Drugs and Drug-Like Compounds: Discriminating Approved Pharmaceuticals from Screening-Library Compounds

Amanda C. Schierz[1] and Ross D. King[2]

[1] Software Systems Research Group, Bournemouth University, Poole House, Talbot Campus,
Poole, BH12 5BB
[2] Computational Biology Research Group, Aberystwyth University, Penglais Campus,
Aberystwyth, SY23 3DB
aschierz@bournemouth.ac.uk, rdk@aber.ac.uk

Abstract. Compounds in drug screening-libraries should resemble pharmaceuticals. To operationally test this, we analysed the compounds in terms of known drug-like filters and developed a novel machine learning method to discriminate approved pharmaceuticals from "drug-like" compounds. This method uses both structural features and molecular properties for discrimination. The method has an estimated accuracy of 91% in discriminating between the Maybridge HitFinder library and approved pharmaceuticals, and 99% between the NATDiverse collection (from Analyticon Discovery) and approved pharmaceuticals. These results show that Lipinski's Rule of 5 for oral absorption is not sufficient to describe "drug-likeness" and be the main basis of screening-library design.

Keywords: Inductive Logic Programming, drug-likeness, machine learning, Rule of 5, compound screening library.

1 Introduction

The successful development and application of Virtual Screening methods for the drug-discovery process has provided a new area of interest to the computer community. With High-Throughput Screening (HTS) technology becoming more accessible together with several commercially-available compound screening-libraries, computer scientists have been given an opportunity to confirm their theoretical observations in *wet* laboratory experiments. The selection of the most appropriate compound screening-library to purchase for these experiments is a difficult task: there are several ready-built libraries that are commercially-available, libraries may be diversity-based or target-based and the storage and purchase of the libraries is costly. This paper reports an analysis of two commercially-available screening libraries and details an Inductive Logic Programming (ILP) discriminant analysis approach to library design: Which library most closely resembles approved pharmaceuticals?

The two main criteria for selecting compounds for screening libraries are: they are similar to existing pharmaceutically-active compounds, and they are structurally diverse. Both criteria can be interpreted as maximising the *a priori* probability that a

V. Kadirkamanathan et al. (Eds.): PRIB 2009, LNBI 5780, pp. 331–343, 2009.

compound will be found in the screening-library that is both drug-like and non-toxic. The requirement for diversity is usually explained by the fact that structurally similar compounds tend to exhibit similar activity. The goal is to find compounds that have a similar activity but have a dissimilar structure. In this way, a structurally diverse set of compounds covers the activity search space but with fewer redundant compounds [1]. Ideally, screening-library compounds should have a low molecular weight and be of low complexity in order to maximise the chance of binding to a target. These compounds should also be amenable to medicinal chemistry optimisation to increase the chance of the primary-screening *hit* being developed further and becoming a *lead* for a specific target. As several hit compounds may never be suitable as a lead compound for a target, some researchers such as Hann et al [2] claim that virtual screening methods should focus on lead-likeness and not drug-likeness. As our interest is on the primary-screening process, the focus here is on the drug-likeness (hit-likeness) of the compounds in the screening-libraries.

Drug-like properties are usually defined in terms of ADME - Absorption, Distribution, Metabolism, and Excretion - and describe the action of the drug within an organism, such as intestinal absorption or blood-brain-barrier penetration. One of the first methods, and still the most popular, to model the absorption property was the “Rule of 5” developed by Lipinski et al [3] which identifies the compounds where the probability of useful oral activity is low. The "rule of 5" states that poor absorption or permeation of a compound is more likely when:

1. There are more than 5 Hydrogen-bond donors
2. The Molecular Weight is over 500.
3. The LogP (partition coefficient) is over 5 (or MLogP is over 4.15).
4. There are more than 10 Hydrogen-bond acceptors

Though these rules were never meant to describe the drug-likeness of compounds, their negation is usually used as the main selection filter for the compounds to include in a screening-library. For example, chemical companies such as Maybridge, Chembridge, Analyticon, TimTec, amongst others, all describe their screening-libraries in terms of the number of Lipinski rules covered by the compounds. Though these rules are not definitive, the properties are simple to calculate and not only provide a guideline for good oral absorption of the compound but also for general drug-likeness of that compound.

