TẠP CHÍ KHOA HỌC VÀ GIÁO DỤC
ISSN: 1859-1612
Journal homepage: https://tckhgd.dhsphue.edu.vn

Áp dụng biến đổi Hilbert và mô hình học tổng hợp vào tiên đoán tương tác giữa hai protein

Mai Xuân Văn, Trần Võ Hoàng Nguyên, Trần Hoài Nhân, Nguyễn Tương Tri

Khoa Tin học, Trường Đại học Sư phạm, Đại học Huế

THÔNG TIN BÀI BÁO

Quá trình xử lý:
Ngày nhận bài: 04/7/2024
Ngày nhận bản chỉnh sửa: 27/12/2024
Ngày nhận đăng: 02/1/2025
Ngày xuất bản: 20/4/2026

Từ khóa:
Tương tác protein-protein;
Học tổng hợp;
Biến đổi Hilbert.

TÓM TẮT

Xác định tương tác protein-protein (PPI) có vai trò quan trọng trong việc hiểu biết về cấu trúc và chức năng sinh học của tế bào. Việc dự đoán chính xác các tương tác vẫn đang là một thách thức. Trong bài báo này, chúng tôi giới thiệu một phương pháp kết hợp giữa biến đổi Hilbert trên thông tin tiến hóa cho bước trích xuất đặc trưng và mô hình phân loại Deep Forest (DF) cho bước dự đoán tương tác. Việc sử dụng biến đổi Hilbert là nhằm tăng cường thông tin tiến hóa được lưu trữ trong trình tự protein. Bên cạnh đó, DF, một mô hình phân loại tổng hợp mạnh, được sử dụng để cải thiện hiệu suất dự đoán. Phương pháp đề xuất đã được đánh giá trên bộ dữ liệu chuẩn Yeast và so sánh với các phương pháp hiện có khác. Kết quả cho thấy phương pháp của chúng tôi đã đạt hiệu quả cao trong dự đoán PPI. Các bộ dữ liệu và mã nguồn của phương pháp được cung cấp tại https://github.com/thnhanhub/FeatPSSM.git.

1. GIỚI THIỆU

Protein, hay còn gọi là chất đạm, là đại phân tử sinh học có vai trò quan trọng trong nhiều chức năng sinh học và là một phần quan trọng của cấu trúc tế bào và mô, cũng như tham gia vào nhiều quá trình sinh học khác nhau trong cơ thể. Chúng được cấu tạo nên từ nhiều amino acid liên kết lại với nhau bằng liên kết peptide (Copeland, 1994). Tương tác protein – protein (PPI) là sự liên kết và tác động qua lại giữa hai hay nhiều protein. Các PPI này có thể xảy ra một cách trực tiếp hoặc gián tiếp thông qua các phân tử trung gian và đóng vai trò trung tâm trong sự tiến hóa và tiến triển của các quá trình sinh học khác nhau để hiểu các quá trình của tế bào như phiên mã DNA, chu trình trao đổi chất và tái tạo (De Las Rivas & Fontanillo, 2010). Trong thời hiện đại của nghiên cứu sinh học phân tử, việc hiểu rõ về tương tác giữa hai protein là một thách thức lớn để mang lại những tri thức quan trọng về cơ chế hoạt động của các hệ thống sinh học. Vì vậy, dự đoán PPI đóng vai trò quan trọng trong việc mở rộng kiến thức về mối quan hệ chức năng giữa chúng, là điều cần thiết để làm sáng tỏ các chức năng của protein và hiểu cấu trúc sinh học trong tế bào. Ngoài ra, dự đoán PPI không chỉ giúp kiểm tra sâu hơn về cách các protein phát huy các chức năng của chúng, mà còn cung cấp thông tin quan trọng cho việc phát triển các loại thuốc. Vào năm 2019, đại dịch COVID-19 bùng phát bởi loại virus mang tên SARS-CoV-2, gây ra bệnh viêm đường hô hấp và có thể nhanh chóng truyền nhiễm ngay trong thời gian ủ bệnh. Một số tổ chức và viện nghiên cứu bắt đầu làm việc để tạo ra một số loại vacxin bằng việc nghiên cứu cách thức tương tác protein chủ chốt liên quan đến sự xâm nhập của SARS-CoV-2, từ đó thiết kế và phát triển các hợp chất có thể ức chế hoặc điều hòa các tương tác protein quan trọng (Chakraborty et al., 2023). Trong thập kỷ qua, đã có nhiều phương pháp thực nghiệm sinh học đã được nghiên cứu rộng rãi như: Mass Spectrometry (Yakubu et al., 2019), Affinity Purification (Carnes et al., 2019) và kỹ thuật lai hai loại nấm men (Castel et al., 2021). Tuy nhiên, những nghiên cứu này có một số nhược điểm, chẳng hạn như: chi phí cao, tốn nhiều thời gian, tỷ lệ dương tính giả và âm tính giả cao. Do đó, việc phát triển các phương pháp tính toán mới để dự đoán các cặp PPI tiềm năng sẽ có giá trị to lớn đối với các nhà sinh học.

