Journal of Molecular Graphics and Modelling 28 (2010) 420–426

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

PubChem BioAssays as a data source for predictive models

Bin Chen, David J. Wild *

Indiana University School of Informatics, 901 East Tenth Street, Bloomington, IN 47408, United States

ARTICLE INFO

Article history:
Received 13 July 2009
Received in revised form 1 October 2009
Accepted 2 October 2009
Available online 12 October 2009

Keywords:
Predictive models
PubChem
BioAssay
Cheminformatics
Bayesian

ABSTRACT

Predictive models are widely used in computer-aided drug discovery, particularly for identifying potentially biologically active molecules based on training sets of compounds with known activity or inactivity. The use of these models (amongst others) has enabled "virtual screens" to be used to identify compounds in large data sets that are predicted to be active, and which would thus be good candidates for experimental testing. The PubChem BioAssay database contains an increasing amount of experimental data from biological screens that has the potential to be used to train predictive models for a wide range of assays and targets, yet there has been little work carried out on using this data to build models. In this paper, we take an initial look at this by investigating the quality of naive Bayesian predictive models built using BioAssay data, and find that overall the predictive quality of such models is good, indicating that they could have utility in virtual screening.

1. Introduction

PubChem [1] is a public repository of chemical information including structures of small molecules and various molecular properties. It is administered as part of the NIH Molecular Libraries Initiative (MLI) [2]. At the time of writing, the database contained structure and property information on over 40 million compounds from a variety of sources including chemical vendors, assay providers, journals, and the NIH themselves, in the *Substance* and *Compound* databases. A third database, the *BioAssay* database, contains experimental results for some of the compounds in PubChem that have been tested in MLI screening centers or elsewhere for activity against particular biological targets. Only recently have significant numbers of assay results been submitted to this database. As of July 2008, the repository contained the results of 1133 biological assays, with 662,908 compounds tested, of which 139,326 compounds have a positive result for at least one bioassay. The MLI has created several high HTS centers collectively called the Molecular Libraries Screening Centers Network (MLSCN), whose goal is to test large compound collections in a variety of investigator-defined biological assays. The chemical structures of the tested compound and the assay results are routinely deposited in PubChem and thus are made publicly available. While there is no general rule or guideline on what biological system or target is assayed, the broad goal of this initiative is to identify chemical probes of biological function. Because of the emergence of these screening centers, the amount of bioassay data deposited in PubChem is expected to increase significantly in the coming years. The PubChem BioAssay collection thus contains an increasingly rich body of information that has the potential to be computationally analyzed to reveal relationships between chemical structure and various biological activities. In particular, if accurate predictive models can be built using the experimental data already in the database for particular biological targets, virtual screens may be carried out using other PubChem compounds (or compounds from elsewhere) to indicate other molecules that could be active against the target. This procedure is termed *imputation* and has been extensively studied in various fields such as genetics [3], microarray analysis [4], and proteomics [5]. A wide variety of methods are available for imputation including nearest neighbor methods [6,7], least squares [8], random forest [9] and the use of external meta-data [10]. For a more in depth review of imputation the reader is referred to Zhang [11]. Of course, imputation is only useful if the method of prediction is good.

However, there are potential problems with this methodology. The experimental technique used to generate much of the BioAssay data, *high throughput screening* (*HTS*), is known to be prone to errors [12]. A recent study carried out at NIH on a small set of bioassay results in PubChem did indicate however that predictive models created on this data perform well [13–16]. For this study, we employed Bayesian modeling [17] as it performs well in high levels of noise [18,19], although it does assume that the descriptors used to build the model are independent. Bayesian modeling has been widely applied for this kind of predictive modeling: for example, classification of kinase inhibitors [20], use of a chemogenomic

* Corresponding author. Tel.: +1 812 856 1848.
E-mail address: djwild@indiana.edu (D.J. Wild).

1093-3263/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2009.10.001

database to predict biological targets [21], and predictive pharmacology models using large-scale SAR data to fill the pharmacological space [22].

For this work, we employed the naive Bayes component in the Pipeline Pilot package. Workflows and pipelining tools have become widely used for a variety of computational tasks that require the solution of complex problems by coupling computational building blocks. In drug discovery and cheminformatics, Pipeline Pilot is one of the most widely used tools, and has been applied to a variety of tasks including data analysis, QSAR, lead optimization and clustering [23]. In this work, we have investigated the use of such Bayesian predictive models in building predictive models for assays that have been submitted to the PubChem bioassay collection, using compounds with submitted actual bioassay results for training and validation.

2. Experimental method

2.1. Data preparation and access

PubChem provides a variety of tools to access its compound and Bioassay data including online search, FTP access, and direct automated access to the data through the Entrez Utilities [24]. This latter method is especially suited to use in workflow tools as one can create components to read the data directly without manual intervention. In order to maximize the flexibility of this approach, we created a local web service interface to the Entrez utilities. Web services are an emerging way of integrating data sources and software that we have previously employed for deployment of data and computation [25] and specifically of predictive models [26]. Our interface consists of local JAVA classes mapping to the *EFetch*, *ELink*, *ESearch*, *ESummary*, *ESpell*, *EInfo* and *EGQuery* functions in Entrez. These classes plus some other internal Java classes are used to create various new functions such as conversion of PubChem identifiers (CIDs) to SMILES notation, retrieval of active compounds and inactive compounds tested in a given Bioassay, and identification of the bioassays in which a given compound is active or inactive. The web services that we created are deployed in the Tomcat web server using Axis2.