To assess how well the compounds in the screening-libraries resemble existing pharmaceutically-active compounds, two types of analysis was carried out:

- The comparison of the compounds in the screening-libraries and the set of approved pharmaceuticals in terms of the number of Lipinksi rules covered by the compounds (Hydrogen bond donors and acceptors, molecular weight and LogP).
- Machine learning techniques have been used to discriminate between each screening-library and the set of approved pharmaceuticals. 3 decision trees per screening-library have been learnt based on a differing molecular representation – substructures only, quantitative properties only, and both substructures and quantitative properties.

This discriminatory approach is not novel and similar work has been carried out using neural networks [4], [5], [6] and decision trees [7] with relatively good prediction success for drug-likeness. In related work, the success of the Lipinski rules has encouraged research on refining and improving them. For example, Oprea [8], [9] has shown that the "Rule of 5" alone is not sufficient to distinguish between drugs and non-drugs, and proposes other quantitative filters such as rotatable bonds, rigid bonds and ring counts; Veber et al [10] claim that molecular Polar Surface Area is also important when describing drug-likeness and Baurin et al [11] include filters such as tractability and aqueous solubility, amongst others. One important way in which our approach differs from this previous work is that these methods all used the Available Chemicals Directory (ACD) as the dataset of non-drugs, and either the World Drug Index (WDI), MDL Drug Data Report (MDDR) or Medicinal Chemistry database (CMC) as the dataset for drugs (and drugs in development). In our approach, we use approved pharmaceuticals as the drug dataset and commercially-available compound screening-libraries as the non-drug dataset. This adds difficulty to the discrimination task as all the compounds in the screening-libraries are already identified as having drug-like properties.

The second significant way that our approach differs is in the representation of molecules. Almost all chemoinformatics is based around using tuples of attributes to describe molecules. An attribute is a proposition which is true or false about a molecule, for example having a Log P of 0.947, the existence of a benzene ring, etc. This representational approach typically results in a matrix where the examples are rows and the columns are attributes. This attribute-based form of data is assumed by standard statistical and machine learning analysis methods. This way of representing molecules has a number of important disadvantages. Perhaps the most important of these is that it is chemically unnatural. Chemists think of molecules as structured objects (atom/bond structures, connected molecular groups, 3D structures, etc.). Such structured objects cannot easily be represented using attributes, and therefore their use forces chemists to use a language that cannot express their most basic concepts. Another important disadvantage of the attribute-based approach is that it is computationally inefficient in terms of space, i.e. to fully capture molecular structure requires an exponential number of attributes to be created. This is the fundamental reason that it is not unusual in chemoinformatic applications to see molecules described using hundreds if not thousands of attributes.

A more natural and spatially efficient way to represent molecular structure is to use relations: atom1 *bonded to* atom2; a benzene ring *connected to* an amide group, etc. The main disadvantage of using such relational representations is that it requires more complex machine learning methods which are often slower than attribute-based approaches. One machine learning method that can use relational data is Inductive Logic Programming (ILP). The first representation was based on atoms, bonds and some quantitative attributes [12] and a more recent representation has added attributes derived from Richard Bader's Atom in Molecules (AIM) quantum topology theory [13], [14]. ILP enables the usage of background knowledge by defining high-level concepts, e.g. functional groups, aromatic rings, etc and the output of an ILP method is rich, relational rules such as "A compound is active if it has an aliphatic carbon atom attached by a single bond to a nitrogen atom which is in a six-membered aromatic monocycle".

2 Materials and Methods

2.1 Data Sets

Two compound-screening libraries were chosen for the research – the target-based NatDiverse collection from Analyticon Discovery (Version 070914) and the diversity-based HitFinder (Version 5) collection from Maybridge. The libraries from these companies are publicly available and therefore computational analysis could be carried out: this was the sole reason for their inclusion in this research. We would like to thank Analyticon Discovery and Maybridge for their data.

The HitFinder collection includes 14,400 compounds representing the drug-like diversity of the Maybridge Screening Collection (approximately 60,000 compounds). Compounds have generally been selected for inclusion in the library if they are known to be non-reactive and meeting 2 or more of Lipinski's Rule of 5 (www.maybridge.com). AnalytiCon Discovery (www.ac-discovery.com) currently offers 13 NatDiverse libraries which are tailor-made synthetic nitrogen-containing compounds. The libraries are template / target-based and include collections containing quinic acid and shikimic acid, hydroxyproline, santonine, dianhydro-D-glucitol, hydroxypipecolinic acid, andrographolide, piperazine-2-carboxylic acid, cytosine, quinidine, quinine, indoloquinolizidine, cyclopentene and ribose. The total number of compounds is 17,402.