Đến nay, nhiều phương pháp thực nghiệm sinh học đã được giới thiệu để dự đoán PPI. Các phương pháp

Tác giả liên hệ: Nguyễn Tương Tri
Địa chỉ e-mail: ntuongtri@hueuni.edu.vn
DOI: https://doi.org/10.26459/jse.065.2026

này có thể được nhóm lại thành ba loại: Phương pháp dựa trên trình tự, phương pháp dựa trên cấu trúc và phương pháp dựa trên phối tử (Uetz et al., 2008). Trong những năm gần đây, sau sự tiến bộ của công nghệ gen, một lượng lớn dữ liệu trình tự protein đã được thu thập và nhập vào cơ sở dữ liệu. Do đó, các phương pháp dựa trên trình tự để xác định PPI được quan tâm hơn, phần lớn các phương pháp tính toán truyền thống hiện có dựa trên các thuật toán học máy, bao gồm: Rừng xoay chiều (Wang et al., 2018), Máy vector hỗ trợ (Chakraborty et al., 2021; Romero-Molina et al., 2019) và Naive Bayes (Lin & Chen, 2013). Ví dụ, Huang và cộng sự (2015) đã áp dụng các đặc trưng từ Biến đổi Cosine rời rạc và mô hình Biểu diễn thưa thớt có trọng số (Weighted Sparse Representation) để dự đoán PPI từ chuỗi protein; You và cộng sự (2013) đã đề xuất một phương pháp gọi là PCA-EELM, sử dụng bốn loại thông tin trình tự khác nhau để dự đoán PPI; Li và cộng sự (2017) đã đề xuất một phương pháp gọi là PSIPEL kết hợp phương pháp trích xuất tính năng Low Rank Approximation vào rừng xoay chiều để dự đoán PPI từ các chuỗi protein; Zeng và cộng sự (2020) đã phát triển một khung học sâu để dự đoán PPI, sử dụng mạng nơ-ron tích chập văn bản và cửa sổ trượt để nắm bắt các đặc điểm trình tự toàn cục và ngữ cảnh cục bộ từ các protein mục tiêu tương ứng; Chen và cộng sự (2019) đã áp dụng biến đổi Fourier nhanh để thu thập các đặc trưng của protein và đưa chúng vào phép chiếu ngẫu nhiên (Random Projection) để huấn luyện và phát hiện các protein tự tương tác (Pan et al., 2021). Tuy nhiên, đối mặt với sự phức tạp và đa dạng của dữ liệu protein, các phương pháp truyền thống thường gặp khó khăn trong việc dự đoán PPI một cách hiệu quả. Điều này thách thức tìm kiếm những phương pháp mới và tiến bộ để nâng cao độ chính xác và độ nhạy của các mô hình. Vào năm 2022, Jie Pan và công sự đã giới thiệu một phương pháp dự đoán PPI mới dựa trên học tổng hợp và sử dụng phép biến đổi Hilbert từ chuỗi protein sơ cấp. Biến đổi Hilbert là một phương pháp toán học đã chứng minh sức mạnh trong việc xử lý dữ liệu không gian chiều cao, như dữ liệu liên quan đến cấu trúc và đặc trưng của protein, trong khi mô hình học tổng hợp giúp tự động hóa quá trình học và tối ưu hóa các tham số mô hình. Vì vậy, trong nghiên cứu này, chúng tôi giới thiệu một phương pháp kết hợp giữa biến đổi Hilbert và mô hình phân loại tổng hợp DF. DF sử dụng các cây quyết định với cấu trúc tương tự mạng nơ-ron để học tập mô hình dự đoán, cải thiện hiệu suất dự đoán. Bên cạnh đó, biến đổi Hilbert được sử dụng để trích xuất thông tin tiến hóa từ trình tự protein giúp tăng cường khả năng biểu diễn dữ liệu và cung cấp cái nhìn toàn diện hơn về mối quan hệ giữa các amino acid, từ đó cải thiện độ chính xác trong công việc dự đoán PPI.

2. PHƯƠNG PHÁP

2.1. Ma trận PSSM

PSSM là ma trận chứa xác suất của mỗi amino acid và nucleotide ở mỗi vị trí, được sử dụng để thể hiện trình tự protein, đã được trình bày bởi Gribskov và cộng sự (1987) để phân tích sự tương đồng về trình tự của protein. PSSM tạo ra kết quả xuất sắc trong nhiều lĩnh vực, chẳng hạn như dự đoán cấu trúc thứ cấp protein (Wang et al., 2016), dự đoán vùng rối loạn (Zhao et al., 2013) và dự đoán chức năng DNA (Gelfand, 1995).