2.2. Bayesian model building

The Bayesian method is based on Bayes rule that in its general form states that

$$P(A|B) = P(B|A)\frac{P(A)}{P(B)}$$

where, $P(A|B)$ is the probability of A given B, $P(A)$ and $P(B)$ are the probabilities of A and B and $P(B|A)$ is the probability of B given A. In the context of predictive modeling, A generally represents an activity classified (e.g. active or inactive) and B will represent one of a number of molecular features. Separate probabilities are created for each descriptor. The method, like many others, assumes that the molecular features are not correlated: though this is generally not the case, the models derived from this assumption work surprisingly well in real world scenarios. The actual implementation of Naive Bayes in Pipeline Pilot employs the Laplacian correction [20] which addresses errors arising from sampling and is described in detail in Nidhi et al. [21]. For a more detailed discussion of Naive Bayes, the reader is referred to Hastie et al. [27].

The Naive Bayes method has three features which make it suitable for the current study [20]. First, it is efficient and can be applied to large datasets. Furthermore the speed of the method is independent of the number of descriptors employed. Second, it is fairly robust to irrelevant features. Third, it weights descriptors by assigning greater significance to those that discriminate categories—thus it does not need preselection of relevant descriptors. For our work, we decided to use descriptors that encoded structural features as well as common molecular properties. Specifically, we selected the Molecular_Weight, ALogP, number of hydrogen bond acceptors and donors, the number of rotatable bonds and the FCFP_6 circular substructural fingerprints [28].

Creation of the Bayesian models in Pipeline Pilot is straightforward, and requires only specification of the descriptors for compounds, a training set where the outcome is known for the compounds and is used in building the model, and a validation set where the outcome is known but is not presented using model building. Fig. 1 shows a workflow for building a Bayesian model. Since the Bayesian model component provided by PP has been

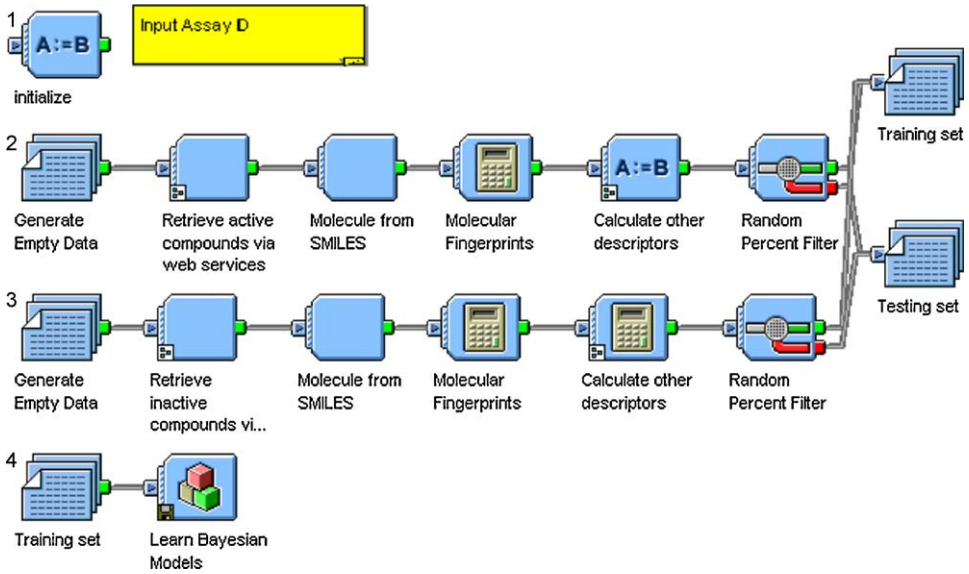

Fig. 1. Bayesian model building. Active/inactive compounds are read to PP and 80% of them are used as training samples, the rest are used as testing samples. Learning Bayesian models components are provided by PP. The number of active compounds is as much as the number of inactive compounds. The ratio of training set and testing set is 4:1. Descriptors are FCFP_6, Molecular_Weight, ALogP, Num_H_Acceptors, Num_H_Donors, Num_RotatableBond.

shown previously to be good for building predictive models of screening data [20], we used this component directly.

2.3. Model validation

To assess the models, we employed internal validation primarily using Leave-One-Out cross-validation (LOO) and external validation using rational division of a dataset into training set and testing set for all the models [29]. To facilitate the external evaluation of models, we retained 20% of the bioassay results to create a validation set. For both internal training and external validation, we generated *ROC Scores* (the area under the curve of true positive rate versus false positive rate, known as the ROC curve [30]), we shall refer to these as ROCT and ROCV, respectively. All the models were investigated by analyzing assay description as well as training set properties. Measures were also used to evaluate the diversity of training set. The relation between diversity and accuracy was examined.

