DỰ ĐOÁN PROTEIN CARBONYLATION SITES

Sinh viên thực hiện

- Mạc Quang Huy, 20173169, CNTT-06, K62,
  huy.mq173169@sis.hust.edu.vn, 0947123810

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:
Cacbonyl hóa được coi là một biến đổi sau dịch mã không thể đảo ngược và được coi là một dấu ấn sinh học của stress oxy hóa. Nó đóng một vai trò quan trọng không chỉ trong việc điều phối các quá trình sinh học khác nhau mà còn liên quan đến một số bệnh như bệnh Alzheimer, tiểu đường và bệnh Parkinson. Tuy nhiên, vì các công nghệ thử nghiệm tốn kém và tốn thời gian để phát hiện các vị trí cacbonyl hóa trong protein, một phương pháp tính toán chính xác để dự đoán các vị trí cacbonyl hóa là một vấn đề cấp bách có thể hữu ích cho việc phát triển thuốc. Trong bài báo gốc, tác giả đề xuất một công cụ để xác định nhanh, hiệu quả các vị trí Carbon hóa được gọi là iCarPS dựa trên sequence information. Phương pháp mã hóa mới này được gọi là tọa độ hình nón của residues (chất cặn) kết hợp với 9 đặc tính hóa lý của chúng, từ đó giúp ta xác định các mẫu carbonylated protein và non-carbonylated protein. Nhận thấy thuật toán Random Forest mà tác giả dùng để phân loại chỉ dừng lại ở mức đề cập và được đóng gói thành một gói jar để sử dụng. Vì vậy, dựa trên bài báo gốc và thực hiện kết hợp với các bài báo khác, project đã sử dụng các đặc trưng được trích xuất theo phương pháp gốc của tác giả, tiến hành phân loại không chỉ dừng ở Random Forest mà bài báo đã đề xuất mà tiến hành phân loại với một số thuật toán khác.

Từ khóa: Reactive Oxygen Species, Protein Carbonylation, Residues Conical Coordinates, Support Vector Machine, Random Forest,..

1. Giới thiệu (Introduction)

Stress oxi hóa là hiện tượng khi mà các loại phản ứng với oxy liên tục xảy ra trong và ngoài cơ thể. Sự đa dạng về cấu trúc và chức năng của protein cũng như tính dẻo và tính động của tế bào sống bị chi phối đáng kể bởi các biến đổi sau dịch mã (PTM) [1]. Không chỉ vậy, PTM còn có nhiệm vụ mở rộng mã di truyền và điều

hòa sinh lý tế bào. Một loạt các PTM như hydroxyl hóa, nitrat hóa, sulfhydryl hóa, cacbonyl hóa và glutathionyl hóa đã được tạo ra từ stress oxy hóa là kết quả trực tiếp của sự mất cân bằng trong sản xuất và phân hủy các loại phản ứng oxy (reactive oxygen species-ROS) và các loại phản ứng nitơ (reactive nitrogen species-RNS). Stress oxy hóa có thể xảy ra khi sinh dư thừa các loại phản ứng oxy (ROS) với khả năng giải độc của tế bào và làm suy yếu khả năng sửa chữa tổn thương .

Trong số nhiều loại PTM do stress oxy hóa gây ra, sự Cacbonyl hóa Protein được coi là dấu ấn sinh học cho stress oxy hóa , Carbonyl hóa Protein thường được đặc trưng là tính bền vững- stability, không thể đảo ngược- irreversibility và hình thành sớm tương đối- relative early formation, được coi là các đặc điểm để tính toán mức độ stress oxy hóa. Một số nhà nghiên cứu đã xác nhận rằng quá trình carbonyl hóa protein có tác động xấu đến chức năng của protein, sự gấp protein, sự phân giải protein và rối loạn chức năng tế bào, thường dẫn đến sự phân hủy protein của protein.

Bên cạnh đó, cũng có rất nhiều các loại bệnh khác nhau như bệnh phổi mãn tính, bệnh Parkinson, bệnh đục thủy tinh thể, suy thận mãn tính, nhiễm trùng huyết,… có liên quan đến mức độ cacbonyl hóa protein. Vì vậy, việc có thể xác định được các vị trí cacbonyl hóa protein có thể cung cấp cho ta các manh mối quan trọng của quá trình chuyển hóa tế bào và protein bị ảnh hưởng.

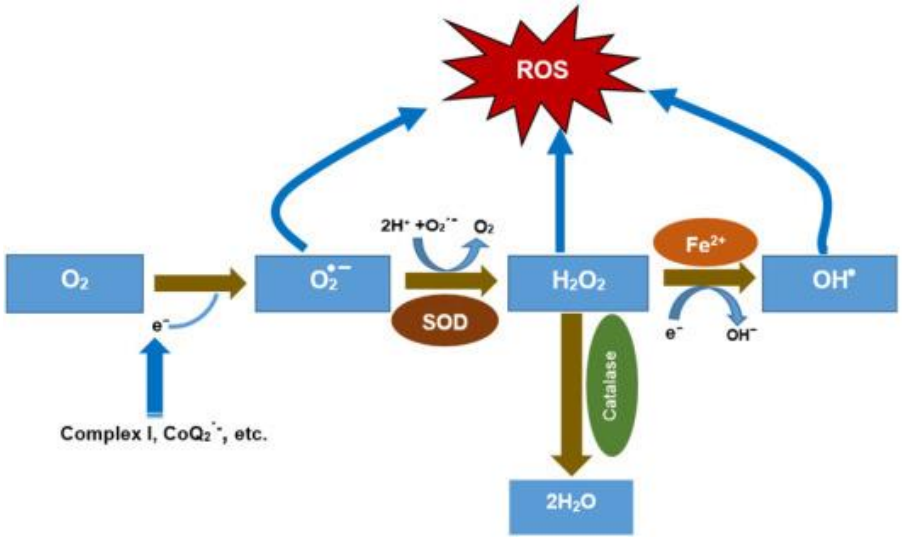

Figure 1: Reactive Oxygen Species – ROS

Một số nghiên cứu chỉ ra rằng một số gốc axit amin, đặc biệt là R, K, P và T, có khả năng bị cacbonyl hóa và có ảnh hưởng đến các gốc lân cận. Hơn nữa, các vị trí cacbonyl hóa nhiều hơn được tìm thấy ở các vùng giàu RKPT và các vị trí cacbonyl hóa có xu hướng tập hợp mạnh.