2.2. Biến đổi Hilbert

Biến đổi Hilbert là một phép biến đổi tín hiệu sang miền thời gian phức được giới thiệu lần đầu tiên bởi nhà toán học người Đức, David Hilbert, vào năm 1905. Ông đã sử dụng biến đổi Hilbert để nghiên cứu các vấn đề trong giải tích toán học. Biến đổi Hilbert lấy một tín hiệu thực làm đầu vào và tạo ra một tín hiệu phức làm đầu ra gọi là tín hiệu phân tích. Tín hiệu phân tích có cùng độ lớn với tín hiệu đầu vào, nhưng pha của nó được thay đổi. Pha của tín hiệu phân tích được tính toán bằng cách sử dụng tích phân Cauchy. Phương trình chuẩn của biến đổi Hilbert cho một tín hiệu thực $x(t)$ là:

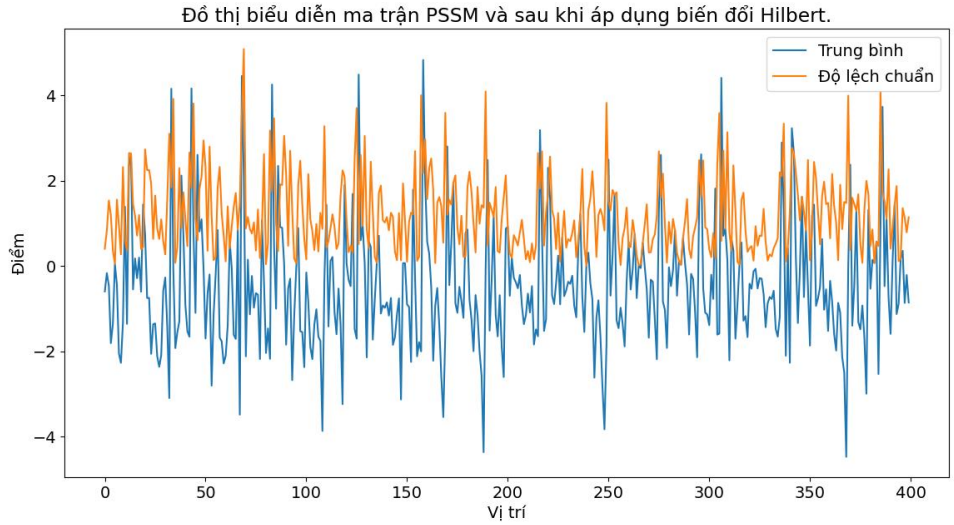
$$z(t) = x(t) * h(t) \tag{1}$$

trong đó, $*$ là phép tích chập, $z(t)$ là tín hiệu phức và hàm $h(t)$ được định nghĩa như sau:

$$h(t) = \frac{1}{\pi t} \tag{2}$$

Mục tiêu của biến đổi Hilbert là tạo ra tín hiệu phân tích từ ma trận PSSM, giúp cung cấp các đặc trưng cần thiết làm đầu vào cho các phương pháp học máy như biên độ và pha của tín hiệu. Trên cơ sở đó, biến đổi Hilbert giúp nâng cao khả năng dự đoán, tăng độ chính xác và hiệu suất trong việc dự đoán tương tác giữa hai protein. Có thể tính toán biến đổi Hilbert thông qua biến đổi Fourier (Kido, 2015). Hình 1 biểu diễn giá trị trung bình và độ lệch chuẩn của ma trận PSSM sau khi được áp dụng biến đổi Hilbert. Tín hiệu phân tích được biểu diễn như sau:

$$z* = DFT^{-1}\left(X[i] \times \left(1 + sign\left(\frac{N}{2} - k\right) sign(k)\right)\right)$$



Hình 1. Đồ thị biểu diễn giá trị trung bình và độ lệch chuẩn của ma trận PSSM sau khi được áp dụng biến đổi Hilbert.

