ISSN: 1859-2171
e-ISSN: 2615-9562
TNU Journal of Science and Technology 204(11): 149 - 154

PHÂN LỚP VỊ TRÍ PROTEIN FARNESYLATION VỚI MÁY VECTOR HỖ TRỢ (SVM) VÀ CÂY QUYẾT DỊNH

Trần Thị Xuân[1], Nguyễn Văn Núi[2*]
[1]Trường Đại học Kinh tế và Quản trị kinh doanh – ĐH Thái Nguyên
[2]Trường Đại học Công Nghệ Thông Tin và Truyền Thông – ĐH Thái Nguyên

TÓM TẮT

Protein Prenylation sự bổ sung của các phân tử kháng nước tới một protein hoặc một hợp chất hóa học. Nó là một quá trình biến đổi hậu dịch mã (PTM: Post Translational Modification) đóng vai trò rất quan trọng, ảnh hưởng đến nhiều quá trình phân tử cũng như ảnh hưởng đến nhiều chức năng tế bào khác. Protein S-Farnesyl Cysteine Prenylation là một trường hợp đặc biệt của Prenylation liên quan đến sự dịch chuyển của một phân nửa (moiety) farnesyl tới một cysteine tế bào chất tại hoặc gần khu vực đầu cuối-C (C-turminus) của protein mục tiêu. Những phát hiện gần đây cho thấy vai trò rất quan trọng của S-Farnesyl Cysteine Prenylation (SFCP) ảnh hưởng đến nhiều quá trình sinh học cũng như có liên quan đến rất nhiều căn bệnh phổ biến hiện nay. Cho đến nay, có khá nhiều nghiên cứu về SFCP, đồng thời một vài công cụ tính toán cũng đã được đề xuất cho việc phân lớp, dự đoán vị trí SFCP. Tuy nhiên, hầu hết các nghiên cứu và công cụ dự đoán này hoặc chưa đáp ứng được các yêu cầu về kiến thức sâu rộng liên quan, hoặc hiệu năng dự đoán chưa đáp ứng được kỳ vọng. Vì vậy, trong nghiên cứu này chúng tôi đề xuất cách tiếp cận phân lớp vị trí protein SFCP trên cơ sở kết hợp sử dụng các phương pháp học máy và cây quyết định. Nhiều đặc trưng được tiến hành thử nghiệm để xây dựng mô hình dự đoán có hiệu năng tốt nhất. Kết quả cho thấy mô hình mà chúng tôi đề xuất có tính khả thi cao trong việc dự đoán vị trí SFCP. Điều này có thể sẽ là gợi ý về một hướng tiếp cận có thể giúp ích hữu hiệu cho các nhà nghiên cứu liên quan đến việc SFCP.

Từ khóa: *Biến đổi hậu dịch mã; máy vector hỗ trợ; cây quyết định; phân loại dữ liệu; protein S-Farnesyl Cysteine Prenylation.*

Ngày nhận bài: 23/7/2019; Ngày hoàn thiện: 15/8/2019; Ngày đăng: 19/8/2019

CLASSIFYING PROTEIN S-FARNESYLATION SITES WITH SUPPORT VECTOR MACHINE AND DECISION TREE

Thi-Xuan Tran[1], Van-Nui Nguyen[2*]
[1]University of Economics and Business Administration – TNU
[2]University of Information and Communication Technology - TNU

ABSTRACT

Protein prenylation is the addition of hydrophobic molecules to a protein or a chemical compound. It is a post-translational modification that plays very important roles affecting to many cellular processes as well as many other cellular functions. Protein S-farnesyl cysteine prenylation is a specific kind of prenylation related to the transfer of a farnesyl moiety to a cytoplasmic cysteine at or near the C-terminus of the target protein. Recent findings have exhibited the very important roles of S-Farnesyl Cysteine Prenylation (SFCP) that affect to many biological processes as well as have involed in many current common diseases. So far, there has been some researches on SFCP, and several computational tools have been proposed for the classification, prediction of SFCP sites. However, almost of them have not met our demand on related extensive knowlegde, or the predictive performance has not met the requirements. Therefore, in this work, we are motivated to propose an approach to classify protein SFCP based on the incorporation of support vector machine and decision tree. Various features have been investigated to generate the optimal model that has highest predictive performance. The obtained results have demonstrated its ability and feasiblity in the classification of SFCP sites. This could be a suggestion on an approach that can useful for researchers regarding to SFCP.

