Hoai-Huong Emilie DOAN
Brenda ENRIQUEZ
Birsu GUZEL
Isabelle WU

MEET-U – Summary

Proposing potential inhibitors for the Sars-CoV-2 Helicase (Nsp13) by Molzip approach

Supervised by :
Juliana Silva Bernardes
Elodie Laine
Vaitea Opuu

2023-2024 Academic Year

Our goal is to test Molzip, a new prediction approach that is unconventional as it does not take into account three-dimensional structures to predict (Probst, 2023). Molzip is a classification method based on information theory and uses probabilities to quantify the average information content of a set of messages, including computer coding. And promising results on the prediction of physicochemical variables, but also on the prediction of ligand-protein affinity were obtained with this algorithm with only the sequence in the input, which is little.

Also, we used this notion of compression that we entered into machine learning methods known as Random Forest, neural network and others in order to see if using the same idea as Molzip would give us interesting results in different algorithms with the same data.

Molzip method

We chose to study the technique "Molzip" as we wanted to test the reliability of the prediction of this tool via different tests and then use it to predict potential inhibitory ligands.

First, we tested the prediction of variables such as molecular weight with a database of over 5,000 ligands. After plotting the results, we see clusters forming, the low values group together and the same for the high values. This test allows us to confirm that Molzip succeeds in differentiating and predicting ligand variables from only their SMILES sequences. Second, we used a database (Li et al., 2023) containing ligands and pockets with an affinity score. We did a cross validation to predict the affinity scores of the ligands, and to check the performance of the model. Each split gave us similar plots showing the consistency of the model. These two steps allowed us to validate the potential of the Molzip method.

Therefore, in order to predict ligands for the ARN and ATP pockets of NSP13, we used the previous dataset (Li et al., 2023) containing the protein-ligand binding affinity scores as a training set. And then, we predicted the affinity of potential ligands (dataset of 5000 potential ligands). We considered that ligands with the highest affinity value are the best predicted ligands.

Method with known Machine Learning Methods

We wanted to test different supervised machine learning techniques on the PDB Bind dataset in order to compare our results with the Molzip method. To achieve this, we used the same idea of "zip" the ligand with the possible pocket. Then, we tested a Neural Network, Random Forest, AdaBoost, KNeighbors, GradientBoosting, and SVM. According to their performance on the PDB Bind dataset, we used some of these methods on the dataset with 5000 possible ligands to select the best ligands found by supervised machine learning algorithms and Molzip.

Results :

Molzip method

To validate the performance of the Molzip technique, we tested the method on different predictions like we said earlier in the approach. For each pocket, we predicted the potential ligand's affinity value and selected those with the highest affinity scores. To do so, we used "LP_PDBBind.csv" as a training set and we only took as entries lines which have a Kd value so that our training base has an affinity score which is based on the same constant which is here the equilibrium dissociation constant (K_D). Therefore, we used around 6000 samples and took "value" as

a prediction value. Molzip has no problem in predicting affinity scores as it can manage continuous values. The highest values represent the best affinities and we took the ranking of the 20 best potential ligands (Figure S2). These ligands can be found with their EOS numbers on the data file "pilot_library.csv" and also on the internet. It should be noted that certain ligands have been predicted to bind with both pockets, which is not normal. Our results with Molzip are therefore not relevant.

Method with known Machine Learning Methods

When we predicted the "kd/ki" and "values" columns of the PDB Bind file, the supervised machine learning algorithms didn't show a good performance so we decided to make categories according to literature.

After categorizing the values of the "kd/ki" columns, we observed accuracy scores around 0.5. This indicates that the classifiers obtained by the tested methods perform similarly to a random classifier, achieving approximately 50% accuracy. These results suggest that the classifiers have a tendency to give incorrect results around half of the time.

According to our methodology, we got better results when we categorized the "values" column than when using the "kd/ki" feature (categorized) to classify the fixation index (Table 1). However, while testing with the Dataset 5000, we realized that it was hard to predict and classify them as this value is in logarithm scale and so a very small variance in the dataset caused a lot of divergence in the results. On the other hand, "kd/ki" proved a lower accuracy level but since we have information about the performance of this parameter in the fixation, we decided to use this feature to classify the possible fixation ligands on the dataset of 5000 ligands, and we set a very small range for class 0 so we are very confident that we're proposing highly good performing ligands.

Comparison of Molzip and machine learning methods

By comparing the results obtained with Molzip and supervised machine learning algorithms, we observe that Molzip predicts better the "values", meaning affinities of different ligands, when we test the "LP_PDBBind.csv" data. While comparing the results given by the different supervised learning methods, we didn't find a common set of proposed inhibitors, as we expected, there were only three common inhibitors found for ATP Pocket (Highlighted in Tables 2 and 3). However, we don't observe any ligands in common between the ligands found with Molzip and the ligands found with supervised learning algorithms.

Conclusion:

We don't have any results that are relevant at the moment with Molzip. Through our predictions, we can notice that this approach can surprisingly achieve some good results but it is not completely precise and fair. This may depend on the database. It would therefore be interesting to continue with it but by applying several other methods in addition to improve the results and reliability. This method can also make it possible to make an initial selection of potential ligands in order to reduce the number of ligands that could be docked with more traditional methods.

References

1. Probst D. Parameter-Free Molecular Classification and Regression with Gzip. ChemRxiv. 2023; doi:10.26434/chemrxiv-2023-v1s2s.
2. Li J, Guan X, Zhang O, Sun K, Wang Y, Bagni D, Head-Gordon T. Leak Proof PDBBind: A Reorganized Dataset of Protein-Ligand Complexes for More Generalizable Binding Affinity Prediction. Preprint. 2023.