HelmholtzZentrum münchen
Deutsches Forschungszentrum für Gesundheit und Umwelt

Marie Curie Initial Training Network

Environmental Chemoinformatics (ECO)

Final report

24 July 201312

Mining ChEMBL database to model target cross-reactivity of chemical compounds

Duration of Short Term fellowship:
20. September – 20. December

Early stage researcher:
Ctibor Škuta

Project supervisor:
Dr. Igor V. Tetko

Research Institution:
Helmholtz Zentrum München

ECO fellowship - final report

Introduction

The use of computational toxicology has receive a lot of attention in environmental toxicology. These methods are important to substitute animal testing. There is also a wide interest and hope to substitute the animal testing using the HTS screening. ToxCast™ project[1, 2] initiated by US EPA represent one of the most widely known initiatives in this area while similar projects, such as eTOx (http://www.etoxproject.eu) and OpenPHACTS (http://www.openphacts.org/) are initiated by Academy and Industry in Europe in respect to human toxicology studies. Moreover, a lot of data about experimental measurements have been already produced and collected in public databases. The idea of this project is to explore such publicly available data with respect of their modelling potential by convenient QSAR methods. The QSAR models could be used to identify potential targets for chemical compounds and allow to understand their mechanism of action and expected toxicity.

In the last few years we have seen a steep rise of open databases of biological and chemical data. Thanks to them we can use and analyse millions of values based on drugs, drug-leads and toxicity measurements without the necessity to do our own experimental measurements. The data are vast and free, but there are also downsides to them. More sources often introduce the heterogeneity accompanying by heteroscedasticity to the data.[3] That means, that the data are not consistent and usually lack the needed continuous coverage of the chemical space. Different sources of the data can also affect the accuracy and there always can be errors on the way from the laboratory to the database. We should always take into account all the possibilities and be prepared when working with this kind of data.

There are two distinct approaches we can use when taking the advantage of open databases in QSAR/QSPR modelling. In individual approach we usually work only with data for one specific target or few related we are interested in. We are trying to understand the important aspects which determine examined type of response. The main goal is to create a model with the best possible prediction capability.

The other approach represents the introduction of the raw computational power to the process. Now, when computers enable us to analyze millions of structures in a very short time we can actually calculate models for immense number of targets but without the possibility of treating them on the individual level. In most cases the resulted models lack the high precision,

but their quantity can serve for different purposes. In our study we use models with suitable correlation between measured and predicted values to find the trends across target spectrum and detect subsets of targets which are likely to be influenced by the similar compounds. Such subsets may suggest, e.g. possible side effects of analysed drugs.

Data

We chose to work with ChEMBL database[4, 5]. ChEMBL is an open data database containing binding, functional and ADMET information for a large number of drug-like bioactive compounds. These data are in most cases manually extracted from the primary published literature (peer-reviewed scientific publications in a variety of journals, such as *Journal of Medicinal Chemistry*, *Bioorganic Medicinal Chemistry Letters* and *Journal of Natural Products*). ChEMBL curates the data and also tries to convert them to standard activity types and units. This year EBI (*European Bioinformatics Institute*) released the 14th version of ChEMBL database with more than 10 000 000 activity measures for about 9000 protein targets.

We selected small organic molecules from the database regardless the activity type and unit. From them we extracted only records with activity types pIC50, pEC50, pKi and pKd. ChEMBL contains up to 5000 of different activity types and they are very often connected with more than one standard unit. Usually there are also multiple names for one activity type. For example if one is searching for pIC50 data they can be found under names pIC50, Log IC50, -Log IC50, log(1/IC50) etc. These activity types were merged together to get as large dataset as possible.

