BiRDS - Binding Residue Detection from Protein Sequences using Deep ResNets. Supporting Information.

Vineeth Chelur and U. Deva Priyakumar*

Center for Computational Natural Sciences & Bioinformatics

International Institute of Information Technology

Hyderabad - 500032, India

E-mail: deva@iiit.ac.in

Contents

List of Tables

Evaluation Metrics

Confusion Matrix

A confusion matrix is a table that allows for the visualisation of the performance of a supervised learning algorithm. The following terminologies can be defined in the binary classification of a residue as a binding residue (BR) or non-binding residue (NBR).

- True Positive (TP): Number of BRs predicted correctly as BRs.
- True Negative (TN): Number of NBRs predicted correctly as NBRs.
- False Positive (FP): Number of NBRs predicted incorrectly as BRs.
- False Negative (FN): Number of BRs predicted incorrectly as NBRs.

The following metrics can be derived from the confusion matrix

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

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

Recall: $TPR = \frac{TP}{TP+FN}$

F1 score: $F_1 = \frac{2TP}{2TP+FP+FN}$

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

MCC

The Matthew's Correlation varies from $[-1, +1]$, with $+1$ representing a perfect prediction, 0 representing no better than a random prediction and -1 representing total disagreement between the prediction and the observation.

DCC

DCC is the distance between the centre of the predicted binding pocket and the centre of the actual binding pocket. It is commonly used for evaluating 3D-structure based models.

The success rate of DCC is defined as the fraction of predictions below a given threshold. Predicted pockets with DCC below 4Å are considered to be correctly located.

Table 1: Train Set - Obseleted PDB IDs along with replacement PDB IDs

Obseleted PDB IDs	**Replacement PDB IDs**
1HWZ	6DHD
1QY5	6D28
1U0Y	6D1X
2CMJ	5YZH
2CMV	5YZI
2PDT	6CNY
3G07	5UNA
3KWN	5QC4
3LNS	3LNS
3LV1	3LV1
3MPE	5QBY
3MVQ	6DHL
3MW9	6DHM
3N3N	5SXQ
3Q9K	6L9E
3QL6	6LAQ
3TUW	5ZGS
4DGO	6QS5
4EGB	6BI4
4GDC	5VWT
4GDD	5VWU
4KA6	5SYI
4KG1	5H5O
4KNZ	6NNR
4N3L	6EO8
4N7A	6LF7
4NT3	6LCO
4NZE	6EO9
4OA9	5LPV
4OAC	5LPB
4OTW	6OP9
4P7P	5I0K
4PT0	5GTK
4PT3	5GTL
4UTD	6CPF
4WBN	6I5C
4Y9Q	5MF5
5AAJ	5OMO
5CTO	6J63
5LI5	5LX2

Table 2: Test Set Removed - Obseleted PDBs

5MWR	5X8O	6C2L	6ELH	6I2L
5MYH	5YV4	6C2P	6EMC	6IQ7
5MZ9	5Z3T	6C38	6EME	6IQ8
5N7R	5Z4I	6CB4	6ESO	6IRM
5NWO	5ZBV	6CO0	6ETX	6IRN
5OLI	5ZD5	6CSY	6FJR	6JLW
5UF9	5ZD7	6CU4	6FMM	6JLX
5UUR	6A5V	6D4J	6FO4	6JU2
5VB1	6ABB	6D5I	6FOT	6JU3
5VB4	6ABD	6DM2	6FOU	6MPU
5VFL	6ABE	6DTJ	6FOV	6MSZ
5VG5	6ABF	6DU1	6FPE	6N3T
5WVQ	6ABG	6DUA	6FQC	6NLC
5WVS	6BW7	6DX6	6G9Y	6NLD
5WVT	6C0C	6E5O	6GJX	6NPG
5WW1	6C2E	6E79	6GKP	6NXH
5WW2	6C2J	6E7A	6H7K	6UHD
5WWA	6C2K	6EDO	6HBO	

Table 3: Test Set Removed - Reindexing Errors

6DC9
6DCA
6EOD
6FE0
6FE1
6IEY
6MQC
6MQE
6N16
6NCP
6RUL
6SNC
6SND
6SNE