Large-Scale Benchmarking of Protein Descriptors for Protein Ligand Prediction in Target-Based Modelling and Proteochemometrics

Heval Ataş, Ahmet Rifaioglu, Rengul Atalay, Volkan Atalay, Maria Martin, Tunca Doğan

Introduction

The computational prediction of drug/compound-target interactions (DTIs) using machine learning has become an attractive approach in drug discovery due to slow and costly process of conventional approaches. To generate a DTI prediction model, input ligands and/or proteins are converted into quantitative feature vectors (i.e., descriptors) using their molecular properties, and these vectors are usually fed to supervised machine learning algorithms to construct predictive models. Therefore, the selection of descriptor sets is crucial to generate models with high predictive performance. Ligand-based DTI prediction approaches such as QSAR studies represent only the chemical space in modelling (i.e., system input = compound descriptors), where the target proteins are used just as labels in models. However, an important factor in DTIs is the properties of the target proteins. Proteochemometric modelling approach takes both target and compound properties as a pair at the input level, and predicts the binding affinity of the input compound-target pair. While there are many studies regarding the benchmarking of ligand descriptors (due to the abundance of ligand-based DTI prediction methods), protein descriptor performance comparison studies are scarce.

Here, we perform a large-scale benchmark analysis of 42 sequence-based protein descriptors considering various physicochemical amino acids characteristics, sequence composition, pssm profiles and functional characteristics of proteins; using random forests (RF) and support vector machine (SVM) algorithms. To investigate the protein descriptors, we assumed 2 different modelling approaches: *(i)* the target-based approach, in which an individual predictive model was generated for each compound cluster (composed of structurally very similar compounds); and *(ii)* the proteochemometric approach, in which both the compound-target feature pairs are fed to the system. This study will help to identify the protein feature types with better representation capabilities to be used both in DTI prediction and for other types of automated protein annotation.

Methodology

Target-based modelling:

In the target-based DTI prediction approach, each (drug candidate) compound comprise a predictive model and the protein feature vectors are given as input to the model to predict if the query protein could be the target of the corresponding compound. We generated predictive models using SVM and RF classifiers for 42 different types of protein descriptors, 10 of which are shown in Table 1. The task here is the binary classification of input proteins as active or inactive against the corresponding compounds.

In this part, the training bioactivity dataset was obtained from the ChEMBL (v23) database, where compound-target pair bioactivity data points with pChEMBL values (i.e., -log(IC50, EC50, Ki, Kd, Potency, …)) > 5 considered for the positive set (actives) and instances with pChEMBL <= 5 considered for the negatives set (inactives). The size of training datasets for most of the compounds were not sufficiently large to generate reliable models. To overcome this limitation, we clustered compounds based on their molecular similarities and generated nine independent datasets by merging bioactivity information of compound clusters. Hence, each compound cluster comprises a model.

Proteochemometric modelling:

We utilized both compound and target space by feeding the RF regression models with concatenated compound-protein feature vectors as the input. For the representation of compounds, we used ECFP4 fingerprints. To convert proteins into feature vectors, we selected 10 different protein descriptors out of the total 42, each taking a different aspect of proteins into account as described in Table 1. Here, the task is predicting the actual binding affinity values of the input compound-target pairs in terms of pChEMBL values.

For the training and test dataset, we used Davis kinase benchmark set (doi:10.1038/nbt.1990). This dataset includes ~30,000 data points; however, bioactivity values of ~20,000 of them is pChEMBL = 5, which may cause a bias in the predictive models. To prevent this situation, we removed these instances from both train and test sets. For the train set, we also applied three additional filters targets to prevent data memorization. For this, we removed all bioactivities of compounds and targets if the compound or target: (*i)* only have active or inactive bioactivities based on the threshold pChEMBL=6.2, which is the median value of all dataset, (*ii)* have an active-to-inactive ratio > 4 or < 1/4 considering its all bioactivities, and (*iii)* has a bioactivity distribution standard deviation < 0.3. After the filtering, we trained our models on 6,706 data points and tested on 1,542.

Table 1. Properties of protein descriptors.

Descriptor Name	Descriptor Type	Dimension
pfam	Domain profiles	2519
spmap	Subsequence-based feature map	544
combo	k_sep + apaac	480
dde	Dipeptide composition deviation	400
k_sep	Column transformation based position specific scoring matrix (pssm) profiles	400
ctriad	Triad frequency of residues classified on dipoles and volumes of aa side chains	343
geary	Autocorrelation regarding the distribution of physicochemical properties of aas	240
random	Randomly generated numbers [0-1]	200
ctd_d	Chain length based distribution of aas for selected physicochemical properties	195
qso	Sequence order effect based on physicochemical distances between coupled residues	100
apaac	Amino acid composition regarding the sequence order correlated factors computed from hydrophobicity and hydrophilicity indices of aas	80
taap	Summation of corresponding residue values for selected physicochemical properties	10

Results

Target-based modelling:

The performance of generated models was evaluated by 5-fold external validation on a test set based on accuracy, MCC, F1-score, ROC-AUC and AUC-PR metrics. Only ten of 42 different types of protein descriptors that reflect the

actual distribution of the results are represented to make the results more apparent. According to Figure 1, combo descriptor (k_sep + apaac) was the best according to MCC values averaged over the 9 target-based models, and dde was a close second. Also, RF models performed slightly but consistently better than SVM models in general (Figure 1). Therefore, we generated proteochemometric based models only using the RF algorithm.