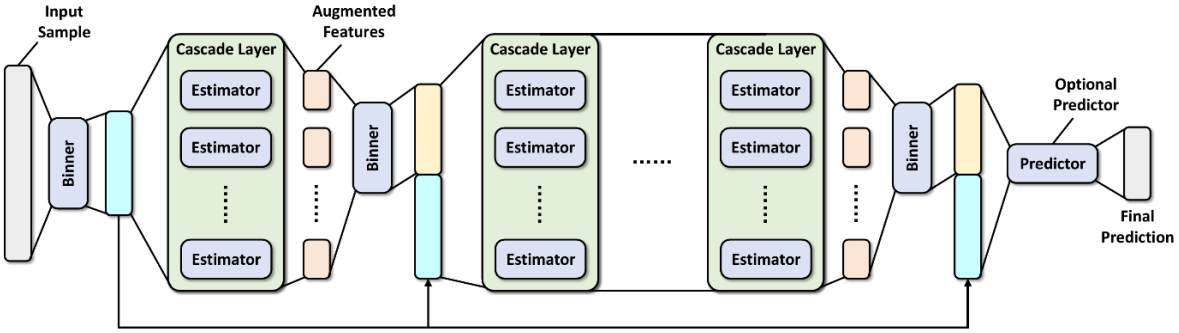
2.3. Rừng xoay chiều

Rừng xoay chiều (Rotation Forest – RoF) là một phương pháp phân loại tổng hợp được giới thiệu lần đầu tiên bởi Juan J. Rodriguez và cộng sự (2006) để giải quyết một trong những điểm yếu lớn của các mô hình thành phần, các mô hình mà cây quyết định của nó chỉ có thể phân vùng không gian đối tượng một cách trực giao. Để thể hiện các ranh giới không trực giao, cây quyết định yêu cầu số lượng phân chia lớn, dẫn đến tình trạng quá khớp (overfitting) và hiệu suất kém. RoF giảm thiểu điều này bằng cách xoay ngẫu nhiên từng cây trong rừng. Trong thực tế, điều này được thực hiện bằng cách chia ngẫu nhiên các đặc trưng của mô hình thành một số phân vùng, lấy mẫu từ các phân vùng này và phương pháp phân tích thành phần chính (Principal Component Analysis – PCA) được áp dụng trên các tập con của đặc trưng đã được chia ngẫu nhiên đó. Quá trình này tạo ra các ma trận xoay bằng cách nhân các tập con đặc trưng ban đầu với ma trận xoay thu được từ PCA. Tính ngẫu nhiên trong phương pháp này được đảm bảo bởi quy trình lấy mẫu bootstrap và phân chia ngẫu nhiên tập đặc trưng. Kết quả là mỗi cây trong rừng quyết định sẽ hoạt động theo một không gian đặc trưng được xoay, giúp tạo ra một tập hợp các mô hình mạnh mẽ, đồng thời vẫn giữ lại được các đặc điểm quan trọng nhất của dữ liệu.

2.4. Rừng sâu

Rừng sâu (Deep Forest – DF), hay còn gọi là gcForest, là thuật toán học máy được đề xuất bởi Zhou và Feng (2019). DF được thiết kế như là lựa chọn thay thế cho mạng nơ-ron sâu (DNN), nhằm mục đích giải quyết các vấn đề phân loại và học sâu mà không cần đến mạng nơ-ron. DF sử dụng các cây quyết định trong một cấu trúc tầng lớp xếp tầng (cascade), thay vì theo từng lớp như DNN, với mục tiêu giảm độ phức tạp của mô hình. Ngoài ra, DF sử dụng các mô hình không tham số như cây quyết định, với tính ngẫu nhiên được cung cấp thông qua phép lấy mẫu và phân vùng nên một trong những ưu điểm lớn của DF là nó có tính siêu tham số thấp hơn so với DNN và độ phức tạp của mô hình được cải thiện dựa trên dữ liệu đầu vào.

Về kiến trúc của mô hình DF, đầu tiên dữ liệu được đưa qua bộ phân loại thành các nhóm (hay bộ chia nhóm-binner) để rời rạc hóa đầu vào. Sau đó, dữ liệu được sử dụng để huấn luyện các cây quyết định tại các lớp xếp tầng. Đầu ra của mỗi lớp sẽ được kết hợp với dữ liệu gốc, tạo thành đặc trưng bổ sung (augmented features) làm đầu vào cho các lớp tiếp theo. Mô hình tiếp tục được xếp tầng đến khi không còn cải thiện hiệu suất. Hình 2 mô tả kiến trúc của mô hình DF.



Hình 2. Kiến trúc của mô hình DF.

3. THỬ NGHIỆM

3.1. Độ đo hiệu suất của mô hình

Để đánh giá hiệu suất của mô hình, chúng tôi sử dụng các số liệu đánh giá bao gồm: độ chính xác tổng quát (acc), độ nhạy (sen), độ đặc hiệu (spec), độ chính xác (pre), F1 và hệ số tương quan Matthews (MCC). Các độ đo này được xác định bởi các công thức sau:

$$Acc = \frac{TP + TN}{TN + FP + TP + FN} \tag{3}$$

$$Sen = \frac{TP}{FN + TP} \tag{4}$$

$$Spec = \frac{TN}{TN + FP} \tag{5}$$

$$Pre = \frac{TP}{FP + TP} \tag{6}$$

$$F1 = 2 \times \frac{Pre \times Sen}{Pre + Sen} \tag{7}$$

$$MCC = \frac{TN \times TP - FN \times FP}{\sqrt{(TP + FN) \times (TP + FP) \times (TN + FP) \times (TN + FN)}} \tag{8}$$

trong đó TN, TP, FN, FP lần lượt là số lượng mẫu dương tính (cặp protein tương tác) được dự đoán là dương tính, số lượng mẫu âm tính (cặp protein không tương tác) được dự đoán là âm tính, số lượng mẫu dương tính được dự đoán là âm tính, và số lượng mẫu âm tính được dự đoán là dương tính.

Ngoài ra, diện tích dưới đường cong ROC (AUROC) và diện tích dưới đường cong Precision-Recall (AUPRC) cũng được sử dụng để đánh giá hiệu suất của mô hình. Diện tích càng lớn thể hiện hiệu suất càng cao của mô hình.