Keywords: *Post-translational modification; support vector machine; decision tree; data classification; S-Farnesyl Cysteine Prenylation.*

Received: 23/7/2019; Revised: 15/8/2019; Published: 19/8/2019

* Corresponding author. *Email: nvnui@ictu.edu.vn*

1. Giới thiệu chung

Protein prenylation (còn được biết đến với các tên gọi khác: isoprenylation or lipidation), được phát hiện lần đầu tiên ở nấm vào năm 1978 [1], là việc bổ sung các phân tử kháng nước vào protein hoặc hợp chất hóa học. Protein prenylated đầu tiên trong các tế bào động vật có vú, lamin B, được phát hiện khoảng mười năm sau đó [2, 3]. Trong các loài nhân chuẩn (eukaryote), prenylation protein là một PTM quan trọng, ảnh hưởng đến nhiều quá trình tế bào [4]. Quá trình prenyl hóa được thực hiện và thúc đẩy bởi 3 enzymes với đặc tính bề mặt chồng chéo 1 phần: Farnesyl Transferase, Caax protease and geranylgeranyl transferase [5]. Protein S-farnesyl cysteine prenylation (SFCP) liên quan đến sự dịch chuyển của một phân nửa (moiety) farnesyl tới một cysteine tế bào chất tại hoặc gần khu vực đầu cuối-C (C-turminus) của protein mục tiêu [6].

Do vai trò rất quan trọng gây ra bởi SFCP, số lượng nghiên cứu để tìm hiểu sâu rộng về đặc tính của SFCP đã tăng nhanh trong những năm qua [5, 7-9]. Gần đây, có một vài mô hình phân lớp được nghiên cứu, đề xuất để hỗ trợ các nhà nghiên cứu trong việc phân lớp, dự đoán vị trí SFCP [10-12]. Tuy nhiên, ở thời điểm hiện tại, vẫn còn thiếu các mô hình tính toán phù hợp và công cụ dự đoán với độ chính xác cao có thể hỗ trợ hiệu quả hỗ trợ cho việc đặc tả, dự đoán vị trí SFCP. Bên cạnh đó, do sự tiến bộ của khoa học kỹ thuật và ảnh hưởng của cách mạng công nghiệp 4.0, dữ liệu SFCP đã kiểm chứng thực nghiệm đang ngày càng được bổ sung nhiều hơn. Chính vì vậy, việc thiếu hụt mô hình phân lớp dự đoán vị trí SFCP là một vấn đề cấp thiết cần được quan tâm giải quyết.

Tiếp tục phát triển các ý tưởng nghiên cứu trước đây [13-16], trong nghiên cứu này chúng tôi đề xuất một cách tiếp cận khác giải quyết bài toán phân lớp dự đoán vị trí SFCP với sự kết hợp của SVM và cây quyết định.

