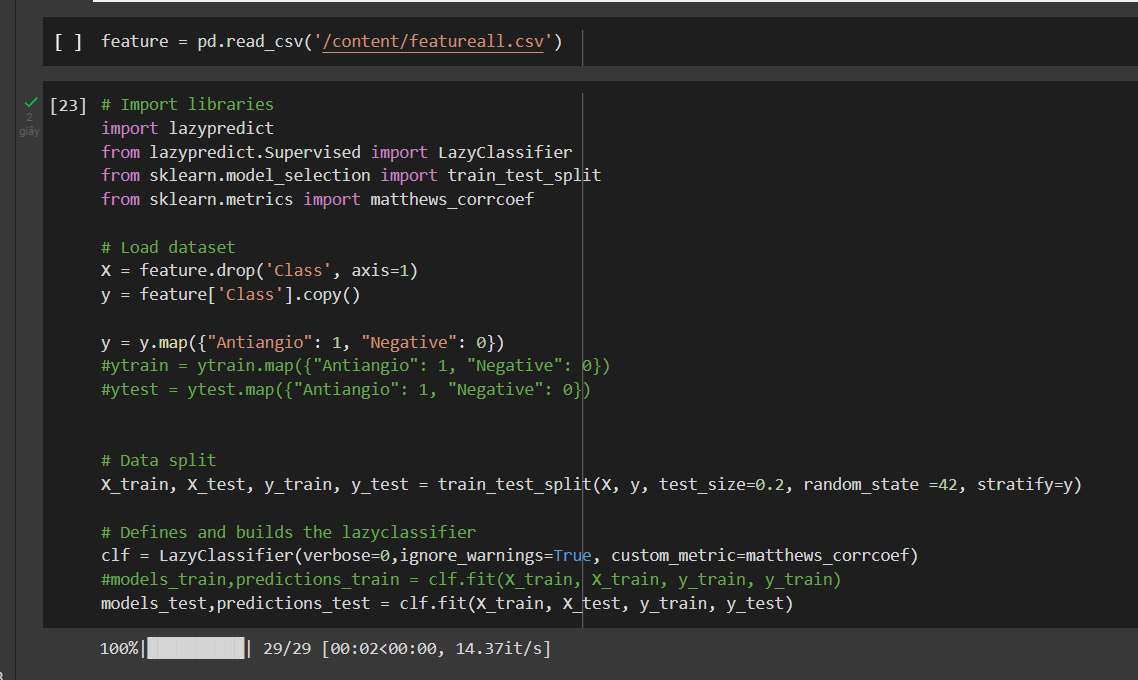


Image 8 :xử lý đầu vào, chia tập dữ liệu và huấn luyện nhanh 29 mô hình

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 :

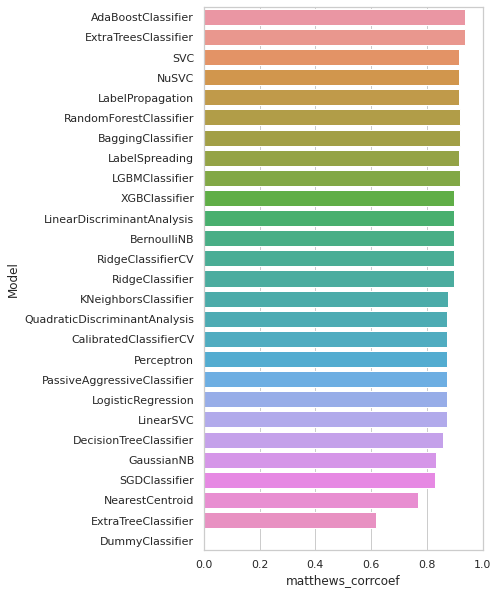


Image 9 :top những mô hình có độ chính xác cao nhất -1

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:

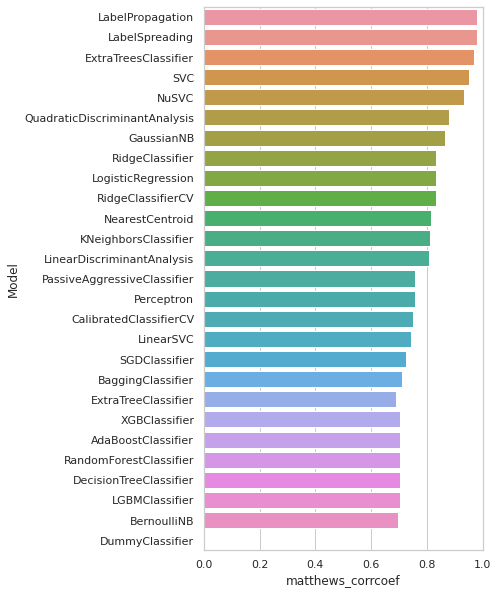


Image 10 : top những mô hình có độ chính xác cao nhất - 2

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



Image 11 : Tính toán các độ đo đánh giá

Ta được bảng sau :

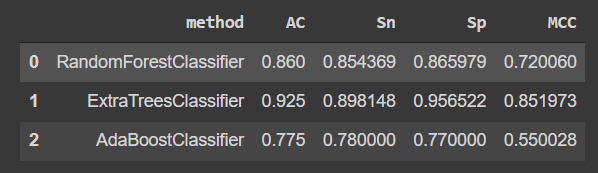


Image 12 : So sánh các độ đo đánh giá của 3 mô hình

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 phướng pháp còn lại 2 tỷ lệ này khá cân bằng.