3.2. Tập dữ liệu

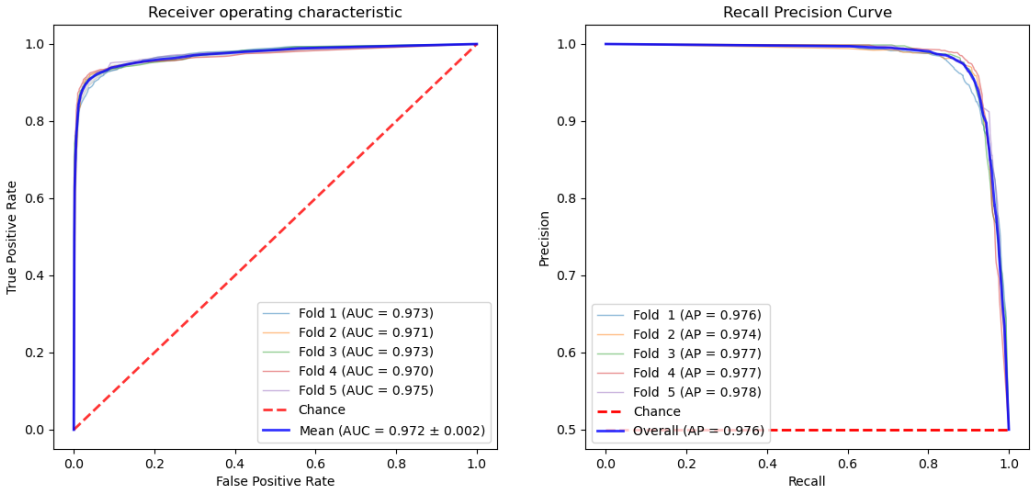
Để đánh giá hiệu suất mô hình dự đoán PPI, cần xây dựng hoặc chọn một bộ dữ liệu chuẩn để đánh giá yếu tố dự đoán một cách chính xác. Trong bài báo này, bộ dữ liệu Yeast (Shen et al., 2007) được sử dụng để làm tập dữ liệu thử nghiệm, bao gồm các cặp tương tác protein được chọn từ cơ sở dữ liệu DIP (Xenarios et al., 2002). Trong đó, có 5.594 cặp tương tác và 5.594 cặp không tương tác sau khi đã loại bỏ các cặp protein có độ dài trình tự nhỏ hơn 50 và có độ đồng nhất lớn hơn 40% bằng công cụ CD-HIT (Fu et al., 2012).

3.3. Kết quả thử nghiệm

Giá trị trung bình và độ lệch chuẩn đã được tính toán từ số liệu thu được, được sử dụng để đánh giá hiệu suất dự đoán của mô hình. Bảng 1 liệt kê hiệu suất dự đoán của phương pháp đề xuất thông qua xác thực chéo 5 lần. Cụ thể, giá trị trung bình của độ chính xác tổng quát, độ nhạy, độ đặc hiệu, độ chính xác, điểm F1 và hệ số tương quan của Matthews lần lượt là 93.42%, 90.45%, 96.39%, 96.17%, 93.22%, 87.00% cùng với độ lệch chuẩn lần lượt là 0.55%, 0.83%, 0.92%, 0.92%, 0.57%, 1.12%. Đồng thời, diện tích dưới đường cong ROC và đường cong Recall-Precision đã được tính toán và hiển thị trong hình 3.

Bảng 1. Hiệu suất mô hình đề xuất thông qua xác thực chéo 5 lần

	Acc (%)	**Sen (%)**	**Spec (%)**	**Pre (%)**	**F1 (%)**	**MCC (%)**	**AUC (%)**	**AUPR (%)**
1	92.49	90.26	94.73	94.48	92.32	85.07	0.9733	0.9757
2	93.79	91.51	96.07	95.88	93.64	87.67	0.9709	0.9737
3	93.39	89.63	97.14	96.91	93.13	87.02	0.9735	0.9775
4	94.14	91.33	96.96	96.78	93.98	88.43	0.9702	0.9766
5	93.29	89.53	97.05	96.81	93.03	86.83	0.9746	0.9779
Trung bình	**93.42 ± 0.55**	**90.45 ± 0.83**	**96.39 ± 0.92**	**96.17 ± 0.92**	**93.22 ± 0.57**	**87.00 ± 1.12**	**0.9725 ± 0.0017**	**0.9763 ± 0.0015**



Hình 3. Diện tích đường cong ROC và Recall-Precision của mô hình đề xuất thông qua xác thực chéo 5 lần.

3.4. So sánh với các phương pháp phân loại khác

Tiếp theo, chúng tôi sẽ so sánh khả năng dự đoán của phương pháp đề xuất với các phương pháp phổ biến khác. Bảng 2 liệt kê kết quả dự đoán với một số tiêu chí đánh giá như độ chính xác tổng quát, độ nhạy, độ đặc hiệu và độ chính xác. Kết quả cho thấy khả năng dự đoán của phương pháp đề xuất cao hơn so với các phương pháp phổ biến khác, trong đó độ chính xác cao hơn 1.49%; độ nhạy, độ đặc hiệu và độ phân tán cao hơn lần lượt là 0.19%, 2.35% và 1.86%.

Bảng 2. So sánh hiệu suất của phương pháp đề xuất với các phương pháp hiện có. Kết quả được tác giả công bố, giá trị đậm thể hiện hiệu suất cao nhất

Phương pháp	Bộ phân loại	Acc (%)	Sen (%)	Pre (%)	MCC (%)
Guo và cộng sự (2008)	ACC + SVM	89.33	88.87	89.93	N/A[1]
Yang và cộng sự (2020)	LD + KNN	86.15	81.30	90.24	N/A
Yang và cộng sự (2020)	3-MER + CNN	90.26	88.14	91.65	82.38
Zhou và cộng sự (2011)	LD + SVM	88.56	87.37	89.50	77.15
An và cộng sự (2019)	PSSMMF + ELLM	90.48	90.26	90.58	82.84
You và cộng sự (2013)	PCA + ELLM	87.00	86.15	87.59	77.36
Jie Pan và cộng sự (2022)	DHT + RoF	91.93	89.78	93.82	85.14
Phương pháp đề xuất	DHT + DF	**93.42**	**90.45**	**96.17**	**87.00**