2. Xây dựng, huấn luyện mô hình

2.1. Thu thập, tiền xử lý dữ liệu

Trong nghiên cứu này, dữ liệu đã kiểm chứng thực nghiệm SFCP được thu thập từ nhiều nguồn khác nhau: [6] [17], [18] [19] [20]. Thực tế các nguồn dữ liệu này có thể công bố dữ liệu trùng lặp/chồng chéo nhau, vì vậy cần phải tiến hành thực hiện một số bước tiền xử lý để loại bỏ dữ liệu trùng lặp/dư thừa. Sau quá trình loại bỏ dữ liệu trùng lặp/dư thừa, chúng tôi thu được 718 dữ liệu vị trí SFCP từ 670 proteins khác nhau. Để xây dựng dữ liệu huấn luyện (training data) và dữ liệu kiểm thử (testing data), trong nghiên cứu này, chúng tôi tiến hành lấy ngẫu nhiên 70 proteins từ tổng số 670 proteins đã thu được trước đó làm dữ liệu kiểm thử. Phần còn lại gồm 600 proteins sẽ được sử dụng để xây dựng dữ liệu huấn luyện.

Tại nghiên cứu này, chúng tôi tiến hành xây dựng mô hình dựa trên việc phân tích các đặc tính protein nền (substrate protein) dựa trên dạng chuỗi fasta (cấu trúc bậc 1 của protein). Theo dạng biểu diễn này, mỗi protein được biểu diễn như là một chuỗi gồm các ký tự đại diện cho 20 amino acid, trong đó protein S-Farnesylated cysteine được hiểu là tồn tại một amino acid Cysteine (C) đã được khẳng định là S-Farnesyl cysteine. Để chuyển đổi từ dữ liệu thô sang dạng vector ứng dụng được với máy vector hỗ trợ (SVM) và cây quyết định (Decision Tree), quá trình tiền xử lý dữ liệu cần được tiến hành. Trên cơ sở các phương pháp đã được triển khai từ những nghiên cứu tương tự trước đó [13, 14], một biến window size =13 được sử dụng để cắt các đoạn chuỗi nhỏ với vị trí trung tâm là Cysteine (C). Ngoài ra, để tránh trường hợp hiệu năng mô hình bị đánh giá quá cao hoặc quá thấp do dữ liệu huấn luyện trùng lặp hoặc tương đồng quá nhiều, bộ công cụ CD-HIT [21] cũng được áp dụng. Với việc sử dụng giá trị tương đồng 40%, sau khi chạy CD-HIT, dữ liệu huấn luyện (training data) thu được gồm 296 positive data and 1051 negative data; dữ liệu kiểm thử độc lập (independent testing) thu được bao gồm 28 positive and 332 negative data.

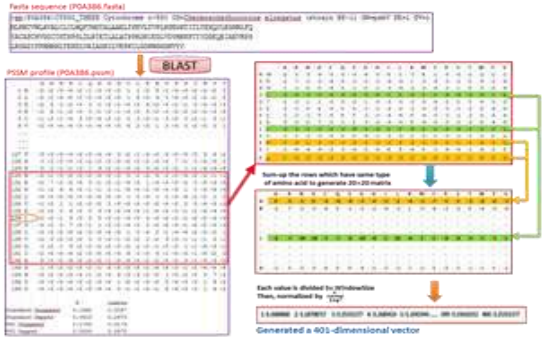
2.2. Trích chọn và mã hóa đặc trưng

Để phục vụ cho việc xây dựng và huấn luyện mô hình phân lớp SFCP, chúng tôi tiến hành kết hợp sử dụng SVM và Decision Tree. Trước tiên, các đặc trưng phổ biến thường được sử dụng phục vụ cho xây dựng, huấn luyện mô hình, gồm: AAC (Amino Acid Composition), AAPC (Amino Acid Pairwise Composition), PSSM (Evolutionary information). Các đặc trưng này được trích xuất và mã hóa như sau:

AAC: Sử dụng một vector 21 chiều v=(class, x_1, x_2, …, x_{20}) để biểu diễn, trong đó: Giá trị class thường được chọn bằng 1 (SFCP site) hoặc bằng 2 (non-SFCP site); Mỗi giá trị x_i (i=1..20) được tính bằng số lần xuất hiện của 1 trong số 20 amino acids tương ứng chia cho tổng số amino acid của chuỗi.