The approved pharmaceuticals dataset was obtained from the KEGG Drug database and contains 5,294 approved drugs from the United States and Japan. The compounds were not filtered to remove reactive functionalities [8] or any other undesirable properties. The datasets were randomly split into a training and validation dataset and an independent test set. 20% of the compound libraries and 8% of approved pharmaceuticals were used for the independent testing**.**

2.2 Molecular Descriptors

The software PowerMV [15] was used to generate the molecular properties for the compounds. The four properties associated with Lipinski's Rule of 5 - Molecular weight, LogP, hydrogen bond acceptors, and hydrogen bond donors were calculated, together with the number of rotatable bonds, polar surface area, blood-brain indicator (if the compound penetrates the brain or not) and the number of chemically reactive or toxic groups in the compound.

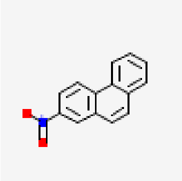
2.3 Data Preprocessing

The OpenBabel suite [16] was used to convert the SDF datasets to the MOL2 chemical format so that the aromatic bonds could be identified and hydrogens added. A text-processing script parsed the MOL2 file into a Prolog-readable format containing data on atoms, bonds and aromaticity. The data is fully normalised according to relational database design standards [17] so each compound and atom are assigned a unique identifier. For example, *atom(2,4,c)* means that atom number 4 in compound number 2 is a carbon; *bond(2,4,5,2)* means that in compound number 2, atoms 4 and 5 are bonded by a double bond (the final digit 2).

2.4 Molecular Structure Generator

A bespoke Molecular Structure Generator (MSG) program, written in Prolog, uses this atom and bond information to generate descriptions of substructures by referring to a pre-coded library of over 200 chemical rings, functional groups, isomers and analogues. Figure 1 shows a fragment of the normalised relational data representation generated for the illustrated compound. The numbers represent the unique identifiers, for example,

ring(compound_id, structure_id, ring_name),
ring_length(compound_id,structure_id,ring_length).



ring_length(1,1,6).	fused_pair_name(1,4,naphthalene).	poly_no_rings(1,6,3).	nextto(1,1,2,fused).
aromatic_ring(1,1).	carbon_fused_pair(1,4).	group(1,7,nitro).	nextto(1,6,7,bonded).
carbon_ring(1,1).	polycycle(1,6,phenanthrene)	group(1,8,aryl_nitro).	count_ring(1,benzene,3).
ring(1,1,benzene).	carbon_poly(1,6).	parent(1,8,nitro).	

Fig. 1. A fragment of the background knowledge generated for 2-nitrophenanthrene using the Prolog Molecular Structure Generator. Image from Pubchem.

The relational facts can be read as, for example,

- For compound number 1, the first substructure identified is a benzene ring of length 6. It is a carbon ring and it is aromatic.
- For compound number 1, the fourth substructure identified is naphthalene which is a fused pair of rings and is only carbon.
- For compound number 1, the eighth substructure identified is an aryl-nitro which is a type of (has parent) nitro
- For compound number 1, the sixth substructure (phenanthrene) is bonded to the seventh substructure (nitro)

2.5 Decision Trees

The data mining software Tilde [18] is available as part of the ACE data mining system (http://www.cs.kuleuven.ac.be/~dtai/ACE/) which provides a common interface to several relational data mining algorithms. Tilde is an upgrade of the popular C4.5 decision tree learner [19] and can be used for relational data mining: facts represented in Prolog can be both the input and the output of Tilde. For all experiments, the minimal cases allowed for a tree node was set to 5, the search heuristic employed was *gain*, and the Tilde mode was set to *classify*. All other options were kept as the default values. The complete datasets were split into a training and validation set and an independent test set. A ten-fold cross-validation was used for Tilde to learn the decision trees. A cross-validation is a standard statistical technique where the training and validation data set is split into several parts of equal size, in this case 10% of the compounds. For each run of Tilde, 10% of the data is excluded from the training set and put in a corresponding validation set. Each training set is used to construct a classification tree which is then used to make predictions for the corresponding validation set. For each of the three scenarios (structural information only, quantitative information only, and both structural and quantitative information), the ten-fold cross-validation was carried out with identical training and validation sets. The classification tree that performed the best in the training and validation stage was then applied to the independent test set.

