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

BÁO CÁO PROJECT
DỰ ĐOÁN PROTEIN CARBONYLATION

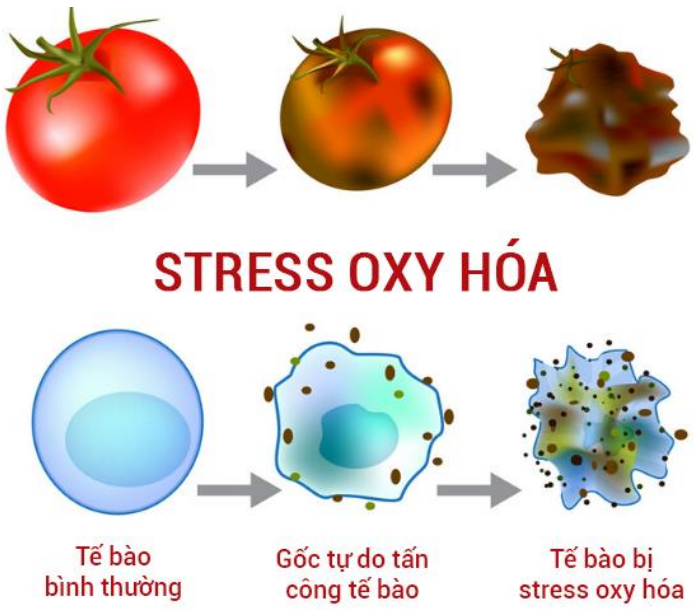
GV: TS. NGUYỄN HỒNG QUANG

SINH VIÊN THỰC HIỆN
MẠC QUANG HUY 20173169

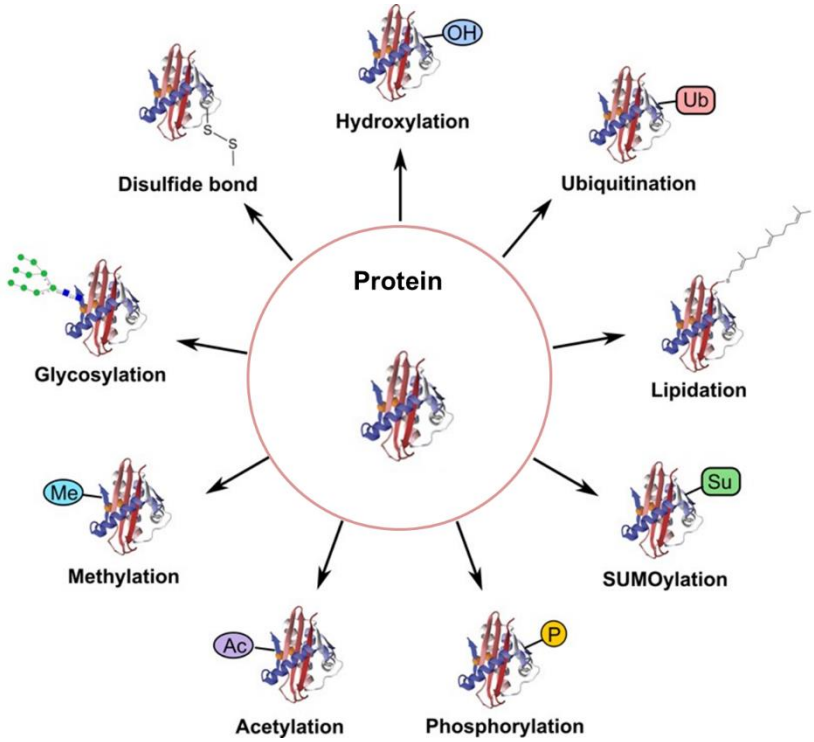
Nội dung trình bày

1. Giới thiệu
2. Phương pháp đề xuất
3. Kết quả và thảo luận