AAPC: Sử dụng một vector 401 chiều v=(class, x_{ij}); i,j=1..20 để biểu diễn, trong đó mỗi giá trị x_{ij} (i,j=1..20) được tính bằng số lần xuất hiện của 1 cặp trong số 20 amino acids tương ứng chia cho tổng số cặp amino acid của chuỗi.

PSSM: Sử dụng một vector 401 chiều v=(class, x_{ij}); i,j=1..20 để biểu diễn. Các bước chi tiết để mã hóa đặc trưng PSSM được hiển thị như ở Hình 1 dưới đây.



Hình 1. *Các bước trích xuất và mã hóa đặc trưng PSSM*

Ngoài các đặc trưng riêng lẻ, chúng tôi còn tiến hành kết hợp lai ghép các đặc trưng sau đây trong việc xây dựng, đánh giá và tìm kiếm mô hình phân lớp tối ưu nhất, bao gồm: AAC_AAPC, AAC_PSSM, AAPC_PSSM, và AAC_AAPC_PSSM.

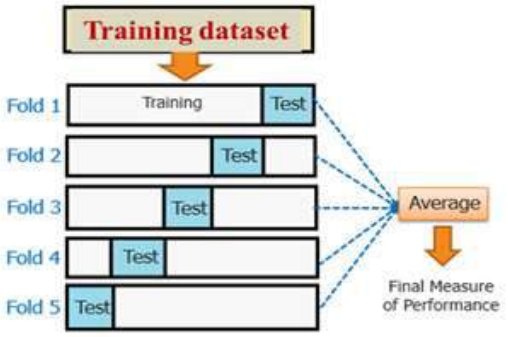
2.3. Xây dựng và huấn luyện mô hình

Máy vector hỗ trợ được sử dụng kết hợp với cây quyết định để xây dựng mô hình phân lớp. Trong nghiên cứu này, bộ công cụ Weka cùng với thuật toán máy vector hỗ trợ và cây quyết định được sử dụng để phân tích, đánh giá hiệu năng của mô hình. Cây quyết định (decision tree) là một mô hình học máy thuộc nhóm thuật toán học có giám sát (supervised learning). Nó là một phương pháp học máy mạnh và phổ biến đã được biết đến và áp dụng thành công cho bài toán khai phá dữ liệu và phân lớp. Cây quyết định chính là cây mà mỗi nút biểu diễn một đặc trưng, mỗi nhánh (branch) biểu diễn một quy luật (rule), mỗi nút lá biểu diễn một kết quả (giá trị cụ thể hoặc một nhánh tiếp tục). Cây quyết định có thể được dùng cho bài toán phân lớp dữ liệu bằng cách xuất phát từ gốc của cây và di chuyển theo các nhánh cho đến khi gặp nút lá. Một ví dụ về cây quyết định được mô tả quyết định **CHƠI** hay **HỌC** của 1 sinh viên được minh họa như ở Hình 2. (Quy tắc để cậu SV này đưa ra quyết định học hay chơi như sau: Nếu còn nhiều hơn hai ngày nữa mới tới ngày thi, cậu sẽ **CHƠI**. Nếu còn không quá hai ngày và đêm hôm đó có một trận bóng đá hay, cậu sẽ sang nhà bạn **CHƠI** và cùng xem bóng đêm đó. Cậu sẽ chỉ **HỌC** trong các trường hợp còn lại)