In order to further study the robustness of the good models, we applied Y-randomization [29] as another validation method. In Y-randomization, the dependent variable vector (activity outcome) is randomly shuffled and a new predictive model is built using the original independent-variable matrix. For each tested assay, we repeated the experiments for five times and compared the models with the original one. We also built models using a 2:8 split for training set and testing set (20% of samples assigned to the training set, the remaining 80% assigned to the test set) and compared the results with that with a 8:2 split. In addition, we also looked at whether there was a correlation between the accuracy of the model, and the size of the training set used to build the model or the ratio of active to inactive compounds. To perform this latter study, we varied the number of active and inactive compounds used to build the models, and investigated the effect on accuracy.

As a final step, we compared the accuracy of different classes of experiment, to see, for example, if models based on primary screening data were less accurate than those built using follow-up assays. The classes employed are: NCI human tumor cell line; NCI anticancer drug screening (presumed accurate experiments originally taken from the NCI Tumor Cell Line Database); primary screening (the results from High Throughput Screening runs); follow-up screening (including secondary screening, dose response screening and confirmatory screening); and "Unclear Screening" (screens for which a primary or follow-up classification cannot be determined). A further classification was models built from five standard QSAR datasets external to PubChem [31] ("Other QSAR") to use as a benchmark. The results from these tests led us to carry

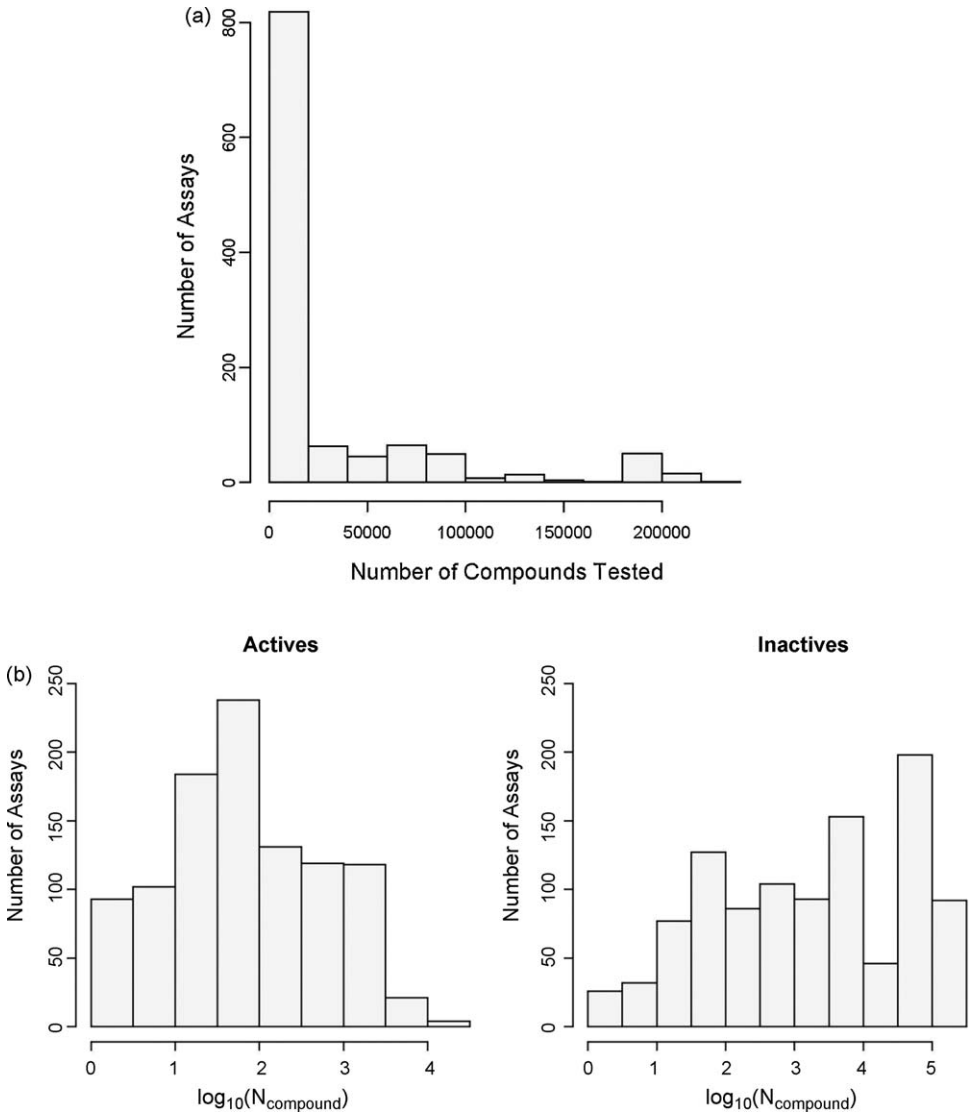

Fig. 2. (a) Distribution of assay size (including actives and inactives) and (b) numbers of actives and inactives in the PubChem bioassay collection.

out further experiments investigating the diversity of these sets in order to account for the differences in accuracy.