3 Results

3.1 Lipinski Attribute Analysis

The datasets were first analysed according to the Lipinski Rule of 5. This analysis was carried out to see how well the two commercially-available screening-libraries matched the set of approved pharmaceuticals in terms of the Lipinksi rule properties (Hydrogen bond donors and acceptors, molecular weight and LogP). Each combination of the rules has been allocated an identifier tag as in Table 1. For example, Lip4 is compounds that have all the Lipinski drug-like properties; Lip2b is compounds that have 2 Lipinski drug-like properties (Less than or equal to 5 hydrogen bond donors and a molecular weight less than or equal to 500).

Each compound in the two screening-libraries and the set of approved pharmaceuticals was allocated a tag according to the Lipinski Rule combinations shown in Table 1. Table 2 shows the percentages of compounds from each dataset for each identifier tag.

Table 1. Identifier tags for the combination of Lipinski Rules

Lipinski Rule ID	H-bond donors ≤5	Mol. Weight ≤500	LogP ≤5	H-bond acceptors ≤ 5
Lip4	✓	✓	✓	✓
Lip3a	✓	✓		✓
Lip3b	✓		✓	✓
Lip3c	✓	✓	✓	
Lip3d		✓	✓	✓
Lip2a	✓			✓
Lip2b	✓	✓		
Lip2c	✓		✓	
Lip2d		✓		✓
Lip2e			✓	✓
Lip2f		✓	✓	
Lip1a	✓			
Lip1b				✓
Lip1c		✓		
Lip1d			✓	
Lip0				

Table 2. Percentage of compounds with each combination of Lipinski Rules in the compound screening-libraries and approved pharmaceuticals (App)

Lipinski Rule ID	NATDiverse	HitFinder	App
Lip4	82.3%	88.9%	74.29%
Lip3a	2.67%	10.85%	8.05%
Lip3b	1.08%	0.02%	1.19%
Lip3c	1.26%	0.02%	2.26%
Lip3d	9.85%	0.01%	3.53%
Lip2a	0.51%	0.17%	1.19%
Lip2b	0	0.02%	0.02%
Lip2c	1.47%	0	3.53%
Lip2d	0	0	0
Lip2e	0.12%	0	0.02%
Lip2f	0.36%	0	1.10%
Lip1a	0.03%	0	0.36%
Lip1b	0	0	0.09%
Lip1c	0	0	0
Lip1d	0.33%	0	3.82%
Lip0	0.02%	0	0.15%

The majority of the compounds in all datasets meet at least 3 of Lipinksi's 4 drug-like properties. The most diverse combinations are in the set of approved pharmaceuticals with just over 10% of compounds meeting 2 or less of the Rule of 5 properties. Interestingly, nearly 4% of approved pharmaceuticals only meet the LogP filter. The HitFinder diversity-library has the least diverse coverage with 0.19% of compounds having 2 or less combinations. According to this attribute-based analysis, the NATDiverse targeted-library is more closely matched to the set of approved pharmaceuticals dataset than the HitFinder library in terms of the Lipinski drug-like properties. Interestingly, no dataset has a compound that just satisfies the molecular weight and hydrogen bond acceptor criteria (Lip2d) or just the molecular weight criteria (Lip1c). Essentially this tells us that if the compound violates rules on LogP and hydrogen bonding it doesn't matter what the molecular weight is, the compound is not likely to be a potential drug.

3.2 Discrimination Analysis

Three tests were carried out per dataset pairing (screening-library : approved pharmaceuticals) – one based on structural information only using the relations generated by the MSG Prolog program, another on quantitative attributes only (Molecular weight, LogP, hydrogen bond acceptors, hydrogen bond donors, the number of rotatable bonds, polar surface area, blood-brain indicator and the number of chemically reactive or toxic groups in the compound), and the third based on both structural information and the quantitative attributes. Please note that as the datasets are of uneven size (approximately 3:1, screening-library: approved pharmaceuticals), we have shown the results in terms of True Positives (approved pharmaceuticals correctly classified as such) and False Positives (screening-library compounds that have been incorrectly classified as approved pharmaceuticals). Table 3 shows the average accuracy of the 10 classification models when applied to the validation set together with the size of the most accurate decision tree produced.

Table 3. Average accuracy of the classification trees when applied to the validation set. For each screening-library, the results of the 3 data representation results are shown.

