**iCarPS: a computational tool for identifying protein**

**carbonylation sites by novel encoded features**

# Abstract

Các biến đổi sau dịch mã do ứng suất oxy hóa gây ra (PTM), đặc trưng bởi:

* Tính bền vững
* Không thể đảo ngược
* Cấu thành sớm tương đối

Nó liên quan đến nhiều loại bệnh khác nhau.Các công nghệ hiện nay: tốn kém, mất thời gian, k xử lý nhiều proteon 1 lúc

=> Việc xác định nhanh, hiệu quả các vị trí cacbonyl hóa bằng pp tính toán cung cấp dấu hiệu của nguồn bệnh

Phương pháp dự báo iCarPS: để xác định carbonylation sites dự trên sequence information:

Một schema encoding feature mới được gọi là tọa độ hình nón chất cặn + đặc tính của chúng -> tạo ra các mẫu carbonylated protein + non-carbonylated protein.

# Introduction

Các loại phản ứng với O2 (ROS) có thể liên tục xảy ra trong tế vào với nhiều nguồn endogenous(nội) và exogenous(ngoại). Và khi **ROS mất cân bằng** với khả năng chống Oxi hóa của cơ thể, **sinh ra stress oxi hóa**. Nó có thể tạo ra nhiều biến đổi sau dịch mã - post-translational modifications (**PTMs**) trên protein (VD: nitrat hóa, cacbonyl hóa, sulfhydryl hóa, hidroxyl hóa,…). Trong đó, **cacbonyl hóa**, dạng biến đổi không thể đảo ngược (irreversible) và không phải enzym (non-enzymatic) rất phổ biến và được nhiều chú ý.

Carbonyl hóa protein (Hình 1) (Rauniyar, et al., 2010), thường được đặc trưng là tính bền vững- **stability**, không thể đảo ngược- **irreversibility** và hình thành sớm tương đối- relative early **formation**, từ lâu đã được nghiên cứu kỹ lưỡng và được coi là dấu ấn sinh học để **đo 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.

Hơn nữa, nhiều nghiên cứu đã chỉ ra rằng mức độ cacbonyl hóa protein cao đã được quan sát thấy trong rất nhiều loại bệnh chính của con người, chẳng hạn như **bệnh phổi mãn tính, bệnh Parkinson, bệnh đục thủy tinh thể, suy thận mãn tính, nhiễm trùng huyết** và nhiều bệnh liên quan đến tuổi tác khác. Do đó, việc nghiên cứu cacbonyl hóa protein có ý nghĩa lớn đối với nghiên cứu y sinh và phát triển thuốc. Đặc biệt, việc **identification** các **protein carbonylation sites** có thể cung cấp manh mối quan trọng **để hiểu process** **và** **consequences** (hậu quả) của **quá trình chuyển hóa tế bào và protein bị ảnh hưởng**.

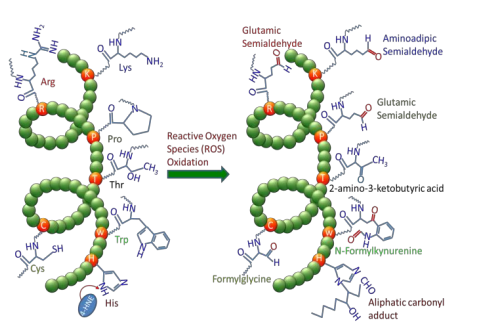


Figure 1 Cấu trúc hóa học của các sản phẩm oxy hóa cacbonyl hóa, bao gồm Aminoadipic Semialdehyde, Glutamic Semialdehyde và axit 2- amino-3-ketobutyric, Formylglycine, N-Formylkynurenine, và hợp chất cacbonyl Aliphatic, được tạo ra bằng cách cacbonyl hóa K, R hoặc P, T, C, W, và H tương ứng, trong đó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

Những năm gần đây, nhiều phương pháp thử nghiệm đã được sử dụng để định lượng quá trình cacbonyl hóa protein và xác định vị trí cacbonyl hóa. Một số kết quả quan tâm đã 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.