2.4. Use of models for virtual screening

One of our intended applications of the models is virtual screening. Other than good overall classification ability (as indicated by a high ROC score), a good model for virtual screening requires the ability to identify the potential hits as high in the ranked list of tested compounds as possible. We therefore measured the *enrichment factor* (EF) of some of our models. This measure shows how many active compounds were found in the top given percentage of the list of compounds ranked by a predictive model, relative to what would be returned at random. In our case, we use a 10% cutoff, and the scoring is the probability of activity output by the Bayesian models. To create a "blind test" for our models, the drugs known to interact with the target were retrieved from DrugBank [32] to construct the active set of a library, in which the NCI diversity set [33] composed of 1364 compounds was used as "decoy" inactives. The compounds in the library were all screened and ranked by the probability score assigned by the predictive model.

3. Results and discussion

3.1. Data collection

We retrieved 1133 bioassays from the bioassay collection as it stood in July 2008. Fig. 2 shows the distribution of number of compounds tested, number of actives and number of inactives, in this bioassay collection. 478 bioassays had less than 20 active or less than 20 inactive compounds. Given that the size of the whole dataset as well as the size of the individual classes within a dataset can play an important factor in the reliability and predictivity of a model, we excluded these, leaving 655 bioassays which contained more than 20 actives and 20 inactives. However, a bioassay with few inactive compounds can be made to produce an accurate model by introducing other random compounds into training sample as presumed inactives: this will be discussed later. Prior to modeling the structures were cleaned by removing salts.

3.2. Bayesian model building

Models for all 655 bioassays were built within 2 hours on a server with Intel CoreDuo 1.6 GHz processor and 2 GB RAM. Although the Naive Bayes method is robust to noisy data [18], the overall performance is dependent on the nature of the dataset, and in particular, good results tend to be found when there are approximately equal numbers of actives and inactives in the training set. We conducted two tests of this by changing the ratio of active compounds and inactive compounds in training set. In the first case we considered an equal number of inactives and actives. If the size of the inactive set is greater than active set, we randomly removed some inactive compounds to make the size of inactive set equal to the active set. If the size of the active set was greater than the inactive set, then we randomly removed some active compounds. In the second case we used all the actives and inactives that were present in the original assay. As expected, models built using the former approach exhibit better predictive ability (Fig. 3). However, there is significant variation in the level of improvement among models, with the greatest improvement for those models that had a lower performance (for example, the ROCV of BioAssay 248 increases from 0.91 to 0.94, but the ROCV of BioAssay 372 increases from 0.80 to 0.99). Consequently, in this paper, we use the ratio 1:1 as default.

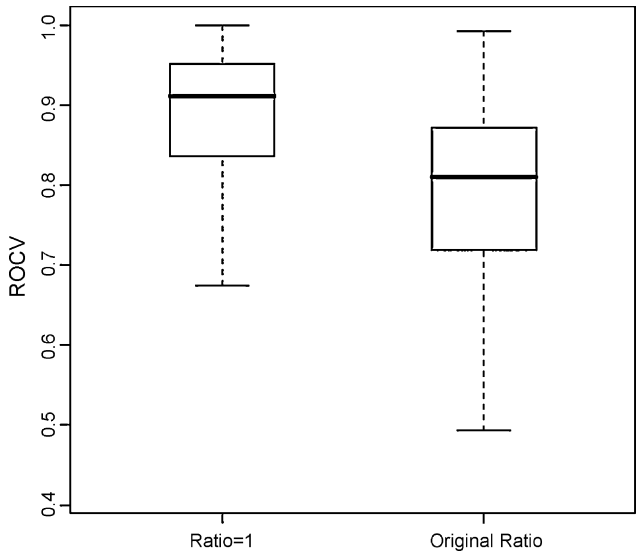

Fig. 3. Range of ROCV values for models built with a 1:1 ratio of actives to inactives ("Ratio = 1") versus the natural ratio of actives to inactives in the set ("Original ratio"). The middle line of the box is median. The whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the box. Outliers are not drawn in the plot.

In these experiments, we were concerned that removing random compounds could introduce bias to the set (i.e. that if we were removing a small number of randomly chosen compounds, their properties might skew the models). We thus replicated building models of three bioassays five times using the ratio 1:1 of active set and inactive set and 8:2 split of training set and testing set. The results and associated low standard deviation shown in Table 1 indicate that random selection does not result in bias.