Validation Dataset	Accuracy	False Positives	True Positives	Tree size
HitFinder/App structures only	87.68%	7%	75%	367
NATDiverse/App structures only	98.62%	1%	96%	119
HitFinder/App properties only	83.53%	8%	64%	423
NATDiverse/App properties only	90.31%	5%	76%	348
HitFinder/App structures & properties	88.29%	7%	78%	389
NATDiverse/App structures & properties	97.75%	1%	95%	138

The results of the cross-validation are promising with high accuracy figures. The classification system has had more difficulty discriminating the approved pharmaceuticals from the HitFinder library than the NATDiverse library – this has resulted in larger decision trees with lower accuracy rates for the HitFinder library. The best result for the HitFinder / Approved Pharmacaeuticals data has been achieved when the data is represented by both structures and quantitative properties; the least accurate is when the data is represented by quantitative properties only. For the NATDiverse / Approved Pharmaceuticals data the best result is achieved by representing the data by structural information only and the least accurate result is when the data is represented by quantitative properties only. As the datasets are of uneven distribution, the ROC (Receiver Operating Characteristics) points which illustrate the trade-off between the hit-rate and false-alarm rate have been shown in Figure 2.

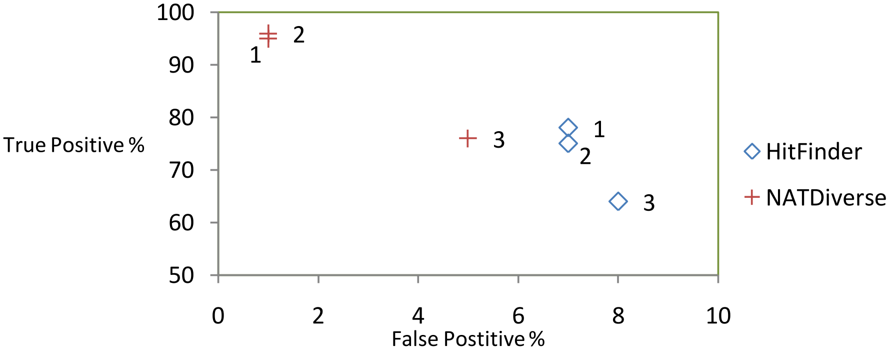


Fig. 2. The ROC points of the classifiers when applied to the validation data. The numbers are the data representation: 1 is structural and quantitative, 2 is structural only and 3 is quantitative only.

For each scenario, the classification tree that provided the lowest True Positive : False Positive ratio was applied to the independent test set, see Table 4.

Table 4. Accuracy of the best classification tree when applied to the independent test set. For each screening-library, the results of the 3 data representation results are shown.

Testing Dataset	Accuracy	False Positives	True Positives
HitFinder / App structures only	89.53%	8%	74%
NATDiverse / App structures only	99.00%	1%	96%
HitFinder / App properties only	83.43%	10%	62%
NATDiverse / App properties only	89.29%	8%	74%
HitFinder / App structures & properties	90.75%	7%	75%
NATDiverse / App structures & properties	98.98%	1%	97%

The independent test results are very good and even show a slight improvement over the validation results in some scenarios. This shows us that our model has not been over-fitted to the training data. The results also show that the inclusion of quantitative attributes resulted in a slight increase in the classification accuracy for the HitFinder / Approved Pharmaceuticals data but actually decreased the overall accuracy for the NatDiverse / Approved Pharmaceutical data (though there is an increase in the True Positive rate). Figure 3 shows the ROC points of the classifier.

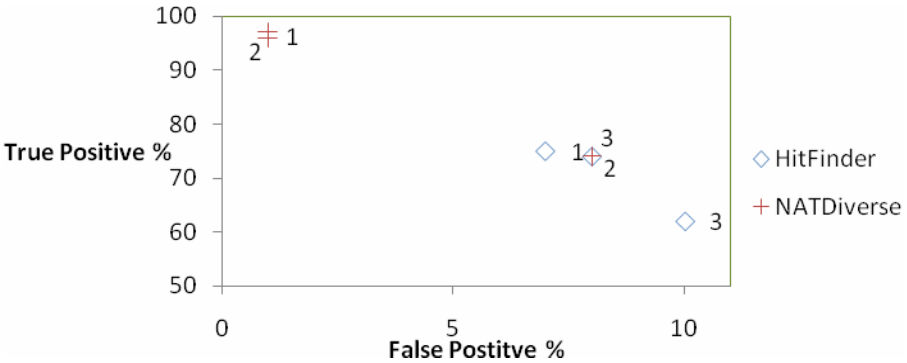


Fig. 3. The ROC points of the classifier when applied to the test data. The numbers are the data representation: 1 is structural and quantitative, 2 is structural only and 3 is quantitative only.