//Đoạn văn dưới nói về các nghiên cứu liên quan

**Xác định chính xác vị trí sửa đổi** có thể cung cấp manh mối quan trọng **để hiểu ROS và các loại bệnh**. Tuy nhiên, sẽ tốn kém nếu chỉ sử dụng kỹ thuật thí nghiệm sinh hóa thuần túy 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 thông minh nhân tạo cung cấp cho chúng tôi 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. **Lv và cộng sự**. (Lv, et al., 2014) 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** (Jia, et al., 2016) 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 (Hasan, và cộng sự, 2017) đã xây dựng một công cụ dự đoán dựa trên SVM được gọi là **predCar-site** . Hơn nữa, **Xu** và cộng sự (Xu, và cộng sự, 2014) 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 (Lv, et al., 2016) 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 (Kao, et al., 2017) 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**](http://tratu.soha.vn/dict/en_vn/Pattern)**).** Trong cùng năm đó, **Weng** và cộng sự (Weng, et al., 2017) đã **xây dựng các mô hình dự đoán để xác định vị trí cacbonyl hóa của protein ở người.**

//Đoạn này nêu các điểm yếu của các phương pháp trên

Mặc dù các công trình trước đã đạt được những kết quả to lớn nhưng rất tiếc vẫn còn một số hạn chế, thiếu 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 **(caused by** **an over-fitting of the training dataset because the proposed method was constructed based on high dimensional feature set).** 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 đã 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

//Bài này: Tối ưu 4 nhược điểm trên.

-> Sử dụng tập train và test chất lượng cao

-> Sử dụng **residues conical coordinates** với 9 đặc điểm của axit amin

-> Trích ra đặc trưng của mẫu cacbonyl hóa và không cacbonyl hóa

-Điểm F: tối ưu hóa đặc trưng

-Random Forest algorithm: thực hiện phân loại

-> Test phương pháp: sử dụng cross-validation test + independent data test

-> Web server: iCarPS

Trong **nghiên cứu này**, chúng tôi dành để **cải thiện khả năng dự đoán** trong việc xác định các vị trí cacbonyl hóa protein **từ bốn nhược điểm nói trên.** Ban đầu, tập train và tập test đã được xác minh bằng thực nghiệm. Sau đó, một sơ đồ mã hóa mới được gọi là tọa độ hình nón chất cặn (**residues conical coordinates**) dựa trên cấu trúc toán học đã được đưa vào 9 đặc tính hóa lý của axit amin để đặc trưng cho các mẫu cacbonyl hóa và không cacbonyl hóa. Trong khi đó, điểm F được sử dụng để tối ưu hóa các đặc trưng. Và sau đó, thuật toán RF được sử dụng để thực hiện phân loại. Chúng tôi đã sử dụng kiểm tra xác nhận chéo và kiểm tra dữ liệu độc lập để đánh giá phương pháp của mình. Dựa trên mô hình được đề xuất, một máy chủ web có tên iCarPS đã được thành lập.

# Method

## Bộ dữ liệu Benchmark

**Bộ dữ liệu điểm chuẩn của CarSPred’s** (Lv, et al., 2014) đã được sử dụng. Nó được lấy từ **230 chuỗi protein cacbonyl hóa từ người** + **20 chuỗi protein cacbonyl hóa từ các động vật có vú** khác. Các vị trí sửa đổi chứa bốn loại vị trí cacbonyl hóa, K, P, R và T. Xem xét 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**, chúng tôi chỉ xây dựng mô hình **dự đoán về dư lượng K, P, R, T** trong nghiên cứu này.

**Tập dữ liệu chuẩn bao gồm bốn tập con ký hiệu là 𝕊 ⊛** (với ⊛ biểu thị duy nhất 1 chất cặn K, P, R hoặc T)



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

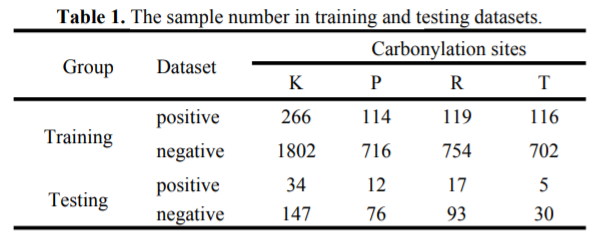
Để xây dựng tập dữ liệu với công thức trên, trước hết, **cửa sổ trượt (2𝜉 + 1)-mer** được sử dụng để **trích xuất các mẫu dương và âm** với 𝕌 = ⊛ ở tâm dọc theo mỗi phân đoạn trình tự protein Do đó, một mẫu trình tự protein chứa vị trí cacbonyl hóa tiềm năng có thể được biểu thị bằng