Fig. 4 shows the results of our comparison of the quality of models (measured by AUC) built using different classes of experiment. The Bayesian model performs extremely well in the external QSAR sets. None of the models built using BioAssay data are as accurate as these, although the tumor cell line set and primary screening set have overall higher performance, and could be considered reasonable quality models. The accuracy of models based on follow-up screening varies greatly. The lower accuracy of models built for follow-up screens relative to primary screens is a counter-intuitive result, as one would expect the data in follow-up screens will have better results due to less errors in the training set. We hypothesized that the follow-up screening sets might have contained a less diverse set of compounds and this lack of diversity might have accounted for a lack of generalizability of the models built using the compounds. However, a comparison of mean intermolecular distances calculated using tanimoto similarity coefficient and FCFP6 as its descriptor (Fig. 5) indicates that this is not

Table 1
ROCT and ROCV scores of three BioAssays by replicating five times randomly, where the ratio of active set and inactive set is 1:1 and the ratio of training set and testing set is 8:2. STDEV is the standard deviation of results.

	Assay 1		Assay 53		Assay 823	
	ROCT	ROCV	ROCT	ROCV	ROCT	ROCV
1	0.935	0.928	0.922	0.913	0.92	0.924
2	0.93	0.933	0.923	0.907	0.924	0.924
3	0.93	0.936	0.92	0.925	0.914	0.934
4	0.932	0.927	0.92	0.927	0.932	0.934
5	0.929	0.935	0.922	0.923	0.917	0.943
STDEV	0.0024	0.0041	0.0013	0.0086	0.0070	0.0080

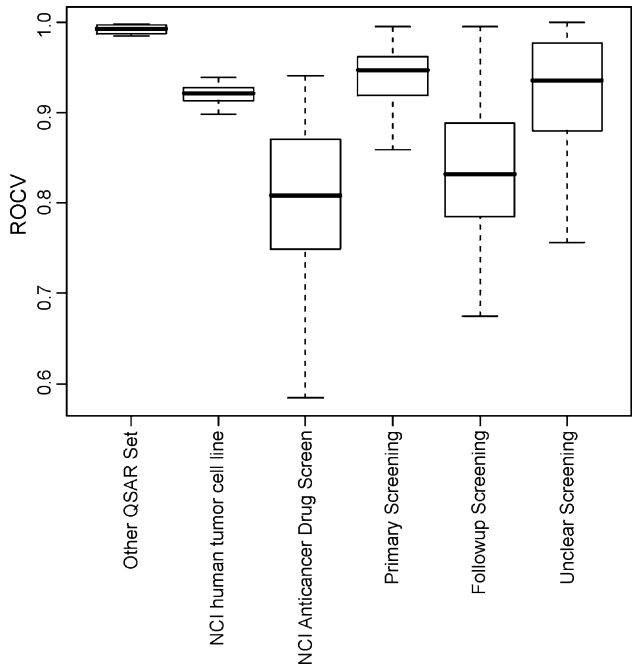

Fig. 4. Range of ROCV values from different classes of BioAssay data set. The middle line of the box is median. The whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the box. Outliers are not drawn in the plot.

the case; the diversity of the sets is broadly similar. However, we did find a difference when we looked at the mean distance between the active and inactive sets for each class (Fig. 6). In this case, there is a significantly greater distance between actives and inactives for the follow-up class than the primary class, possibly indicating that there is not enough inactive information in the follow-up sets to

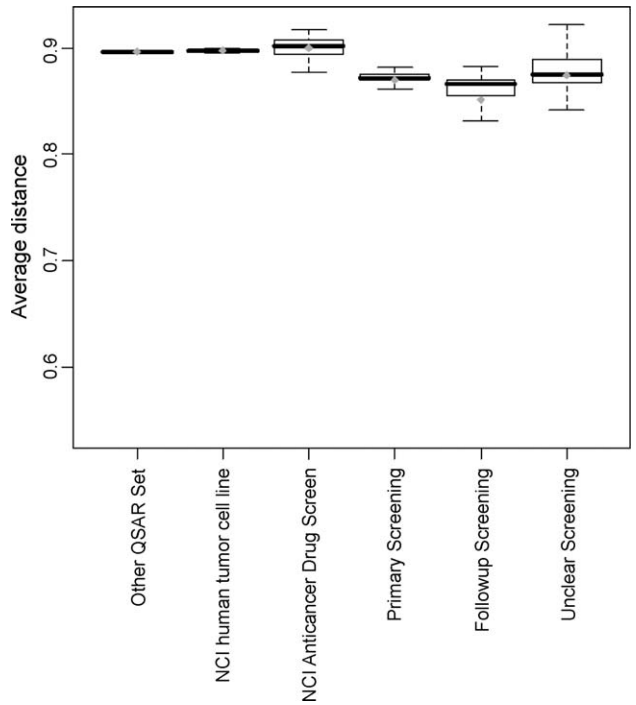

Fig. 5. Mean inter-molecular distance as a measure of diversity for different classes of BioAssay data set. The middle line of the box is median, the dot is mean. The whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the box. Outliers are not drawn in the plot.