Hình 2. *Cây quyết định về việc học hay chơi của 1 SV*

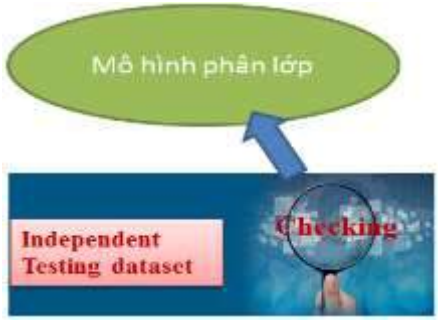
Để đánh giá hiệu năng của mô hình, 2 phương pháp phổ biến được sử dụng đó là: đánh giá chéo 5-mặt (5-fold cross-validation) và kiểm thử độc lập (Independent testing) sử dụng bộ

dữ liệu độc lập (independent testing dataset với bộ dữ liệu huấn luyện (training dataset). Với phương pháp đánh giá chéo 5 mặt (Như hiển thị ở Hình 3), tập dữ liệu huấn luyện sẽ được chia ngẫu nhiên thành 5 tập con bằng nhau, lần lượt mỗi tập con sẽ được dùng cho vai trò kiểm thử trong khi 4 tập còn lại được dùng làm dữ liệu huấn luyện.



Hình 3. *Mô hình đánh giá kiểm tra chéo 5-mặt*

Như hiển thị ở Hình 4, theo phương pháp đánh giá kiểm thử độc lập, hiệu năng của mô hình sẽ được xác định bằng việc sử dụng một bộ dữ liệu kiểm thử hoàn toàn khác biệt và không trùng lặp với bộ dữ liệu huấn luyện đã dùng cho việc huấn luyện mô hình (Independent testing dataset). Việc sử dụng bộ dữ liệu kiểm thử độc lập này sẽ giúp ta kiểm tra, đánh giá một cách khách quan nhất hiệu năng phân lớp của mô hình.



Hình 4. *Mô hình kiểm thử độc lập*

Các đại lượng thông dụng được sử dụng để đo lường và đánh giá hiệu năng của mô hình bao gồm: SEN (Tỷ lệ phân lớp đúng dữ liệu SFCP), SPE (Tỷ lệ phân lớp đúng dữ liệu non-SFCP), ACC (Tỷ lệ phân lớp chính xác nói chung), và MCC (Giá trị tương quan theo công thức của Matthews - Matthews Correlation Coefficient):

$$SEN = \frac{TP}{TP+FN}; \quad SPE = \frac{TN}{TN+FP}; \quad ACC = \frac{TP}{TP+FN};$$

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

Trong đó các đại lượng TP, TN, FP và FN biểu diễn số lượng phân lớp tương ứng TRUE SFCP, TRUE non-SFCP; FALSE SFCP và FALSE non_SFCP.

3. Kết quả và một số thảo luận

3.1. Kết quả huấn luyện và đánh giá mô hình phân lớp theo phương pháp đánh giá chéo 5-mặt

Như đã trình bày trước đó, trong nghiên cứu này, chúng tôi tiến hành sử dụng kết hợp thuật toán của máy vector hỗ trợ và cây quyết định để xây dựng và huấn luyện mô hình trên cơ sở 3 đặc trưng riêng lẻ cơ bản AAC, AAPC và PSSM. Theo thông tin tổng hợp ở Bảng 1, với đặc trưng AAC, mô hình đạt hiệu năng phân lớp với độ chính xác là 91,91%, giá trị MCC = 0,80. Tương tự, mô hình được xây dựng dựa trên đặc trưng AAPC đạt độ chính xác 88,27%, giá trị MCC = 0,74. Mô hình xây dựng dựa trên đặc trưng PSSM đạt độ chính xác 92,68%, giá trị MCC = 0,81.

Bảng 1. *Bảng kết quả đánh giá mô hình bằng phương pháp đánh giá chéo 5-mặt*

Feature	SEN	SPE	ACC	MCC
AAC	96,95%	90,49%	91,91%	0,80
AAPC	98,31%	85,44%	88,27%	0,74
PSSM	96,28%	91,76%	92,68%	0,81
AAC_AAPC	96,66%	92,96%	93,78%	0,84
AAC_PSSM	95,33%	93,62%	94,00%	0,84
AAPC_PSSM	95,33%	93,52%	93,93%	0,84
AAC_AAPC _PSSM	**98,31%**	**92,96%**	**94,14%**	**0,85**