Trong đó:

* 𝕌 = ⊛ (K, P, R hoặc T)
* 𝜉 : là 1 số nguyên
* : **upstream** (nơi phiên mã sớm hơn) axit amin thứ 𝜉 tính từ trung tâm
* : **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 ,** **nếu các tâm** của chúng là **các vị trí cacbonyl hóa** đã được **xác nhận bằng thực nghiệm.** Nếu không, các phân đoạn trình tự protein được coi là mẫu âm và vào tập âm **.** Các tập dữ liệu chuẩn được xây dựng cuối cùng được tóm tắt trong Bảng 1. Nhận dạng trình tự của tập dữ liệu chuẩn đã được giảm xuống 30% bằng cách sử dụng chương trình CD-HIT.



## Cấu trúc vector đặc trưng

### 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 (độ bền, 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 sự 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:



Trong đó:

* : giá trị của đặc tính hóa lý axit amin cục bộ thứ v đối với cặn- **residue** 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) ở phương trình (2) có thể được ký hiệu là vectơ 𝑛 × 𝐿 chiều, được hiển thị như sau:



Trong đó:

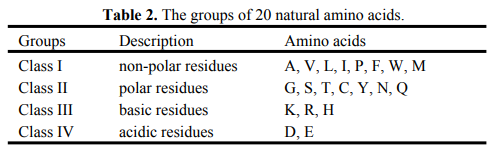
* n: là số đặc tính hóa lý
* L: là độ dài của trình tự protein, ký hiệu là toán tử chuyển vị "𝑇"
* x: phần tử biểu thị các giá trị của đặc tính hóa lý trên vị trí tương ứng của gốc axit amin dọc theo trình tự protein.

Do đó, giá trị n trong phương trình. (4) bằng 9 (9 đặc tính hóa lý) và chiều dài của mỗi chuỗi protein L bằng 27(27 chất cặn), như được sử dụng trong CarSPred. Sau đó, mỗi axit amin được xây dựng thành 9 đặc điểm. Đối với một đoạn peptit, vectơ đặc trưng 243 chiều (27 × 9 = 243) thu được từ sơ đồ mã hóa này.

VD: 1 chuỗi : XXXXXXXXXXXXMKWVTFISLLFLFSS (27 chất cặn, mỗi chất 9 đặc tính)

### Một lược đồ mã hóa mới dựa trên cấu trúc toán học

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:



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:

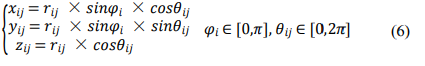


Để nắm bắt các đặc điểm chính của protein 1 cách đơn giản và hiệu quả, 2 giả thuyết được đưa ra:

1. Amino axit cùng nhóm -> phân bố trên cùng 1 mặt nón vì nó có các đặc điểm giống nhau
   * Class I – chất cặn không phân cực - A, V, L, I, P, F, W, M ( ) được cố định với mặt nón với .
   * Class II – chất cặn phân cực - G, S, T, C, Y, N, Q ( ) được cố định với mặt nón với .
2. Để thể hiện sự khác biệt giữa các amino axit, r = trọng lượng phân tử của amino axit.

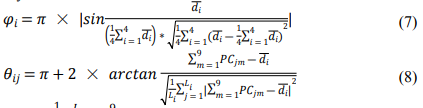
<http://lin-group.cn/server/iCarPS/download.html>

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



Trong đó:

* : Khối lượng phân tử của amino axit ở class I
* : Số lượng amino axit ở class i
* , :



* + với là giá trị tiêu chuẩn của đặc tính hóa lý thứ m của amino axit thứ j của class i.
  + **: 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)

* 𝑃𝜉(𝕌) ở phương trình 2 sẽ được chuyển thành vector 3 × 𝐿 = 3 × (2𝜉 + 1) chiều



Trọng tâm của hình: của mẫu 𝑃𝜉(𝕌) sẽ có dạng:



Trọng tâm của hình tích lỹ: của mẫu 𝑃𝜉(𝕌) sẽ có dạng:



Trọng tâm của hình: của class i (1,2,3,4) trong mẫu 𝑃𝜉(𝕌):



Cuối cùng, 𝑃𝜉(𝕌) có dạng:



* vectơ 18 chiều, được sử dụng làm vectơ đặc trưng để mô tả định lượng các đặc tính nội tại của mẫu protein 𝑃𝜉 (𝕌).