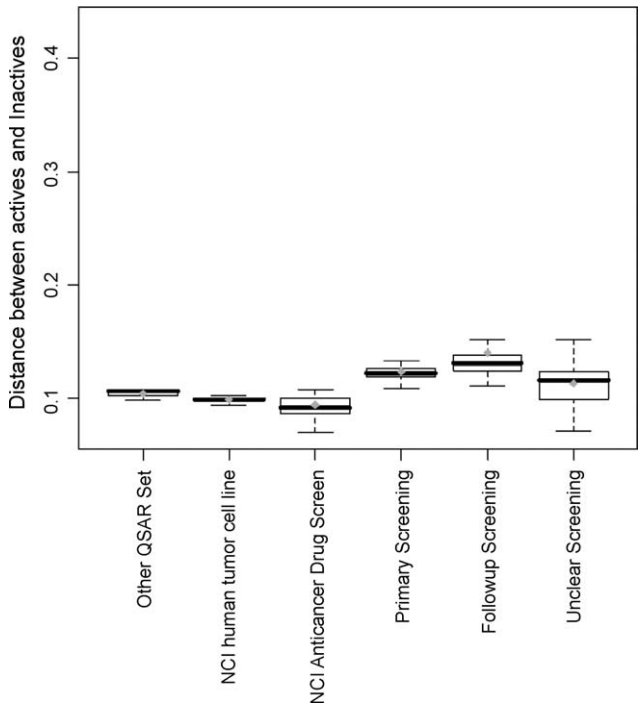

Fig. 6. Mean inter-molecular distance between compounds in the active and inactive sets for different classes of BioAssay data set. The middle line of the box is median, the dot is mean. The whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the box. Outliers are not drawn in the plot.

build a good model. We therefore built new models for three of the classes in which the inactive set was increased by adding 2422 compounds which were known to be inactive in any of the screens. This significantly improves the performance of the follow-up screen (Fig. 7).

As described earlier, bioassays with small numbers of inactive compounds relative to the actives are balanced by the addition of random compounds as inactives. Of the 48 bioassays created in this way, 44 have ROCT greater than 0.8. The overall good results show that Bayesian models can distinguish active compounds from other background compounds.

For the good models, the number of active compounds in training set ranges from 20 to 16,316. Interestingly, for these the ROCT does not vary greatly, indicating that accuracy is not related to the number of active compounds in an assay. To further investigate this, we built different models for three assays (129, 256 and 573) by decreasing the number of active compounds in training set and keeping inactive set constant. Table 2 shows that the accuracy does not significantly change with the size of active set although the imbalanced data do lead to worse results. Furthermore, when only 20% of the samples are used for training

Table 2
Models for three bioassays by built with different active set sizes.

BioAssay 129			BioAssay 256			BioAssay 573		
Active set	Inactive set	ROCT	Active set	Inactive set	ROCT	Active set	Inactive set	ROCT
172	1309	0.867	97	759	0.85	150	1151	0.921
327	1309	0.897	195	759	0.89	285	1151	0.934
460	1309	0.901	267	759	0.896	397	1151	0.941
623	1309	0.899	371	759	0.901	542	1151	0.94
795	1309	0.914	475	759	0.904	703	1151	0.937
971	1309	0.915	574	759	0.904	854	1151	0.941
1126	1309	0.918	662	759	0.904	987	1151	0.944
1309	1309	0.911	759	759	0.896	1151	1151	0.948

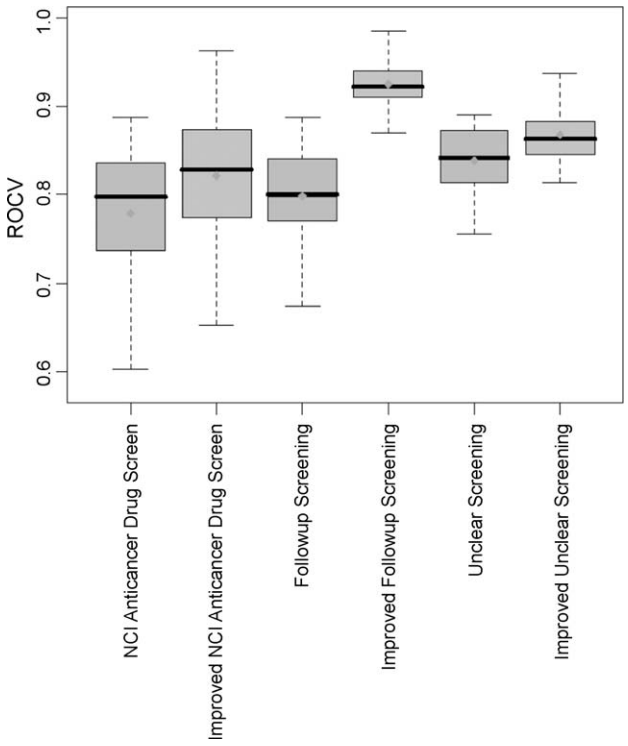

Fig. 7. Range of ROCV values from three different classes of BioAssay data set for original models and models built with additional inactive compounds ("improved"). The middle line of the box is median. The whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the box. Outliers are not drawn in the plot.

(vs. 80%), all of six assays in Fig. 8 also keep high performance although the overall ROCV score is lower than that with 8:2 split.

The robustness of good models is further demonstrated in Table 3 using Y-randomization. All the randomized ones perform much worse than the original dataset with respect to ROCV. This indicates that their good performance is neither due to the chance of correlation nor due to structural redundancy of training set.

