Biases of Drug–Target Interaction Network Data

Twan van Laarhoven and Elena Marchiori

Institute for Computing and Information Sciences, Radboud University Nijmegen, The Netherlands
{tvanlaarhoven,elenam}@cs.ru.nl

Abstract. Network based prediction of interaction between drug compounds and target proteins is a core step in the drug discovery process. The availability of drug–target interaction data has boosted the development of machine learning methods for the *in silico* prediction of drug–target interactions. In this paper we focus on the crucial issue of data bias.

We show that four popular datasets contain a bias because of the way they have been constructed: all drug compounds and target proteins have at least one interaction and some of them have only a single interaction. We show that this bias can be exploited by prediction methods to achieve an optimistic generalization performance as estimated by cross-validation procedures, in particular leave-one-out cross validation. We discuss possible ways to mitigate the effect of this bias, in particular by adapting the validation procedure. In general, results indicate that the data bias should be taken into account when assessing the generalization performance of machine learning methods for the *in silico* prediction of drug–target interactions.

The datasets and source code for this article are available at
http://cs.ru.nl/~tvanlaarhoven/bias2014/

1 Introduction

An important problem in pharmacology is to find interactions between drug compounds and target proteins in order to understand and study their effects. The *in silico* prediction of such interactions is crucial for improving the efficiency of the laborious and costly experimental determination of drug–target interaction, see e.g. [5].

Drug-target interaction data are publicly available for various classes of pharmaceutically useful target proteins including enzymes, ion channels, GPCRs (G Protein-Coupled Receptors) and nuclear receptors [13]. Various databases have been built and maintained, such as KEGG BRITE [16], DrugBank [29], GLIDA [23], SuperTarget and Matador [12], BRENDA [26], and ChEMBL [24], containing drug–target interaction and other related sources of information, like chemical and genomic data.

The availability of these data stimulated the development of machine learning methods for the *in silico* prediction of drug-target interactions [8]. The current state-of-the-art for the *in silico* prediction of drug–target interaction involves

M. Comin et al. (Eds.): PRIB 2014, LNBI 8626, pp. 23–33, 2014.

methods that employ similarity measures for drug compounds and for target proteins in the form of kernel functions, e.g., Bleakley et al. [2], Chen et al. [4], Gönen [11], vanLaarhoven et al. [21], Mei et al. [22], Wassermann et al. [28], Yamanishi et al. [30, 31].

One can distinguish between prediction for 'known' drug compounds or targets, for which at least one interaction is present in the training set; and prediction for 'unseen' drug compounds or targets, for which no interaction is available in the training set. This results in four possible settings for predicting drug-target interaction, depending on whether the drug compounds and/or targets are known or unseen [30].

In our recent work on predicting drug–target interactions [20] we discovered that a positive bias was implicity introduced in a published method. This motivated the two main research questions we will address in this paper.

1. How does data bias affect the results of procedures used to estimate the generalization performance of a method?
2. Can we quantify and avoid such bias?

Cross-validation (CV) [19] is typically used to assess the generalization performance of methods in the above mentioned settings. The dataset is repeatedly partitioned into two disjoint parts, a training set and a hold-out set. For each partition, the training set is used to construct the predictor and the hold-out set is used for testing. Popular variants are 10-fold CV, where the data is partitioned into ten folds, and each fold is used once as the hold-out set, and leave-one-out cross-validation (LOOCV), where each example constitutes one hold-out set. In the context of drug–target interaction various cross-validation settings can be defined, depending on what is considered an example (e.g. a single drug–target pair or all interactions with a single drug compound) and on the selected CV procedure.

We consider the four popular drug–target interaction datasets in humans involving enzymes, ion channels, G-protein-coupled receptors (GPCRs) and nuclear receptors from Yamanishi et al. [30]. These data have been used as benchmark datasets in recent works, e.g. Bleakley et al. [2], Chen et al. [4], Gönen [11], vanLaarhoven et al. [21, 20], Mei et al. [22], Wassermann et al. [28].

In this paper we show experimentally that these datasets contain a bias which may lead to optimistic CV generalization results. Furthermore, the extent to which this bias affects the results can differ for different methods. As a result, it is unclear whether a method with better CV results on these datasets will also have better performance in real applications.

