ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

Dự đoán các biến đổi của protein sau khi tổng hợp sử dụng các kỹ thuật khai phá dữ liệu

GVHD: TS. **Đặng Thanh Hải**

SV: Phạm Quốc Hưng

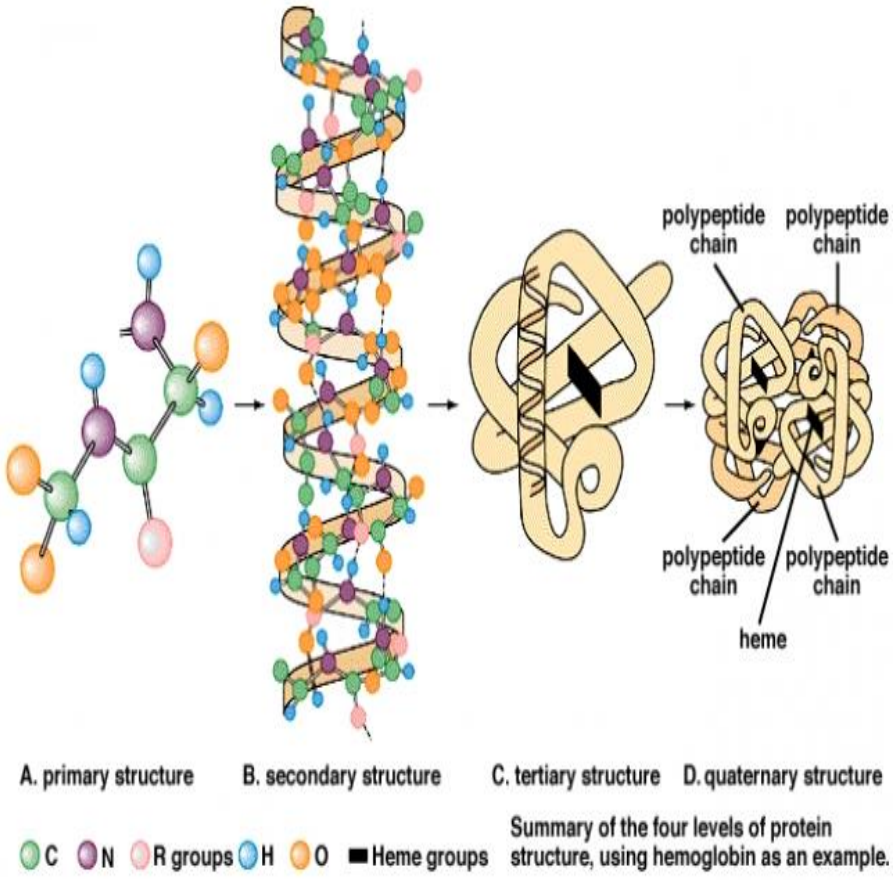
Lớp: K56CB

Nội dung

- Protein và các biến đổi sau tổng hợp
  - Quá trình phospho hoá
- Các mô hình dự đoán vị trí protein bị phospho hoá
- Tổng quan về khai phá dữ liệu
  - Khai phá tập phổ biến và luật kết hợp
  - Thuật toán Máy Vector hỗ trợ (SVM)
- Mô hình dự đoán phospho hoá triển khai
- Thực nghiệm
- Kết luận và hướng phát triển

Protein là gì ?

- Protein là một đại phân tử trong tế bào
- Protein gồm 4 cấu trúc
  - Cấu trúc bậc 1: là các axit amin liên kết với nhau bằng liên kết peptit
  - Cấu trúc bậc 2: là sự sắp xếp đều đặn các chuỗi polypeptide trong không gian
  - Cấu trúc bậc 3: protein cuộn với nhau thành từng búi có hình dạng đặc trưng
  - Cấu trúc bậc 4: là cấu trúc của nhiều protein liên kết với nhau



Tầm quan trọng của protein

- Protein tham gia vào tất cả các quá trình hoạt động của tế bào sống
  - Kháng thể
  - Enzyme
  - Thông tin
  - Thành phần cấu trúc
  - Vận chuyển dự trữ