Trong học máy, hướng tiếp cận kết hợp hai hay nhiều phương pháp khác nhau để khai thác lợi thế của chúng được hiểu như là một cách tiếp cận tự nhiên, dễ hiểu và khá phổ biến. Chính vì vậy, trong nghiên cứu này, chúng tôi cũng tiến hành kết hợp lai ghép các đặc trưng riêng lẻ để xây dựng các đặc trưng phức tạp hơn hỗ trợ trong việc huấn luyện mô hình phân lớp vị trí SFCP. Cụ thể, 4 đặc trưng lai ghép: AAC_AAPC, AAC_PSSM, AAPC_PSSM, và AAC_AAPC_PSSM đã được xây dựng từ việc kết hợp 3 đặc trưng riêng lẻ trước đó.

Kết quả đánh giá chéo 5-mặt (Bảng 1) cho các mô hình xây dựng dựa trên các đặc trưng lai ghép có hiệu năng phân lớp SFCP tốt hơn các đặc trưng riêng lẻ. Trong đó, đặc trưng lai ghép AAC_AAPC_PSSM được coi là đặc trưng tốt nhất khi mô hình phân lớp tương ứng có hiệu năng tốt nhất, với độ chính xác đạt 94,14% và giá trị MCC=0,85. Kết quả này chỉ ra rằng đặc trưng lai ghép AAC_AAPC_PSSM giúp tạo ra mô hình có hiệu năng tốt nhất trong việc phân lớp, dự đoán vị trí SFCP.

3.2. Kết quả đánh giá mô hình sử dụng phương pháp kiểm thử độc lập

Như đã đề cập trước đó, phương pháp đánh giá độc lập giúp kiểm chứng khả năng thực nghiệm của mô hình trong trường hợp thực tế, khách quan nhất. Để thực hiện được việc này, một bộ dữ liệu kiểm thử độc lập đã được xây dựng bao gồm 28 dữ liệu positive và 332 dữ liệu negative.

Kết quả kiểm tra đánh giá hiệu năng của mô hình khi tiến hành bởi phương pháp kiểm thử độc được thể hiện chi tiết ở Bảng 2. Qua các con số thể hiện ở Bảng 2, ta thấy rằng mô hình đạt độ chính xác tương đối cao và có tính khả thi tốt trong việc dự đoán vị trí SFCP. Đặc biệt, mô hình xây dựng bởi thuộc tính lai ghép AAC_AAPC_PSSM cũng mang lại hiệu năng phân lớp cao nhất, với độ chính xác đạt 95,00% và giá trị MCC=0,75. Kết quả này cho thấy tính khả thi và hiệu quả phân lớp dự đoán của mô hình mà chúng tôi đề xuất. Bên cạnh đó, kết quả thu được cũng gợi ý rằng cách tiếp cận lai ghép các đặc trưng riêng lẻ có thể được coi là một cách tiếp cận hiệu quả và hứa hẹn trong việc xây dựng mô hình phân lớp, dự đoán vị trí protein SFCP sites.

Bảng 2. *Bảng kết quả đánh giá mô hình bằng phương pháp kiểm thử độc lập*

Feature	SEN	SPE	ACC	MCC
AAC	85,71%	92,47%	91,94%	0,61
AAPC	89,29%	93,98%	93,61%	0,67
PSSM	89,29%	94,28%	93,89%	0,68
AAC_AAPC	92,86%	94,58%	94,44%	0,72
AAC_PSSM	89,29%	94,28%	93,89%	0,68
AAPC_PSSM	85,71%	94,28%	93,61%	0,66
AAC_AAPC _PSSM	96,43%	94,88%	95,00%	0,75