4. Kết luận và định hướng phát triển

1. Giới thiệu



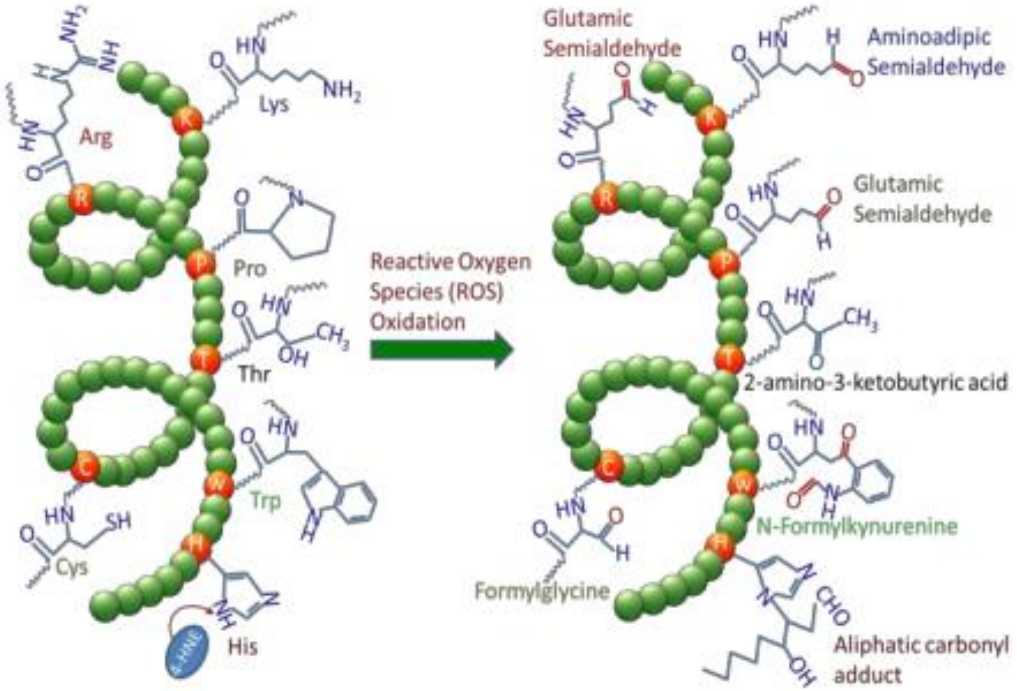
Nguyên nhân

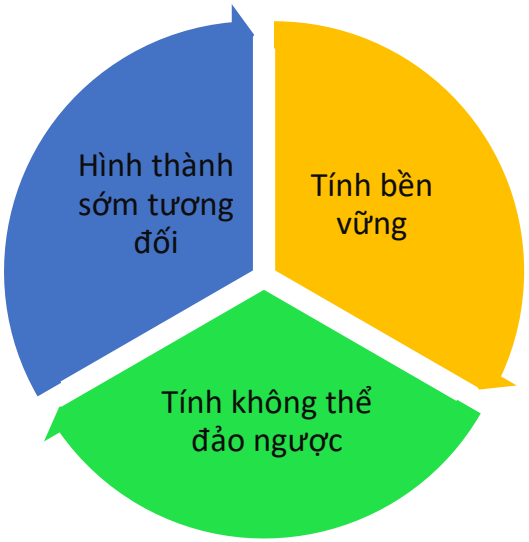


Reactive oxygen species (**ROS**) mất cân bằng với khả năng chống oxy hóa của cơ thể

Các biến đổi sau dịch mã post-translational modifications (**PTMs**) trên protein