Biến đổi protein sau khi tổng hợp

- Là quá trình biến đổi một vị trí nào đó trên protein sau khi được tổng hợp (dịch mã)
  - Có hơn 200 loại biến đổi đã được xác định
  - Các nhà hoá/sinh học vẫn chưa hiểu hết được phần lớn các loại biến đổi này
  - Được xác định bằng thí nghiệm hoá sinh
    - Kỹ thuật Phổ khối lượng (Mass Spectrometry) là một kỹ thuật điển hình
    - Tốn kém và mất thời gian, thậm chí là rất khó
  - Protein phospho hóa là quá trình biến đổi quan trọng nhất của protein
    - Khóa luận này tập trung vào bài toán liên quan đến quá trình Phospho hoá

Quá trình phospho hoá

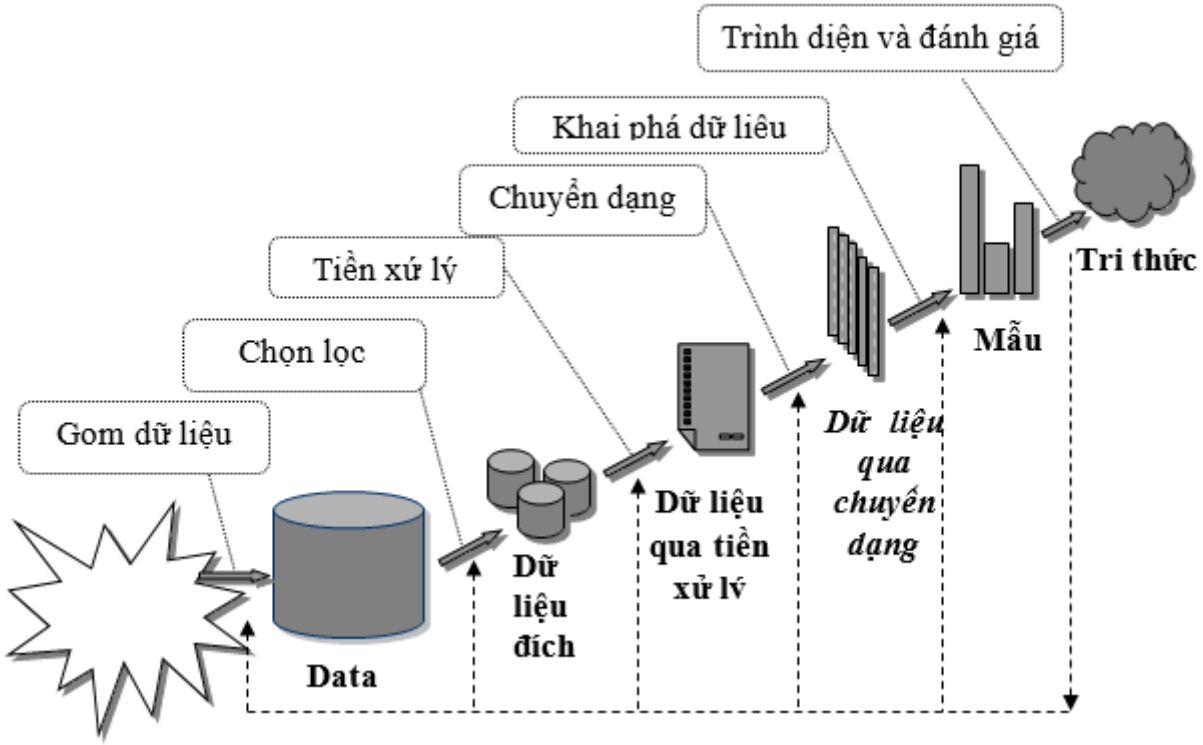
- Là quá trình thêm một nhóm P04 vào một vị trí cụ thể trên chuỗi protein sau khi được tổng hợp
  - Các axit amin serine (S), threomine (T), tyrosine (Y)
  - Được chứng minh là đóng một vai trò vô cùng quan trọng trong việc quyết định đến chức năng của protein
  - Nhận được rất nhiều sự quan tâm nghiên cứu của cộng đồng hoá/sinh học lẫn Tin sinh học

Một số mô hình dự đoán phospho hóa hiện có

- Cách tiếp cận dựa trên trên kỹ thuật học máy khai phá dữ liệu
  - Máy vector hỗ trợ - SVMs
  - Cây quyết định
  - Các thuật toán di truyền
- Cách tiếp cận dựa trên thông tin đầu vào
  - Sử dụng các số lượng các vị trí xung quanh axit amin bị phospho hóa để dự đoán
- Cách tiếp cận sử dụng hay không sử dụng đến thông tin cấu trúc
- Cách tiếp cận có kinase đặc hiệu hay không có kinase đặc hiệu