3.3. Use of models for virtual screening

As only a small number of assays have associated target information (i.e. enzymatic assay), and not all the targets had related compounds in DrugBank, it's not possible to carry out "blind" tests on all the models. Thus we picked up two assays (ID 376, 607)

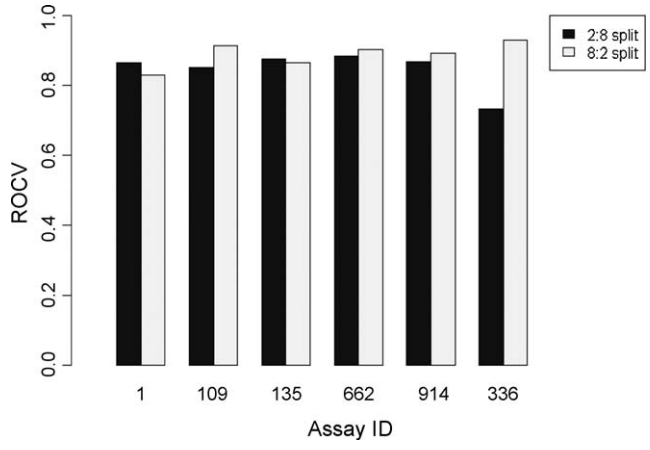

Fig. 8. ROCV scores for six bioassays using 8:2 and 2:8 split of samples.

Table 3
Y-randomization of original model versus randomized models in ROCV.

	Assay 1	Assay 53	Assay 823
Original	0.935	0.926	0.934
Y-randomization 1	0.604	0.605	0.696
Y-randomization 2	0.611	0.601	0.636
Y-randomization 3	0.608	0.602	0.654
Y-randomization 4	0.585	0.575	0.608
Y-randomization 5	0.604	0.607	0.616

with relative higher number of corresponding drugs involved for our experiments. We built the models using equal size of active set and inactive set using the PubChem BioAssay data. The ROCV of the models for 376 and 607 are 0.84 and 0.93, respectively. In bioassay 376, 14 DrugBank compounds were found that were tagged as active against the bioassay target (voltage-gated potassium channel), all of them less than 0.70 tanimoto similar (FCFP6 as fingerprint) to the active compounds in the training set. In bioassay 607, 17 DrugBank compounds were found that were tagged as active against the target (phosphodiesterase), 3 of which are greater than 0.70 similar to the training set (compounds quite similar to the training set might bring bias, thus they were excluded). All the drugs associated with NCI diversity set were screened and the top 10% results ordered by probability score were further analyzed. The EF of 376 and 607 is 3.6 and 5.7, respectively. Although the compounds in the library are not similar to the training set, high number of hits could be identified through the predictive model. This is much better than either similarity search or random selection.

4. Conclusions

We have shown that Bayesian predictive models generated using data from the PubChem database are reasonably accurate although the variability in their accuracy (ROCV = 0.582–0.995, mean 0.881) is much greater than that for models built using high quality QSAR data (ROCV = 0.985–0.998, mean 0.992). Specifically, we found that models built using less inactive compound information (as measured by the mean inter-molecular distance between the active and inactive sets) such as follow-up screens produced less accurate models than those built with more inactive information (such as primary screens). The accuracy of these former models can be greatly improved by introducing a more diverse inactive set as baseline. We suggest that PubChem BioAssay can be used as a rich source for Bayesian predictive models, but care should be taken not to expect the accuracy to be as high as for high quality experimental sets, and to make sure that models are built using a rich enough diversity of compounds to ensure a generalizable model. These models would be appropriate for virtual screening (where the main concern is to reduce the number of compounds that need to be experimentally screened), but not for prediction on small sets where very high levels of accuracy are required.

Acknowledgements

We would like to thank Dr. Rajarshi Guha for assistance in the work reported here and for commenting on drafts of the paper. We thank Accelrys for providing the Pipeline Pilot software and associated assistance.

References

[1] Pubchem. http://pubchem.ncbi.nlm.nih.gov (accessed October 14, 2008).
[2] C.P. Austin, L.S. Brady, T.R. Insel, F.S. Collins, MLI: molecular libraries initiative, Science 306 (2004) 1138–1139.
[3] Z. Yu, D.J. Schaid, Methods to impute missing genotypes for population data, Human Genetics 122 (2007) 495–504.