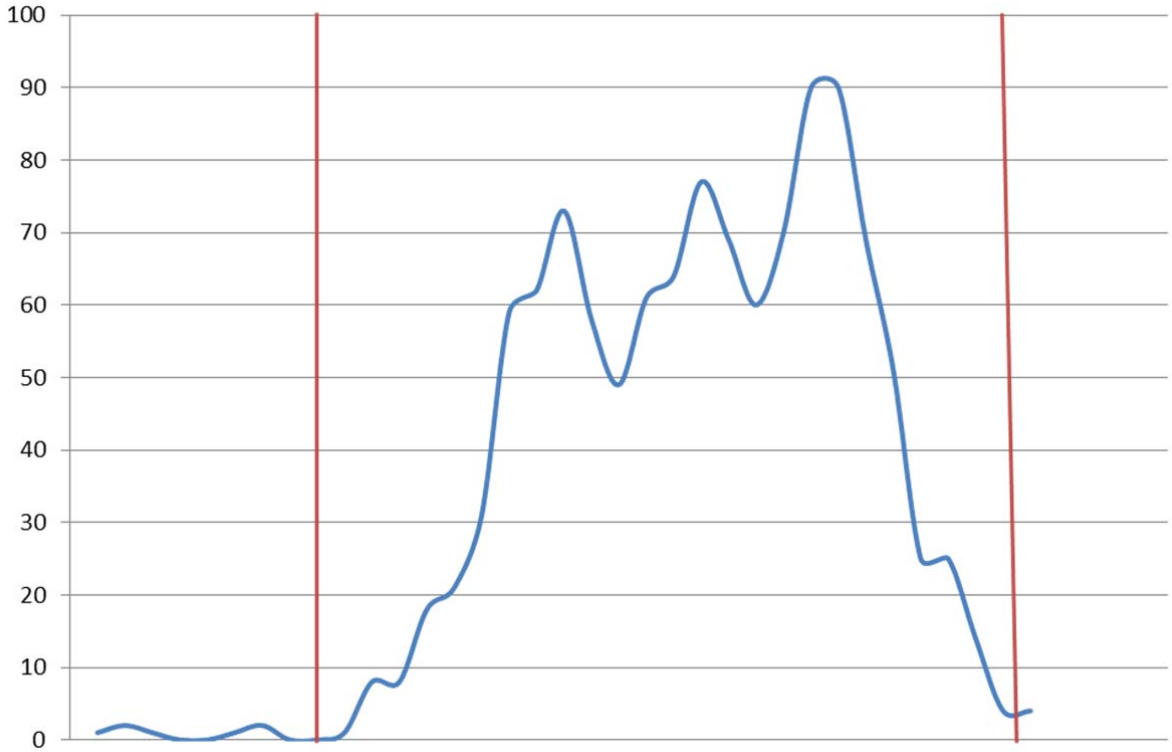
Duplicate detection

The identical records were removed on the basis of ChEMBL molecular registration number, canonical smiles, activity value and id of an article from which was the record extracted (if available). When the article id was missing - the records from different sources were also excluded.

Outliers detection

In case if data are generated with normal distribution, it is a common approach to detect outliers as a data points lying more than 3 standard deviations from the mean average of the dataset. We found that the analyzed data did not represent the normal distribution and were also not equally distributed (graph 2.1) (all data not represented by gauss curve are skewed). We decided to remove outliers by the calculation of a medcouple coefficient.[6] Medcouple

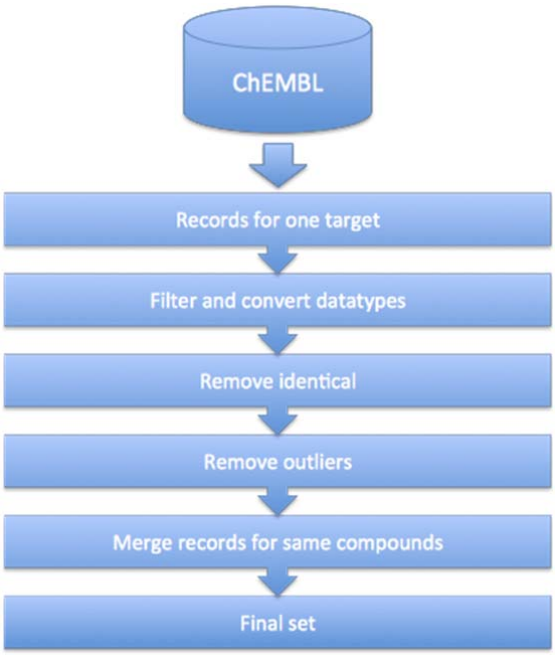
coefficient is a robust measure (using median and median deviation) of the skewness of the data. The skewness determines the size of in most cases asymmetric neighbourhood around dataset median. The data points lying behind the borders of this neighbourhood were considered as outliers and were excluded from further analysis.



2.1 Distribution of skewed data to higher values for one of the targets

Merging data for identical compounds

At this point in the data still were multiple activity values for some of the compounds. Common approach to deal with them is to merge them when the values are quite close together, or to exclude them when the values differ too much, because it is not obvious which of the extremes is right.[7] We decided to use a threshold of of 0.5 [-log(M)] unit to keep the datapoints in the set. When the standard deviation was higher we excluded all the data points, when it was lower we merged the records with value set to the mean average of the original values.



3.1 Data preprocessing workflow

Model training and validation

For the model calculation the ChEMBL data were randomly split to training set and test set (80% : 20%). The minimum threshold of records was set to 100, which means that all models were based at least on 80 records. For testing was used the 5-fold cross-validation method.

External validation set

Even when the data in the test set were unique and not contained in the training set the records didn´t have to be completely independent. Very often the compounds originated from the same compound family and values from same experiment (same article). For these reasons we decided to validate the models with compounds from external set (when possible). For that purpose we chose to use the BindingDB database.

BindingDB[8] is a public, web-accessible database of measured binding affinities, focusing chiefly on the interactions of protein considered to be drug-targets with small, drug-like molecules. BindingDB contains over 900 000 binding data, for over 6 000 protein targets and almost 400 000 small molecules.

In the matching procedure 1102 identical targets were found in the ChEMBL and BindingDB. Record sets from the BindingDB were filtered against the training set to exclude identical

compounds from validation sets. In this way we expected that molecules from the same source will be filtered out.

Methods and tools

All prediction models were calculated in the OCHEM software[9]. The OCHEM is an online database of experimental measurements integrated with the modeling environment. It offers many machine learning methods and molecular descriptors which users can take advantage of when building QSAR models. It also has molecular structure and descriptor normalization capabilities. The process of model calculation (including upload of molecules) was managed through the OCHEM web services interface.

Machine learning method and molecular descriptors were chosen on the basis of test calculation on subset of 150 models for targets from ChEMBL. These models were calculated with the different combination of tools - from machine learning methods: associative neural network (ASNN)[10,11], fast stagewise multiple linear regression (FSMLR), K-nearest neighbours (KNN) and support vector machines (SVM) - from descriptors: ESTATE[12,13]/ALOGPS[14,15,16], DRAGON6[17], CDK[18] or Chemaxon[19] descriptors. According to the statistics of the models we chose to work with the ASNN and the combination of ESTATE and ALOGPS molecular descriptors. In fact ESTATE/ALOGPS descriptors were actually slightly outperformed by DRAGON6 descriptors, but much longer time was required to compute the latter descriptors and develop new models.

ASNN

ASNN represents a combination of an ensemble of feed-forward neural networks and the k-nearest neighbor technique. This method uses the correlation between ensemble responses as a measure of distance amid the analyzed cases for the nearest neighbor technique. This provides an improved prediction by the bias correction of the neural network ensemble.[11]

ESTATE descriptors

Electro-topological state indices combine electronic and graph-topological information about a compound. The E-state representation provides an arrangement of molecules in a space which gives organization to the electronic character of the molecules. They have proven useful in establishment of QSAR and QSPR models. They were introduced by Lowel H. Hall and Lemont B. Kier in the 1990s[12,13].

ALogPS descriptors

ALogPS descriptors are two descriptors calculated by ALOGPS software: octanol/water partition coefficient (logP) and solubility in water (logS). ALOGPS proved to be accurate in predicting lipophilicity and aqueous solubility of molecules.[20]

Model parameters

Resulted models are characterized by four parameters: squared correlation coefficient (r2), cross-validated squared correlation coefficient (q2), root mean square error (RMSE) and mean average error (MAE). The decision whether the QSAR model is usable was determined by the q2 coefficient with threshold of 0.5.

Results

First we used test set of 150 models to measure the difference between models built on preprocessed data and raw data from ChEMBL. This step was necessary for the validation of the proposed averaging technique. The record count of cleaned data against raw data was in average lower by 24.4% (6.7% excluded in process of outliers detection, 17.7% by merging of the records). Models based on cleaned data had in average lower mean average error (MAE) by 0.153 thus indicating an importance of the used data cleaning procedure.

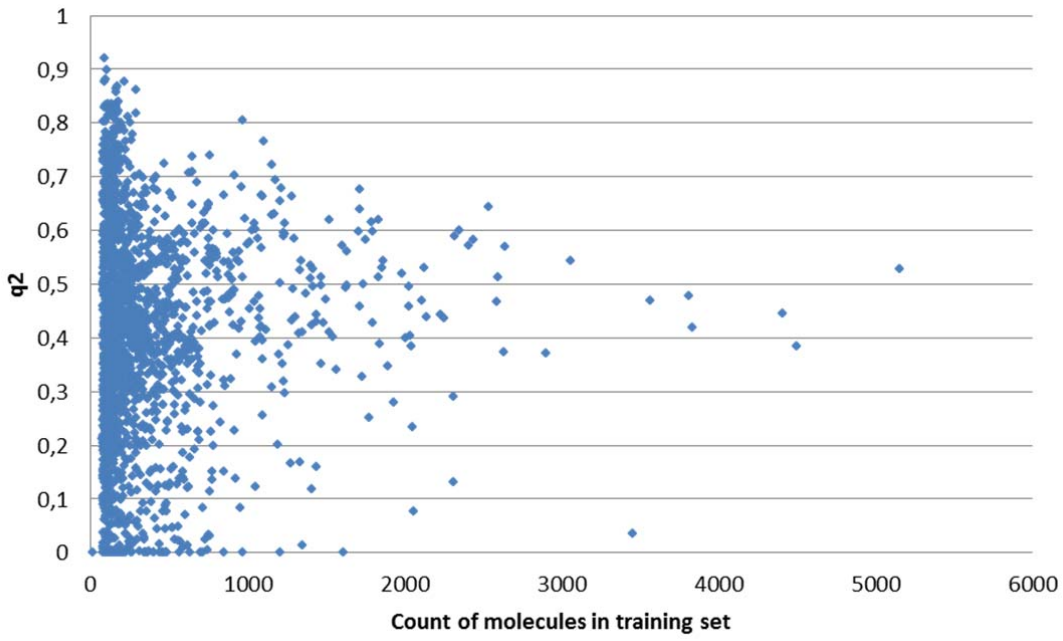
In ChEMBL there were 1810 dataset with >100 records, i.e. above the threshold we set to build models. These sets covered 1298 unique out from 8998 total targets (since there can be multiple models for different activity type for one target). Many of these sets produced models with low prediction accuracy and thus were not suitable for further usage. The poor performance could be accounted to several factors: wrong data, narrow coverage of the chemical space, not suitable structural descriptors etc.

The results for the external validation sets from BindingDB were not much conclusive. In the end there were only 96 identical targets, for which the models were calculated. 29 of validation sets resulted with q2 above 0.5 and 48 of them with q2 equal to 0. The causes resided from: initially bad model, only few molecules in validation set or excessive distance from model application domain (uncovered parts of chemical space). Further we used training set cross-validation and validation sets separated from the ChEMBL data.

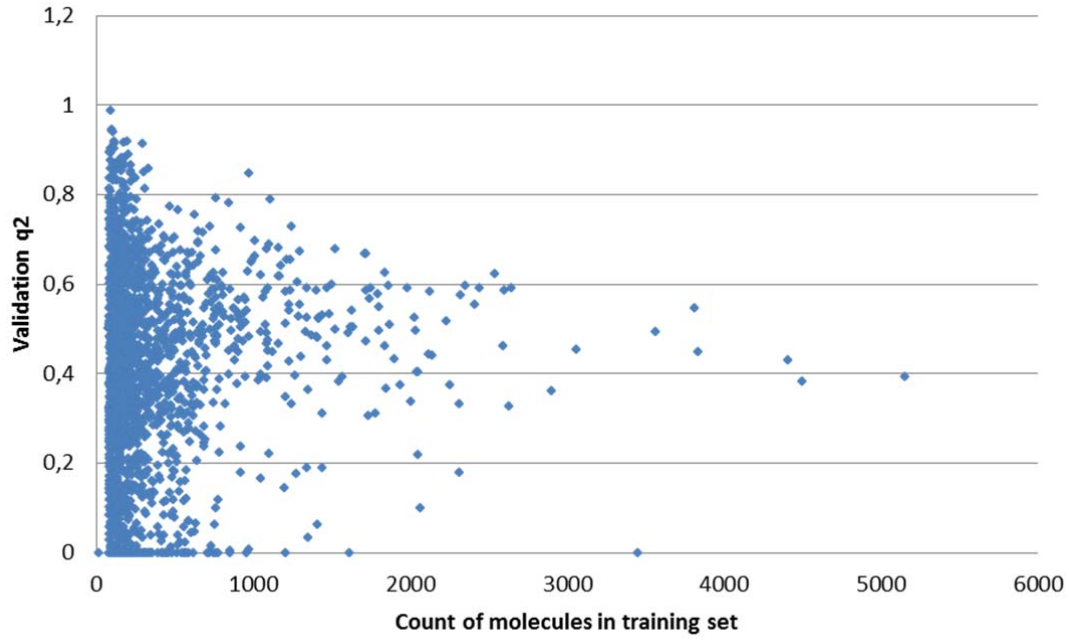
There is a high correlation between validation and training mean average error (MAE) (graph 6.4). The relation between cross-validated R2 (q2) is much more variant. The root of variance is

mainly the difference of size of training and validation set. Graphs 6.1, 6.2 and 6.3 show the large differences especially for smaller sets.

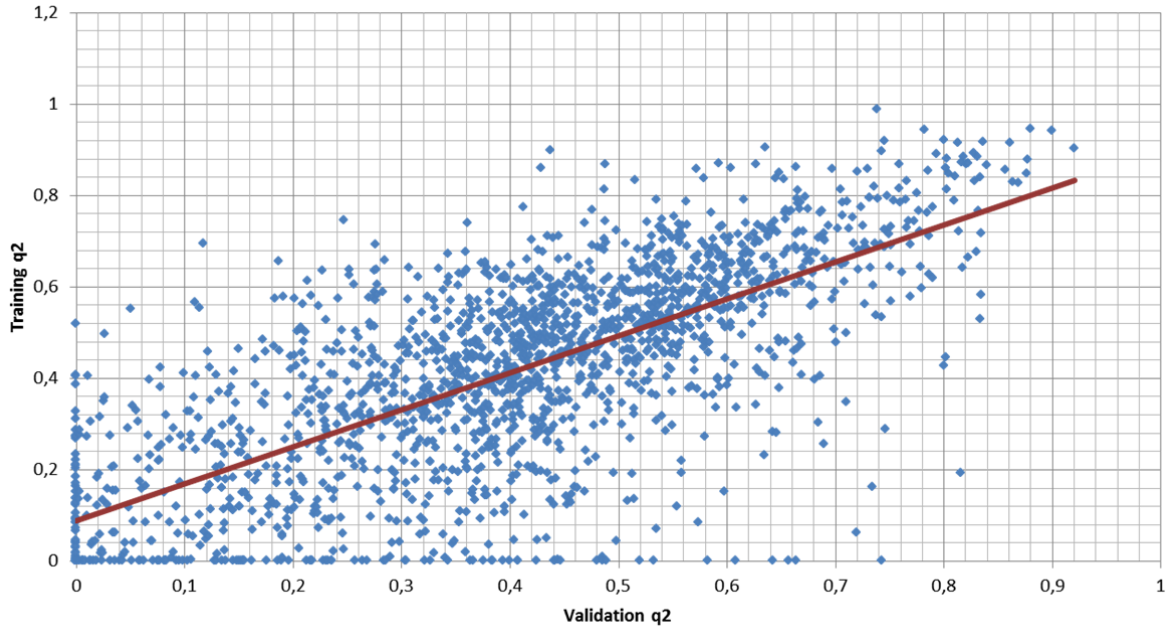
The models for further usage were chosen on the basis of cross-validated R2 (q2) with threshold 0.5 for both training and validation sets. This value fit for 393 models and 354 unique targets, which will be further used to identify cross-reactivity (correlation) between targets.



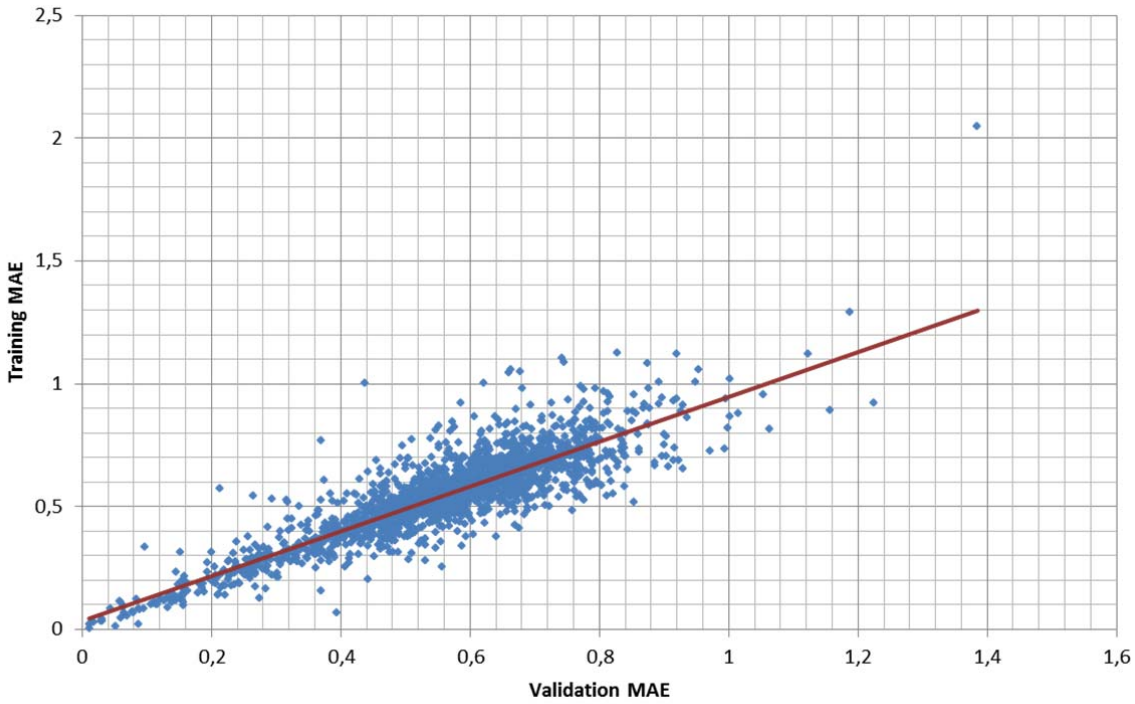
6.1 The dependency of q2 on the size of the training set



6.2 The dependency of validation q2 on the size of the training set



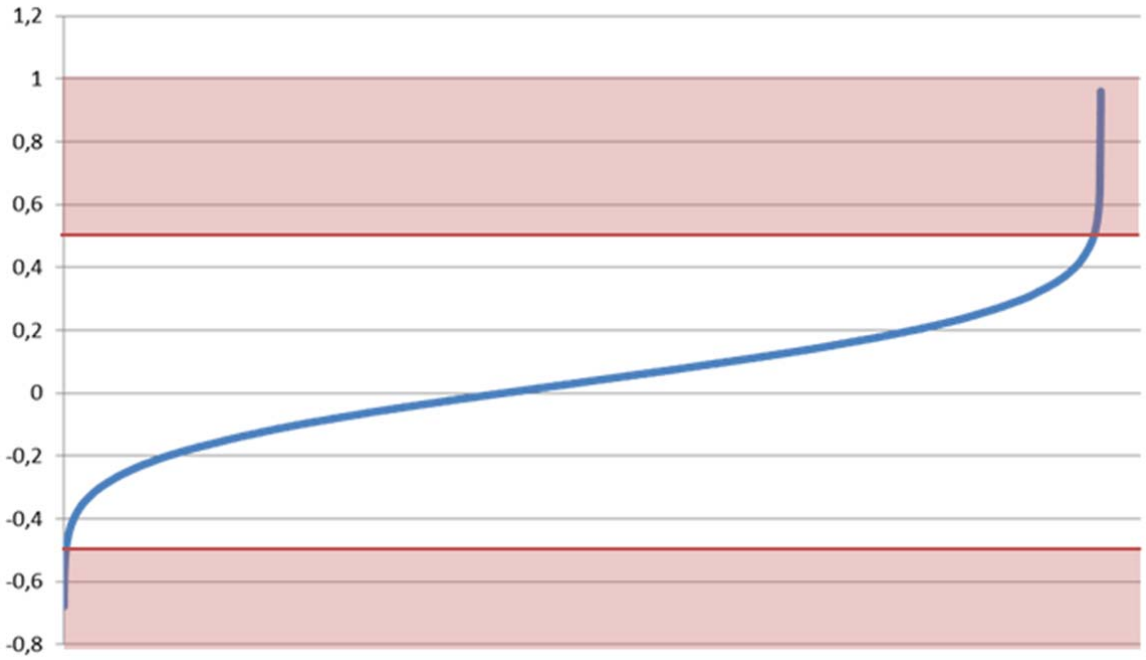
6.3 Relation between training and validation cross-validated error (q2)



6.4 Relation between training and validation mean average error (MAE)

Our goal was to analyze relationships in target set in a way to find such that respond to the similar compounds. To get needed data we applied the chosen models on the set of 9000 molecules from ChemNavigator iResearch Library and analyzed the predictions. We assessed the target relations from the calculation of correlation coefficient of all possible couples of targets. Correlation coefficient (number between -1 and 1) represents positive or negative measure of dependency of couple of variables. The closer is the absolute value of coefficient to 1 the more dependent the variables are and vice versa. Absolute value above 0.5 is considered as a strong relation.

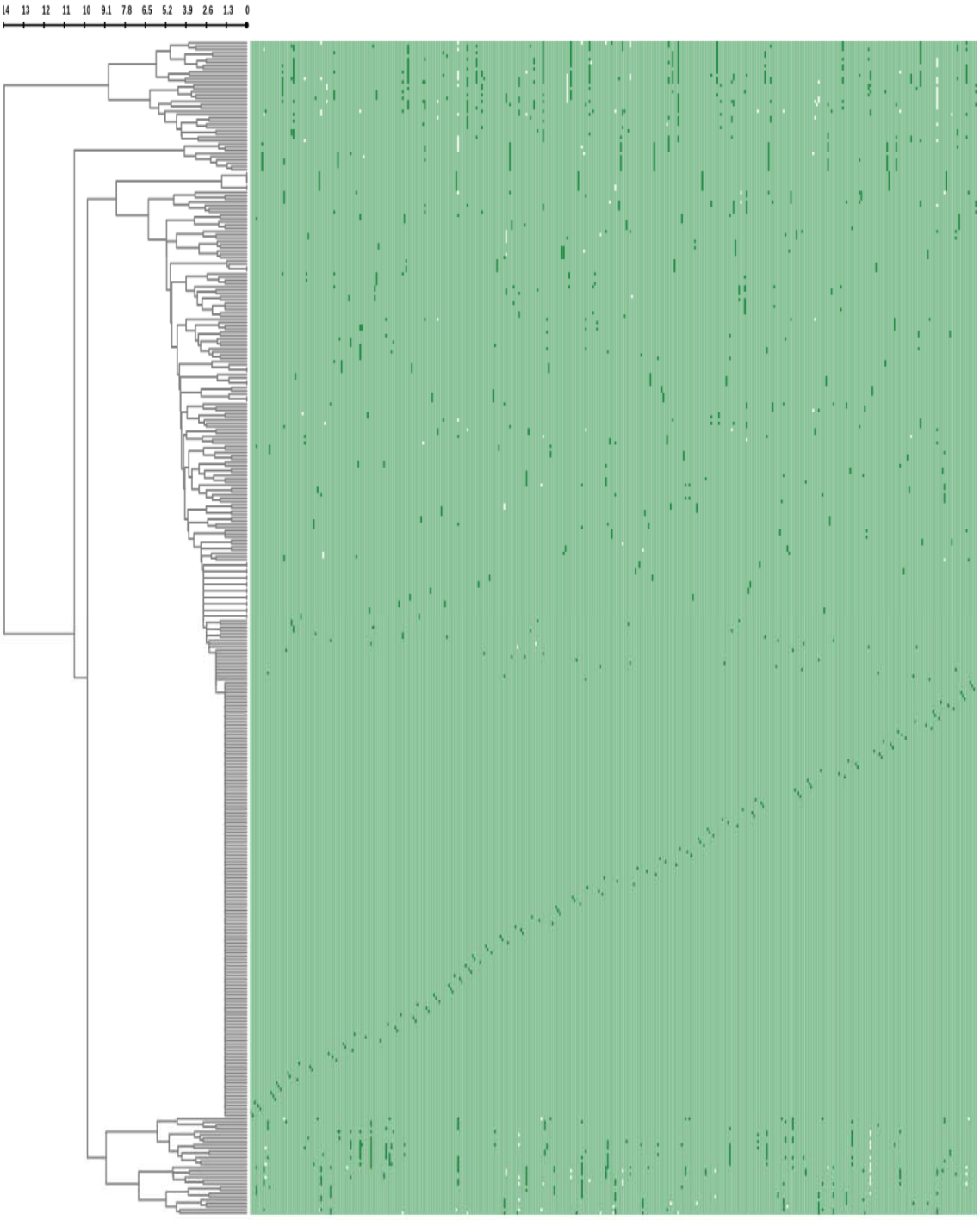
According to correlation coefficient we found 692 couples of strongly correlated (absolute value of CC > 0.5) receptors (177 negative, 515 positive) (*graph 8.1*). The results for the highest positive correlated receptors indicated that the method was working properly. In most correlated 100 there was a majority of targets which belong to the protein family. For example: *GABA receptor alpha-6 subunit - GABA receptor alpha-2 subunit, Protein kinase C delta - Protein kinase C eta*).



8.1 Correlation coefficient distribution for all possible couples of targets

However, for targets from different functional groups was hard to assess the correctness of results without thorough knowledge of biological/biochemical background. Some of them could be proved/disproved by the scientific literature, but the necessary time to find all of them exceeds our time possibilities. We can also assume that by far not all the relations are mentioned in the literature, but that of course doesn't mean the lack of it.

To see a bigger picture we accomplished the cluster analysis of receptors with at least one strong activity correlation to the other receptor. For that we used hierarchical clustering algorithm and ward's linkage (assuming there are more smaller clusters). The values for clustering were adjusted to see the correlated receptors clearly. For CC above 0.5 we set the value to 1, for CC below -0.5 to -1 and for others we set the value to 0. The result of the clustering is displayed in the dendrogram and heatmap graph (8.2). We can see there few very compact clusters with up to ten members. Some of the clusters are shown in the *table 8.2*. There is also large area of 134 non-correlated receptors.



8.2 The clustering dendrogram with heatmap of correlation matrix of 393 receptors

	Receptors in cluster
A	**Sodium channel protein type I alpha subunit** **Sodium channel protein type III alpha subunit** **Sodium channel protein type II alpha subunit** Alpha-1a adrenergic receptor FK506 binding protein 12
B	**Protein farnesyltransferase beta subunit** **Protein farnesyltransferase/geranylgeranyltransferase type I alpha subunit** Serine/threonine-protein kinase AKT Endothelin receptor ET-A Human immunodeficiency virus type 1 protease
C	**Kappa opioid receptor** **Mu opioid receptor** Inhibitor of apoptosis protein 3 Acetylcholinesterase Matrix metalloproteinase 9 Sodium/glucose cotransporter 1 Somatostatin receptor 4

8.3 Three of the clusters found in the cluster analysis of activity correlation of the receptors (receptors assessed to be clustered properly in the same cluster are in bold)

The situation could become clearer by comparing the functional relations between receptors with the similarity measure of their protein sequences, its 3D structures or possibly best 3D structures of their active places. The proper visualization of all the relations between the described functional clusters and their distance from other ones could help to see the bigger picture.

Conclusion

We used the concept of chemogenomic space to find the functional relations between biological targets contained in the ChEMBL database. The relations resulted from the in silico activities of a large library of compounds predicted by target QSAR models. The method proved its concept and found strong relations among biologically close receptors. The correctness of the identified connection between diverse pairs of targets will need experimental confirmation.

The target models are constructed on the basis of ligands. Thus, the similarity of receptors were based on the the similarity of their ligand sets. For further work the comparison of this similarity should be accompanied by structural spatial similarities of proteins and both these similarities will be interesting to compare with respect to their complementarity. It should be mentioned that the ligand based similarity methods have been already successfully used in the field of polypharmacology to search for the new targets for old drugs (drug repurposing).[21-29] The results of the current study can be also explored for similar purposes and/or identification of side effects and toxicity of chemical compounds. Moreover, similar methodology can be used to explore ToxCast™ data and prediction of the environmental toxicity of chemical compounds.

Acknowledgement

This study was supported by FP7 MC ITN project “Environmental Chemoinformatics” (ECO), grant agreement number 238701.

References

1. Dix, D. J.; Houck, K. A.; Martin, M. T.; Richard, A. M.; Setzer, R. W. et al. The ToxCast program for prioritizing toxicity testing of environmental chemicals. Toxicol. Sci. 2007, 95, 5–12.
2. ToxCast. Computational Toxicology Research. http://www.epa.gov/ncct/toxcast/ (accessed June 28, 2013).
3. Varnek, A.; Baskin, I. Machine Learning Methods for Property Prediction in Chemoinformatics: Quo Vadis?. *J. Chem. Inf. Model.* 2012, *52* (6), 1413–1437.
4. *ChEMBL*; European Bioinformatics Institute (EBI): Cambridge, 2011. http:/www.ebi.ac.uk/chembl/ (accessed December 19, 2012)
5. Gaulton, A.; Bellis, L. J.; Bento, A. P.; et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 2012, *40* (D1), D1100–D1107.