Tổng quan về khai phá dữ liệu

- Phát hiện tri thức là quá trình tìm ra các dữ liệu mới hữu ích trong dữ liệu và khai phá dữ liệu là một bước quan trọng.
- Quá trình phát hiện tri thức từ CSDL gồm 6 bước:



Tổng quan về khai phá dữ liệu

- Ở mức cao – tổng quát: hai mục tiêu chủ yếu của bài toán khai phá dữ liệu là dự báo và mô tả
- Ở mức chi tiết – cụ thể:
  - Mô tả khái niệm
  - Quan hệ kết hợp
  - Phân lớp
  - Phân cụm
  - Hồi quy
  - Phát hiện dữ liệu bất thường/ngoại lai

Luật kết hợp và tập phổ biến

- Cho I là một tập các item (mục)
- Cho X = $\{i_1,i_2,...,i_k\} \subseteq$ I được gọi là một itemset (tập mục) hoặc là tập k-item nếu X có tất cả k mục.
- Một giao dịch là một cặp T=(tid, i), một CSDL giao dịch D gồm tất cả các giao dịch T.
- Độ hỗ trợ của một tập mục X trong D bao gồm tất cả các giao dịch trong D có hỗ trợ bởi X:

$$support\ (X, D) \coloneqq \{tid \mid (tid, i), X\ \subseteq\ i\}$$

Luật kết hợp và tập phổ biến

- Cho D là CSDL giao dịch trên một tập mục I, và σ là ngưỡng độ hỗ trợ tối thiểu. Các tập mục phổ biến trong D với độ hỗ trợ σ được ký hiệu là

$$F(D,\sigma) \coloneqq \{X \subseteq I \mid support(X,D) \geq \sigma\}$$

- Độ tin cậy của luật X ⇒ Y trong D là xác suất giao dịch chứa cả X và Y trên tổng những giao dịch có X:

$$confidence(X \Rightarrow Y, D) \coloneqq P(Y|X) = \frac{support(\mathrm{X} \cup \mathrm{Y}, D)}{support(X,D)}$$

- Cho D là một tập CSDL giao dịch trên tập mục I, σ là độ hỗ trợ tối thiểu,γ là độ tin cậy tối thiểu. Tập những luật phổ biến với σ và γ được ký hiệu

$$R(D,\sigma,\gamma) \coloneqq \{X \Rightarrow Y | X,Y \subseteq I, X \cap Y = \{\emptyset\}, X \cup Y \in F(D,\sigma), confident(X \Rightarrow Y, D) \geq \gamma\}$$

Thuật toán fpgrowth

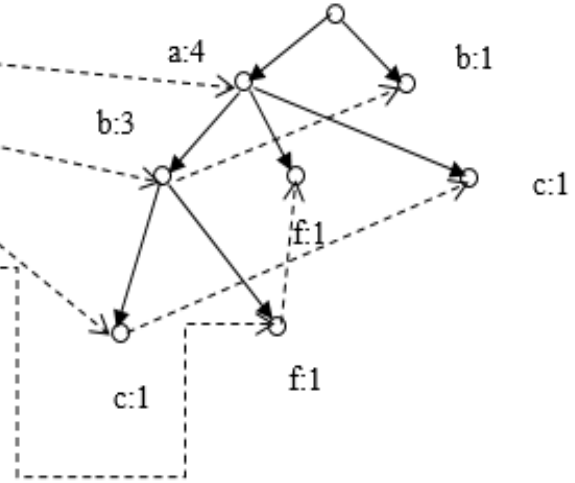
- Được giới thiệu bởi Jiawei Hai Jian và Yiwen Yin năm 2000.
- Thuật toán gồm 3 bước:
  - Duyệt CSDL lần thứ nhất để tính tất cả độ hỗ trợ của tất cả 1-itemsets. Loại bỏ những tập mục có độ hỗ trợ tối thiểu nhỏ hơn σ. Các mục còn lại sắp xếp theo thứ tự giảm dần của độ hỗ trợ.
  - Duyệt CSDL lần thứ 2 , với mỗi tác vụ t, loại bỏ các mục không đủ độ hỗ trợ các mục còn lại sắp xếp theo thứ tự giảm dần và được đưa vào cây FP-tree.
  - Tìm các tập mục phổ biến trên cây FP-tree đã xây dựng mà không cần duyệt lại CSDL nữa