Khối phổ- Mass spectrometry và sắc ký lỏng- liquid chromatography là những kỹ thuật phổ biến nhất để phân tích tính nhạy cảm với protein của PTMs oxy hóa và

xác định vị trí cacbonyl hóa chính xác của nó gần đây. Cần lưu ý rằng trong số tất cả các phần còn lại của các phân tử protein, bốn loại dư lượng axit amin, cụ thể là lysine (K), proline (P), arginine (R) và threonine (T), đã được tìm thấy dễ bị cacbonyl hóa. Tuy nhiên, sẽ tốn kém nếu chỉ sử dụng kỹ thuật thí nghiệm sinh hóa thuần túy đã kể trên để xác định các vị trí cacbonyl hóa protein chính xác, đặc biệt là đối với các bộ dữ liệu quy mô lớn. Việc tích lũy dữ liệu protein và phát triển các kỹ thuật AI cung cấp cho chúng ta cơ hội để tạo ra một mô hình mạnh mẽ cho các vị trí cacbonyl hóa. Cho đến nay, một số bộ phân loại tính toán đã được xây dựng để xác định các vị trí cacbonyl hóa protein. Như một công cụ dự đoán gọi là preCar-site [2] hay MDD-carb [3]. Tuy nhiên tác giả cũng giải thích rằng, mặc dù hiệu quả nhưng các phương pháp trước đó có nhiều hạn chế. Đặc biệt phải kể đến như:

1. Ít máy chủ web hoạt động và khó sử dụng
2. Hiệu suất của các phương pháp đó có thể được cải thiện
3. Một số tập dataset test hiện nay đã được xây dựng để kiểm chứng
4. Hầu hết các phương pháp trước chưa tối ưu bằng các sử dụng các kỹ thuật trích chọn đặc trưng.

Bên cạnh đó, để đưa ra một công cụ dự đoán thống kê dựa trên trình tự hữu ích cho một hệ thống sinh học như đã được chứng minh trong một loạt các ấn phẩm, các quy tắc năm bước của Chou [4] cần được tuân thủ:

- (i) cấu trúc hoặc chọn một tập dữ liệu chuẩn hợp lệ để train và test công cụ dự đoán
- (ii) xây dựng các mẫu trình tự sinh học với một biểu thức toán học hiệu quả có thể phản ánh thực sự mối tương quan nội tại của chúng với mục tiêu được dự đoán
- (iii) giới thiệu hoặc phát triển một thuật toán mạnh mẽ để vận hành dự đoán
- (iv) thực hiện đúng các thử nghiệm xác thực chéo để đánh giá khách quan độ chính xác dự đoán của nó
- (v) thiết lập một máy chủ web thân thiện với người dùng hoặc gói phần mềm có thể truy cập được cho công chúng

Trong bài báo gốc , tác giả đề xuất ra một phương pháp mã hoá vector mới được gọi là residues conical coordinates, dựa trên toán học và kết hợp 9 đặc tính hoá lý của amino axit để nhận dạng carbonylated và non-carbonylated. Sau đó, F-score được sử dụng để tối ưu hóa các đặc trưng, và sử dụng thuật toán Random Forest để phân loại. Bên cạnh đó, sinh viên Mạc Quang Huy cũng đề xuất thử nghiệm thêm các thuật toán phân loại khác như SVM hay Decision Tree dựa trên sách [5].

2. Mô tả bài toán

2.1. Chi tiết bài toán

+ Đầu vào (Input): Một chuỗi protein gồm 27 kí tự.

- Ví dụ: Mẫu P âm của loài trâu
  >**gi**|3336842|emb|CAA76847.1|**bovine**
  KKQTALVELLKHK**P**KATEEQLKTVMEN
- Ví dụ: Mẫu K dương của con người – gen vẫn được tạo ra cho mẫu protein K bằng cách chèn thêm các ký tự X để K nằm giữa
  >**sp**|Q8N3Y7|**RDHE2_HUMAN**
  GILHAMDGFVDQK**K**KLXXXXXXXXXXX

+ Đầu ra (Output): 0 hoặc 1 đại diện cho gốc (kí tự ở giữa trong chuỗi 27 kí tự) có bị carbonylation hay không.

- Ví dụ: Dự đoán các residues là dương tính ở loại K với mẫu

>sp|Q8IZT6|ASPM_HUMAN
YRMHVQQKKWKIMKKAALLIQKYYRAY

Cho ra:

Carbonylation Site:
>sp|Q8IZT6|ASPM_HUMAN Postion:14 is K YRMHVQQKKWKIMKKAALLIQKYYRAY

Non-carbonylation Site:
>sp|Q8IZT6|ASPM_HUMAN Postion:8 is K XXXXXXYRMHVQQKKWKIMKKAALLIQ
>sp|Q8IZT6|ASPM_HUMAN Postion:9 is K XXXXXYRMHVQQKKWKIMKKAALLIQK
>sp|Q8IZT6|ASPM_HUMAN Postion:11 is K XXXYRMHVQQKKWKIMKKAALLIQKYY
>sp|Q8IZT6|ASPM_HUMAN Postion:22 is K KWKIMKKAALLIQKYYRAYXXXXXXXX

2.2. Tập dữ liệu sử dụng

Bộ dữ liệu được sử dụng là bộ chuẩn của CarSPred, bao gồm từ 230 chuỗi protein cacbonyl hóa từ người và 20 chuỗi protein cacbonyl hóa từ các động vật có vú khác. Vì số lượng vị trí cacbonyl hóa trên dư lượng axit amin H, C và W là rất nhỏ và không có nguồn dữ liệu công khai đáng tin cậy, nên ta tập chung vào xây dựng mô hình dự đoán về dư lượng K, P, R, T.

Tập dữ liệu chuẩn bao gồm bốn tập con ký hiệu là $\mathbb{S}_{\circledast}$ (với $\circledast$ biểu thị cho 1 trong các chất cặn K, P, R, T)

$$\mathbb{S}_{\circledast} = \mathbb{S}_{\circledast}^{+} \cup \mathbb{S}_{\circledast}^{-}$$

- Dấu "+": tập con dương gồm các mẫu của true carbonylation site cho chất cặn(residue)
- Dấu "-": tập con âm gồm các mẫu của false carbonylation site cho chất cặn(residue)

Để xây dựng tập dữ liệu với công thức trên, trước hết, cửa sổ trượt $(2\xi + 1)$-mer được sử dụng để trích xuất các mẫu dương và âm với $\mathbb{U} = \circledast$ ở tâm dọc theo mỗi phân đoạn trình tự protein Do đó, một mẫu trình tự protein chứa vị trí cacbonyl hóa tiềm năng có thể được biểu thị bằng

$$P_{\xi}(\mathbb{U}) = P_{-\xi}P_{-(\xi-1)}\cdots P_{-2}P_{-1}\mathbb{U}P_{+1}P_{+2}\cdots P_{+(\xi-1)}P_{+\xi}$$

Trong đó:

- $\mathbb{U} = \circledast$ (K, P, R hoặc T)

- ξ : là 1 số nguyên
- $P_{-\xi}$: upstream (nơi phiên mã sớm hơn) axit amin thứ ξ tính từ trung tâm
- $P_{+\xi}$: downstream (nơi phiên mã muộn hơn) axit amin thứ ξ tính từ trung tâm

Theo thông tin vị trí của các vị trí cacbonyl hóa, các phân đoạn trình tự protein được coi là các mẫu dương và được đưa vào tập con $\boldsymbol{S}^{+}_{\circledast}$, nếu các tâm của chúng là các vị trí cacbonyl hóa đã được xác nhận bằng thực nghiệm. Ngược lại, được đưa vào tập âm $\boldsymbol{S}^{-}_{\circledast}$. Cụ thể:

Group	Dataset	Carbonylation sites			
		K	P	R	T
Train	positive	226	114	119	116
	negative	1802	716	754	702
	% positive	11.14%	13.73%	13.63%	14.18%
Test	positive	34	12	17	5
	negative	147	76	93	30
	% positive	18.78%	13.64%	15.45%	14.29%

2.3. Các độ đo đánh giá

Kiểm tra chéo 10 lần để kiểm tra hiệu suất của model. Hơn nữa, các thước đo truyền thống như sensitivity-độ nhạy (Sn), specificity-độ đặc tính (Sp), overall accuracy-độ chính xác (ACC) và Matthews correlation coefficient –hệ số tương quan Matthews(MCC) đã được thông qua để đánh giá hiệu suất dự đoán của model:

Overall accuracy:

$$\text{ACC} = \frac{\text{số điểm dự đoán đúng}}{\text{tổng số dự đoán}} = \frac{\text{TP} + \text{TN}}{\text{TP+FP +TN+FN}}$$

Specificity:

$$S_p = \frac{\text{TN}}{\text{TN+FP}}$$

Sensitivity:

$$S_n = \frac{\text{TP}}{\text{TP+FN}}$$

Matthews correlation coefficient:

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

F1-score:

$$\frac{2}{\text{F1}} = \frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}$$

Trong đó:

- TP: true positives – dự đoán chính xác là protein cacbonyl hóa
- TN: true negatives - dự đoán chính xác là protein không cacbonyl hóa
- FP: false positives – dự đoán ko chính xác là protein cacbonyl hóa
- FN: false negatives - dự đoán không chính xác là protein không cacbonyl hóa

Giá trị của ACC, Sn và Sp càng cao thì công cụ dự báo càng hiệu quả. Ngoài ra, -1 <MCC <1, giá trị MCC = 1 cho biết dự đoán tốt nhất có thể xảy ra trong khi MCC = - 1 cho biết dự đoán xấu nhất có thể xảy ra (hoặc phản tương quan). MCC = 0 sẽ được mong đợi cho một schema dự đoán ngẫu nhiên. Do bài toán xác định protein carbonylation, do đó 2 chỉ số quan tâm là accuracy và recall. Trong đó, TP là True Possitive, FP là False Positive, FN là False Negative.

3. Các hướng nghiên cứu liên quan (Related Works)

Qua đó ta thấy việc tích lũy dữ liệu protein và phát triển các kỹ thuật thông minh nhân tạo cung cấp cho chúng ta cơ hội để tạo ra một mô hình mạnh mẽ cho các vị trí cacbonyl hóa là rất quan trọng. Cho đến nay, một số bộ phân loại tính toán đã được xây dựng để xác định các vị trí cacbonyl hóa protein. Lv và cộng sự [6] đã thu thập 250 trình tự protein cacbonyl hóa được xác minh bằng các thí nghiệm sinh hóa, và xây dựng một bộ dữ liệu chuẩn có chứa các dư lượng biến đổi R, K, P và T ở người và động vật có vú khác (chuột, thỏ và bò). Một công cụ dự đoán trực tuyến có tên CarSPred được xây dựng bằng cách sử dụng bốn loại đặc trưng và kết hợp kỹ thuật trích chọn đặng trưng mRMR - minimum Redundancy Maximum Relevance (mRMR). Dựa trên tập dữ liệu của Lv, Jia và cộng sự [7] đã phát triển một công cụ dự đoán có tên iCar-PseCp bằng cách sử dụng thuật toán random forest (RF). Sau đó, Hasan và cộng sự [8] đã xây dựng một công cụ dự đoán dựa trên SVM được gọi là predCar-site với kết quả rất ấn tượng . Hơn nữa, Xu và cộng sự [9] đã phát triển một phần mềm độc lập dựa trên SVM được gọi là PTMPred để dự đoán tất cả các loại vị trí PTM bao gồm cả vị trí cacbonyl hóa protein. Bên cạnh đó, Lv và cộng sự [10] cũng đã phát triển một công cụ dự đoán tính toán có tên là CarPred.Y dựa trên SVM để dự đoán các vị trí cacbonyl hóa trong nấm men với một số loại đặc trưng. Kết hợp với profile hidden Markov model, Kao và cộng sự [11] đã phát triển một mô hình tích hợp tên là MDD-carb bằng cách sử dụng các đặc trưng đa dạng để xác định các sites cacbonyl hóa protein trong protein của động vật có vú với substrate(cơ chất) motifs(= pattern). Trong cùng năm đó, Weng và cộng sự [12] đã xây dựng các mô hình dự đoán để xác định vị trí cacbonyl hóa của protein ở người. Tác giả của bài báo “iCarPS: a computational tool for identifying protein carbonylation sites by novel encoded features” [13], Dan Zhang đưa ra một mô hình với cách trích chọn đặc trưng mới lạ dựa trên toạ độ cực và phân loại với thuật toán random forest.

Một số mô hình như CarSPred, PTMPred and CarPred.Y đã đạt được độ chính xác rất cao (AUC ~ 1), nhưng theo tác giả của bài báo "iCarPS: a computational tool for identifying protein carbonylation sites by novel encoded features", Dan Zhang, những mô hình này có thể bị overfitting do những phương pháp đó được xây dựng dựa trên số chiều đặc trưng lớn và không có tập data độc lập để đánh giá.
Bên cạnh đó, các phương pháp này mặc dù đem lại những kết quả rất tốt nhưng tác giả cũng chỉ ra nó còn có một số hạn chế, thiết sót:

- Ít máy chủ web được thành lập: chỉ có ba máy chủ web là iCar-PseCp, predCarsite và MDD-carb (hai công cụ dự đoán trực tuyến cuối hiện không hoạt động). Những công ty khác như CarSPred, PTMPred và CarPred.Y chỉ cung cấp các công cụ dự đoán độc lập mà hầu hết các học giả sinh hóa khó sử dụng chúng
- Hiệu suất của các yếu tố dự báo này có thể được cải thiện hơn từ các chỉ số đánh giá khác nhau. Mặc dù một số công trình báo cáo độ chính xác cao (AUC ~ 1), nó có thể là do tập dữ liệu đào tạo quá phù hợp. Hơn nữa, không có dữ liệu độc lập nào được sử dụng để thực hiện kiểm tra
- Một số bộ dữ liệu thử nghiệm độc lập được xây dựng để xác nhận hiệu suất của các yếu tố dự báo
- Hầu hết các phương pháp này không chọn được các đối tượng tối ưu bằng cách sử dụng các kỹ thuật trích chọn đặc trưng, đây là một trong những bước quan trọng nhất của việc xây dựng một mô hình dự đoán hiệu quả và mạnh mẽ

Tại đây, ta cố gắng để tối ưu 4 nhược điểm nói trên bằng cách sử dụng tập train và tập test có chất lượng cao kết hợp sử dụng residues conical coordinates với 9 đặc điểm của axit amin. Trích ra đặc trưng của mẫu cacbonyl hóa và không cacbonyl hóa bằng cách sử dụng F-score để tối ưu hóa đặc trưng và Random Forest để thực hiện phân loại. Cuối cùng, ta kiểm tra phương phát bằng cách sử dụng cross-validation test và các independent data test.