3.5. So sánh về cách sử dụng đặc trưng

Trong phần này, chúng tôi áp dụng phương pháp đề xuất để so sánh hai phương pháp sử dụng đặc trưng. Bảng 3 liệt kê hiệu suất của mô hình đề xuất sử dụng đặc trưng thông thường và sử dụng đặc trưng đề xuất. Kết quả, phương pháp lấy đặc trưng đề xuất đạt kết quả cao hơn so với cách lấy thông thường được thể hiện qua các độ đo: độ chính xác tổng quát cao hơn 0.49%; độ nhạy, độ đặc hiệu, độ chính xác đều cao hơn lần lượt là 1.01%, 0.01%, 0.93%. Ngoài ra, tốc độ huấn luyện trung bình ở mỗi tầng của mô hình DF với cách lấy đặc trưng đề xuất nhanh gấp 5.72 lần so với cách sử dụng đặc trưng thông thường.

Bảng 3. So sánh tốc độ của mô hình DF qua các cách sử dụng đặc trưng từ biến đổi Hilbert

Phương pháp sử dụng đặc trưng	Thời gian trung bình mỗi tầng của DF (s)	Acc (%)	Sen (%)	Pre (%)	MCC (%)
Chuỗi protein (độ dài cố định 50 và được áp dụng biến đổi Hilbert)	3713	92.93	89.44	96.16	86.07
Phương pháp sử dụng đặc trưng đề xuất	**649**	**93.42**	**90.45**	**96.17**	**87.00**

3.6. Thảo luận và đánh giá

3.6.1. Đánh giá hiệu suất mô hình đề xuất

Từ kết quả thử nghiệm cho thấy phương pháp đề xuất không những đạt được khả năng dự đoán với độ chính xác tổng quát lên đến 93%, mà còn thể hiện được sự mạnh mẽ, khả năng phân biệt tốt giữa các lớp dữ liệu, được thể hiện qua giá trị trung bình AUPRC và AUROC đạt được lần lượt là 0.98 và 0.97. Khi so sánh

[1] N/A: không được cung cấp.

với các phương pháp phổ biến khác, phương pháp đề xuất cho thấy khả năng dự đoán có sự cải thiện không chỉ nằm ở độ chính xác tổng quát, mà còn là sự cân bằng giữa độ nhạy, độ đặc hiệu và độ chính xác. Ngoài ra, kỹ thuật sử dụng đặc trưng đề xuất giúp cải thiện hiệu suất dự đoán của mô hình với thời gian thực hiện trên mỗi tầng của mô hình DF giảm xuống còn 17% so với cách sử dụng đặc trưng truyền thống. Điều này giúp mô hình có thể được áp dụng linh hoạt hơn trong các tình huống thực tế, nơi mà phản hồi nhanh chóng chính là yếu tố then chốt.

3.6.2. Thảo luận về kết quả đánh giá

Chúng tôi đã thực hiện đánh giá hiệu suất của phương pháp đề xuất và so sánh với các phương pháp khác. Kết quả thực nghiệm đã cho thấy phương pháp đề xuất có hiệu quả, hiệu suất tốt dựa trên các giá trị độ đo và thời gian thực hiện tính toán. Để có được kết quả đó, một phần là nhờ vào các điều kiện đạt được trong quá trình thực hiện thao tác với dữ liệu cũng như chọn các đặc trưng làm đầu vào cho mô hình học máy. Trong đó, việc áp dụng thuật toán Hilbert để trích xuất các đặc trưng của protein từ ma trận PSSM giúp mô hình nắm bắt được các thông tin phức tạp và chiều cao một cách rõ ràng, đặc biệt là khi sử dụng biên độ và pha của tín hiệu phân tích làm đặc trưng, có thể bỏ qua bước chuẩn hóa độ dài chuỗi, qua đó, lượng thông tin từ dữ liệu được trích xuất tổng quát, rõ ràng và đầy đủ hơn, đảm bảo mô hình có thể triển khai đạt hiệu quả tốt hơn. Ngoài ra, mô hình DF, được biết đến với khả năng xử lý dữ liệu không cân đối và có cấu trúc phức tạp mà không cần điều chỉnh tham số như các mô hình học sâu khác và RF được sử dụng để huấn luyện ở mỗi lớp của DF cho phép mô hình học được các mức độ phức tạp khác nhau từ dữ liệu, khả năng nắm bắt được các mối quan hệ phi tuyến tính giữa các đặc trưng và PPI.

4. KẾT LUẬN

Trong bài báo này, chúng tôi đã giới thiệu một phương pháp làm tăng hiệu quả và hiệu suất dự đoán PPI bằng cách kết hợp phép biến đổi Hilbert vào mô hình phân loại DF. Cụ thể, RoF được áp dụng như một mô hình huấn luyện ở mỗi lớp của DF nhằm mục đích tăng cường khả năng phân loại. Cùng với đó, phép biến đổi Hilbert khi được sử dụng lên ma trận PSSM giúp làm nổi bật các mối quan hệ phi tuyến tính và phức tạp giữa các amino acid trong chuỗi protein, qua đó nâng cao độ chính xác của mô hình phân loại. Kết quả thử nghiệm đã chứng minh phương pháp đề xuất mang lại hiệu quả và hiệu suất tốt. Tuy nhiên, kết quả hiện tại chỉ mới thử nghiệm với mô hình RoF và những tham số mặc định, do đó chưa phản ánh được hết tiềm năng của phương pháp đề xuất. Trong tương lai, chúng tôi sẽ nghiên cứu sâu hơn và tinh chỉnh tham số của mô hình để cải thiện, và mở rộng ứng dụng của phương pháp vào thực tiễn.