Thuật toán fpgrowth

Cách dựng cây fp-tree:

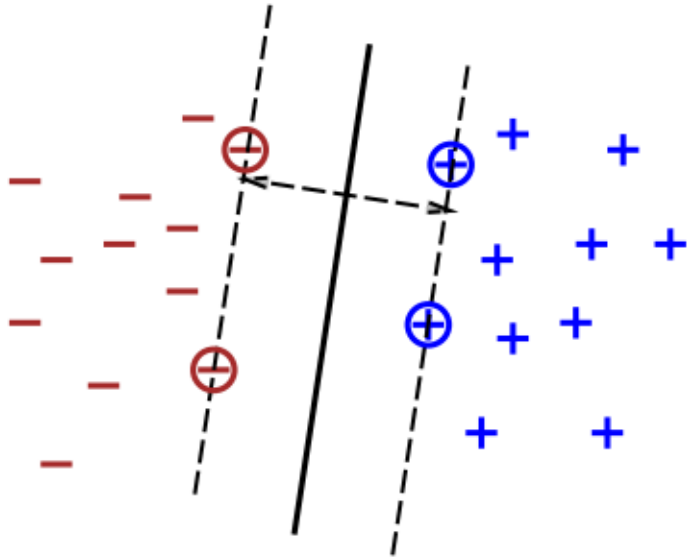
TID	Tập mục trong giao dịch
1	{a, b, c, d, e}
2	{b, c}
3	{a, b, f}
4	{a, b, g}
5	{a, f, h}

Mục tiêu	Tần suất	Node link
A	4	
B	4	
C	2	
F	2	
D	1	
E	1	
G	1	
H	1	



Thuật toán máy vector hỗ trợ

- Là thuật toán được Vapnik và Chervonekis giới thiệu năm 1995
- Tìm một siêu phẳng $y(x) = W \cdot \Phi(\mathrm{x}) + \mathrm{b}$ phân chia dữ liệu thành 2 phần
- Phân lớp dữ liệu mới bằng cách xác định dấu của: $y(x) = W \cdot \Phi(\mathrm{x}) + \mathrm{b}$
  - Thuộc lớp dương nếu y(x) > 0
  - Thuộc lớp âm nếu y(x) < 0



Mô hình dự đoán phospho hoá triển khai

- Bước 1: Chuyển đổi chuỗi protein xung quanh vị trí S, Y, T một cửa sổ n vị trí thành các giao dịch
  - Các vị trí axit amin xung quanh vị trí S, Y, T một cửa sổ n vị trí cần được đánh các chỉ số (từ -n,…, n) để không làm mất mát thông tin về vị trí của các axit amin.
- Bước 2: Áp dụng thuật toán FP-growth để phát hiện các tập phổ biến và các luật kết hợp
- Bước 3: Biểu diễn các chuỗi protein xung quanh này thành các vector
  - Mỗi một trường tương ứng với sự xuất hiện của một luật kết hợp
- Bước 4: Áp dụng thuật toán SVM để dự đoán phospho hoá

Thực nghiệm

- Dự đoán phospho hoá sử dụng cửa sổ 5 axit amin xung quanh
- Thực nghiệm này sẽ nghiên cứu 3 kinase là PKA_group, PKC_group, CK2_group

R-5	M-4	R-3	R-2	N-1	S0	F1	T2	P3	L4	S5
K-5	L-4	R-3	G-2	R-1	S0	F1	M2	N3	N4	W5
P-5	F-4	R-3	R-2	H-1	S0	W1	G2	P3	G4	K5
S-5	P-4	K-3	R-2	N-1	S0	I1	S2	R3	T4	H5
F-5	H-4	M-3	R-2	S-1	S0	M1	S2	G3	L4	H5
T-5	L-4	N-3	R-2	M-1	S0	F1	A2	S3	N4	L5
L-5	K-4	L-3	R-2	R-1	S0	S1	S2	V3	G4	Y5
K-5	L-4	R-3	R-2	S-1	S0	S1	V2	G3	Y4	I5
L-5	R-4	R-3	S-2	S-1	S0	V1	G2	Y3	I4	S5
L-5	T-4	R-3	R-2	A-1	S0	F1	S2	A3	Q4	S5
Q-5	K-4	K-3	R-2	V-1	S0	M1	I2	L3	Q4	S5
K-5	S-4	K-3	K-2	Y-1	S0	D1	V2	E3	V4	P5