4. Đề xuất mô hình

4.1. Các đặc tính hóa lý (PCPs)

Mỗi gốc axit amin có nhiều đặc tính lý hóa và sinh học cụ thể, có thể ảnh hưởng đến tính chất của protein và đóng một vai trò quan trọng trong việc xác định cấu trúc và chức năng của protein. Trong nghiên cứu này, chúng tôi đã sử dụng chín đặc tính hóa lý được sử dụng trong tài liệu tham khảo trước đó, bao gồm tính kỵ nước, tính ưa nước, khối lượng, pK1, pK2, pl, độ bền, tính linh hoạt, không thể thay thế, trong đó sáu tính chất đầu tiên đã được phổ biến rộng rãi. Sau đây chúng tôi sẽ giới thiệu sơ lược về ba đặc tính lý hóa cuối cùng (tính cứng, tính linh hoạt, tính không thể thay thế):

- Rigidity và flexibility của chuỗi bên axit amin đã được ước tính đối với polypeptit và các miền protein cục bộ liên quan đến sự thay đổi tính chất của protein, hai đặc tính này cũng được sử dụng để dự đoán sự thay đổi cấu trúc và nếp gấp của protein.
- Bên cạnh đó, trong quá trình phát triển, một số chất cặn(residue) rất dễ được thay thế, trong khi những chất khác rất khó - và sự suy giảm đột biến trung bình (AMD) của các axit amin có thể được sử dụng để mô tả khả năng không thể thay thế của chúng. Do đó, tính không thể thay thế (irreplaceability) phản ánh sự suy thoái mang tính đột biến trong quá trình tiến hóa của sự sống

⇨ 3 đặc tính này có thể đóng một vai trò bổ sung quan trọng cho các đặc tính khác để mô tả các tính năng của protein hoặc peptit. Đáng chú ý, các giá trị của đặc tính hóa lý của axit amin giả X được xác định bằng 0 trong công trình của chúng tôi. Tất cả các giá trị ban đầu của các đặc tính hóa lý này có thể được tìm thấy trong: http://lingroup.cn/server/iCarPS/download.html

Quá trình xử lý không thứ nguyên(dimensionless) của tất cả các giá trị ban đầu phải được thực hiện trước khi sử dụng các giá trị của 9 đặc tính hóa lý của axit amin, được trình bày như sau:

$$P_v(R_i) = \frac{P_v(R_i) - < P_v >}{SD(P_v)}$$

Trong đó:

- $P_v(R_i)$: giá trị của đặc tính hóa lý axit amin cục bộ thứ v đối với cặn-residue R_i ở vị trí i
- <> nghĩa là giá trị trung bình của các axit amin
- SD biểu thị độ lệch chuẩn

Theo đó, mỗi mẫu trình tự protein cacbonyl hóa (hoặc không cacbonyl hóa) $P_\xi(\mathbb{U})$ ở phương trình (2) có thể được ký hiệu là vectơ $n \times L$ chiều, được hiển thị như sau:

$$P_\xi(\mathbb{U}) = [x_1, x_2, \cdots, x_n, \cdots, x_{n \times L}]^T$$

Trong đó:

- n: là số đặc tính hóa lý
- L: là độ dài của trình tự protein, ký hiệu là toán tử chuyển vị "T"
- x: phần tử biểu thị các giá trị của đặc tính hóa lý trên vị trí tương ứng của gốc axit amin dọc theo trình tự protein.

Do đó, giá trị n trong phương trình bằng 9 và chiều dài của mỗi chuỗi protein L bằng 27, như được sử dụng trong CarSPred. Sau đó, mỗi axit amin được xây dựng thành 9 đặc điểm khác nhau. Đối với một đoạn peptit, vectơ đặc trưng 243 chiều ($27 \times 9 = 243$) thu được từ sơ đồ mã hóa này.

4.2. Lược đồ mã hóa sử dụng biểu đồ nón (CC)

Ai cũng biết rằng một số axit amin có những đặc điểm tương tự. Nên ta sẽ chia nó vào các nhóm. Trong bài này, 20 axit amin sẽ được chia vào 4 class, như ở bảng 2:

Group	Description	Amino axit
Class I	non-polar residutes	A, V, L, I, P, F, W, M
Class II	polar residues	G, S, T, C, Y, N, Q
Class III	basic residues	K, R, H
Class IV	acidic residues	D, E

1. Chất cặn (residues) không phân cực: A, V, L, I, P, F, W, M
2. Chất cặn phân cực: G, S, T, C, Y, N, Q
3. Chất cặn cơ bản: K, R, H
4. Chất cặn có tính axit: D, E

Và mỗi amino axit đó sẽ được thể hiện trong 1 không gian 3 chiều P(x,y,z), và sự dụng tọa độ hình nón để thể hiện trình tự protein. Từ Oxy -> P(x,y,z) nón:

$$\begin{cases} x = r \times sin\varphi \times cos\theta \\ y = r \times sin\varphi \times sin\theta \\ z = r \times cos\theta \end{cases} \quad \varphi \in [0,\pi], \theta \in [0,2\pi]$$

Để nắm bắt các đặc điểm chính của protein 1 cách đơn giản và hiệu quả, 2 giả thuyết được đưa ra:

1. Amino axit cùng nhóm -> phân bố trên cùng 1 mặt nón vì nó có các đặc điểm giống nhau
   - Ví dụ 1: Class I – chất cặn không phân cực - A, V, L, I, P, F, W, M ($P_{1j}, j = 1,2,3,..,8$) được cố định với mặt nón với $\varphi = \varphi_1$.
   - Ví dụ 2: Class II – chất cặn phân cực - G, S, T, C, Y, N, Q ($P_{2j}, j = 1,2,3,..,7$) được cố định với mặt nón với $\varphi = \varphi_2$.
2. Để thể hiện sự khác biệt giữa các amino axit, r = trọng lượng phân tử của amino axit.
   http://lin-group.cn/server/iCarPS/download.html

⇨ Hệ trục tọa độ nón + 9 tính chất hóa lý của axit amin, mỗi axit amin sẽ có dạng:

$$\begin{cases} x_{ij} = r_{ij} \times sin\varphi_i \times cos\theta_{ij} \\ y_{ij} = r_{ij} \times sin\varphi_i \times sin\theta_{ij} \\ z_{ij} = r_{ij} \times cos\theta_{ij} \end{cases} \quad \varphi_i \in [0,\pi], \theta_{ij} \in [0,2\pi]$$

Trong đó:

- r_{ij}: Khối lượng phân tử của amino axit ở class I
- L_i: Số lượng amino axit ở class i
- φ_i, θ_{ij}:

$$\varphi_i = \pi \times \left| sin \frac{\overline{d_i}}{\left(\frac{1}{4}\sum_{i=1}^{4}\overline{d_i}\right) * \sqrt{\frac{1}{4}\sum_{i=1}^{4}\left(\overline{d_i} - \frac{1}{4}\sum_{i=1}^{4}\overline{d_i}\right)^2}} \right|$$

$$\theta_{ij} = \pi + 2 \times arctan \frac{\sum_{m=1}^{9} PC_{jm} - \overline{d_i}}{\sqrt{\frac{1}{L_i}\sum_{j=1}^{L_i}\left|\sum_{m=1}^{9} PC_{jm} - \overline{d_i}\right|^2}}$$

  - $\overline{d_\iota} = \frac{1}{L_i}\Sigma_{j=1}^{L_i}\Sigma_{m=1}^{9} PC_{jm}$ với PC_{jm} là giá trị tiêu chuẩn của đặc tính hóa lý thứ m của amino axit thứ j của class i.
  - $\overline{\boldsymbol{d_\iota}}$: là tổng 9 giá trị đặc tính hóa lý của tất cả các amino axit trong nhóm đó

Do đó, mỗi amino axit được biểu diễn ở dạng vector 3 chiều (x,y,z)

⇨ $P\xi(\mathbb{U})$ ở phương trình 2 sẽ được chuyển thành vector $3 \times L = 3 \times (2\xi + 1)$ chiều

$$\boldsymbol{P_\xi(\mathbb{U})} = [x_1, y_1, z_1, x_2, y_2, z_2, \cdots, x_L, y_L, z_L]^T$$

Trọng tâm của hình: $(\bar{x}, \bar{y}, \bar{z})$ của mẫu $P\xi(\mathbb{U})$ sẽ có dạng:

$$\overline{x} = \frac{1}{L}\sum_{n=1}^{L} x_n, \overline{y} = \frac{1}{L}\sum_{n=1}^{L} y_n, \quad \overline{z} = \frac{1}{L}\sum_{n=1}^{L} z_n$$

Trọng tâm của hình tích lỹ: $(\bar{X}, \bar{Y}, \bar{Z})$ của mẫu $P\xi(\mathbb{U})$ sẽ có dạng:

$$\overline{X} = \frac{1}{L}\sum_{h=1}^{L} X_h, \quad \overline{Y} = \frac{1}{L}\sum_{h=1}^{L} Y_h, \overline{Z} = \frac{1}{L}\sum_{h=1}^{L} Z_h$$

in which, $X_h = \sum_{n=1}^{h} x_n, \quad Y_h = \sum_{n=1}^{h} y_n, \quad Z_h = \sum_{n=1}^{h} z_n.$

Trọng tâm của hình: $(\overline{x_\iota}, \overline{y_\iota}, \overline{z_\iota})$ của class i (1,2,3,4) trong mẫu $P\xi(\mathbb{U})$:

$$\overline{x_i} = \frac{1}{L_i}\sum_{n=1}^{L_i} x_n, \overline{y} = \frac{1}{L_i}\sum_{n=1}^{L_i} y_n, \quad \overline{z} = \frac{1}{L_i}\sum_{n=1}^{L_i} z_n$$

Cuối cùng, $P\xi(\mathbb{U})$ có dạng là một vectơ 18 chiều, được sử dụng làm vectơ đặc trưng để mô tả định lượng các đặc tính nội tại của mẫu protein.

$$P_\xi(\mathbb{U}) = [\overline{x_1}, \overline{y_1}, \overline{z_1}, \overline{x_2}, \overline{y_2}, \overline{z_2}, \overline{x_3}, \overline{y_3}, \overline{z_3}, \overline{x_4}, \overline{y_4}, \overline{z_4}, \overline{x}, \overline{y}, \overline{z}, \overline{X}, \overline{Y}, \overline{Z}]^T$$

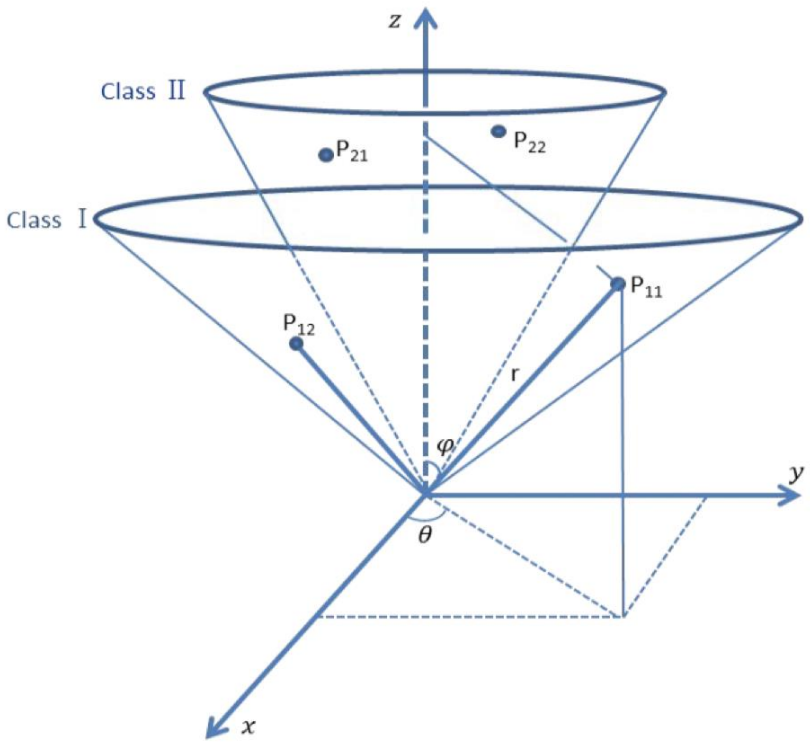

Figure 2 Sơ đồ minh họa để thể hiện biểu diễn hình nón 3 chiều để đặc trưng cho dư lượng axit amin. Class I, II lần lượt là viết tắt của nhóm cặn không phân cực và nhóm cặn phân cực. Pi,j đại diện cho mỗi axit amin của nhóm tương ứng, trong đó i biểu thị nhóm thứ i và j biểu thị axit amin thứ j của nhóm tương ứng. φ đại diện cho bề mặt hình nón được hình thành bằng cách chiếu các axit amin của nhóm tương ứng. Ngoài ra, đại diện cho trọng lượng phân tử r của axit amin

4.3. Thuật toán Random forest (RF)

Rừng ngẫu nhiên (RF) là một phương pháp học tập tổng hợp bao gồm nhiều cây quyết định riêng lẻ, chủ yếu được sử dụng trong hồi quy và phân loại. Nó sử dụng kỹ thuật lấy mẫu lại bootstrap để tạo tập dữ liệu huấn luyện mới được lấy mẫu ngẫu nhiên từ tập dữ liệu huấn luyện ban đầu và được sử dụng để đánh giá tại mỗi nút của cây quyết định. Sau đó, quyết định cuối cùng được đưa ra bằng quyết định hợp nhất tất cả các cây bằng biểu quyết đa số. Ngoài ra, RF được coi là một công cụ phân loại thích hợp để xử lý tập dữ liệu ở quy mô lớn, đặc biệt là đối với tập dữ liệu không cân bằng. Do đó, RF có một số ưu điểm riêng như khả năng chống nhiễu tốt và dễ dàng song song hóa nên nó được sử dụng rộng rãi để xây dựng các mô hình dự đoán tính toán để giải quyết các vấn đề tin sinh học.