For both screening-libraries, there has been a decrease (5 to 10%) in performance when using physicochemical quantitative properties only. Interestingly, this may mean that even though the screening-library compounds are similar to approved pharmaceuticals in terms of certain *drug-likeness* filters, they are dissimilar in terms of certain substructures.

These results are converse to the attribute-based Lipinski rules analysis carried out previously. According to Lipinski's criteria, the target-based NATDiverse library more closely resembles approved pharmaceuticals than the diversity-based HitFinder library. Here the opposite is true – it has been harder to discriminate between the HitFinder compounds and approved pharmaceuticals. This means that the compounds in the HitFinder library resemble approved pharmaceuticals closer than the NATDiverse compounds when more molecular background knowledge is added.

3.3 Pruning the Trees

One of the advantages of using Tilde is that the decision trees may be represented as a set of Prolog rules, each of which represents a decision tree node. The most accurate rules, i.e. those with the maximum positive coverage and minimal negative coverage, were extracted to build a probabilistic decision list. The aim was to find a decision list that had a minimum overall accuracy of 85% and had less than 10 rules.

For the HitFinder / Approved Pharmaceuticals datasets, a pruned decision list of 10 rules was found that had an overall accuracy of 85% and can correctly classify 63% of approved pharmaceuticals with only 7% false positives. Table 5 shows the resulting decision list rules together with their confidence probabilities. The rules may be read as **If** the compound has a molecular weight greater than 500.502 **then** there is a 99.9% probability the compound is an approved pharmaceutical, **else if** the compound has a molecular weight smaller than 150.133 **then** there is a 99.6% probability the compound is an approved pharmaceutical, and so on.

Table 5. The ten best rules for discriminating between the HitFinder library and the set of approved pharmaceuticals. These rules can successfully classify 63% of the approved pharmaceuticals and 93% of the HitFinder compounds.

1. If molecular weight > 500.502 then approved pharmaceutical (99.9%)
2. else if molecular weight < 150.133 then approved pharmaceutical (99.6%)
3. else if there's more than 1Hydroxyl then approved pharmaceutical (93%)
4. else if there's a Sulphur-containing Aromatic Monocycle then HitFinder (91%)
5. else if there's a Thiophene then Hitfinder (89%)
6. else if there's more than 2 Methylenes then approved pharmaceutical (75%)
7. else if there's a Cyclohexane next to a cyclopentane and there's a Methyl then approved pharmaceutical (95%)
8. else if there's an Aromatic ring and an Azetine next to an Amide then approved pharmaceutical (97%)
9. else if there's a Cyclohexane next to a Methyl and molecular weight > 269.388 then approved pharmaceutical (86%)
10. else the compound is from the HitFinder library (67%)

The rules generated are simple to understand and provide insight into the structural differences between the HitFinder library and approved pharmaceuticals. Apart from molecular weight, no other physicochemical property has been employed as a discriminatory feature; this is probably due to the library being designed using these types of filters.

For the NatDiverse (NAT) / Approved Pharmaceuticals (App) datasets, a pruned decision list with just 8 rules can classify the compounds with 90% accuracy, with 84% of approved pharmaceuticals classified correctly with 8% False Positves. The rules here are longer and include more structural relations than those for the HitFinder library, see Table 6.

Table 6. The eight best rules for discriminating between the NATDiverse library and the set of approved pharmaceuticals. These rules can successfully classify 84% of the approved pharmaceuticals and 92% of the NATDiverse compounds.

1. If there's a non-aromatic ring and less than 6 Amides and a Hetero ring with length < 5 then approved pharmaceutical (100%)
2. else if there's a non-aromatic ring and less than 6 Amides and a fused-pair of Hetero rings then NATDiverse (94%)
3. else if there's a non-aromatic ring and less than 6 Amides, a Piperidine bonded to an Amide and Hydrogen Bond Donors = 1 or 2 then NATDiverse (91%)
4. else if there's a non-aromatic ring and an aromatic monocycle and a Nitrogen-containing ring and an Oxygen-containing ring and any ring with length of 5 then NATDiverse (79%)
5. else if there's a non-aromatic ring and less than 6 Amides and more than one 1H-Quiolizine then NATDiverse (100%)
6. else if there's a non-aromatic ring and less than 6 Amides and a Cyclohexane bonded to an Alcohol then NATDiverse (94%)
7. else if there's a non-aromatic ring and less than 6 Amides and Hydrogen Bond Donors > 1 then NATDiverse (62%)
8. else the compound is an approved pharmaceutical (91%)