[4] M. Ouyang, W.J. Welsh, P. Georgopoulos, Gaussian mixture clustering and imputation of microarray data, Bioinformatics 8 (2004) 917–923.
[5] R. Pedreschi, M.L. Hertog, S.C. Carpentier, J. Lammertyn, J. Robben, J. Noben, B. Panis, R. Swennen, B.M. Nicolai, Treatment of missing values for multivariate statistical analysis of gel-based proteomics data, Proteomics 8 (2008) 1371–1383.
[6] N.L. Crookston, A.O. Finley, yaImpute: an R package for kNN imputation, J. Stat. Soft. 20 (2008) 1–16.
[7] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, R.B. Altman, Missing value estimation methods for DNA microarrays, Bioinformatics 17 (2001) 520–525.
[8] T.H. Bo, J. Dysvik, I. Jonassen, LSimpute: accurate estimation of missing values in microarray data with least squares methods, Nucl. Acids Res. 32 (2004) e34.
[9] B. Nonyane, A. Foulkes, Multiple imputation and random forests (MIRF) for unobservable, high-dimensional data, Intl. J. Biostat. 3 (2007) 1049–11049.
[10] J. Tuikkala, L. Elo, O.S. Nevalainen, T. Aittokallio, Improving missing value estimation in microarray data with gene ontology, Bioinformatics 22 (2006) 566–572.
[11] P. Zhang, Multiple imputation: theory and method, Intl. Stat. Rev. 71 (2003) 581–592.
[12] K. Babaoglu, A. Simeonov, J. Irwin, M. Nelson, B.Y. Feng, C.J. Thomas, L. Cancian, M.P. Costi, D. Maltby, A. Jadhav, J. Inglese, C.P. Austin, B.K. Shoichet, A comprehensive mechanistic analysis of hits from high-throughput and docking screens against beta-lactamase, J. Med. Chem. 51 (2008) 2502–2511.
[13] L. Han, Y. Wang, S.H. Bryant, Developing and validating predictive decision tree models from mining chemical structural fingerprints and high throughput screening data in PubChem, BMC Bioinform. 9 (2008) 401.
[14] D.C. Weis, D.P. Visco, J.L. Faulon, Data mining PubChem using a support vector machine with the signature molecular descriptor: classification of factor XIa inhibitors, J. Mol. Graph. Model. 27 (2008) 466–475.
[15] R. Guha, S.C. Schürer, Utilizing high throughput screening data for predictive toxicology models: protocols and application to MLSCN assays, J. Comput. Aided Mol. Des. 22 (2008) 367–384.
[16] H. Rao, Z. Li, X. Li, X. Ma, C. Ung, H. Li, X. Liu, Y. Chen. Identification of small molecule aggregators from large compound libraries by support vector machines. J. Comput. Chem., in press. Available online at: http://www3.interscience.wiley.com/journal/122474666/abstract.
[17] P. Labute, Binary, QSAR: a new method for the determination of quantitative structure activity relationships, Pac Symp. Biocomput. 4 (1999) 444–455.
[18] M. Glick, A.E. Klon, P. Acklin, J.W. Davies, Enrichment of extremely noisy high-throughput screening data using a naive Bayes classifier, J. Biomol. Screen. 9 (2004) 32–36.
[19] M. Glick, J.L. Jenkins, J.H. Nettles, H. Hitchings, J.W. Davies, Enrichment of high-throughput screening data with increasing levels of noise using support vector machines, recursive partitioning, and Laplacian-modified naive bayesian classifiers, J. Chem. Inf. Model. 46 (2006) 193–200.
[20] X. Xia, E.G. Maliski, P. Gallant, D. Rogers, Classification of kinase inhibitors using a Bayesian model, J. Med. Chem. 47 (2004) 4463–4470.
[21] Nidhi, M. Glick, J.W. Davies, J.L. Jenkins, Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases, J. Chem. Inf. Model. 46 (2006) 1124–1133.
[22] G.V. Paolini, R.H. Shapland, W.P. van Hoorn, J.S. Mason, A.L. Hopkins, Global mapping of pharmacological space, Nat. Biotechnol. 24 (2006) 805–815.
[23] M. Hassan, R.D. Brown, S. Varma-O'Brien, D. Rogers, Cheminformatics analysis and learning in a data pipelining environment, Mol. Divers. 10 (2006) 283–299.
[24] Entrez Utilities Help Page. http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html. Accesed November 3, 2008.
[25] X. Dong, K.E. Gilbert, R. Guha, R. Heiland, J. Kim, M.E. Pierce, G.C. Fox, D.J. Wild, Web service infrastructure for chemoinformatics, J. Chem. Inf. Model. 47 (2007) 1303–1307.
[26] R. Guha, A flexible web service infrastructure for the development and deployment of predictive models, J. Chem. Inf. Model. 48 (2008) 456–464.
[27] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer, New York, 2001.
[28] D. Rogers, R.D. Brown, M. Hahn, Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high throughput screening follow-up, J. Biomol. Screen. 7 (2005) 682–686.
[29] P. Gramatica, Principles of QSAR models validation: internal and external, QSAR & Combinatorial Sci. 26 (2007) 694–701.
[30] T. Fawcett, ROC Graphs: Notes and Practical Considerations for Researchers. HP Laboratories Technical report: Palo Alto, 2004.
[31] R.N. Jorissen, M.K. Gilson, Virtual screening of molecular databases using a support vector machine, J. Chem. Inf. Model. 45 (2005) 549–561.
[32] DrugBank Web Page. http://www.drugbank.ca (accessed November 3, 2008).
[33] Diversity Set Information. http://dtp.nci.nih.gov/branches/dscb/div2_explanation.html (accessed August 20, 2009).