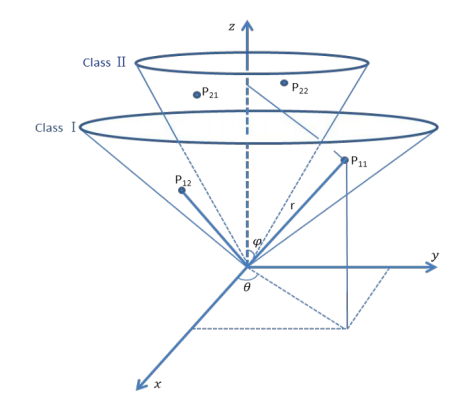


Figure 2 Sơ đồ minh họa để thể hiện biểu diễn hình nón 3 chiều để đặc trưng cho dư lượng axit amin. Loại I, II lần lượt là viết tắt của nhóm cặn không phân cực và nhóm cặn phân cực. Pi, j đại diện cho mỗi axit amin của nhóm tương ứng, trong đó i biểu thị nhóm thứ i và j biểu thị axit amin thứ j của nhóm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

## Ramdom forest (RF)

Random forest (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** (L, 2001). Nó sử dụng kỹ thuật **bootstrap resampling** để tạo ra tập trainning dataset mới – được lấy ngẫu nhiên từ tập train dataset lúc đầu, **để đánh giá tại mỗi node của cây**. Sau đó, **quyết định cuối cùng** được đưa ra bằng quyết định hợp nhất tất cả các cây bằng **voting**. Do khả năng cung cấp phương pháp tiếp cận theo kinh nghiệm đối với các tương tác biến đổi theo đường mòn**, 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** (Livingston, 2005; Zeng, et al., 2019). 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** (Manavalan, et al., 2019; Manavalan, et al., 2019). Các quy trình chi tiết của thuật toán RF và công thức của nó đã được trình bày rất rõ ràng trong nghiên cứu kỹ lưỡng, và do đó không cần phải lặp lại ở đây.

**Trong nghiên cứu này**, mô hình dự đoán để xác định các vị trí cacbonyl hóa protein dựa trên thuật toán **RF được xây dựng bằng application programming interface – 1 thư viện của chương trình RF, được tích hợp trong package WEKA data mining**. Các tham số mặc định của random forest đã được sử dụng để xây dựng model.

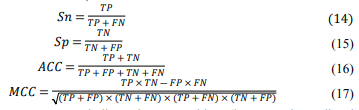
## F-score và trích chọn đặc trưng

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 tổng quát của các mô hình 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**

## Các chỉ số 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:



* 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**.

# Kết quả và thảo luận

## 3.1. Phân tích sự khác biệt theo vị trí cụ thể

Các đặc điểm trình tự theo vị trí cụ thể - **position-specific sequence characteristics**- có thể được công nhận là các **đặc trưng quan trọng** được giữ lại. Để **thể hiện sự khác biệt về vị trí cụ thể giữa các mẫu dương tính và âm tính**, một công cụ dựa trên web được gọi là **Two Sample Logos** (TSL)(Vacic, et al., 2006) đã được sử dụng trong công việc hiện tại **để xác định sự khác biệt về mặt thống kê của các chất cặn – residues xung quanh true(false) carbonylation sites**. Sự khác biệt về phân bố (t test, p value <0,05) của bốn loại chất cặn (Lysine, Proline, Arginine, Threonine) được biểu diễn bằng đồ thị**, trong Hình 3, trong đó residues enriched or depleted (chất cặn đc làm giàu hoặc cạn) xung quanh các vị trí cacbonyl hóa ở trên hoặc dưới trục hoành, tương ứng**.