1. Giới thiệu



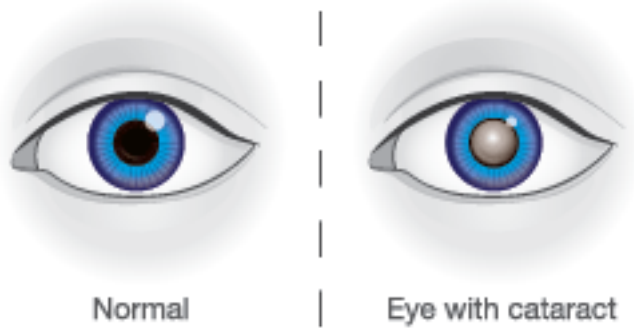


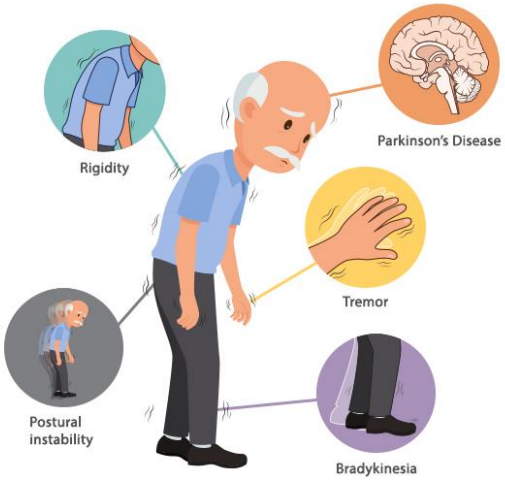
Cacbon hóa protein

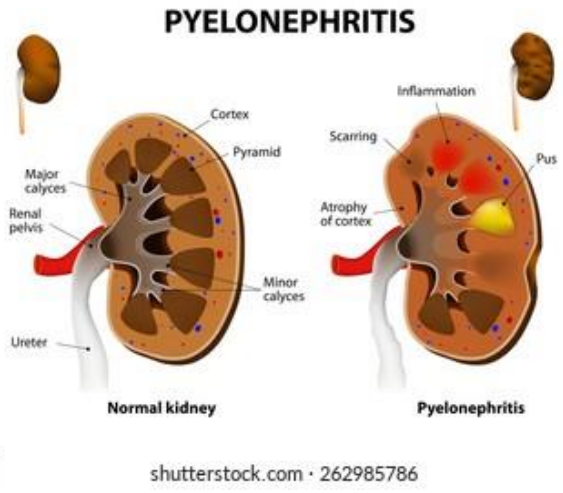
Figure 1 Cấu trúc hóa học của các sản phẩm oxy hóa cacbonyl hóa, bao gồm Aminoadipic Semialdehyde, Glutamic Semialdehyde và axit 2- amino-3-ketobutyric, Formylglycine, N-Formylkynurenine, và hợp chất cacbonyl Aliphatic, được tạo ra bằng cách cacbonyl hóa K, R hoặc P, T, C, W, và H tương ứng, trong đó 4-HNE là đối tượng của phản ứng cộng Michael với chuỗi bên axit amin của H

1. Giới thiệu

Mức độ cacbonyl hóa protein cao trong rất nhiều loại bệnh chính của con người







1. Giới thiệu

Một số nghiên cứu liên quan

CarsPred - Lv và cộng sự 2014

- Xây dựng bộ dữ liệu chuẩn cho chất cặn biến đổi K, P, R, T
- Sử dụng trích chọn đặc trưng mRMR

iCar-PseCp – Jia và cộng sự 2016

- Thuật toán Ramdom Forest

CarPred.Y - Lv và cộng sự 2016

- Dựa trên SVM dự đoán cabon hóa trong nấm men

predCar-site – Hasan và cộng sự 2017

- Kết hợp thông tin liên kết trình tự vào thành phần axit amin giả chung sử dụng SVM

MDD-carb - Kao và cộng sự 2017

- Kết hợp profile hidden Markov model
- Vùng cacbon hóa trong protein động vật có vú

Weng và cộng sự 2017

- Mô hình để dự đoán vùng carbon hóa ở protein người

1. Giới thiệu

Các hạn chế của các công trình trước đó:

1. Ít máy chủ web đang hoạt động
2. Hiệu suất có thể được cải thiện hơn
3. Một số bộ dữ liệu thử nghiệm độc lập đã được xây dựng
4. Chọn các đối tượng tối ưu bằng cách sử dụng trích chọn đặc trưng

⇒***Tập trung vào việc:***

1. Sử dụng tập train và test chất lượng
2. Sử dụng biểu đồ nón (CC) kết hợp với 9 đặc tính hóa lý của axit amin (9_PCPs)
3. Sử dụng F-score để tối ưu hóa đặc trưng + Ramdom Forest để phân lớp
4. Sử dụng 10 cross-validation test + independent data test
5. Triển khai 1 web server

2. Phương pháp đề xuất

2.1. Bộ dữ liệu Benchmark

Bộ dữ liệu điểm chuẩn của CarSPred's (Lv, et al., 2014) đã được sử dụng.

- 230 chuỗi protein cacbonyl hóa từ người
- 20 chuỗi protein cacbonyl hóa từ các động vật có vú

Xây dựng model để dự đoán cho 4 loại residues K, P, R, T tương ứng với 4 tập con:

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

- +: Tập con **dương** gồm các mẫu của **true** carbonylation site cho chất cặn
- -: Tập con **âm** gồm các mẫu của **false** carbonylation site cho chất cặn

Cửa sổ trượt (2ξ + 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

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

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

2. Phương pháp đề xuất

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 $S^{+}_{\circledast}$, 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 $S^{-}_{\circledast}$. 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. Phương pháp đề xuất

1. **Các đặc tính hóa lý (9_PCPs)**

- 9 đặc tính hóa lý bao gồm: tính kỵ nước, tính ưa nước, khối lượng, pK1, pK2, pI, **độ bền, tính linh hoạt, không thể thay thế**
- Quá trình xử lý **không thứ nguyên(dimensionless)**

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

Trong đó:

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

⇒Mỗi mẫu trình tự protein có thể được ký hiệu là vectơ $n \times L$ chiều

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

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.

Ví dụ: GILHAMDGFVDQK**K**KLXXXXXXXXXXX (27 chất cặn, mỗi chất 9 đặc tính=243 chiều)

2. Phương pháp đề xuất

2. Mã hóa dự trên biểu đồ nón (CC)

- 20 axit amin chia vào 4 nhóm

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

- Mỗi axit amin được thể hiện trong 1 không gian 3 chiều P(x,y,z), và sử dụng tọa độ hình nón để thể hiện trình tự protein.