TÀI LIỆU THAM KHẢO

An, J.-Y., Zhou, Y., Zhao, Y.-J., & Yan, Z.-J. (2019). An efficient feature extraction technique based on local coding PSSM and multifeatures fusion for predicting protein–protein interactions. *Evolutionary Bioinformatics, 15*, 1176934319879920. https://doi.org/10.1177/1176934319879920

Carnes, R. M., Kesterson, R. A., Korf, B. R., Mobley, J. A., & Wallis, D. (2019). Affinity purification of NF1 protein–protein interactors identifies keratins and neurofibromin itself as binding partners. *Genes, 10*(9), 650. https://doi.org/10.3390/genes10090650

Castel, P., Holtz-Morris, A., Kwon, Y., Suter, B. P., & McCormick, F. (2021). DoMY-Seq: A yeast two-hybrid–based technique for precision mapping of protein–protein interaction motifs. *Journal of Biological Chemistry, 296*, 100023. https://doi.org/10.1074/jbc.RA120.014284

Chakraborty, A., Mitra, S., Bhattacharjee, M., De, D., & Pal, A. J. (2021). Determining protein–protein interaction using support vector machine: A review. *IEEE Access, 9*, 12473–12490. https://doi.org/10.1109/ACCESS.2021.3051006

Chakraborty, A., Mitra, S., Bhattacharjee, M., De, D., & Pal, A. J. (2023). Determining human-coronavirus protein–protein interaction using machine intelligence. *Medicine in Novel Technology and Devices, 18*, 100228. https://doi.org/10.1016/j.medntd.2023.100228

Chen, Z.-H., You, Z.-H., Li, L.-P., Wang, Y.-B., Wong, L., & Yi, H.-C. (2019). Prediction of self-interacting proteins from protein sequence information based on random projection model and fast Fourier transform. *International Journal of Molecular Sciences, 20*(4), 930. https://doi.org/10.3390/ijms20040930

Copeland, R. A. (1994). Introduction to protein structure. In J. P. J. M. van Os (Ed.), *Methods for protein analysis* (pp. 1–16). Springer. https://doi.org/10.1007/978-1-4757-1505-7_1

De Las Rivas, J., & Fontanillo, C. (2010). Protein–protein interactions essentials: Key concepts to building and analyzing interactome networks. *PLoS Computational Biology, 6*(6), e1000807. https://doi.org/10.1371/journal.pcbi.1000807

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics, 28*(23), 3150–3152. https://doi.org/10.1093/bioinformatics/bts565

Gelfand, M. S. (1995). Prediction of function in DNA sequence analysis. *Journal of Computational Biology, 2*(1), 87–115. https://doi.org/10.1089/cmb.1995.2.87

Gribskov, M., McLachlan, A. D., & Eisenberg, D. (1987). Profile analysis: Detection of distantly related proteins. *Proceedings of the National Academy of Sciences, 84*(13), 4355–4358. https://doi.org/10.1073/pnas.84.13.4355

Guo, Y., Yu, L., Wen, Z., & Li, M. (2008). Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. *Nucleic Acids Research, 36*(9), 3025–3030. https://doi.org/10.1093/nar/gkn159

Huang, Y.-A., You, Z.-H., Gao, X., Wong, L., & Wang, L. (2015). Using weighted sparse representation model combined with discrete cosine transformation to predict protein–protein interactions from protein sequence. *BioMed Research International, 2015*, 902198. https://doi.org/10.1155/2015/902198

Kido, K. (2015). Discrete Fourier transform. In *Digital Fourier analysis: Fundamentals* (pp. 77–105). Springer. https://doi.org/10.1007/978-1-4614-9260-3_4

Li, J.-Q., You, Z.-H., Li, X., Ming, Z., & Chen, X. (2017). PSPEL: In silico prediction of self-interacting proteins from amino acid sequences using ensemble learning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 14*(5), 1165–1172. https://doi.org/10.1109/TCBB.2017.2649529

Lin, X., & Chen, X. (2013). Heterogeneous data integration by tree-augmented naïve Bayes for protein–protein interactions prediction. *Proteomics, 13*(2), 261–268. https://doi.org/10.1002/pmic.201200326

Pan, J., Li, L.-P., Yu, C.-Q., You, Z.-H., Ren, Z.-H., & Tang, J.-Y. (2021). FWHT-RF: A novel computational approach to predict plant protein–protein interactions via an ensemble learning method. *Scientific Programming, 2021*, Article 1607946. https://doi.org/10.1155/2021/1607946

Pan, J., Wang, S., Yu, C., Li, L., You, Z., & Sun, Y. (2022). A novel ensemble learning-based computational method to predict protein–protein interactions from protein primary sequences. *Biology, 11*(5), 775. https://doi.org/10.3390/biology11050775