Specifically, these datasets have been constructed in such a way that each drug compound and target protein has at least one interaction. Furthermore, some drug compound and/or targets have only a single interaction.

We show how this bias can be incorporated into a baseline prediction method in such a way that it significantly increases the LOOCV generalization performance. We investigate how this bias can be reduced and quantified. We show experimentally that 5- or 10-fold CV reduces (but does not eliminate) the bias. Furthermore, the presence of this bias can be quantified by separating

the performance metrics for drug compounds and targets with just one interaction from that for other drug–target interaction pairs. This provides an alternative procedure to assess the generalization performance of a method by highlighting the effect of the data bias.

In general, our results provide a contribution towards the understanding of CV procedures in the presence of data bias in the context of drug-target interaction networks.

1.1 Related Work

Dataset bias has been investigated in different domains, e.g. in ligand based virtual screening [1], where local clustering and global spread of the considered benchmark datasets were identified influencing validation results, and in object recognition [27], where current state of recognition datasets have been comparatively analyzed and evaluated based on criteria including relative data bias and cross-dataset generalization. To the best of our knowledge, this is the first time that drug–target interaction network data bias is analyzed.

The dangers of CV have been studied by the machine learning community in various contexts. For instance, in Isaksson et al. [14] CV and bootstrapping in small sample classification are investigated. A fundamental problem is that the uncertainty in a point estimate obtained with these procedures is unknown and may be quite large. The authors therefore suggest that the final classification performance should be reported in the form of a Bayesian confidence interval or using some other method that yields conservative measures of the uncertainty. Furthermore, in Rao et al. [25] it was empirically shown that when the number of algorithms is large, LOOCV is not an effective estimate of generalization performance for the algorithm that has the best cross-validation performance. The authors showed that this behavior worsens as the sample size decreases, and as the dimensionality and number of algorithms increase. The phenomenon of under-estimating cross validation error was also demonstrated on some benchmark data sets, and was shown to be worse for datasets with higher dimensionality.

2 Materials

In Yamanishi et al. [30] datasets were introduced for the drug–target prediction problem. These datasets are based on four different domains: enzymes, ion channels, GPCRs and nuclear receptors. The datasets are constructed in such a way that only the proteins that have an interacting drug are included, and for each domain only the drugs that interact with at least one protein are included. It turns out that this property introduces problems for validation.

In Table 1 we give an overview of the four datasets as they are used in recent publications. As can be seen in the last column, a large fraction of the drug compounds and target have just one interaction in the dataset. Or equivalently, there are many interactions which are the only interaction for a drug–target. We call such interacting pairs *unique*.

The interactions in a dataset can be encoded in a matrix y_{dt}, such that $y_{dt} = 1$ if drug compound d interacts with target protein t, and $y_{dt} = 0$ otherwise. Besides this interaction information, there is also other information available on the drugs and targets themselves. Usually this is encoded in kernel matrices that give a similarity score between two drugs or two targets.

Table 1. The number of drug compounds, the number of target proteins, the number of interactions and the number of unique interaction pairs (interactions which are the only one for a drug or target) in the drug–target datasets from Yamanishi et al. [30]

Dataset	Drugs	Targets	Interactions	Unique
Enzyme	445	664	2926	451 (15%)
Ion Channel	210	204	1476	103 (7%)
GPCR	223	95	635	132 (21%)
Nuclear Receptor	54	26	90	44 (49%)

3 Methods

3.1 Validation Procedures

There are two main ways in which these datasets of interactions can be used by machine learning methods:

1. To train a model to predict with which targets a previously unseen drug will interact. We call this the 'unseen drug' setting.
2. To find new putative interactions between drugs and targets already in the dataset. We call this the 'pairs' setting.

An overview of the prediction setting and type of CV used in state-of-the-art methods applied to these datasets are shown in Table 2. In this work we focus primarily on the 'pairs' setting, which is used by most of the methods listed in the table.

Usually methods are compared by looking at the ranking of interactions they produce in a cross-validation setting. That is, each drug–target pair is assigned a score by each method, where only other interacting pairs are shown to the method. Then the pairs are ranked based on these scores and the quality of the ranking is compared using AUC, AUPR or other summary statistics. Specifically, the ROC curve of true positives as a function of false positives is computed, and the area under the ROC curve (AUC) is considered as quality measure, see for instance [10]. Furthermore, the precision–recall curve is computed, that is, the plot of the ratio of true positives among all positive predictions for each given recall rate. The area under this curve (AUPR) is a more informative quality measure than the AUC, as it punishes much more the existence of false positive examples found among the top ranked prediction scores [6].

Table 2. A list of papers that used the interaction data in Table 1, showing the type of prediction setting ('unseen drug' or 'pairs') and type of CV procedure used

	Unseen drug	Pairs	CV procedure
Yamanishi et al. [30]	✓	✓	10-fold CV
Bleakley et al. [2]	✓	✓	LOOCV, 10-fold CV
vanLaarhoven et al. [21]	-	✓	LOOCV, 10-fold CV
Chen et al. [4]	-	✓	LOOCV
Gönen [11]	✓	-	5-fold CV
Mei et al. [22]	-	✓	LOOCV, 10-fold CV
vanLaarhoven et al. [20]	✓	✓	LOOCV, 5-fold CV

3.2 Biases

Suppose that a method is tested using LOOCV. Then if a unique interaction (d, t) is left out, the method will see a row (or column) of zeros in the matrix. But we know that the dataset does not have such rows or columns, since each drug and target has at least one interaction. We can therefore know with certainty that this pair interacts. This process is illustrated in Fig. 1.

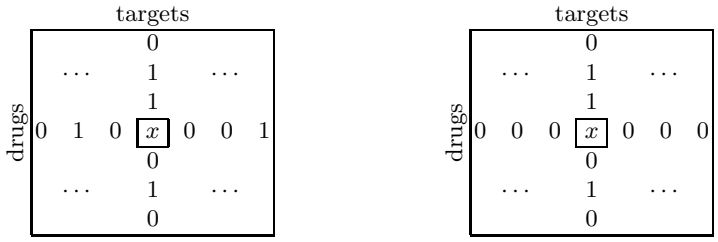

Fig. 1. In the LOOCV procedure, the task is to predict a single unknown drug–target interaction, assuming all other interactions are known. This is indicated by x in the matrix of drug–target interactions. Because of the construction of the dataset, we can know with certainty that in the second matrix $x = 1$, otherwise this drug compound would not be included in the dataset.

Consider a simple baseline method, that ranks drug–target pairs by the number of adjacent pairs that are known to interact, where two drug–target pairs are adjacent if they share a drug or a target. This number of adjacent interacting pairs for the pair (d, t) is

$$a_{dt} = a_{dt}^{\text{drug}} + a_{dt}^{\text{targ}}, \quad \text{where} \quad a_{dt}^{\text{drug}} = \sum_{d' \neq d} y_{d't}, \quad a_{dt}^{\text{targ}} = \sum_{t' \neq t} y_{dt'}.$$

At first glance we would expect drugs or targets that already have many known interactions to be more promiscuous, and therefore also more likely to interact

with other drugs and targets. But as explained in the previous paragraph that is not the case when LOOCV is used.

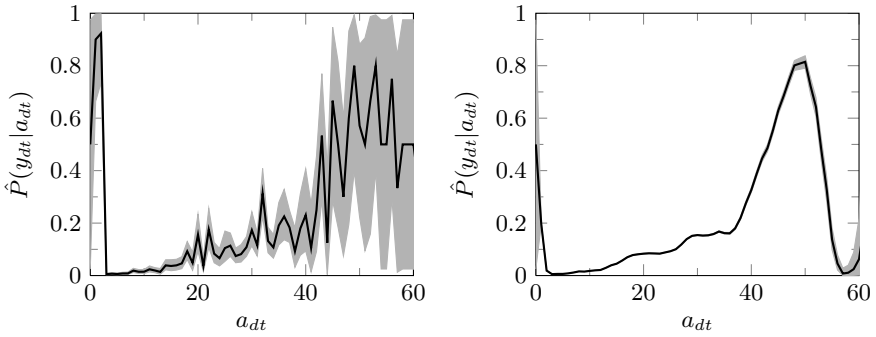

Fig. 2. Probability of a drug–target pair interacting given the number of adjacent interactions. The first plot shows this probability for LOOCV in the GPCR dataset, the second plot for 10-fold cross validation. The shaded area indicate a 95% confidence interval based on a uniform prior.

To test this effect, in Fig. 2 we have plotted the fraction of pairs that interact against a_{dt}. This is an empirical estimate $\hat{P}(y_{dt}|a_{dt})$ of the probability that d and t interact given the number of adjacent pairs for (d, t). Overall there is indeed a trend for larger a_{dt} to correspond to a higher probability of interacting. But for very low a_{dt} we see the bias in action: the probability of such pairs interacting is very high, since many of them are unique interactions.

A method can exploit this knowledge as follows. Consider the biased variant of the baseline method, which is the same as the baseline, except that it ranks the pairs with no observed adjacent pairs sharing a drug or with no pairs sharing a target before all other drug–target pairs. More precisely, instead of ranking pairs by a_{dt}, they are ranked by

$$a_{dt}^{\text{unique}\mapsto\infty} = \begin{cases} \infty & \text{if } a_{dt}^{\text{drug}} = 0 \text{ or } a_{dt}^{\text{targ}} = 0 \\ a_{dt} & \text{otherwise.} \end{cases}$$

In Table 3 we compare the LOOCV performance of this biased method to the unbiased baseline.

To estimate the statistical significance of the AUC results we used the method described in DeLong et al. [7]. To determine significance of the AUPR results we used bootstrapping.

The difference between the unbiased and the biased methods is purely due to the unique interactions. In Table 3 we also show the AUC and AUPR split up for just the unique and non-unique interactions. With the unbiased baseline method, the AUC for unique interactions is barely above random chance level, while the biased baseline method achieves a perfect AUC. The overall AUC is

Table 3. Performance of the unbiased baseline method and the biased variant when tested with LOOCV. The best results for each dataset are indicated in bold.

Dataset	Method	AUC			AUPR		
		overall	unique	other	overall	unique	other
Enzyme	Baseline	0.880	0.668	0.919	0.101	0.006	0.102
	Biased	**0.931**	1.000	0.919	**0.301**	1.000	0.102
Ion Channel	Baseline	0.850	0.528	0.874	0.244	0.003	0.254
	Biased	**0.883**	1.000	0.874	**0.355**	1.000	0.254
GPCR	Baseline	0.796	0.542	0.863	0.157	0.009	0.168
	Biased	**0.891**	1.000	0.863	**0.420**	1.000	0.168
Nuclear Receptor	Baseline	0.703	0.511	0.887	0.152	0.044	0.143
	Biased	**0.942**	1.000	0.887	**0.682**	1.000	0.143

a weighted average of the AUCs for unique and non-unique interactions, where the weight corresponds to the fraction of unique interactions. For example, for the GPCR dataset, $79\% \cdot 0.863 + 21\% \cdot 1.000 = 0.891$. Such a relation does not hold for AUPR scores, but the overall picture is similar.

4 Avoiding the Bias

It seems that the biased results stem from the use of LOOCV. And so one would hope to avoid this problem by using 10 fold CV instead. As the right part of Fig. 2 shows, this indeed reduces the bias, but it does not completely eliminate it.

We have repeated the experiment from the previous section with 10-fold CV instead of LOOCV. This is the setting used by, for instance Yamanishi et al. [30]. As seen in the Table 4, exploiting the data bias still improves the AUC and AUPR scores for unique interactions, but this comes at the cost of the performance for non-unique interactions. In general, with k-fold cross-validation on a dataset with n drugs/targets, for each unique interacting pair, there are on the order of n/k non-interacting pairs that will be excluded in the same fold. These pairs will appear similar to the unique interaction ones. As the dataset becomes larger, there will be more such pairs.

However, it is still possible to beat the baseline method by making a trade-off between the increased performance on unique interactions and decreased performance on other interactions. For example, one can introduce the 'slight bias' method that ranks pairs which appear to be unique as if they have k adjacent interactions. So it ranks pairs by $a_{dt}^{\text{unique}\mapsto k}$ for some $k < \infty$. By tuning this parameter k we can tune the trade-off. In our experiments we chose k with cross validation. As shown in Table 4, this method achieves best AUC and AUPR on all but the smallest dataset; and in all cases shows a significant improvement over the baseline method.

Table 4. Performance of the unbiased baseline method and the biased variants when tested with 10 fold CV. The best results for each dataset are indicated in bold, results in italic do not differ significantly from the best (at $\alpha = 0.05$).

		AUC			AUPR		
Dataset	Method	overall	unique	other	overall	unique	other
Enzyme	Baseline	0.879	0.669	0.917	0.098	0.006	0.099
	Slight bias	**0.900**	0.818	0.915	**0.101**	0.012	0.097
	Biased	0.862	0.982	0.840	0.056	0.135	0.027
Ion Channel	Baseline	0.849	0.530	0.873	0.246	0.003	0.254
	Slight bias	**0.859**	0.695	0.871	**0.248**	0.005	0.252
	Biased	0.836	0.987	0.824	0.123	0.128	0.098
GPCR	Baseline	0.795	0.543	0.859	0.154	0.009	0.163
	Slight bias	**0.841**	0.801	0.853	**0.168**	0.025	0.155
	Biased	0.827	0.975	0.788	0.116	0.180	0.057
Nuclear Receptor	Baseline	0.697	0.533	0.885	0.154	0.047	0.155
	Slight bias	0.857	0.884	0.846	0.247	0.177	0.124
	Biased	**0.878**	0.967	0.781	**0.351**	0.473	0.070

So far we have considered the bias in the pairs setting. Results suggest that perhaps this validation setting should not be used. An alternative is the unseen drug setting, where one or more rows are left out in their entirety from the drug–target interaction matrix. This means that it becomes impossible to see if a pair is unique for a certain drug. But there are still interactions that are unique for a target. As shown in Table 5, this bias can still be exploited for improving CV performance, even when using 5- or 10-fold cross-validation.

Another option is to separate the unique interactions from the non-unique interactions when doing validation. As shown in our experiments, the non-unique interactions are not sensitive to the same bias. A good solution would be to only consider the AUC and AUPR scores for the non-unique interactions when comparing different methods. This still introduces a bias of a different kind, however, since some drug compounds and targets will be unnecessarily excluded.

A different way to validate a method is to seek confirmation of the predictions in other datasets. This is done by for instance Yamanishi et al. [31], van-Laarhoven et al. [21], Gönen [11], where the 10 highest rank predictions are looked up in the literature, and in newer versions of the KEGG BRITE, DrugBank Chembl, SuperTarget and Matador databases. A problem with such validation is that it is hard to quantify the performance, because only a few interactions are verified, and because these databases are extended between the publication of different papers.

Perhaps the most principled way of avoiding biases in validation is to act on the data and construct more realistic datasets. For this problem, that means that the dataset should also include compounds that interact with none of the targets,

Table 5. Performance of the baseline method and biased variants in the unseen drug setting, when validated with 5-fold CV. The best results for each dataset are indicated in bold, results in italic do not differ significantly from the best (at $\alpha = 0.05$).

		AUC			AUPR		
Dataset	Method	overall	unique	other	overall	unique	other
Enzyme	Baseline	0.723	0.320	0.802	0.040	0.003	0.039
	Slight bias	**0.772**	0.637	0.814	**0.041**	0.003	0.040
	Biased	0.747	0.868	0.743	0.023	0.018	0.016
Ion Channel	Baseline	0.699	0.602	0.710	*0.079*	0.010	0.075
	Slight bias	**0.701**	0.677	0.707	**0.080**	0.010	0.075
	Biased	*0.698*	0.797	0.694	0.064	0.017	0.059
GPCR	Baseline	0.766	0.562	0.819	0.094	0.012	0.088
	Slight bias	**0.782**	0.664	0.813	**0.095**	0.013	0.087
	Biased	0.750	0.747	0.747	0.062	0.025	0.047
Nuclear Receptor	Baseline	0.616	0.585	0.650	*0.140*	0.067	0.109
	Slight bias	*0.647*	0.633	0.653	**0.144**	0.070	0.109
	Biased	**0.670**	0.699	0.626	*0.126*	0.084	0.059

or targets for which there is no known interacting compound. The question then becomes which other drug compounds and proteins to include in the dataset. This possibility remains to be explored.

5 Conclusions

We have shown that popular benchmark data for the drug–target interaction problem are biased because they include only drug compounds and target proteins with at least one interaction. This bias can be quantified by looking at the CV performance on these unique interactions separately from non-unique interactions. The bias is the largest with leave-one-out cross-validation in the pairs setting. But even with 5- or 10-fold cross-validation and in the unseen drugs setting there is still a significant bias. Our analysis indicates that results of CV procedures to assess the predictive performance of methods for drug–target interaction networks should be interpreted with care because they could be possibly positively affected by bias contained in the considered datasets.

The baseline method discussed in this paper does not use the similarity information of drug compounds or target proteins at all. Hence, the performance is far below the state of the art. However, the effects of the bias carry over to other methods. For any ranking method r_{dt} we can define a variant $r_{dt}^{\text{unique}\mapsto k}$ that exploits the dataset bias and thereby boosts the performance on unique interacting pairs.

We have not performed an empirical study of the prevalence of biases in published methods. Of course none of the methods in Table 2 exploit the bias

in quite such a blatant way as our 'biased baseline' method. Still, there could be methods that inadvertently take more advantage of the bias than others, for example in the choice of parameter values or in the way they handle specific types of drug–target pairs.

In this work we have focused on a single group of datasets, with a specific type of interaction, drug–target interaction. It remains to be investigated whether other datasets for the drug–target interaction prediction problem and datasets for other similar problems have the same bias. It would also be interesting to consider other interaction datasets, such as the drug–target, enzyme–motabolite and protein–ligand datasets from [17, 3, 9, 15, 18].

References

1. Baumann, K., Rohrer, S.: Exploring benchmark dataset bias in ligand based virtual screening. Chemistry Central Journal 2(suppl. 1), P1 (2008)
2. Bleakley, K., Yamanishi, Y.: Supervised prediction of drug-target interactions using bipartite local models. Bioinformatics 25(18), 2397–2403 (2009)
3. Campillos, M., Kuhn, M., Gavin, A.-C., Jensen, L.J., Bork, P.: Drug target identification using side-effect similarity. Science 321(5886), 263–266 (2008)
4. Chen, X., Liu, M.-X., Yan, G.-Y.: Drug-target interaction prediction by random walk on the heterogeneous network. Mol. Biosyst. 8(7), 1970–1978 (2012)
5. Csermely, P., Korcsmáros, T., Kiss, H.J., London, G., Nussinov, R.: Structure and dynamics of molecular networks: A novel paradigm of drug discovery: A comprehensive review. Pharmacology & Therapeutics 138(3), 333–408 (2013)
6. Davis, J., Goadrich, M.: The relationship between Precision-Recall and ROC curves. In: ICML 2006: Proceedings of the 23rd International Conference on Machine Learning, pp. 233–240. ACM (2006)
7. DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L.: Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics 44(3), 837–845 (1988)
8. Ding, H., Takigawa, I., Mamitsuka, H., Zhu, S.: Similarity-based machine learning methods for predicting drug–target interactions: a brief review. Briefings in Bioinformatics (2013)
9. Faulon, J.-L., Misra, M., Martin, S., Sale, K., Sapra, R.: Genome scale enzyme–metabolite and drug–target interaction predictions using the signature molecular descriptor. Bioinformatics 24(2), 225–233 (2008)
10. Fawcett, T.: An introduction to ROC analysis. Pattern Recognition Letters 27(8), 861–874 (2006)
11. Gönen, M.: Predicting drug-target interactions from chemical and genomic kernels using Bayesian matrix factorization. Bioinformatics 28(18), 2304–2310 (2012)
12. Günther, S., Kuhn, M., Dunkel, M., Campillos, M., Senger, C., Petsalaki, E., Ahmed, J., Urdiales, E.G.G., Gewiess, A., Jensen, L.J.J., Schneider, R., Skoblo, R., Russell, R.B., Bourne, P.E., Bork, P., Preissner, R.: SuperTarget and Matador: resources for exploring drug-target relationships. Nucleic Acids Res. 36(Database issue), D919–D922 (2008)
13. Hopkins, A.L., Groom, C.R.: The druggable genome. Nature reviews. Drug Discovery 1(9), 727–730 (2002)
14. Isaksson, A., Wallman, M., Göransson, H., Gustafsson, M.G.: Cross-validation and bootstrapping are unreliable in small sample classification. Pattern Recognition Letters 29(14), 1960–1965 (2008)

15. Jacob, L., Hoffmann, B., Stoven, B., Vert, J.-P.: Virtual screening of GPCRs: an in silico chemogenomics approach. BMC Bioinformatics 9, 363 (2008)
16. Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., Hirakawa, M.: From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. 34(Database issue), D354–D357 (2006)
17. Keiser, M.J., Roth, B.L., Armbruster, B.N., Ernsberger, P., Irwin, J.J., Shoichet, B.K.: Relating protein pharmacology by ligand chemistry. Nat. Biotechnol. 25(2), 197–206 (2007)
18. Keiser, M.J., Setola, V., Irwin, J.J., Laggner, C., Abbas, A.I., Hufeisen, S.J., Jensen, N.H., Kuijer, M.B., Matos, R.C., Tran, T.B., Whaley, R., Glennon, R.A., Hert, J., Thomas, K.L., Edwards, D.D., Shoichet, B.K., Roth, B.L.: Predicting new molecular targets for known drugs. Nature 462(7270), 175–181 (2009)
19. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, IJCAI 1995, vol. 2, pp. 1137–1143. Morgan Kaufmann Publishers Inc., Montreal (1995)
20. van Laarhoven, T., Marchiori, E.: Predicting Drug-Target Interactions for New Drug Compounds Using a Weighted Nearest Neighbor Profile. PLoS One 8(6), e66952 (2013)
21. van Laarhoven, T., Nabuurs, S.B., Marchiori, E.: Gaussian interaction profile kernels for predicting drug–target interaction. Bioinformatics 27(21), 3036–3043 (2011)
22. Mei, J.-P., Kwoh, C.-K., Yang, P., Li, X., Zheng, J.: Drug-target interaction prediction by learning from local information and neighbors. Bioinformatics 29(2), 238–245 (2013)
23. Okuno, Y., Tamon, A., Yabuuchi, H., Niijima, S., Minowa, Y., Tonomura, K., Kunimoto, R., Feng, C.: GLIDA: GPCR ligand database for chemical genomics drug discovery database and tools update. Nucleic Acids Research 36(suppl. 1), D907–D912 (2008)
24. Overington, J.: ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). Interview by Wendy A. Warr. Journal of Computer-Aided Molecular Design 23(4), 195–198 (2009)
25. Rao, R.B., Fung, G.: On the Dangers of Cross-Validation. An Experimental Evaluation. In: SDM, pp. 588–596. SIAM (2008)
26. Schomburg, I., Chang, A., Ebeling, C., Gremse, M., Heldt, C., Huhn, G., Schomburg, D.: BRENDA, the enzyme database: updates and major new developments. Nucleic Acids Res. 32(suppl. 1), D431–D433 (2004)
27. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, pp. 1521–1528. IEEE Computer Society, Washington, DC (2011)
28. Wassermann, A.M., Geppert, H., Bajorath, J.: Ligand prediction for orphan targets using support vector machines and various target-ligand kernels is dominated by nearest neighbor effects. J. Chem. Inf. Model 49, 2155–2167 (2009)
29. Wishart, D.S., Knox, C., Guo, A.C.C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B., Hassanali, M.: DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res. 36(Database issue), D901–D906 (2008)
30. Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W., Kanehisa, M.: Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. Bioinformatics 24, i232–i240 (2008)
31. Yamanishi, Y., Kotera, M., Kanehisa, M., Goto, S.: Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. Bioinformatics 26(12), i246–i254 (2010)