Trong nghiên cứ tại bài báo gốc của tác giả, mô hình dự đoán để xác định các vị trí cacbonyl hóa protein dựa trên thuật toán RF được xây dựng bằng giao diện lập trình ứng dụng java gọi thư viện chương trình RF, được tích hợp trong gói khai thác dữ liệu WEKA [14].Các tham số mặc định của rừng ngẫu nhiên đã được sử dụng để xây dựng mô hình trong Weka 3.8.
Do hạn chế trong về mặt kiến thức khi tiếp cận một lĩnh vực mới, nên hiện tại báo cáo đang dừng ở việc đọc hiểu bài báo gốc của tác giả, research thêm một số bài báo trong lĩnh vực liên quan đông thời thử với một số thuật toán Machine Learning khác để thử nghiệm các các đặc trưng được trích chọn như SVM, Decision Tree.

4.4. Support Vector Machine

Xét bài toán tách tập các training vectors vào hai lớp riêng biệt , $(x_1, y_1, z_1),\ldots (x_n, y_n, z_n)$ trong đó $x_i \in R^p$ và $y_i \in \{-1,1\}$ là nhãn lớp tương ứng 1<=i<=n. Nhiệm vụ chính của bài toán này là tìm một bộ phân loại có hàm quyết định f (x, θ) sao cho y = f (x, θ), trong đó y là nhãn lớp của x và θ là vectơ chưa biết tham số của chức năng quyết định. Máy vector hỗ trợ là một máy phân loại nổi tiếng và nó đã được ứng dụng rộng rãi trong nhiều bài toán phân loại. Về mặt hình học, thuật toán mô hình hóa SVM tìm thấy một siêu phẳng tối ưu với biên lớn nhất để tách hai lớp, điều này yêu cầu giải quyết vấn đề ràng buộc sau:

$$\text{minimize}_{w,b}\frac{1}{2}\|w\|^2$$

Subject to:

$$y_i(w^T x_i + b) \geq 1, i = 1,2,3,\ldots\ldots,n$$

Sử dụng phương pháp nhân Lagrange, chúng ta có thể thu được công thức kép được biểu diễn dưới dạng các biến α_i. Cuối cùng, hàm phân biệt tuyến tính có dạng sau :

$$f(x) = \Sigma_i^n \alpha_i x_i^T x + b$$

Trong nhiều ứng dụng, bộ phân loại phi tuyến tính cung cấp độ chính xác tốt hơn. Trong SVM, cách đơn giản để tạo bộ phân loại phi tuyến tính ra khỏi bộ phân loại tuyến tính là ánh xạ dữ liệu của chúng ta từ không gian đầu vào X sang không gian đặc trưng F bằng cách sử dụng hàm phi tuyến tính ∅: X → F. Trong không gian F, tối ưu hóa có dạng sau bằng cách sử dụng hàm nhân, hàm phân biệt sẽ có dạng:

$$f(x) = \Sigma_i^n \alpha_i k(x, x_i) + b$$

Radial basis function kernel đã được sử dụng để xây dựng bộ phân loại SVM được định nghĩa dưới đây:

$$K(x_i, x_j) = \exp(-\frac{\|x_i - x_j\|^2}{\sigma}),\ \sigma \text{ is the width of the function}$$

4.5. Decision Tree

Cây quyết định (Decision Tree) là một cây phân cấp có cấu trúc được dùng để phân lớp các đối tượng dựa vào dãy các luật. Các thuộc tính của đối tượng có thể thuộc các kiểu dữ liệu khác nhau như Nhị phân (Binary) , Định danh (Nominal), Thứ tự (Ordinal), Số lượng (Quantitative) trong khi đó thuộc tính phân lớp phải có kiểu dữ liệu là Binary hoặc Ordinal.

5. Thực hiện hệ thống

5.1. Xử lý dữ liệu đầu vào

Xem xét số lượng vị trí cacbonyl hóa trên dư lượng axit amin H, C và W trên bộ dữ liệu điểm chuẩn của CarSPred's là rất nhỏ và không có nguồn dữ liệu công khai đáng tin cậy, tại đây ta chỉ xây dựng mô hình dự đoán về dư lượng K, P, R, T trong nghiên cứu này.

Dữ liệu đầu vào bao gồm 4 tập dữ liệu tương ứng với K, T, P, R. Một chuỗi protein như đã đề cập ở bên trên bao gồm 27 phần tử được trích xuất bằng 1 cửa sổ trượt dài ($2\xi + 1$), dưới đây là ví dụ của 1 chuỗi protein trong tập K (gốc K ở vị trí giữa, vị trí số 14, được tô đậm) – cũng nhắc lại rằng, chất cặn X được thêm vào để chất cặn K nằm vị trí trung tâm và các đặc tính hóa lý của X được đặt bằng 0:

>**sp**|Q8N3Y7|**RDHE2_HUMAN**
GILHAMDGFVDQK**K**KLXXXXXXXXXXX

Với cách mã hóa thứ nhất – sử dụng 9 đặc tính hóa lý PCPs - mỗi axit amin được xây dựng thành 9 đặc điểm. Để cụ thể hơn, một chuỗi XXXXXXXXXXXXMKWVTFISLLFLFSS (27 chất cặn, mỗi chất 9 đặc tính). Vì vậy, đối với một đoạn peptit, vectơ đặc trưng sẽ là 243 chiều ($27 \times 9 = 243$).

Với cách mã hóa thứ hai – sử dụng tọa độ nón 3 chiều CC – một vectơ 18 chiều, được sử dụng làm vectơ đặc trưng để mô tả định lượng các đặc tính nội tại của mẫu protein $P\xi$ ($\mathbb{U}$).

Việc trích chọn đặc trưng này, được lưu dưới dạng file csv để sử dụng cho việc phân lớp ở các bước tiếp theo.

5.2. Thực hiện hệ thống

Với đầu vào là file dạng .fas cho bốn loại K, P, R, T, để thực hiện việc trích chọn đặc trưng bằng việc kết hợp PCPs và CC đã được đề cập tại phần trước, ta sử dụng mã nguồn có tên "featureExtraction.py".

Ban đầu 1 hàm "getSampleSequenceFile" có nhiệm vụ cắt ngắn chuỗi đầu vào thành nhiều seg với độ dài 27, đồng thời kiểm tra xem có có các chất cặn – residues hợp lệ hay không, đầu ra của hàm trả về 3 biến bao gồm: tên file đã tạo-outFile, chú thích của file - noteLine, và sequence.

Tiếp sau đó, hàm "ObtainCCAndPCPFeature" nhận đầu vào là outFile và PCPFile (file chứa thông tin đặc tính hóa lý của từng loại chất). Hàm bao gồm 3 phần:

1. Trích chọn các đặc trưng theo CC bằng cách trích xuất các tọa độ trình tự, tạo giá trị trung bình xyz của từng nhóm axit amin và giá trị trung bình xyz của tất cả các nhóm – 18 chiều
2. Trích xuất các đặc trưng cho 9 đặc tính hóa lý 9_PCPs, bằng cách tính cho từng đặc tính hóa lý và sau đó concat: df_9PCP = pd.concat(df1->df9) – 27x9=243 chiều.
3. Hợp nhất 2 lọa được trưng để tạo giá trị cuối cùng.

Kết quả, mỗi chuỗi sẽ được biểu diễn dưới dạng 1 vector 18+243=261 chiều và được lưu dưới dạng file csv có tên là "test_fscore_sorted".

5.3. Xử lý kết quả đầu ra

Sử dụng các đặc trưng đã lưu ở file csv, triển khai mô hình sử dụng thư viện sklearn với các thuật toán Random Forest mà tác giả đã đề xuất và thêm vào đó là SVM và Decision Tree. Với dữ liệu đầu vào là một chuỗi được mã hoá thành vector 261 chiều, dữ liệu đầu ra là giá trị của y, trong đó y = 1 ứng với gốc đó bị carbonylation hoá và y = 0 ứng với gốc đó không bị carbonylation hoá. Tập train và tập test đã được chia độc lập, sau khi huấn luyện mô trình sử dụng bảng đánh giá classification_report đã được cài đặt sẵn trong mô hình để đưa ra các chỉ số đánh giá precision, recall, f1-score, accuracy.

Chi tiết được thực hiện trong mã nguồn có tên "iCarPS_offline.py". Với đầu vào là dự đoán 1 cho 4 loại K, P, R, T, ứng mỗi mỗi loại hàm "run_rf_predict" sẽ nhận đầu vào lần lượt:

- inSeqFile: chuỗi cần dự đoán
- predType: loại chất cần dự đoán
- PCPFile: file các chỉ số đặc tính hóa lý ứng với mỗi chất
- Fscorefile: file đặc trưng được sắp xếp bằng cách sử dụng fscore dạng [K,P,R,H]_sort.fscore
  Ví dụ: Irreplaceability_4 0.002541
  Mass_10 0.002596
- paraFile: file tham số chứa các top đặc trưng dạng [K,P,R,H].para
- csv_file: file các giá trị fscore đã được sắp xếp
  Ví dụ: -1.498 -0.844 -0.836 0.95 0.241 -0.894

Đầu ra hàm trả về biến "finallyresultfile" gồm 2 phần:

1. Carbonylation Site: Trả về các chuỗi gồm 27 chất cặn và vị trí chất cặn cần được dự đoán bị cacbon hóa.
   Carbonylation Site:
   >P01024_HUMAN Postion:678 is K PAARRRRSVQLTEKRMDKVGKYPKELR
2. Non- carbonylation Site: Trả về các chuỗi gồm 27 chất cặn và vị trí chất cặn cần được dự đoán mà không bị cacbon hóa.

Non-carbonylation Site:
>P01023_HUMAN Postion:3 is K XXXXXXXXXXXMGKNKLLHPSLVLLLL
>P01023_HUMAN Postion:5 is K XXXXXXXXXMGKNKLLHPSLVLLLLVL

6. Thử nghiệm và kết quả

6.1. Môi trường thử nghiệm

- Cấu hình máy tính: Môi trường colab: Intel(R) Xeon(R) CPU @ 2.30GHz,13GB RAM, non-GPU
- Các phần mềm, bộ công cụ toolkits đã sử dụng: sklearn, numpy
- Mã nguồn bài báo gốc: https://github.com/zhangdan0/iCarPS

6.2. Các thử nghiệm

- Thử nghiệm với thuật toán Decision Tree

Tập K

n_estimators	30	100	500	700	1000	2000
accuracy	0,54	0,44	0,54	0,55	0,54	0,53
positive recall	0,42	0,60	0,49	0,42	0,45	0,50

0.53

Tập P

n_estimators	30	100	500	700	1000	2000
accuracy	0,57	0,43	0,48	0,45	0,38	0,41
positive recall	0,73	0,36	0,82	0,64	0,64	0,91

0.64

Tập R

n_estimators	30	100	500	700	1000	2000
accuracy	0,55	0,57	0,45	0,48	0,45	0,55
positive recall	0,56	0,50	0,44	0,56	0,50	0,56

0.66

Tập T

n_estimators	30	100	500	700	1000	2000
accuracy	0,52	0,48	0,39	0,45	0,45	0,30
positive recall	0,75	0,25	0,25	0,50	0,50	0,50

0.67

- Thử nghiệm với thuật toán SVM, C là hằng số phạt trong SVM lề mềm

Tập K

C	100	500	800	1000	1500
accuracy	0,27	0,21	0,19	0,17	0,15

positive recall	0,70	0,70	0,70	0,69	0,68

0.51

Tập P

C	100	500	800	1000	1500
accuracy	0,73	0,73	0,78	0,74	0,74
positive recall	0,45	0,55	0,64	0,55	0,45

0.72

Tập R

C	100	500	800	1000	1500
accuracy	0,67	0,69	0,70	0,69	0,69
positive recall	0,44	0,38	0,38	0,38	0,38

0.56

Tập T

C	100	500	800	1000	1500
accuracy	0,88	0,88	0,88	0,88	0,88
positive recall	0,75	0,75	0,75	0,75	0,75

0.82

7. Thảo luận

Kết quả thống kê trên nhắc chúng ta rằng các vị trí cacbonyl hóa có thể được xác định bằng phương pháp tính toán dựa trên các đặc điểm mà ta đã đề cập ở các mục trên. Để đánh giá hiệu suất dự đoán của các đặc trưng khác nhau, các mô hình dự đoán đã được train và test bằng 10-fold cross-validation dựa trên RF. Lý do tại sao chúng tôi sử dụng giá trị AUC làm tiêu chuẩn là nó có thể cung cấp một đánh giá khách quan hơn trên tập dataset mất cân bằng so với độ nhạy- sensitivity, độ đặc hiệu- specificity và độ chính xác tổng thể- overall accuracy.

Sử dụng 2 loại đặc trưng, 9 đặc tính hóa lý của amino axit(9_PCPs) và tọa độ nón 3 chiều(CC) để tạo mẫu. Dựa vào 10-fold cross-validation test, hiệu suất dự đoán của từng đặc trưng dựa trên trình tự.

- 9_PCPs có thể tạo ra các giá trị AUC tương ứng là **0,741, 0,727, 0,580 và 0,626** cho K, P, R, T carbonylation sites.
- mã hóa CC thu được các giá trị AUC là **0,725, 0,786, 0,661, 0,735** cho các dự đoán vị trí cacbonyl hóa K, P, R và T carbonylation sites.

Hai đặc trưng có thể mô tả các mẫu cacbonyl hóa từ quan điểm thành phần trình tự và các đặc tính hóa lý. Do đó, chúng tôi đoán rằng sự kết hợp giữa hai tính năng có thể cải thiện hiệu suất dự đoán. Tuy nhiên, sau khi thực hiện kiểm tra, chúng tôi nhận thấy rằng việc kết hợp 2 đặc trưng này không thể cải thiện hiệu suất cho tất cả các vị trí cacbonyl hóa:

- Việc kết hợp hiệu quả với K,P,R, T với AUC = **0,775, 0,765, 0,662, 0,745**

Tóm lại, nhiễu hoặc thông tin dư thừa có thể làm giảm hiệu suất, độ mạnh mẽ và hiệu quả của mô hình. Do đó, hiện tượng giảm độ chính xác có thể bắt nguồn từ sự dư thừa thông tin => cần phải chọn ra các đặc trưng tốt nhất để cải thiện độ chính xác của dự đoán. Vì vậy trong bài báo gốc, tác giả đã đề cập đến việc sử dụng F-score để tối ưu hóa các đặc trưng.
Sau khi tối ưu, với 10-fold-cross-validation các giá trị AUC tối đa lần lượt là **0,789, 0,814, 0,726 và 0,790** cho các dự đoán vị trí cacbonyl hóa K, P, R và T. Để đánh giá thêm về độ tin cậy- reliable và mạnh mẽ- robust của các optimal(tối ưu) models xây dựng từ các optimal features, bộ independent testing datasets được sử dụng. Các giá trị AUC trên independent data đạt tới **0,756, 0,752, 0,649 và 0,840**, để dự đoán vị trí cacbonyl hóa K, P, R và T.

Bên cạnh đó, sinh viên Mạc Quang Huy cũng đề xuất và thử nghiệm 2 phương pháp sau khi hoàn thành việc trích chọn đặc trưng của tác giả. Desition Tree được thử nghiệm và đem lại kết quả **0,53, 0,54, 0,67, 0,69** để dự đoán vị trí cacbonyl hóa K, P, R và T. Và SVM đem lại kết quả khá ấn tượng cho việc dự đoán chất cặn T với kết quả lần lượt là **0,51, 0,72, 0,56, 0,82** cho dự đoán vị trí cacbonyl hóa K, P, R và T. Chi tiết được thể hiện tại bảng sau.

Predictor	AUC for Carbonylaytion Sites			
	K	P	R	T
origin -iCarPS	0,75	0,75	0,65	0,84
using-SVM-iCarPS	0,51	0,72	0,56	0,82
using-desition-tree-iCarPS	0,53	0,54	0,67	0,69
CarSpred	0,67	0,78	0,53	0,68

8. Tổng kết và phương hướng phát triển (Conclusions and Perspectives)

Mô hình trong bài tập lớn đề xuất là sử dụng cách trích chọn đặc trưng kết hợp giữa biến đổi toạ độ cực và 9 đặc trưng hoá lý của amino axit, sau đó sử dụng bộ phân lớp SVM và Desition Tree để phân loại nhãn positive và negative.Có thể thấy bằng việc sử dụng SVM hay Desition Tree để phân lớp thay vì sử dụng Ramdom Forest cũng đã đem lại những kết quả khá tích cực, điển hình ở việc sử dụng SVM để dự đoán loại T với AUC = 0,82, hay với Desition Tree để dự đoán loại R cũng đem lại kết quả tích cực hơn so với bài báo gốc với AUC = 0,67. Tuy nhiên, nhận thấy tại bộ dữ liệu mất cân bằng và có nhiều đặc tính, cá nhân nhận thấy việc trích chọn đặc trưng mới là phần quan trọng nhất trong việc dự đoán các chất bị cacbon hóa. Tương lai, project sẽ tập trung vào việc trích chọn đặc trưng bằng cách kết hợp với một số kỹ thuật deep learning hiện đại nhằm tìm ra một phương pháp tối ưu cho việc dự đoán các chất bị cacbon hóa.

9. Tài liệu tham khảo

[1] Y. D. J. W. L. Y. C. K. C. Xu, " iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition," 2013.

[2] Hasan, "Carbonylation sites prediction in proteins using support vector machine with resolving data imbalanced issue," *Analytical biochemistry,* p. 525, 2017.

[3] Kao, "MDD-carb: a combinatorial model for the identification of protein carbonylation sites with substrate motifs," *BMC systems biology,* p. 11, 2017.

[4] K. C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition," *Journal of theoretical biology,* pp. 236-247, 2011.

[5] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition, O'Reilly Media, Inc., 2019.

[6] H. Lv, "CarSPred: a computational tool for predicting carbonylation sites of human proteins," 2014.

[7] J. Jia, "iCar-PseCp: identify carbonylation sites in proteins by Monte Carlo sampling and incorporating sequence coupled effects into general PseAAC," 2016.

[8] M. Hasan, "predCar-site: Carbonylation sites prediction in proteins using support vector machine with resolving data imbalanced issue," *Analytical biochemistry,* pp. 107-113, 2017.

[9] Y. Xu, "Prediction of posttranslational modification sites from amino acid sequences with kernel methods, Journal of theoretical biology," pp. 78-87, 2014.

[10] H. Lv, "A computational method to predict carbonylation sites in yeast proteins," *Genetics and molecular research : GMR,* 2016.

[11] H. Kao, "MDD-carb: a combinatorial model for the identification of protein carbonylation sites with substrate motifs," *BMC systems biology,* 2017.

[12] S. Weng, " Investigation and identification of protein carbonylation sites based on position-specific amino acid composition and physicochemical features," *BMC bioinformatics,* 2017.

[13] Z.-C. X. W. S. Y.-H. Y. H. L. H. Y. H. L. Dan Zhang, "iCarPS: a computational tool for identifying protein carbonylation sites by novel encoded features," vol. 37, no. 2, pp. 171-177, 2020.

[14] T. a. F. E. Smith, "Introducing Machine Learning Concepts with WEKA," *Methods in molecular biology,* pp. 353-378, 2016.

Mục lục

Contents

Lưu ý:

- Không sử dụng lại các hình vẽ của các nghiên cứu khác. Cần vẽ các hình vẽ theo cách hiểu của riêng mình.
- Không copy lại quá 6 từ liên tiếp từ các báo cáo khác (kể cả chuyển ngữ từ tiếng Anh sang tiếng Việt).
- Các hình vẽ và bảng biểu cần có chú giải chi tiết (tối thiểu 2 dòng) và cần có tham chiếu đến hình vẽ trong đoạn văn tương ứng.
- Các hình vẽ ngoài việc đưa vào báo cáo thì cần để ở định dạng pdf ở các file riêng và nộp kèm theo báo cáo.
- Khuyến khích vẽ các sơ đồ ở trang https://app.diagrams.net/ và nộp file mã nguồn vẽ sơ đồ kèm theo báo cáo.