VIỆN CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG

TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI

Phụ Lục Bài Tập Lớn Môn Học Mạng Neural và ứng dụng

ĐỀ TÀI 7:

Predicting Protein-DNA Binding Residues

NHÓM: 27

Giảng viên hướng dẫn: TS. Nguyễn Hồng Quang

Sinh viên thực hiện: Ngô Huy Hoàng 20155637
Trần Hải Đăng 20155357

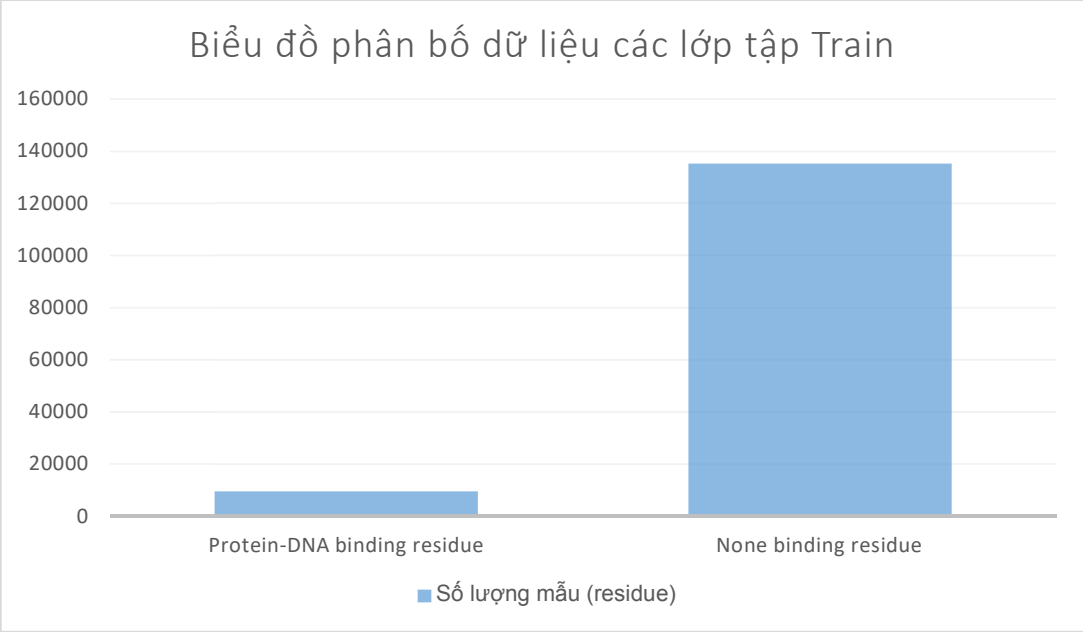
Hà Nội 01/2019

Nội Dung

Sau khi nhận được sự nhận xét của thầy Nguyễn Hồng Quang, nhóm em đã có những thay đổi về tập data cho việc huấn luyện và đánh giá về mô hình tốt nhất dựa trên kết quả thu được từ tập validation.

- **Training Dataset:**

  Tập dữ liệu cho training là tập PDNA-543 gồm 543 chuỗi protein, trong đó có 9549 dư lượng liên kết DNA (dương tính) và 134995 dư lượng không liên kết (âm tính) được thể hiện trong Hình 1 và Bảng 1:



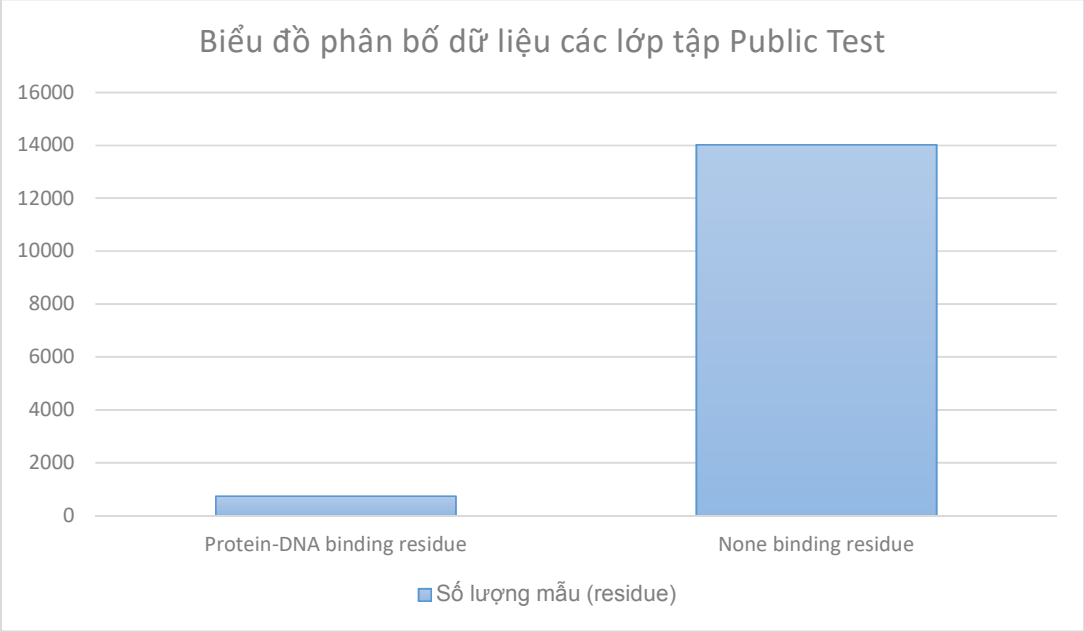
Hình 1. Biểu đồ phân bố dữ liệu giữa Protein-DNA binding residue và None binding residue trong tập train PDNA-543