Hình 3, **residues K, P, R, T được làm giàu ở flanking sequences của true carbonylation sites trong các mẫu dương tính**.

* Các **vị trí cacbonyl hóa có xu hướng mạnh ở các vùng được làm giàu RKPT**.

Thứ hai, trong mỗi biểu đồ con:

* **mức độ làm giàu của residues ở vùng upstreams (nơi phiên mã sớm) của quá trình cacbonyl hóa lớn hơn nhiều so với phần downstreams của các vị trí sửa đổi.** Ví dụ, trong Hình 3 (a), tần suất xuất hiện của cặn K ở upstreams cao hơn downstreams ở modification site

Thứ 3,

* hầu hết các **graphical residues có kích thước và kiểu khác nhau ở cùng một vị trí** cho thấy có sự khác biệt lớn về vị trí cụ thể giữa các tập hợp dương và âm. Chỉ có một số vị trí hiển thị sự khác biệt không đáng kể -> **sự phân bố theo vị trí cụ thể của các residues xung quanh các residues K, P, R, T có thể ảnh hưởng đến quá trình cacbonyl hóa của chúng.**

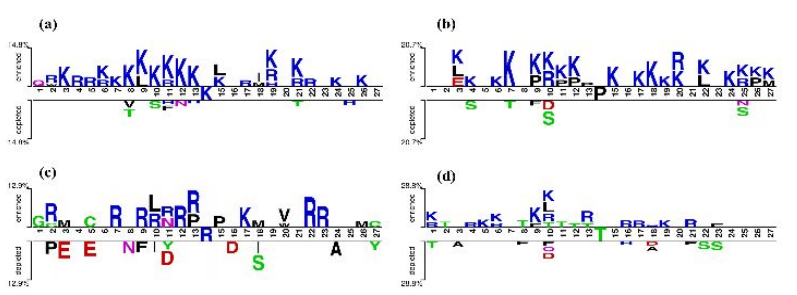


Figure 3 Hình minh họa để cho thấy sự khác biệt đáng kể về vị trí cụ thể của các gốc xung quanh các vị trí cacbonyl hóa và không cacbonyl hóa. Các đồ thị con từ (a) - (d) lần lượt là các vị trí cacbonyl hóa K, P, R, T

Theo quan điểm, sự khác nhau về vị trí cụ thể -> ý nghĩa để thống kê các mẫu dương/âm của mỗi loại carbonylation site để tạo ra 1 method để xác định các carbonylation sites chỉ bằng cách sử dụng thông tin về trình tự đó.

Trên thực tế, tọa độ hình nón của các axit amin được đề xuất trong bài báo này có thể phản ánh thông tin về vị trí và thành phần. Do đó, sự phân bố residues có thể được sử dụng như các đặc điểm để xác định các vị trí cacbonyl hóa.

## 3.2. Khảo sát các đặc tính hóa lý xung quanh vị trí cacbonyl hóa

Để thêm thông tin các đặc tính hóa lý xung quanh vị trí cacbonyl hóa, ta phân tích thống kê sự phân bố của chín đặc tính hóa lý của residues xung quanh vị trí cacbonyl hóa và không cacbonyl hóa dựa trên bộ tranning dataset như ở hình S1. Ta thấy các phân đoạn chứa cacbonyl hóa khác hẳn với các phân đoạn không chứa cacbonyl hóa:

1. Bốn modified residue, các mẫu dương cho sự biến chuyển lớn (về 9 đặc tính hóa lý) hơn các mẫu âm
2. Giá trị trung bình của tính ưa nước- **hydrophilicity**, **pI** và tính linh hoạt- **flexibility** của các **phân đoạn chứa cacbonyl hóa lớn hơn giá trị của các mẫu âm tính ở hầu hết các vị trí**, **đặc biệt là ở các vị trí gần các vị trí đã sửa đổi** (-7 ~ + 7), có thể có liên quan đến sự thay đổi cấu trúc protein do biến đổi cacbonyl hóa gây ra. Tuy nhiên, sự phân bố của tính kỵ nước- **hydrophobicity**, **pk2** và tính không thể thay thế- **irreplaceability** được thể hiện hiện tượng ngược lại. Nó cũng chỉ ra rằng sự **phân bố của chín đặc tính hóa lý ở các vị trí cacbonyl hóa là không cân bằng.**

* Residues ở các vị trí được modified tại upstream và downstream có ảnh hưởng khác nhau đến quá trình cacbonyl hóa.