To summarize the target-based model results, for each descriptor, we calculated the mean performance scores of 9 models for each evaluation metric listed above, and calculated the median of ranks of the descriptors considering the mean performance measures (Figure 2). This way, all metrics are merged into one rank measure. As seen in Figure 2, the combined model performed best in the overall ranking results (lower is better), which is not surprising since it increases the chance of capturing relevant properties of proteins that are included in k-sep and apaac separately. Considering the individual descriptors, k-sep outperformed all others, proving to be a useful homology detector.

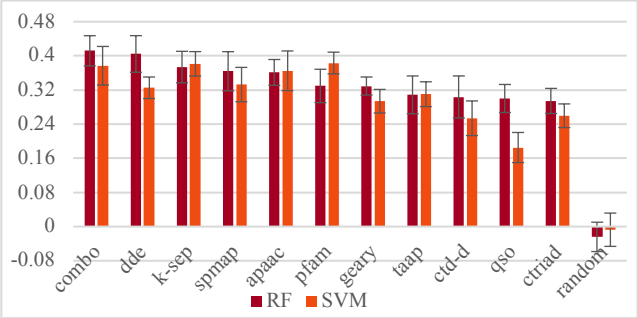

Figure 1. Average MCC test scores for RF and SVM based classification models in the target-based modelling approach.

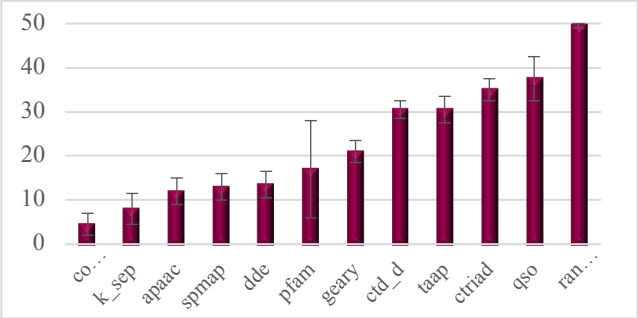

Figure 2. Median ranking of test performance scores (i.e., accuracy, MCC, F1_score, ROC_AUC, PR_AUC) of RF and SVM based classification models in the target-based modelling approach.

Proteochemometric modelling:

For the performance evaluation of models on the test dataset, we used RMSE, Pearson correlation, F1_score and MCC. Here, lower RMSE values mean better performance. We set active/inactive prediction threshold as 7 for the calculation of F1_score and MCC (a generally accepted threshold for kinases).

As shown in Figure 3a and 3b, k-sep performed the best even though overall performance ranks change for different score metrics. Combo (k-sep+apaac) descriptor could not exceed the performance of its constituents. We can also infer that proteochemometric modelling approach provides better performance results compared to the target-based approach, when they are compared regarding MCC.

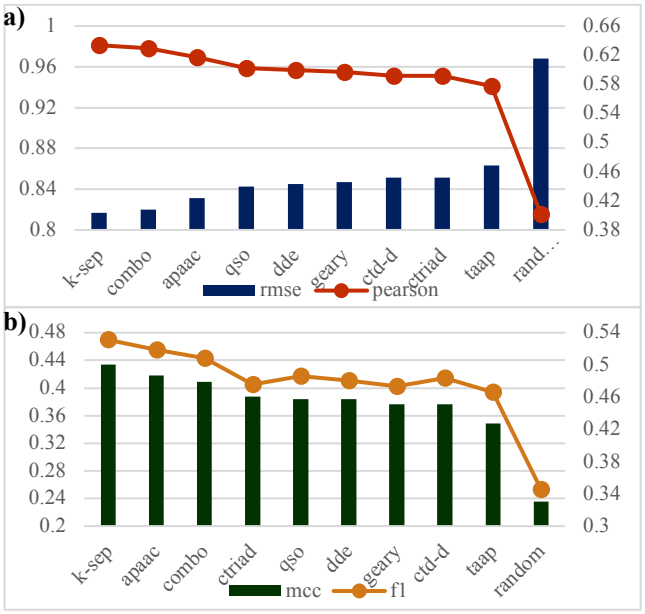

Figure 3. Test performance results of proteochemometric RF regression models based on: **a)** RMSE and Pearson **b)** MCC and F1-score.

Discussion

- Target-based modelling approach suffers from the problem of low number of training instances, thus, not suitable for large-scale DTI prediction, due to the necessity of constructing an individual model for each compound molecule.
- Proteochemometrics is especially suitable for compounds and targets with low number of (or no) training instances, and thus, suitable for the identification of druggability potential of all human proteins. Since the input to the system is compound-target pairs, it is possible to use only one predictive model for all data points.
- Proteochemometrics requires carefully constructed training datasets that contain targets & compounds with balanced and well-distributed (inactive to active) data points. Before the extensive dataset filtering operation described above, our models memorized the data during training and provided low test performance results.
- For the representation of proteins in automated ligand prediction, k-sep could be preferable, which provided consistently best performances in both target-based and proteochemometric modelling analyses. If the computational complexity is considered, apaac could be also a good choice due to its considerably low dimension size.
- With the aim of constructing large-scale gold standard datasets for training and benchmarking proteochemometric DTI prediction models, we are currently extending this dataset filtering approach to bioactivity datasets extracted from the ChEMBL database, categorized into six different target classes: membrane receptors, ion channels, transporters, transcription factors, epigenetic regulators and enzymes (with five subgroups).
- Furthermore, we are extending descriptor benchmarking analyses for all datasets described in the previous item, to observe best protein family specific descriptors for DTI prediction. We also plan to include combinations of the descriptors given above, and possibly feature selection procedures, to maximize the performance.