6. Hubert, M.; Van Der Veeken, S. Outlier detection for skewed data. *J. Chemom.* 2008, *22* (3-4), 235–246.
7. Fourches D.; Muratov E.; Tropsha A. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J. Chem. Inf. Model.* 2010 50 (7), 1189–1204.
8. Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: A web-accessible database of experimentally determined protein−ligand binding affinities *Nucleic Acids Res.* 2007, 35, D198–D201
9. Sushko, I.; Novotarskyi, S.;, Körner, R.; et. al. Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. *J. Comput. Aided. Mol. Des.* 2011; 25(6):533-54.
10. Tetko, I.V. Associative Neural Network, *Neural Processing Letters*, 2002, 16, 187-199.
11. Tetko, I.V. Neural Network Studies. 4. Introduction to Associative Neural Networks, *J. Chem. Inf.* Comput. Sci., 2002, 42, 717 -728.
12. Hall, L. H.; Kier, L. B. An atom-centered index for drug QSAR models. In Bernard Testa (editor): Advance in Drug Design, vol. 22, Academic Press, 1992.
13. Hall, L. H.; Kier, L. B. Electrotopological state indices for atom types: A novel combination of electronic, topological, and valence state information, *J. Chem. Inf. Comput. Sci.,* 1995, 35(6), 1039-1045.
14. Tetko, I. V.; Tanchuk, V. Y. Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 program, *J. Chem. Inf. Comput. Sci.*, 2002, 42, 1136-45.
15. Tetko, I. V.; Tanchuk, V. Y.; Villa, A. E. Prediction of n-octanol/water partition coefficients from PHYSPROP database using artificial neural networks and E-state indices, *J. Chem. Inf. Comput. Sci.*, 2001, 41, 1407-21.
16. Tetko, I. V.; Tanchuk, V. Y.; Kasheva, T. N.; Villa, A. E. Estimation of aqueous solubility of chemical compounds using E-state indices, *J. Chem. Inf. Comput. Sci.*, 2001, 41, 1488-93.
17. Mauri, A.; Consonni, V.; Pavan, M.; Todeschini, R. DRAGON software: An easy approach to molecular descriptor calculations *Commun. Math. Comput. Chem.* 2006, 56, 237–248.
18. Steinbeck, C.; Han, Y. Q.; Kuhn, S.; et al. The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics. *J. Chem. Inf. Comput. Sci.* 2003, *43* (2), 493–500.
19. Chemaxon. Chemical hashed fingerprints. http://www.chemaxon.com/jchem/doc/user/fingerprint.html (accessed June 28, 2013).
20. Rayne, S.; Forest, K. Performance of the ALOGPS 2.1 program for octanol-water partition coefficient prediction with organic chemicals on the Canadian Domestic Substances List. *Nature Preceedings* 2012, DOI: 10.1038/npre.2009.3882.1.
21. Keiser, M.J., et al., *Relating protein pharmacology by ligand chemistry.* Nature Biotechnology, 2007. 25(2): p. 197-206.

22. Lounkine, E., et al., *Large-scale prediction and testing of drug activity on side-effect targets.* Nature, 2012. 486(7403): p. 361-+.
23. Campillos, M., et al., *Drug target identification using side-effect similarity.* Science, 2008. 321(5886): p. 263-266.
24. Keiser, M.J., et al., *Predicting new molecular targets for known drugs.* Nature, 2009. 462(7270): p. 175-U48.
25. DeGraw, A.J., et al., *Prediction and Evaluation of Protein Farnesyltransferase Inhibition by Commercial Drugs.* Journal of Medicinal Chemistry, 2010. 53(6): p. 2464-2471.
26. Moneriz, C., et al., *Multi-targeted activity of maslinic acid as an antimalarial natural compound.* Febs Journal, 2011. 278(16): p. 2951-2961.
27. Mestres, J., S.A. Seifert, and T.I. Oprea, *Linking Pharmacology to Clinical Reports: Cyclobenzaprine and Its Possible Association With Serotonin Syndrome.* Clinical Pharmacology & Therapeutics, 2011. 90(5): p. 662-665.
28. Schlessinger, A., et al., *Structure-based discovery of prescription drugs that interact with the norepinephrine transporter, NET.* Proceedings of the National Academy of Sciences of the United States of America, 2011. 108(38): p. 15810-15815.
29. Gregori-Puigjane, E., et al., *Identifying mechanism-of-action targets for drugs and probes.* Proceedings of the National Academy of Sciences of the United States of America, 2012. 109(28): p. 11178-11183.