5. Kết luận

Protein Prenylation sự bổ sung của các phân tử kháng nước tới một protein hoặc một hợp chất hóa học. Nó là một quá trình biến đổi hậu dịch mã (PTM: Post Translational Modification) đóng vai trò rất quan trọng ảnh hưởng đến nhiều quá trình phân tử cũng như ảnh hưởng đến nhiều chức năng tế bào khác. Protein S-Farnesyl Cysteine Prenylation là một trường hợp đặc biệt của Prenylation liên quan đến sự dịch chuyển của một phân nửa (moiety) farnesyl tới một cysteine tế bào chất tại hoặc gần khu vực đầu cuối-C (C-turminus) của protein mục tiêu. Những phát hiện gần đây cho thấy vai trò rất quan trọng của S-Farnesyl Cysteine Prenylation (SFCP) ảnh hưởng đến nhiều quá trình sinh học cũng như có liên quan đến rất nhiều căn bệnh phổ biến hiện nay. Trong nghiên cứu này chúng tôi đề xuất cách tiếp cận phân lớp vị trí protein SFCP trên cơ sở kết hợp sử dụng các phương pháp học máy và cây quyết định. Nhiều đặc trưng được tiến hành thử nghiệm để xây dựng mô hình dự đoán có hiệu năng tốt nhất. Kết quả cho thấy mô hình mà chúng tôi đề xuất đặt kết quả phân lớp cao nhất với đặc trưng lai ghép AAC_AAPC_PSSM, có tính khả thi cao trong việc phân lớp dự đoán vị trí SFCP. Điều này được kỳ vọng sẽ là một hướng tiếp cận hữu ích, hỗ trợ tốt cho các nhà nghiên cứu phân tích, xử lý dữ liệu có liên quan đến SFCP.

Lời cảm ơn

Nhóm tác giả xin được bày tỏ lòng biết ơn đến Trường Đại học Công nghệ thông tin và Truyền thông đã hỗ trợ một phần tài chính cho nghiên cứu này theo đề tài cấp Đại học Thái Nguyên mã số: DH2018-TN-07.

TÀI LIỆU THAM KHẢO

[1]. Kamiya Y., Sakurai A., Tamura S., Takahashi N: Structure of rhodotorucine A., "A novel lipopeptide, inducing mating tube formation in Rhodosporidium toruloides", *Biochemical and biophysical research communications*, 83(3), pp. 1077-1083, 1978.

[2]. Farnsworth C. C., Wolda S. L., Gelb M. H., Glomset J. A., "Human lamin B contains a farnesylated cysteine residue", *The Journal of biological chemistry*, 264(34), pp. 20422-20429, 1989.

[3]. Wolda S. L., Glomset J. A., "Evidence for modification of lamin B by a product of mevalonic acid", *The Journal of biological chemistry*, 263(13), pp. 5997-6000, 1988.

[4]. Soni R., Sharma D., Patel S., Sharma B., Bhatt T. K., "Structure-based binding between protein farnesyl transferase and PRL-PTP of malaria parasite: an interaction study of prenylation process in Plasmodium", *Journal of biomolecular structure & dynamics*, 34(12), pp. 2667-2678, 2016.

[5]. Novelli G., D'Apice M. R., "Protein farnesylation and disease", *Journal of inherited metabolic disease*, 35(5), pp. 917-926, 2012.

[6]. Maurer-Stroh S., Koranda M., Benetka W., Schneider G., Sirota F. L., Eisenhaber F., Towards complete sets of farnesylated and geranylgeranylated proteins", *PLoS computational biology*, 3(4), pp. e66, 2007.

[7]. Hechinger A. K., Maas K., Durr C., Leonhardt F., Prinz G., Marks R., Gerlach U., Hofmann M., Fisch P., Finke J. et al, "Inhibition of protein geranylgeranylation and farnesylation protects against graft-versus-host disease via effects on CD4 effector T cells", *Haematologica*, 98(1), pp. 31-40, 2013.

[8]. Maurer-Stroh S., Washietl S., Eisenhaber F., "Protein prenyltransferases: anchor size, pseudogenes and parasites", *Biological chemistry* 384(7), pp.977-989, 2003.