Hơn nữa, quan sát thấy rằng sự khác biệt của **hydrophobicity, hydrophilicity, pI và flexibility tại mỗi vị trí giữa các mẫu dương tính và âm tính là lớn hơn khi so sánh với các đặc tính hóa lý khác (pK1, rigidity and irreplaceability).** Chúng tôi phát hiện ra một hiện tượng thú vị là so với T, R và P được sửa đổi**, các chất cặn xung quanh trong chất cặn K hiển thị ít biến động hơn**. Từ Hình 3, chúng tôi cũng phát hiện ra rằng các **trình tự đồng thuận xung quanh modified residue K** được cân bằng hơn so với các trình tự xung quanh các khu vực được sửa đổi khác. Điều này có thể giải thích hiện tượng trên đối với **modified residue K.** Các kết quả này đã minh họa thêm cho 9 đặc tính mã hóa tính hóa lý ở trên có ý nghĩa lớn đối với quá trình cacbonyl hóa và có thể được sử dụng như các tính năng để thực hiện dự đoán

## Đánh giá hiệu suất của các đặc trưng khác nhau

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 đó. 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 được vẽ trong Hình 4.

* 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. Phần sau sẽ tập trung vào tối ưu đặc trưng.

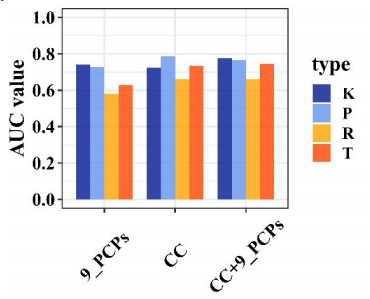


Figure 4 Dự đoán hiệu suất của các mô hình RF được đào tạo với các đặc trưng khác nhau dựa trên 10-fold cross-validation

## Xác định các đặc trưng tối ưu

Tập con các đặc trưng tối ưu có thể được lấy ra bằng cách tìm tất cả các tổ hợp đặc trưng có thể có. Tuy nhiên, không thể làm như vậy vì số lượng quá lớn và ta không thể kiểm tra chúng. Để tiết kiệm thời gian tính toán và cải thiện độ chính xác của dự đoán, chúng tôi đã sử dụng chiến lược tối ưu hóa đặc trưng để trích xuất ra các đặc trưng phân biệt nhất.

Đầu tiên, F-score được sử dụng để tính điểm của từng đặc trưng của tập đặc trưng PCPs và CC, sau đó sắp xếp score giảm dần. Mỗi tập con đặc trưng được thêm vào liên tiếp bằng giá trị lớn nhất của F-score, và tổng số 261 tập con đặc trưng (feature subsets) đã thu được. Tính hiệu xuất của các feature subsets này bằng RF với 10-fold cross-validation, và sau đó vẽ biểu đồ bốn đường cong IFS cho bốn residues (Hình 5) để xác định feature subset tối ưu.

Hình 5 cho thấy đỉnh của giá trị AUC đối với các vị trí cacbonyl hóa K, P, R và T đạt được khi số lượng feature subsets lần lượt bằng 193, 38, 77 và 60. 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. Những kết quả này được ghi lại trong Bảng 2. So sánh kết quả giữa bộ train và bộ test cho thấy rằng model ta đề xuất này mạnh và có thể sử dụng để xác định các vị trí cacbonyl hóa trong protein. Dựa trên mô hình được đề xuất, chúng tôi đã xây dựng một máy chủ web trực tuyến gọi là iCarPS có thể được truy cập miễn phí từ http://lin-group.cn/server/iCarPS. Công cụ này sẽ cung cấp sự tiện lợi cho hầu hết các học giả

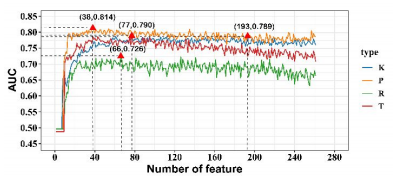


Figure 5 Biểu đồ thể hiện quy trình IFS được tối ưu hóa F-score để xác định các vị trí cacbonyl hóa. Các top đặc trưng trong đó các đỉnh của giá trị AUC của K, P, R và T được đánh dấu bằng tam giác đỏ tương ứng trong 10-fold cross-validation được sử dụng để thực hiện dự đoán.

