Virtual screening of bioassay data

Context

Drug development is is time-consuming and expensive (15 years and 800$ million for a single drug to market). In High-Throughput Screening (HTS), batches of compounds are tested against a biological target to test the compound's ability to bind to the target. Targets might be antibodies for example. If the compound binds to the target then it is active for that target and known as a hit.

Virtual screening is the computational or *in silico* screening of biological compounds and complements the HTS process. It is used to aid the selection of compounds for screening in HTS bioassays or for inclusion in a compound-screening library.

Drug discovery is the first stage of the drug-development process and involves finding compounds to test and screen against biological targets. This first stage is known as primary-screening and usually involves the screening of *thousands of compounds.*

This dataset is a collection of 21 bioassays (screens) that measure the activity of various compounds against different biological targets.

Content

Each bioassay is split into test and train files.

Here are some descriptions of some of the assays compounds. The source, unfortunately, does not have descriptions for every assay. That's the nature of the beast for finding this kind data and was also pointed out in the original study.

Primary screens

- AID362 details the results of a primary screening bioassay for Formylpeptide Receptor Ligand Binding University from the New Mexico Center for Molecular Discovery. It is a relatively small dataset with 4279 compounds and with a ratio of 1 active to 70 inactive compounds (1.4% minority class). The compounds were selected on the basis of preliminary virtual screening of approximately 480,000 drug-like small molecules from Chemical Diversity Laboratories.
- AID604 is a primary screening bioassay for Rho kinase 2 inhibitors from the Scripps Research Institute Molecular Screening Center. The bioassay contains activity information of 59,788 compounds with a ratio of 1 active compound to 281 inactive compounds (1.4%). 57,546 of the compounds have known drug-like properties.
- AID456 is a primary screen assay from the Burnham Center for Chemical Genomics for inhibition of TNFa induced VCAM-1 cell surface expression and consists of 9,982 compounds with a ratio of 1 active compound to 368 inactive compounds (0.27% minority). The compounds have been selected for their known drug-like properties and 9,431 meet the Rule of 5 [19].
- AID688 is the result of a primary screen for Yeast eIF2B from the Penn Center for Molecular Discovery and contains activity information of 27,198 compounds with a ratio of 1 active compound to 108 inactive compounds (0.91% minority). The screen is a reporter-gene assay and 25,656 of the compounds have known drug-like properties.

- AID373 is a primary screen from the Scripps Research Institute Molecular Screening Center for endothelial differentiation, sphingolipid G-protein-coupled receptor, 3. 59,788 compounds were screened with a ratio of 1 active compound to 963 inactive compounds (0.1%). 57,546 of the compounds screened had known drug-like properties.
- AID746 is a primary screen from the Scripps Research Institute Molecular Screening Center for Mitogen-activated protein kinase. 59,788 compounds were screened with a ratio of 1 active compound to 162 inactive compounds (0.61%). 57,546 of the compounds screened had known drug-like properties.
- AID687 is the result of a primary screen for coagulation factor XI from the Penn Center for Molecular Discovery and contains activity information of 33,067 compounds with a ratio of 1 active compound to 350 inactive compounds (0.28% minority). 30,353 of the compounds screened had known drug-like properties.

Primary and Confirmatory

- AID604 (primary) with AID644 (confirmatory)
- AID746 (primary) with AID1284 (confirmatory)
- AID373 (primary) with AID439 (confirmatory)
- AID746 (primary) with AID721 (confirmatory)

Confirmatory

- AID1608 is a different type of screening assay that was used to identify compounds that prevent HttQ103-induced cell death. National Institute of Neurological Disorders and Stroke Approved Drug Program. The compounds that prevent a release of a certain chemical into the growth medium are labelled as active and the remaining compounds are labelled as having inconclusive activity. AID1608 is a small dataset with 1,033 compounds and a ratio of 1 active to 14 inconclusive compounds (6.58% minority class).
- AID644
- AID1284
- AID439
- AID721
- AID1608
- AID644
- AID1284
- AID439
- AID721

Bioassay Descriptors (columns in the data)

As previously mentioned, the software PowerMV [11] was used to generate descriptors for the bioassay SDF files from PubChem. 179 descriptors were generated for each dataset.

• 8 descriptors useful for characterizing the drug-likeness of a compound. These include XlogP (the propensity of a molecule to partition into water or oil), the number of Hydrogen bond donors and acceptors, molecular weight, polar surface area, the number of rotatable bonds, a descriptor to indicate if the compound penetrates the blood-brain barrier and a descriptor for the number of reactive or toxic functional groups in the compound.

• 24 continuous descriptors based on a variation of BCUT descriptors to define a low dimensional chemistry space. The method used by PowerMV differs from BCUT in that PowerMV uses electro-negativity, Gasteiger partial charge or XLogP on the diagonal of the Burden connectivity matrix before calculating the eigenvalues.

• 147 bit-string structural descriptors known as Pharmacophore Fingerprints based on bioisosteric principles - two atoms or functional groups that have approximately the same biological activity are assigned the same class.

For the confirmatory datasets, Fragment Pair Fingerprints were also generated using PowerMV. For fragment-based descriptors, 14 classes of paired functional groups are defined. For example, two phenyl rings separated by two bonds are expressed as AR_02_AR

Acknowledgements

Original study: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2820499/

Data downloaded from UCI ML repository:

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Task

Use this virtual bioassay data to classify compounds as hits (active) against their biological targets.

Answers format:
Target 1 (Formylpeptide Receptor) - [hits], accuracy train/test, logloss, AUC, FP (%), FN (%), F1 score (%)
Target 2 (Rho kinase 2 inhibitors) - [hits], accuracy train/test, logloss, AUC, FP (%), FN (%), F1 score (%)
Target 3 (TNFa induced VCAM-1 cell surface inhibitors) - [hits], accuracy train/test, logloss, AUC, FP (%), FN (%), F1 score (%)
Target 4 (Yeast eIF2B) - [hits], accuracy train/test, logloss, AUC, FP (%), FN (%), F1 score (%)
Target 5 (Sphingolipid G-protein-coupled receptor) - [hits], accuracy train/test, logloss, AUC, FP (%), FN (%), F1 score (%)
Target 6 (Mitogen-activated protein kinase) - [hits], accuracy train/test, logloss, AUC, FP (%), FN (%), F1 score (%)
Target 7 (Coagulation factor XI) - [hits], accuracy train/test, logloss, AUC, FP (%), FN (%), F1 score (%)

Implement the following (for each target):

1) Plot the data:
   t-SNE -> 3D -> matplotlib, active compounds - red, inactive - blue
   t-SNE -> 2D -> matplotlib, active compounds - red, inactive - blue

2) Reduce Dimensionality. Use PCA or SVD of MDA. In general, MDA is better for separation tasks.
   (2.1) Compute reference number of principal components is 3
   (2.2) Plot them, compare with 3D t-SNE
   (2.3) Define number of principal components using value of common explained variance ratio of 80-90%

3) Build histograms for how 2 classes (active/inactive) are distributed along the obtained principal components. Are your new features are consistent?

4) Implement the following approaches for classification using your original feature set (~140), reference 3 PCs from item (2.1) and custom number of PCs from item (2.3)
   - KNeighborsClassifier

- Naive Bayes (choose distribution assumption based on features dominant distribution)
- Random Forest
- Gradient Boosting Machine
- SVM with RBF kernel
- stack all mentioned above classifier predictions and use XGBoost to make second-level predictions.

Compare this ensemble prediction accuracy with other single model accuracies.

5) Take into account that **classes are imbalanced** (number of inactive compounds are much bigger than number of active) and you have to balance the classes using one of the approaches:
   5.1 the reweighting of the training instances according to the total cost assigned to each class in the cost matrix or
   5.2 predicting the class with the minimum expected misclassification cost using the values in the cost matrix. A cost matrix may be seen as an overlay to the standard confusion matrix used to evaluate the results of a predictive modelling experiment.
   5.3 sampling techniques
   read more on this topic in “Cost-Sensitive Classifiers” section of the article

6) Compare all the metrics for the three feature sets (2.1, 2.3, all) for all the models. Which one of them worked better? Can you make any conclusions?