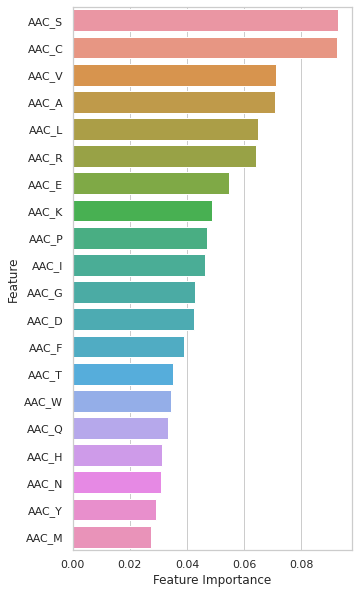


Image 13 : top những đặc trưng AAC quan trọng

## 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 mền, bộ công cụ toolkits sử dụng:

-Google colab.

Các thử nghiệm khác:

Sử dụng đặc trưng DPC thay cho đặc trưng AAC

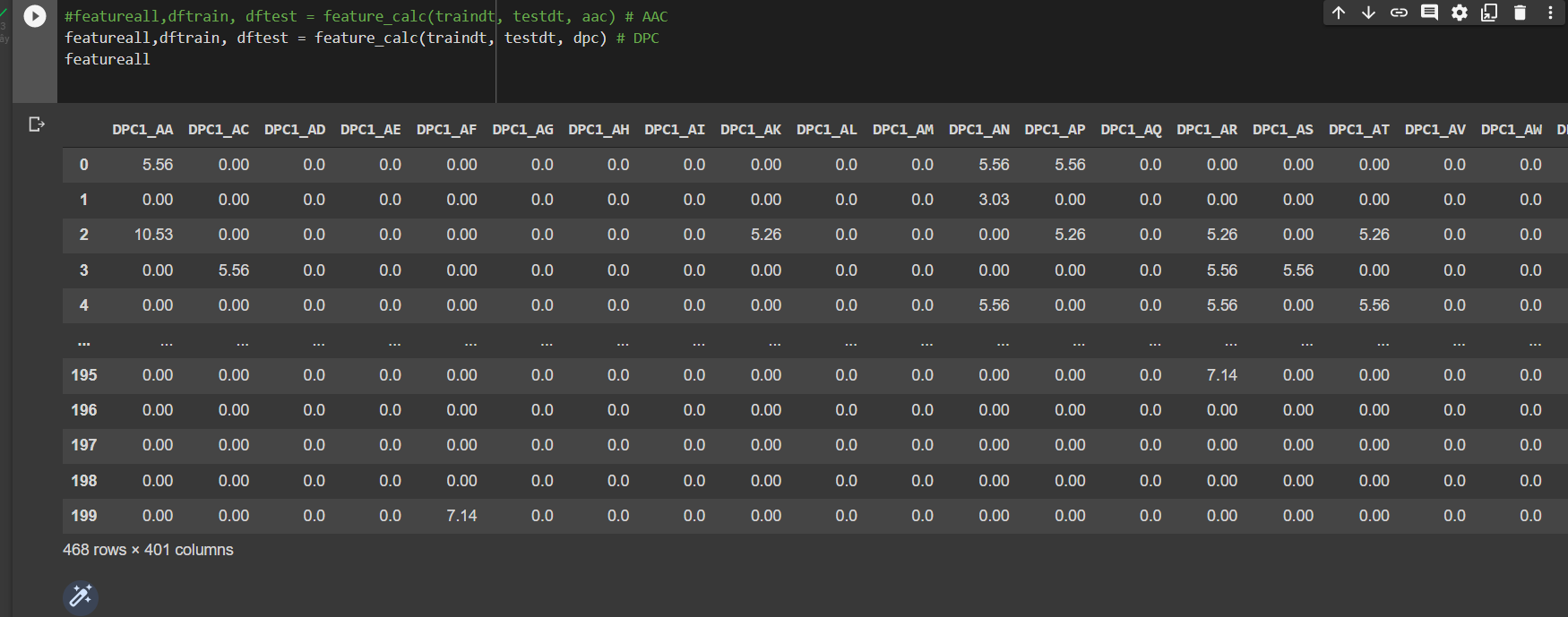


Image 14 : Sử dụng đặc trưng DPC

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) :