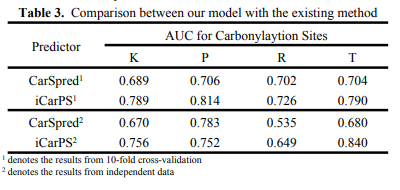
## So với phương pháp khác đã ra báo

Để chứng minh thêm hiệu suất của phương pháp của chúng tôi, chúng tôi phải so sánh phương pháp của chúng tôi với phương pháp đã xuất bản khác. Ở đây, **CarSPred đã được chọn để thực hiện so sánh vì nó sử dụng cùng benchmark dataset**. Kết quả của phương pháp CarSPred trên cùng tập train và tập independent test được thu thập trực tiếp từ các báo cáo của họ. Kết quả so sánh được liệt kê trong Bảng 3.

Như có thể thấy từ Bảng 3, giá trị ACC và AUC trung bình của phương pháp của chúng tôi (iCarPS) cao hơn 1,20% và 7,95% so với phương pháp của CarSPred trên tập train in 10-fold cross-validation. Trên independent data, mặc dù giá trị AUC của iCarPS thấp hơn một chút so với của CarSPred đối với vị trí cacbonyl hóa loại P, giá trị AUC của iCarPS được cải thiện đáng kể 12% đối với các loại khác (K, R, T) của vị trí cacbonyl hóa. Hơn nữa, sự khác biệt trung bình của AUC giữa training and independent data là 0,072 và 0,056, tương ứng đối với CarSpred và mô hình của chúng tôi, cho thấy rằng mô hình đề xuất của chúng tôi ổn định hơn. Những so sánh này chỉ ra rằng phương pháp của chúng tôi vượt trội hơn so với CarSPred.

0,53, 0,54, 0,66, 0,67

0,51, 0,72, 0,56, 0,82



# Kết luận

Carbonyl hóa là một biến đổi post-translational (sau dịch mã) - **import post-translational modification** – một trong những loại chính của stress oxy hóa. Trong nghiên cứu này, chúng tôi đã **đề xuất một công cụ dự báo** mới để xác định các vị trí cacbonyl hóa bằng cách **sử dụng sequence-derived features (đặc trưng dựa trên trình tự).** Một **bộ mô tả thành phần trình tự** mới được gọi là **tọa độ hình nón của residues.** Một số thí nghiệm đã được thực hiện để chứng minh tính mạnh mẽ và hiệu quả của phương pháp của chúng tôi. Ngoài ra, thuật toán trong nghiên cứu này **đã tính đến đầy đủ thông tin thành phần trình tự và thông tin đặc tính hóa lý**. Một **phương pháp trích chọn đặc trưng** được sử dụng là **F-score** để **chọn các informative features và loại bỏ nhiễu**. Quá trình này cải thiện đáng kể sức mạnh dự đoán của mô hình.

Dựa trên mô hình tối ưu, một công cụ dự đoán trực tuyến iCarPS đã được thành lập để xác định các protein cacbonyl hóa. Nó cung cấp sự thuận tiện cho các nhà khoa học thực nghiệm, những người có thể thu được kết quả mong muốn một cách nhanh chóng và chính xác mà không cần lặp lại các chi tiết toán học. Ngoài ra, một gói phần mềm dành cho máy tính cục bộ đã có sẵn trên trang web của chúng tôi. Do đó, chúng tôi chắc chắn rằng công cụ dự báo này sẽ trở thành một công cụ hữu ích để phân tích cacbonyl hóa và các nghiên cứu thực nghiệm tiếp theo. Chúng tôi cũng gợi ý rằng phương pháp được đề xuất trong nghiên cứu này, đặc biệt là đối với các tọa độ hình nón của residues, có thể được khái quát hóa để dự đoán các loại biến đổi sau dịch mã khác trong các nghiên cứu về protein.