June 30, 2022

Dear Peer Review Advisors  
BMC Bioinformatics

Please find submitted our revised manuscript, entitled “Integrated Structure-based Protein Interface Prediction”, by Walder et al. We have compared the sequence identities of our test set and the training sets of the individual classifiers. We have added a paragraph in the manuscript (under the Training and Test Sets Section) to reflect this. We thank the reviewers for their helpful comments, and we respond in detail to the reviewer’s comment below:

Sincerely,

Raji Viswanathan

Reviewer 2

The authors have replied to almost all the comments well.

I have only one comment for the manuscript about the training data and test data. I need to explain more clearly for the first question in my previous comments. Could the authors ensure that there is no overlap between the test dataset of this meta-method and the training dataset of the input methods like PredUS and ISPRED4? If there is some or many overlap between them, some overfitting risks maybe arise.

Sequence identity with ISPRED4 TRAINING SET:

ISPRED4 uses 314 proteins in their training set. We calculated the pairwise sequence identity of each protein in our test set with the 314 in the training set and the following figure shows the distribution of the sequence identities of the 33 test set proteins. As seen in the histogram below, 9/33 proteins in the ISPIP test set have a high similarity to the ISPRED4 training set. The following table lists the sequence identity for each of our test set PDB as well as the PDB ID of the corresponding protein in the ISPRED4 training set. The nine high similarity proteins are colored in RED.

Chart, histogram

Description automatically generated

|  |  |  |
| --- | --- | --- |
| ISPIP TEST PDB | ISPRED4-Training PDB | Seq.Identuty |
| 1PDK\_A | 1E6O\_L | 0.1788 |
| 1CNZ\_A | 2QA9\_E | 0.163 |
| 1KXP\_D | 1KW2\_B | 0.9719 |
| 1COZ\_A | 1JVM\_A | 0.1794 |
| 1CLI\_A | 3QI0\_A | 0.1708 |
| 3BX7\_A | 3BX8\_A | 0.9765 |
| 1DE4\_E | 1EGL\_A | 0.213 |
| 2PCB\_A | 1CCP\_A | 0.9983 |
| 1B8A\_A | 1HCL\_A | 0.1585 |
| 1EG9\_A | 1CCP\_A | 0.1649 |
| 1DOW\_A | 1SYQ\_A | 0.2004 |
| 1STF\_E | 3BPF\_A | 0.349 |
| 2B4J\_C | 1CEW\_I | 0.2 |
| 1E4K\_C | 1FNL\_A | 0.9913 |
| 1DOR\_A | 2C8B\_X | 0.1836 |
| 1SOX\_A | 2YVF\_A | 0.1759 |
| 1PVH\_B | 1EMR\_A | 0.9512 |
| 1YVE\_I | 1A12\_A | 0.1597 |
| 2NAC\_A | 1GJR\_A | 0.1614 |
| 1QFH\_A | 1SUP\_A | 0.1889 |
| 1QOR\_A | 3VL8\_A | 0.1828 |
| 1HSS\_A | 2UUX\_A | 0.2169 |
| 3VLB\_B | 3VL8\_A | 0.9932 |
| 1JTD\_A | 1ZG4\_A | 0.9943 |
| 1TMQ\_B | 1QFK\_L | 0.216 |
| 2SIC\_I | 3SSI\_A | 0.9953 |
| 2UTG\_A | 2J5Y\_A | 0.2443 |
| 2CFH\_C | 2BJN\_A | 0.9416 |
| 1YPI\_A | 3F74\_A | 0.1827 |
| 1I8L\_A | 3RVT\_D | 0.2043 |
| 1CP2\_A | 1HUR\_A | 0.1782 |
| 1CMB\_A | 1UNK\_D | 0.2094 |
| 1TCO\_A | 1IAM\_A | 0.1676 |

Sequence Identity of ISPIP test set with PredUs training set.

PredUs 2.0 uses 2766 hetero complexes to train the predictions of interface propensity. Of these 2766 complexes, since some proteins have more than one complex with different ligands, we found a unique set of 2444 PDBs from this training set. We report below the sequence identity of each test protein (in our model) with the 2444 PDBs in the training set of PredUs2.0 below. As seen in the histogram below, 6/33 proteins in the ISPIP test set have a high similarity to the PredUs2.0 training set. The following table lists the sequence identity for each of our test set PDB as well as the PDB ID of the corresponding protein in the PredUs2.0 training set. The high similarity proteins are colored in RED.

Two of these six are also of high sequence identity to the ISPRED4 training set. So, we removed these 13 PDBs from our test set and recalculated the F-score and MCC scores for the remaining data. The results are given below:

Chart, histogram

Description automatically generated

|  |  |  |  |
| --- | --- | --- | --- |
| *ISPIP TEST PDB* | PredUs-Training PDB | | Seq.Identuty |
| 1PDK\_A | 2W07\_A | 0.8903 | |
| 1CNZ\_A | 1JFI\_B | 0.2569 | |
| 1KXP\_D | 1JFI\_B | 0.2651 | |
| 1COZ\_A | 2B9S\_A | 0.2319 | |
| 1CLI\_A | 1JFI\_B | 0.2521 | |
| 3BX7\_A | 4GH7\_A | 0.6541 | |
| 1DE4\_E | 4H25\_E | 0.2067 | |
| 2PCB\_A | 1JFI\_B | 0.2563 | |
| 1B8A\_A | 1JFI\_B | 0.2444 | |
| 1EG9\_A | 2BMO\_A | 0.7804 | |
| 1DOW\_A | 1JFI\_B | 0.2216 | |
| 1STF\_E | 3IMA\_A | 0.9123 | |
| 2B4J\_C | 2B9S\_A | 0.2086 | |
| 1E4K\_C | 2B9S\_A | 0.2292 | |
| 1DOR\_A | 1JFI\_B | 0.2511 | |
| 1SOX\_A | 1JFI\_B | 0.2607 | |
| 1PVH\_B | 2B9S\_A | 0.24 | |
| 1YVE\_I | 1JFI\_B | 0.2591 | |
| 2NAC\_A | 1JFI\_B | 0.2631 | |
| 1QFH\_A | 2FO1\_A | 0.2651 | |
| 1QOR\_A | 1JFI\_B | 0.2388 | |
| 1HSS\_A | 1JFI\_B | 0.2598 | |
| 3VLB\_B | 1JFI\_B | 0.2438 | |
| 1JTD\_A | 3C7V\_A | 0.9255 | |
| 1TMQ\_B | 1NVP\_B | 0.2348 | |
| 2SIC\_I | 1N6J\_A | 0.2593 | |
| 2UTG\_A | 2B9S\_A | 0.1987 | |
| 2CFH\_C | 1JFI\_B | 0.2123 | |
| 1YPI\_A | 1JFI\_B | 0.2256 | |
| 1I8L\_A | 1JFI\_B | 0.2164 | |
| 1CP2\_A | 1JFI\_B | 0.267 | |
| 1CMB\_A | 2B9S\_A | 0.227 | |
| 1TCO\_A | 4F0Z\_A | 0.8543 | |

**Revised F-score and MCC scores:**

|  |  |  |
| --- | --- | --- |
|  | **Test Set with 33 proteins** | **Test Set with 20 proteins** |
| **Average F-score** | **0.516** | **0.542** |
| **Average MCC score** | **0.487** | **0.503** |

**There is a small increase in the F-score and MCC score of the revised test set.**