Project 01 Phosphoglycerylation

Xác định các vị trí Lysine bị phosphoglyceryl hóa

State-of-the-art: 20 December 2020

https://www.mdpi.com/2073-4425/11/12/1524

Chandra, Abel A., Alok Sharma, Abdollah Dehzangi, and Tatushiko Tsunoda. 2020. "RAM-PGK: Prediction of Lysine Phosphoglycerylation Based on Residue Adjacency Matrix" Genes 11, no. 12: 1524. https://doi.org/10.3390/genes11121524

I. Bài toán

Cho mỗi protein ở dạng chuỗi amino acids. Hãy xác định các vị trí Lysine nào bị phosphoglyceryl hóa.

Sửa đổi sau dịch mã (PTM) là một quá trình sinh học trong đó diễn ra sự thay đổi enzym trong protein và điều này xảy ra sau quá trình dịch mã protein trong ribosome. Sự phát triển của proteomics thông lượng cao từ PTM và các enzym biến đổi protein theo vị trí cụ thể đã dẫn đến sự quan tâm của cộng đồng khoa học [1]. Lysine là một trong những axit amin bị biến đổi mạnh nhất trong số 20 axit amin hình thành mã di truyền [2,3]. Dựa trên các báo cáo [4], lysine có thể dễ dàng trải qua các biến đổi cộng hóa trị. Một số biến đổi cộng hóa trị có thể phát hiện được là metyl [5], punyl [6], succinyl [7], propionyl [8], crotonyl [9], acetyl [10], glycation [11] và glycosyl [12]. Kết quả của những sửa đổi và thay đổi đối với các enzym điều hòa, đã gây ra một số bệnh ở người như bệnh đa xơ cứng, rối loạn thoái hóa thần kinh, bệnh tim mạch vành, viêm khớp thấp khớp, bệnh celiac, cao huyết áp và tăng huyết áp cơ bản [13–16] .

Phosphoglyceryl hóa là một loại biến đổi không phải enzym được tìm thấy trong tế bào người và mô gan có thể được phân loại là một PTM mới được phát hiện [17,18]. Nó có liên quan nhiều đến các bệnh tim mạch như suy tim vì liên kết của nó với quá trình đường phân và chuyển hóa glucose [19,20]. Phosphoglyceryl hóa là một quá trình thuận nghịch được gọi là 3-phosphoglyceryl-lysine (pgK), xảy ra khi chất trung gian glycolytic chính (1,3-BPG) phản ứng với dư lượng lysine [18]. Hợp chất này ảnh hưởng đến các enzym đường phân và tích tụ trên các tế bào tiếp xúc nhiều với glucose, và do đó tạo ra quá trình phản hồi tiềm năng, gây ra sự tích tụ và chuyển hướng của các chất trung gian đường phân thành các con đường sinh tổng hợp thay thế. PTM mới này, là quá trình phosphoglyceryl hóa, cần được nghiên cứu thêm để việc xác định và phân tích nó trở nên rõ ràng hơn nhằm nhận ra cơ chế chọn lọc và vai trò điều chỉnh đối với việc cải thiện quy trình chẩn đoán và điều trị của những người bị ảnh hưởng.

II. Tập dữ liệu

Tập dữ liệu có 91 sequences, với tổng cộng số amino acid LYSINE là 3360: trong đó có 111 vị trí bị phosphoglyceryl và 3249 vị trí không bị.

Tập dữ liệu training:

Tập dữ liệu test:

Website:

Định dạng dữ liệu: EEIRCYVRDKEMNSQVYSRLTSRGTVKVKSSNIQV

Chuỗi này gồm có 3 Lysine (mã hóa bởi ký tự K, còn ký tự L mã hóa cho Leusine), trong đó Lysine đầu tiên bị phosphoglyceryl hóa, còn 2 Lysine còn lại không bị.

III. Đánh giá kết quả

The assessment of our predictor's performance was carried out via the sixfold cross-validation scheme. The statistical measures discussed in the previous section were calculated for each fold and were then finally averaged to determine the predictor's overall performance. The validation scheme of this work is generally known as the n-fold cross-validation scheme. Apart from this, there are two other well-known schemes to determine the effectiveness of a predictor, which are called the independent dataset test and the jackknife test [53,54]. From the three validation schemes, the jackknife scheme is said to be the least arbitrary, whereby it yields distinct results on the dataset [55]. However, we employed the n-fold cross-validation scheme primarily to avoid the high computational time.

The sixfold cross-validation scheme was carried out using the steps highlighted below:

Step 1: Divide the dataset into six similar parts.

Step 2: Combine the five parts and apply the cleaning treatment to balance the positive and negative classes. Train the predictor using this balanced dataset and test it with the part left out.

Step 3: Set the predictor parameters with the train set.

Step 4: Acquire the statistical measures on the test set.

Step 5: Repeat steps 2 to 4 for the remainder of the folds.

III.1. Các độ đo

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Accuracy = \frac{TN + TP}{FN + FP + TN + TP}$$

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

III.2. State-of-the-Arts

Table 2. Comparison of the iPGK-PseAAC, CKSAAP_PhoglySite and Bigram-PGK methods with the RAM-PGK predictor using 6-fold cross-validation scheme. Highest values of the metrics are highlighted in bold. MCC: Mathews correlation coefficient.

Predictor	Sensitivity	Specificity	Precision	Accuracy	MCC
CKSAAP_PhoglySite [17]	0.3494	0.6722	0.0358	0.6616	0.0090
iPGK-PseAAC [43]	0.0185	0.9791	0.0064	0.9473	−0.5048
Bigram-PGK [42]	0.4055	0.6639	0.0428	0.6554	0.0292
RAM-PGK (No Feature Selection)	0.5380	0.6328	0.0472	0.6298	0.0631
RAM-PGK	0.5741	0.6436	0.0531	0.6414	0.0824

IV. Phương pháp State-of-the-Arts

residue adjacency matrix feature set

support vector machine algorithm