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

- Để nắm bắt các đặc điểm chính của protein 1 cách đơn giản và hiệu quả, 2 giả thuyết được đưa ra:
  - 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 $\varphi = \varphi_2$.
  - 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

2. Phương pháp đề xuất

2. Mã hóa dự trên biểu đồ nón (CC)

- Hệ trục tọa độ nón:

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

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}:

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

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

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

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

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

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

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

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

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

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: $(\overline{x_i}, \overline{y_i}, \overline{z_i})$ của class i (1,2,3,4) trong mẫu $P\xi(\mathbb{U})$:

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

2. Phương pháp đề xuất

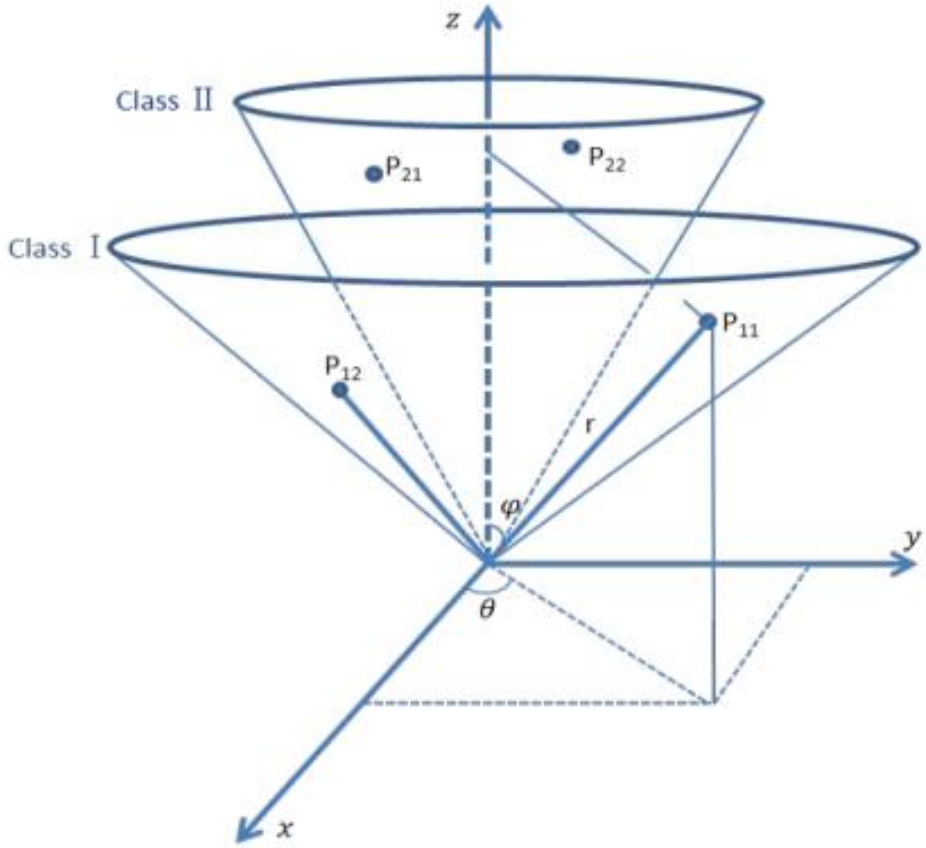


Figure 2 Sơ đồ minh họa để thể hiện biểu diễn hình nón 3 chiều để đặc trưng cho dư lượng axit amin. Loại I, II lần lượt là viết tắt của nhóm cặn không phân cực và nhóm cặn phân cực. Pi, j đại diện cho mỗi axit amin của nhóm tương ứng, trong đó i biểu thị nhóm thứ i và j biểu thị axit amin thứ j của nhóm 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.

2. Phương pháp đề xuất

3. F-score và trích chọn đặc trưng

- Một trích chọn đặc trưng thích hợp không chỉ có **thể khắc phục được kích thước và giảm thời gian đào tạo** mà còn **giảm nguy cơ over-fitting và cải thiện độ chính xác và sức mạnh** của model được đề xuất.

⇒Các trích chọn đặc trưng như**: F-score, mRMR, phân tích phương sai (ANOVA) và phân phối nhị thức (BD), v.v**., áp dụng thành công trong lĩnh vực tin sinh

⇒Cần phát triển **model đơn giản hơn, nhanh** hơn để xác định các vị trí cacbonyl hóa protein bằng cách **sử dụng F-scrore để tối ưu các đặc trưng**

4. Sử dụng Random Forest, SVM, Desition Tree

3. Kết quả

- 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

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

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

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

3. Kết quả

- 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

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

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

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

3. Kết quả

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.
- 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.
- Kết hợp 9_PCPs + CC trả về K,P,R, T với AUC là **0,775, 0,765, 0,662, 0,745**

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.

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

4. Kết luận và định hướng phát triển

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
- 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.

TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI

HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY