1 Constructing QSAR from ChEMBL

1.1 Download as PostgreSQL

Downloaded ChEMBL database from `ftp://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/latest/` and started building the database using pgAdmin4.

1.2 Extract relevant Assays from the SQL Database

Relevant Database tables used in this section are: `activities`, `assays`, `compound_records`, `compound_structures`, `target_dictionary`.

1.2.1 Finding Kinase Assays

The original research used only assays from kinase targets as columns for their QSAR construction. Following the pattern `target_dictionary` was searched for the target name to contain the string `kinase`. This returned 899 different kinases. The `tid` of those were used to obtain the activities in the `activities` table and saved in a separate table.

1.2.2 Filtering Kinase Assays

The main condition to ensure a quality threshold was a minimum of 200 compound entries for each assay to be considered professional and accurate enough to be used. Activities were grouped and entries that didn't meet the threshold were removed. In order to avoid poorly described compounds and dilute the prediction accuracy, compounds were removed that had only one reported activity throughout the whole matrix. This reduced the compound library from around 70,000 to 21,000 more accurately described compounds.

1.2.3 Dealing with Duplicate Entries and other complications

Activity entries for the target/compound pairings were reported in multiple formats and units. As a consequence a couple of assumptions were made:

1. Items that reported standard units '%' were deleted as the reported percentage was the inhibition percentage of the protein when treated with a compound at fixed concentration (10 uM). As this relationship is not linear i.e. the IC50 can't be extrapolated from a percentage concentration, the reported values can't be used for QSAR.

2. Other standard units such as K_i ($\sim$90,000), Potency ($\sim$16,000), EC50 ($\sim$5,000) and AC50 ($\sim$1400) were assumed to be equivalent of IC50s ($\sim$18,000), which is not necessarily true [1]. The inhibition constant K_i denotes the equilibrium constant of the dissociation of the inhibitor-bound enzyme complex, while IC50 quantifies the concentration of inhibitor necessary to halve the reaction rate of an enzyme-catalysed reaction observed under specified assay conditions. The relationship between K_i and IC50

is complex and depends both on the substrate concentration and mode of inhibition. In general IC50 is considered to have a higher value than K_i and only equal for competitive inhibition when the substrate concentration is very small or equal for uncompetitive inhibition when the substrate concentration is very high. However, as no information is available about the assays the assumption was made that IC50s equal K_i as the use of their logarithmic values reduces the propagated error by a factor 10.

3. Around half the items (∼65,000) had reported their `standard_relation` value as either '<' or '>'. In those cases the reported concentration was off-set by a factor of 10 into the respective direction and assumed to have reached their IC50 condition.

4. Duplicates were either deleted if they were lacking necessary information or were an item in assumption 3 (given a remaining duplicate). The remaining duplicates were averaged and kept.

5. Items that had no published data, units or were otherwise missing relevant information were deleted.

1.2.4 Filtering processed data

To ensure that after the data curating the data was still adequate assays below 200 entries were removed again. Further, now that all data points were available an additional condition was imposed removing all assays with pIC50 data points below a standard deviation of 0.5. This made sure that the data points actually contained measured information and are not just off-set assumption values as made in assumption 3 in section 1.2.3. Further, items without corresponding SMILES/Structure were removed.

1.2.5 Saving Data

The resulting 132,180 data points span 196 kinase assays and 21,388 compounds. The data is saved in three files:

1. `ASSAYIDs`: Contains the name of the target protein, compound hit count, ChEMBL IDs, assay type, description and a created assay index (starting from 1).

2. `CMPDIDs`: Contains the ChEMBL ID, three types of structures formats (InChI - International Chemical Identifier, Key and Canonical Smiles) and a created compound index (starting from 1).

3. `ACTIVITIES`: Contains all activities as pIC50s referenced to the assays and compounds both through their database IDs as well as their newly assigned indices.

1.3 Clustering Compounds based on Structural Similarity

In order to accurately predict the accuracy of the QSAR models the test and train sets can't be random but rather be split based on the structural similarity of the compounds in the respective assays. Similar compounds should be assigned to the training set while 'outlier' compounds are assigned to the test set, reflecting the realistic application of QSAR when searching for potential new compound hits.

1.3.1 Defining Structural Similarity

Quantifying structural similarity required converting compounds into a form that allowed them to be compared to each other. This was done using Morgan/Circular Fingerprints [2] with a radius of 2. The algorithm converts molecular structures into a 1024 bit vector based on the direct neighbourhood of the atoms using a hash function. It only retains information of the two dimensional environment. These vectors were then compared against each other using the Tanimoto index as a measure of similarity and constructing a full compound $C \times C$ similarity matrix (the code computes the lower triangular values, transposes the matrix and combines them - the diagonal values equal to 1). The Tanimoto index is considered a powerful tool to quantify similarity [3], ranges from 0 (least similar) to 1 (identical) and is defined as follows:

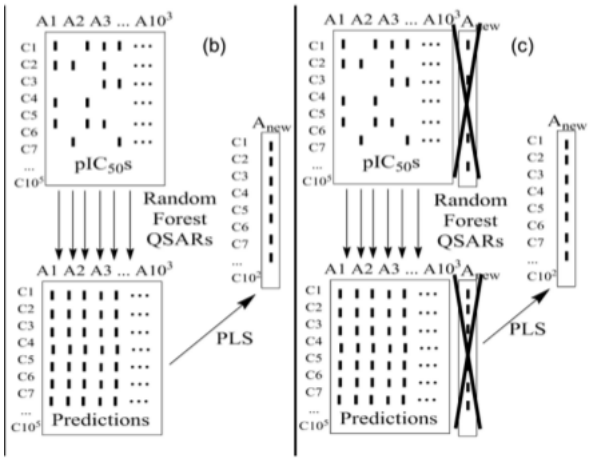
$$S_{A,B} = \frac{\left[\sum_{j=1}^{n} x_{jA} x_{jB}\right]}{\left[\sum_{j=1}^{n} \left(x_{jA}\right)^2 + \sum_{j=1}^{n} \left(x_{jB}\right)^2 - \sum_{j=1}^{n} x_{jA} x_{jB}\right]} \tag{1}$$

1.3.2 Clustering using DBSCAN

For each assay a smaller similarity matrix was constructed containing all compounds with reported activities for that assay by extracting the information from the full matrix. The similarity matrices were then converted into dissimilarity/distance matrices to allow for further processing. The clustering algorithm of choice was the density-based spatial clustering algorithm DBSCAN, because it allows a choice of cluster size based on the maximum distance between two points to be considered in the same neighbourhood (ϵ) and is reasonable fast to iteratively find the desired result for the 1000×1000 sized matrices. The implementation of the algorithm was done by tuning the ϵ parameter iteratively to approximate an average cluster size of 10. Higher ϵ iterations were very fast as all data points were assigned to the same cluster, so that this process was fairly time efficient. For less than 10 cases an additional condition was needed that the largest cluster had to contain less than 75% of data points to allow for the reasonable train/test split. Usually the largest cluster contained around 25-40% of the data points. Clusters were subsequently ordered by size and the largest clusters were assigned to the train data set, while the remaining clusters (often of size 1) were assigned to the test data set.

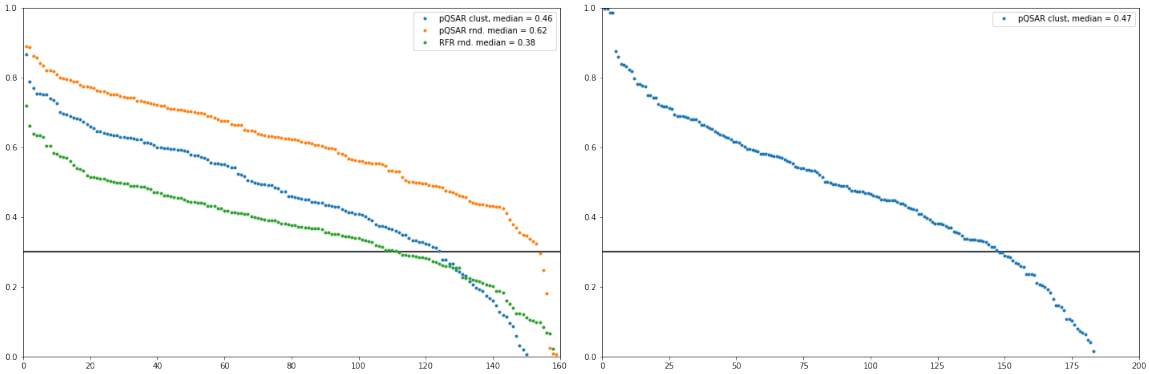
1.4 Constructing QSAR

QSAR was constructed following protocol [4] and the R^2 was computed and slightly improved with increasing assays. While the hit count increased only by around 20%, about double the amount of compounds was being described.



	Hits	Assays	Compounds
Old	114017	159	13190
New	132180	196	21388

1.4.1 Computing R^2



R^2 median only increased slightly, however the number of assays succeeding the 0.3 threshold increased by 25% from 120 to 150 when comparing the realistic splits.

References

[1] B. T. Burlingham and T. S. Widlanski, "An intuitive look at the relationship of Ki and IC50: a more general use for the Dixon plot," *Journal of chemical education*, vol. 80, no. 2, p. 214, 2003.

[2] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *Journal of chemical information and modeling*, vol. 50, no. 5, pp. 742–754, 2010.

[3] D. Bajusz, A. Rácz, and K. Héberger, "Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?," *Journal of cheminformatics*, vol. 7, no. 1, p. 20, 2015.

[4] E. J. Martin, V. R. Polyakov, L. Tian, and R. C. Perez, "Profile-QSAR 2.0: Kinase virtual screening accuracy comparable to four-concentration IC50s for realistically novel compounds," *Journal of chemical information and modeling*, vol. 57, no. 8, pp. 2077–2088, 2017.