Rodriguez, J. J., Kuncheva, L. I., & Alonso, C. J. (2006). Rotation forest: A new classifier ensemble method. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 28*(10), 1619–1630. https://doi.org/10.1109/TPAMI.2006.211

Romero-Molina, S., Ruiz-Blanco, Y. B., Harms, M., Münch, J., & Sanchez-Garcia, E. (2019). PPI-Detect: A support vector machine model for sequence-based prediction of protein–protein interactions. *Journal of Computational Chemistry, 40*(11), 1233–1242. https://doi.org/10.1002/jcc.25780

Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., Li, Y., & Jiang, H. (2007). Predicting protein–protein interactions based only on sequences information. *Proceedings of the National Academy of Sciences, 104*(11), 4337–4341. https://doi.org/10.1073/pnas.0607879104

Uetz, P., Titz, B., & Cagney, G. (2008). Experimental methods for protein interaction identification and characterization. In *Protein–protein interactions: Methods and applications* (pp. 1–32). Humana Press. https://doi.org/10.1007/978-1-84800-125-1_1

Wang, L., Zhang, M., Yang, C., Li, Y., & Wang, Y. (2018). Using two-dimensional principal component analysis and rotation forest for prediction of protein–protein interactions. *Scientific Reports, 8*(1), 12874. https://doi.org/10.1038/s41598-018-30694-1

Wang, Y., Cheng, J., Liu, Y., & Chen, Y. (2016). Prediction of protein secondary structure using support vector machine with PSSM profiles. In *2016 IEEE Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)* (pp. 502–505). IEEE. https://doi.org/10.1109/ITNEC.2016.7560411

Xenarios, I., Salwínski, Ł., Duan, X. J., Higney, P., Kim, S.-M., & Eisenberg, D. (2002). DIP, the database of interacting proteins: A research tool for studying cellular networks of protein interactions. *Nucleic Acids Research, 30*(1), 303–305. https://doi.org/10.1093/nar/30.1.303

Yakubu, R. R., Nieves, E., & Weiss, L. M. (2019). The methods employed in mass spectrometric analysis of posttranslational modifications (PTMs) and protein–protein interactions (PPIs). In *Post-translational modifications in health and disease* (pp. 169–198). Springer. https://doi.org/10.1007/978-3-030-15950-4_10

Yang, X., Yang, S., Li, Q., Wuchty, S., & Zhang, Z. (2020). Prediction of human-virus protein–protein interactions through a sequence embedding-based machine learning method. *Computational and Structural Biotechnology Journal, 18*, 153–161. https://doi.org/10.1016/j.csbj.2019.12.005

You, Z.-H., Lei, Y.-K., Zhu, L., Xia, J., & Wang, B. (2013). Prediction of protein–protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis. *BMC Bioinformatics, 14*(Suppl. 8), S10. https://doi.org/10.1186/1471-2105-14-S8-S10

Zeng, M., Zhang, F., Wu, F.-X., Li, Y., Wang, J., & Li, M. (2020). Protein–protein interaction site prediction through combining local and global features with deep neural networks. *Bioinformatics, 36*(4), 1114–1120. https://doi.org/10.1093/bioinformatics/btz699

Zhao, T.-H., Shi, X.-H., Wang, L., Zhang, Y.-N., & cộng sự. (2013). A novel method of predicting protein disordered regions based on sequence features. *BioMed Research International, 2013*, Article 414327. https://doi.org/10.1155/2013/414327

Zhou, Y. Z., Gao, Y., & Zheng, Y. Y. (2011). Prediction of protein–protein interactions using local description of amino acid sequence. In *Advanced Data Mining and Applications* (pp. 254–262). Springer. https://doi.org/10.1007/978-3-642-22456-0_37

Zhou, Z.-H., & Feng, J. (2019). Deep forest. *National Science Review, 6*(1), 74–86. https://doi.org/10.1093/nsr/nwy108

Applying Hilbert transform and ensemble learning to predicting the protein-protein interactions

Mai Xuan Van, Tran Vo Hoang Nguyen, Tran Hoai Nhan, Nguyen Tuong Tri

Faculty of Informatics, University of Education, Hue University

ARTICLE INFO

Article history:
Received: 04 July 2024
Received in revised form: 27 December 2024
Accepted: 02 January 2025
Published: 20 April 2026

Keywords:
Protein Interactions;
Ensemble Learning;
Hilbert Transform.

Corresponding author:
Nguyen Tuong Tri
E-mail address:
ntuongtri@hueuni.edu.vn

ABSTRACT

Determining protein-protein interactions plays an important role in understanding the structure and biological function of cells. Accurate prediction of interactions remains challenging. In this study, we introduce a method that combine the Hilbert transform on evolutionary information for the feature extraction step and the Deep Forest classification model for the prediction step. The use of Hilbert transformation aims to enhance the evolutionary information stored in the protein sequence. Additionally, Deep Forest, a powerful ensemble classification model, is used to improve prediction performance. The proposed method is evaluated on the benchmark Yeast dataset and compared with other existing methods. The results show that our method is highly effective in predicting PPI. The datasets and source code of the method are provided at https://github.com/thnhanhub/FeatPSSM.git