[9]. Einav S., Glenn J. S., "Prenylation inhibitors: a novel class of antiviral agents", *The Journal of antimicrobial chemotherapy*, 52(6), pp. 883-886, 2003.

[10]. Soni R., Sharma D., Patel S., Sharma B., Bhatt T. K., "Structure-based binding between protein farnesyl transferase and PRL-PTP of malaria parasite: an interaction study of prenylation process in Plasmodium", *Journal of biomolecular structure & dynamics*, 34(12), pp. 2667-2678, 2016.

[11]. Das S., Edwards P. A., Crockett J. C., Rogers M. J., "Upregulation of endogenous farnesyl diphosphate synthase overcomes the inhibitory effect of bisphosphonate on protein prenylation in Hela cells", *Biochimica et biophysica acta*, 1841(4), pp. 569-573, 2014.

[12]. Wojtkowiak J. W., Gibbs R. A., Mattingly R. R., "Working together: Farnesyl transferase inhibitors and statins block protein prenylation", *Molecular and cellular pharmacology*, 1(1), pp. 1-6, 2009.

[13]. Nguyen V. N., Huang K. Y., Huang C. H., Lai K. R., Lee T. Y., "A new scheme to characterize and identify protein ubiquitination sites", *IEEE/ACM transactions on computational biology and bioinformatics/ IEEE*, ACM 2017, 14(2), pp. 393-403, 2017.

[14]. Nguyen V. N., Huang K. Y., Huang C. H., Chang T. H., Bretana N., Lai K., Weng J., Lee T. Y., "Characterization and identification of ubiquitin conjugation sites with E3 ligase recognition specificities", *BMC bioinformatics*, 16 Suppl 1, pp. S1, 2015.

[15]. Lee T. Y., Lin Z. Q., Hsieh S. J., Bretana N. A., Lu C. T., "Exploiting maximal dependence decomposition to identify conserved motifs from a group of aligned signal sequences", *Bioinformatics*, 27(13), pp. 1780-1787, 2011.

[16]. Lee T. Y., Chen Y. J., Lu T. C., Huang H. D., Chen Y. J., "SNOSite: exploiting maximal dependence decomposition to identify cysteine S-nitrosylation with substrate site specificity", *PloS one*, 6(7), pp. e21849, 2011.

[17]. Yubin Xie Y. Z., Hongyu Li, Xiaotong Luo, Zhihao He, Shuo Cao, Yi Shi, Qi Zhao, Yu Xue, Zhixiang Zuo and Jian Ren, "GPS-Lipid: a robust tool for the prediction of multiple lipid modification sites", *Scientific reports*, 6, pp. 28249, 2016.

[18]. Boeckmann B., Bairoch A., Apweiler R., Blatter M. C., Estreicher A., Gasteiger E., Martin M. J., Michoud K., O'Donovan C., Phan I. et al, "The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003", *Nucleic acids research*, 31(1), pp. 365-370, 2003.

[19]. Lu C. T., Huang K. Y., Su M. G., Lee T. Y., Bretana N. A., Chang W. C., Chen Y. J., Chen Y. J., Huang H. D., "DbPTM 3.0: an informative resource for investigating substrate site specificity and functional association of protein post-translational modifications", *Nucleic acids research*, 41(Database issue), pp. D295-305, 2013.

[20]. Keshava Prasad T. S., Goel R., Kandasamy K., Keerthikumar S., Kumar S., Mathivanan S., Telikicherla D., Raju R., Shafreen B., Venugopal A. et al, "Human Protein Reference Database--2009 update", *Nucleic acids research*, 37(Database issue), pp. D767-772, 2009.

[21]. Huang Y., Niu B., Gao Y., Fu L., Li W., "CD-HIT Suite: a web server for clustering and comparing biological sequences", *Bioinformatics*, 26(5), pp. 680-682, 2010.