Project 03 Prediction of Nitrotyrosine

Xác định vị trí Nitrotyrosine trong protein

State-of-the-Arts: Published: 8 March 2021

https://www.mdpi.com/1422-0067/22/5/2704

Nilamyani, Andi N., Firda N. Auliah, Mohammad A. Moni, Watshara Shoombuatong, Md M. Hasan, and Hiroyuki Kurata. 2021. "PredNTS: Improved and Robust Prediction of Nitrotyrosine Sites by Integrating Multiple Sequence Features" International Journal of Molecular Sciences 22, no. 5: 2704. https://doi.org/10.3390/ijms22052704

I. Bài toán

Tyrosine: Y
Nitrotyrosine là sản phẩm của quá trình nitrat hóa tyrosine qua trung gian của các loại nitơ phản ứng. Là một chỉ báo về tổn thương tế bào và tình trạng viêm, protein nitrotyrosine đóng vai trò tiết lộ sự thay đổi sinh học liên quan đến các bệnh khác nhau hoặc oxy hóa do stress. Việc xác định chính xác vị trí nitrotyrosine cung cấp nền tảng quan trọng để làm sáng tỏ thêm cơ chế của quá trình nitro hóa protein.

II. Tập dữ liệu

Tập dữ liệu training: 2382 mẫu với 1191 positive samples and 1191 negative samples

Tập dữ liệu test: 1225 mẫu: 203 positive samples and 1022 negative samples.

Website: http://kurata14.bio.kyutech.ac.jp/PredNTS/index.php

Định dạng dữ liệu:

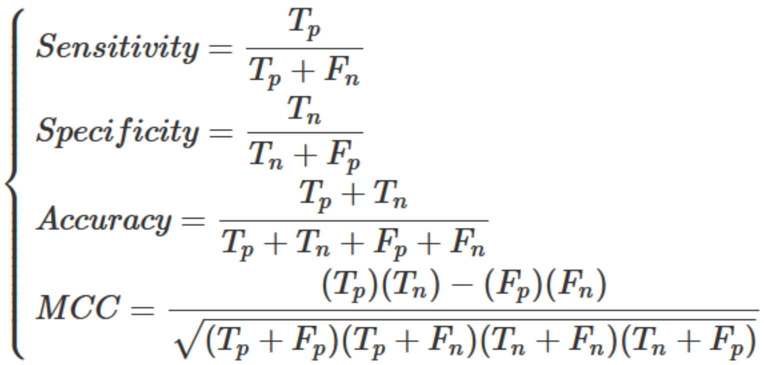
Chuỗi 41 ký tự (amino acids) với Tyrosine (Y) ở giữa.

Ví dụ: ANTETTKVNGSLETKYRWTE**Y**GLTFTEKWNTDNTLGTEITV

Cần dự đoán Y này có bị nitrat hóa hay không: phân lớp nhị phân chuỗi trên

III. Đánh giá kết quả

III.1. Các độ đo



$$\begin{cases} Sensitivity = \dfrac{T_p}{T_p + F_n} \\ Specificity = \dfrac{T_n}{T_n + F_p} \\ Accuracy = \dfrac{T_p + T_n}{T_p + T_n + F_p + F_n} \\ MCC = \dfrac{(T_p)(T_n) - (F_p)(F_n)}{\sqrt{(T_p + F_p)(T_p + F_n)(T_n + F_n)(T_n + F_p)}} \end{cases}$$