Thực nghiệm

- Với kinase ta được các luật kết hợp
  - PKA_group:
    - S0 <- R-3 (60.9091, 89.5522)
    - S0 <- R-2 (57.8788, 89.5288)
    - S0 <- R-2 R-3 (33.0303, 88.9908)
  - PKC_gourp
    - S0 <- R-3 (26.8182, 84.7458)
    - S0 <- K2 (26.3636, 86.2069)
    - S0 <- R-2 (25, 87.2727)
  - CK2_group
    - S0 <- E3 (53, 85.8491)
    - S0 <- D1 (28, 89.2857)
    - S0 <- D3 (26, 82.6923)
    - S0 <- E5 (25.5, 90.1961)
    - S0 <- E2 (23.5, 91.4894)
    - S0 <- E4 (22.5, 91.1111)
    - S0 <- D5 (21, 92.8571)

Thực nghiệm

- Để lượng hoá độ tốt của các luật kết hợp được sinh ra trong việc dự đoán vị trí trên protein bị phospho hoá, chúng ta cần tính các độ đo sau:
  - Độ hồi tưởng (ρ)

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

  - Độ chính xác (π)

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

  - Độ f1 là :

$$f_1 = \frac{2\pi\rho}{\pi + \rho}$$

- Trong đó :
  - TP (True Positive): là ví dụ dương mà luật xác định đúng là dương
  - FP (False Positive): là ví dụ dương mà luật xác định là âm
  - TN (True Negative): là ví dụ âm mà luật xác định đúng là âm
  - FN (False Negative): là ví dụ âm mà luật xác định là dương

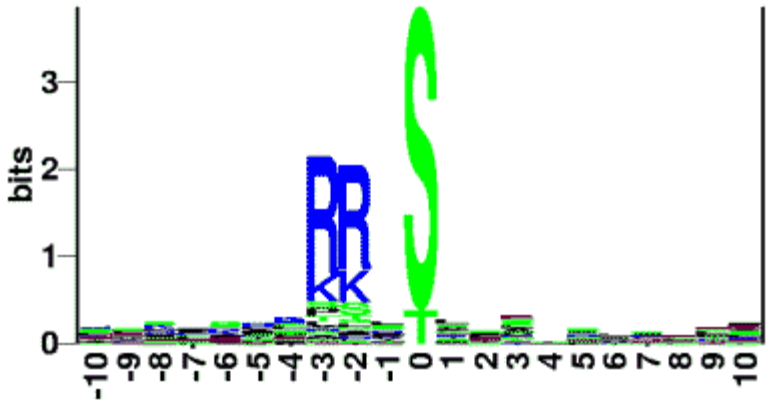
Thực nghiệm

- Dự đoán bằng SVM kết hợp với luật kết hợp
- Tìm các luật tương ứng với mỗi kinase
- Với mỗi chuỗi axit amin xung quanh vị trí bị phospho hoá bởi một kinase sẽ được biểu diễn thành một vector có số chiều bằng số luật kết hợp được sinh ra, trong đó giá trị của mỗi chiều (0 hoặc 1)
- Độ đo trung bình f1 của 2 cách tính:

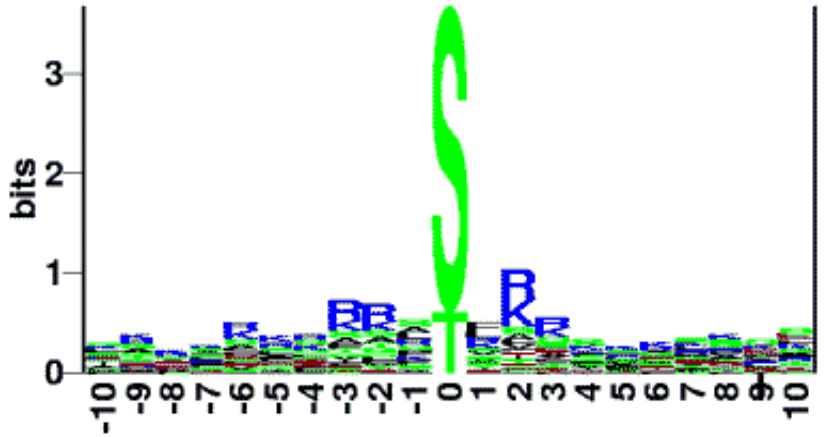
	Fpgrowth	Svmlight
PKA_group	72.53%	81.47%
PKC_group	65.7%	67.01%
CK2_group	74.35%	75.23%