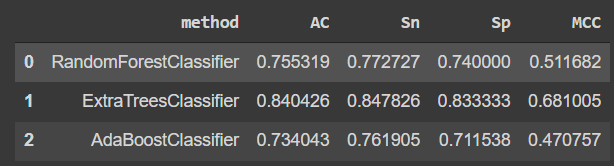


Image 15 : So sánh độ đo đánh giá 3 mô hình khi sử dụng đặc trưng DPC -1

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

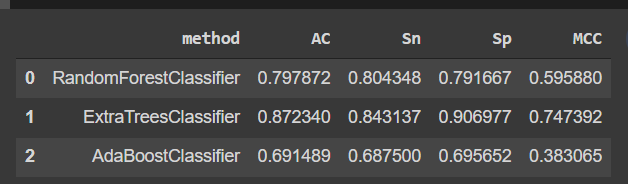


Image 16 : So sánh độ đo đánh giá 3 mô hình khi sử dụng đặc trưng DPC -2

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

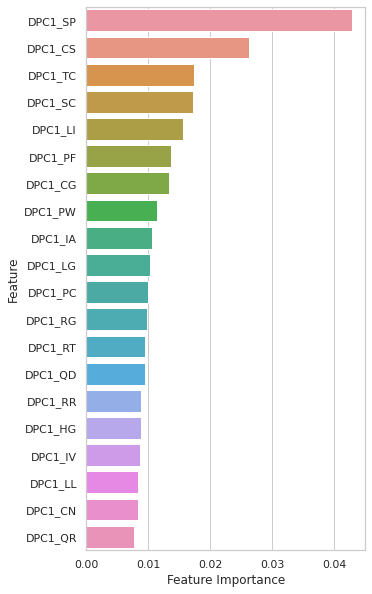


Image 17 : Top những đặc trưng DPC quan trọng

## Thảo luận

Với tập dữ liệu benmark 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 đọ 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.

## Tống 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 choogns 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

## 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, Schaduangrat 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 α1β1 and α2β1 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, Niranjan 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](#_Toc16242)

[Giáo viên hướng dẫn: 2](#_Toc10801)

[Tóm tắt : 3](#_Toc318)

[1. Giới thiệu 4](#_Toc14062)

[2. Mô tả bài toán 6](#_Toc9034)

[2.1. Chi tiết bài toán : 6](#_Toc15394)

[2.2. Tập dữ liệu sử dụng: 6](#_Toc1346)

[2.3. Các độ đo đánh giá: 7](#_Toc9876)

[3. Thực hiện hệ thống 10](#_Toc25980)

[3.1. Tiền xử lý dữ liệu 10](#_Toc13664)

[3.2. Xử lý dữ liệu đầu vào 10](#_Toc76)

[3.3. Xây dựng mô hình 12](#_Toc3697)

[3.4. So sánh các mô hình 14](#_Toc31545)

[4. Thử nghiệm và kết quả 20](#_Toc874)

[5. Thảo luận 22](#_Toc682)

[6. Tống kết và phương hướng phát triển 23](#_Toc29477)

[7. Tài liệu tham khảo 24](#_Toc30195)

[Mục lục 28](#_Toc21500)

[Table 1 : Các nghiên cứu về dự đoán chống tạo mạch đã có 5](#_Toc1376)

[Table 2 : Tập dữ liệu - 1 7](#_Toc997)

[Table 3 : Tập dữ liệu - 2 7](#_Toc15169)

[Image 1 : Tiền xử lý dữ liệu 10](#_Toc31164)

[Image 2 : Tải về các file dữ liệu 12](#_Toc2393)

[Image 3 Hàm tính đặc trưng AAC của trình tự peptides và ghi ra file csv 12](#_Toc15240)

[Image 4 : Loại bỏ đặc trưng có phương sai thấp 13](#_Toc24248)

[Image 5 : map dữ liệu đầu ra 13](#_Toc26563)

[Image 6 : Huấn luyện mô hình 14](#_Toc23792)

[Image 7 : Cài đặt thư viện lazypredict 15](#_Toc21978)

[Image 8 :xử lý đầu vào, chia tập dữ liệu và huấn luyện nhanh 29 mô hình 15](#_Toc8160)

[Image 9 :top những mô hình có độ chính xác cao nhất -1 16](#_Toc21197)

[Image 10 : top những mô hình có độ chính xác cao nhất - 2 17](#_Toc16325)

[Image 11 : Tính toán các độ đo đánh giá 18](#_Toc13518)

[Image 12 : So sánh các độ đo đánh giá của 3 mô hình 18](#_Toc8551)

[Image 13 : top những đặc trưng AAC quan trọng 19](#_Toc19722)

[Image 14 : Sử dụng đặc trưng DPC 20](#_Toc5776)

[Image 15 : So sánh độ đo đánh giá 3 mô hình khi sử dụng đặc trưng DPC -1 20](#_Toc11207)

[Image 16 : So sánh độ đo đánh giá 3 mô hình khi sử dụng đặc trưng DPC -2 21](#_Toc25282)

[Image 17 : Top những đặc trưng DPC quan trọng 21](#_Toc27311)