III.2. State-of-the-Arts

0.522 for Sn, 0.809 for Sp, 0.761 for Acc, and 0.286 for MCC

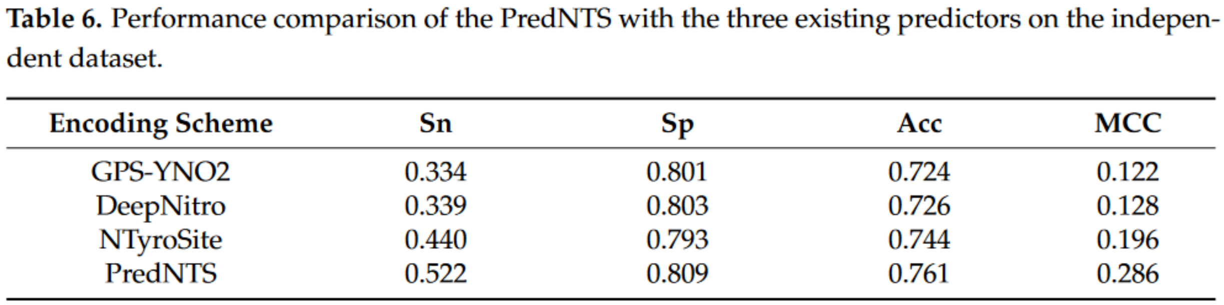


Table 6. Performance comparison of the PredNTS with the three existing predictors on the independent dataset.

Encoding Scheme	Sn	Sp	Acc	MCC
GPS-YNO2	0.334	0.801	0.724	0.122
DeepNitro	0.339	0.803	0.726	0.128
NTyroSite	0.440	0.793	0.744	0.196
PredNTS	0.522	0.809	0.761	0.286

IV. Phương pháp State-of-the-Arts

3.2. Sequence Encoding Scheme

The binary amino acid encoding scheme was used to encode position information from the sequence windows [23–25]. Here, by adopting the binary encoding, we converted a 41 amino acid sequence, including the gap that is represented as (-), into a 861 (=41 x 21) - dimensional feature vector.

The physicochemical properties of amino acids have been extracted from the Aaindex database 24 (version 9.1) [26]. Herein, we used 15 types of AAindex properties to generate a 615 (=41 15)-dimensional vector.

The composition of k-spaced amino acid pairs (CKSAAP) encoding is the composition of the k-spaced residue pairs in the window, which is widely used in a protein bioinformatics field [14,20,27]. In this scheme, k represents the gap length of two amino acids. For example, k = 0 provides 400 amino acid residue pairs (i.e., AA, AC, AD, ..., YY). At k = 0, 1, 2, 3, and 4, it generates a 2000-dimensional feature vector. Details of the CKSAAP encoding are described in our previous studies [14,25].

The K-mer encoding is widely used in the field of genomics and bioinformatics [23,28–31].

We employed the K-mer to minimize the impact of an arbitrary starting point. The K-mer encodes a monopeptide into a 20-dimensional feature vector at K = 1. Similarly, at K = 2 and 3, it encodes dipeptides and tripeptides, which generates an 8020-dimensional feature vector.

3.3. Feature Selection

We considered the RFE as a feature selection approach to remove non-essential features from the dataset [32]. This method was classified as a wrapper method, which started from building a learning model for the entire dataset. We calculated the important scores from each predictor and trimmed the least important features out of the current set of features.

The procedure is repeated until the number of optimal performance features converges.

The ‘rfe’ function from the ‘Caret’ R package was adopted to obtain the important features.

3.4. Machine Learning Algorithm

The RF is a supervised and ensemble machine learning classifier that combines multiple tree-based representations to create a more powerful and interpretable model. It is widely used in protein bioinformatics research [33–43]. It performs as a huge assortment of uncorrelated decision trees, and the votes are carried to decide the final classification from the whole trees. The prediction model of our PredNTS was built using the ‘RandomForest’ R package (https://cran.r-project.org/web/packages/randomForest/ (accessed on 1 March 2021)). In addition, we compared the RF algorithm with the naïve Bayes (NB) and k-nearest neighbor (KNN) algorithms. An R package of an NB algorithm (https://cran.r-project.org/web/packages/naivebayes/ (accessed on 1 March 2021)) was employed to classify the nitrotyrosine proteins, while the R package (https: //rpubs.com/njvijay/16444 (accessed on 1 March 2021)) was used to build the KNN model.