Thực nghiệm

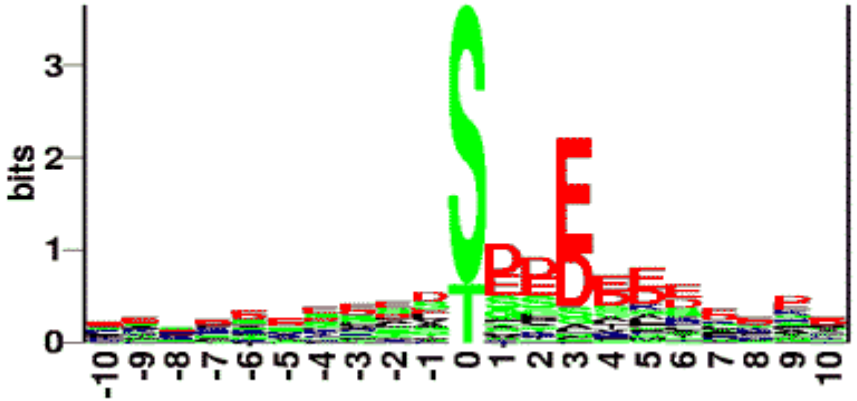
Sequence logos của PKA_group:



Sequence logos của PKC_group:



Sequence logos của CK2_group:



Thực nghiệm

- Sau khi thực hiện với 5 axit amin xung thì tôi bắt đầu thực hiện với cửa sổ là 10 axit amin xung quanh
- Luật nhận được cũng chỉ phụ thuộc vào 5 vị trí xung quanh axit amin

Q-10	N-9	D-8	L-7	K-6	L-5	R-4	R-3	S-2	S-1	S0	V1	G2	Y3	I4	S5	K6	A7	Q8	E9	Y10
G-10	S-9	D-8	G-7	S-6	P-5	G-4	K-3	S-2	P-1	S0	K1	K2	K3	K4	K5	F6	R7	T8	P9	S10
S-10	K-9	K-8	K-7	K-6	K-5	F-4	R-3	T-2	P-1	S0	F1	L2	K3	K4	S5	K6	K7	K8	S9	D10
Y-10	T-9	Y-8	R-7	P-6	W-5	T-4	R-3	G-2	G-1	S0	L1	E2	R3	S4	Q5	S6	R7	K8	D9	S10
T-10	R-9	G-8	G-7	S-6	L-5	E-4	R-3	S-2	Q-1	S0	R1	K2	D3	S4	L5	D6	D7	S8	G9	S10
I-10	A-9	K-8	R-7	R-6	T-5	R-4	V-3	P-2	P-1	S0	R1	R2	G3	P4	D5	A6	V7	A8	A9	P10
L-10	V-9	M-8	L-7	R-6	K-5	R-4	Q-3	Y-2	G-1	T0	I1	S2	H3	G4	I5	V6	E7	V8	D9	P10
Q-10	T-9	E-8	R-7	K-6	S-5	G-4	K-3	R-2	Q-1	T0	E1	R2	E3	K4	K5	K6	K7	I8	L9	A10
L-10	Q-9	L-8	V-7	L-6	C-5	V-4	L-3	A-2	T-1	T0	D1	R2	R3	R4	R5	D6	L7	G8	G9	S10
F-10	I-9	L-8	A-7	P-6	R-5	S-4	S-3	D-2	L-1	T0	D1	R2	V3	K4	V5	W6	T7	S8	G9	Q10
A-10	K-9	D-8	F-7	Q-6	D-5	I-4	Q-3	Q-2	L-1	S0	S1	E2	E3	N4	D5	H6	P7	F8	H9	Q10
S-10	K-9	K-8	K-7	K-6	K-5	F-4	R-3	T-2	P-1	S0	F1	L2	K3	K4	S5	K6	K7	K8	E9	K10

Kết luận và hướng phát triển

- Đã áp dụng được luật kết hợp và tập phổ biến để dự đoán các biến đổi của protein
- Tiếp tục tìm hiểu và áp dụng các kỹ thuật khai phá dữ liệu với các biến đổi khác của protein

ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

Thank You !