Label	#Samples
Protein-DNA binding residue	9,549
None binding residue	134,995
Total	**144,544**

Bảng 1. : Số lượng tệp mẫu được cung cấp để dùng làm tập Train/Validation cho mô hình

- **Testing dataset**

  Tập dữ liệu cho testing gồm 41 chuỗi protein có 743 mẫu dương tính và 14021 mẫu âm tính được biểu diễn trong Hình 2 và Bảng 2:



Hình 2. Biểu đồ phân bố dữ liệu giữa Protein-DNA binding residue và None binding residue trong tập dữ liệu test

Label	#Samples
Protein-DNA binding residue	734
None binding residue	14021
Total	**14755**

Bảng 2: Số lượng tệp mẫu được cung cấp để dùng làm tập dữ liệu Test cho mô hình

- **Kết quả thu được:**

  Kết quả thu được về accuracy và loss cho tập validation khi huấn luyện với tập training set được thay đổi như trên được thể hiện trong Bảng 3 và Bảng 4:

	1	2	3	4	5	6	7	8	9	10	MAX
Fold 1	92.68	92.24	**92.96**	92.63	92.77	**92.96**	92.11	92.88	92.23	92.98	**92.96**
Fold 2	92.61	92.81	93.02	92.70	93.07	92.79	92.33	93.02	92.56	**93.21**	**93.21**
Fold 3	92.24	92.66	92.95	**92.96**	93.18	93.12	91.93	92.42	92.11	91.89	**92.96**
Fold 4	93.21	**93.30**	92.98	93.16	93.18	93.00	92.32	92.32	93.11	93.07	**93.30**
Fold 5	93.13	**93.43**	92.96	93.31	93.25	93.30	92.97	93.21	92.06	92.21	**93.43**
MAX											**93.43**

Bảng 3: Kết quả accuracy(%) trên tập Validation khi huấn luyện mô hình thu được khi chạy 10 random seeds khác nhau với 5 Folds

	1	2	3	4	5	6	7	8	9	10	MIN
Fold 1	0.3598	0.3779	**0.3259**	0.3603	0.3488	0.3357	0.3454	0.3299	0.3431	0.3853	**0.3259**
Fold 2	0.3668	0.3471	0.3431	0.3567	**0.3196**	0.3537	0.3312	0.3712	0.3489	0.3712	**0.3196**
Fold 3	0.3859	0.3576	0.3730	**0.3503**	0.3587	0.3589	0.3611	0.3587	0.3689	0.3445	**0.3503**
Fold 4	0.3489	**0.3391**	0.3645	0.3392	0.3468	0.3668	0.3366	0.3555	0.3645	0.3719	**0.3391**
Fold 5	0.3540	0.3370	0.3497	**0.3153**	0.3528	0.3347	0.3212	0.3412	0.3630	0.3443	**0.3153**
MIN											**0.3153**

Bảng 4: Kết quả Loss trên tập Validation khi huấn luyện mô hình thu được khi chạy 10 random seeds khác nhau với 5 Folds

Từ 2 bảng trên, ta thu được 2 mô hình tốt nhất đó là mô hình có accuracy cao nhất trên tập validation và mô hình có loss thấp nhất trên tập validation.

Sử dụng hai mô hình đó cho tập dữ liệu PDNA-Test ta thu được kết quả như trong Bảng 5:

Model	**Accuracy(%)**
Model 1 (Fold 5 seed 2)	94.80
Model 2 (Fold 5 seed 4)	94.62

Bảng 5: Kết quả accuracy(%) trên tập Test từ 2 mô hình tốt nhất được lựa chọn từ kết quả đã huấn luyện ở tập Valiation