Where as the rules for the HitFinder collection were a mixture of classifying compounds from both the library and the set of approved pharmaceuticals, here the rules seem to be focused on the library compounds – 91% of the compounds left after applying these rules will probably be approved pharmaceuticals. This is probably due to the nature of target-based screening-libraries; they are normally designed around specific molecular structures. Once again, because of the screening-library compounds being close to approved pharmaceuticals in terms of Lipinski rule filters, the rules are mainly based around differing substructures. This time it is only Hydrogen bond donors that have been found in the discriminating rules.

Employing an ILP approach to this discrimination task has produced a rich, relational and small set of rules that provide insightful information about the differences between the compounds in the screening-libraries and approved pharmaceuticals.

4 Discussion and Conclusion

This research exercise has been interesting to us for several reasons. From a technical viewpoint, the Prolog Molecular Structure Generator provided descriptive molecular background knowledge and this has resulted in some clear, easy to understand relational rules. From a screening-library compound perspective, we were surprised that the classifiers provided some very accurate results. It was expected that the HitFinder library would be harder to discriminate than the NATDiverse collection as it is diversity–based rather than target-based. However, neither task was too challenging and this leads back to the concept of *lead-likeness* and the argument that virtual screening methods should focus on lead-likeness and not drug-likeness [2].

The final interesting perspective is that of screening-library design. The properties associated with the Rule of 5 and others such as Polar Surface Area are predominantly used for the design of screening-libraries. These properties are treated as filters and do not consider a lot of compounds that are filtered out and classed as being non-drug-like. This research has shown that even though the compounds in the screening-libraries resemble approved pharmaceuticals with regard to these filters, there are a lot more factors that need to be considered. The filter approach is almost certainly non-optimal because such filters are "soft", i.e. they are only probabilistic and can be contravened under some circumstances.

We have taken a discrimination-based approach to the problem of selecting and designing compound libraries for drug screening. We have demonstrated that by using our ILP machine learning method we can accurately discriminate between approved pharmaceuticals and compounds in state-of-the-art screening-libraries with high accuracy. These discrimination functions are expressed in easy to understand rules, are relational in nature and provide useful insights into the design of a successful compound screening-library.

References

1. Leach, A.R., Gillet, V.J.: An Introduction to Chemoinformatics. Kluwer Academic Publishers, Dordrecht (2003)
2. Hann, M.M., Leach, A.R., Harper, G.: Molecular Complexity and Its Impact on the Probability of Finding Leads for Drug Discovery. Journal of Chemical Information and Computer Sciences 41(3), 856–864 (2001)
3. Lipinski, C.A., Lombardo, F., Dominy, B.W., Feeney, P.J.: Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv. Drug Delivery Rev. 23(1-3), 3–25 (1997)
4. Ajay, W., Walters, W.P., Murcko, M.A.: Can We Learn To Distinguish between "Drug-like" and "Nondrug-like" Molecules? J. Med. Chem. 41(18), 3314–3324 (1998)
5. Sadowski, J., Kubinyi, H.: A scoring scheme for discriminating between drugs and nondrugs. J. Med. Chem. 41, 3325–3329 (1998)
6. Murcia-Soler, M., Pérez-Giménez, F., García-March, F.J., Salabert-Salvador, M.T., Díaz-Villanueva, W., Castro-Bleda, M.J.: Drugs and nondrugs: an effective discrimination with topological methods and artificial neural networks. J. Chem. Inf. Comput. Sci. 43(5), 1688–1702 (2003)
7. Wagener, M., van Geerestein, V.J.: Potential drugs and nondrugs: prediction and identification of important structural features. J. Chem. Inf. Comput. Sci. 40 (2000)
8. Oprea, T.I., Davis, A.M., Teague, S.J., Leeson, P.D.: Is there a difference between leads and drugs? A historical perspective. J. Chem. Inf. Comput. Sci. 41, 1308–1315 (2001)
9. Oprea, T.I.: Lead structure searching: Are we looking at the appropriate property? J. Comput.-Aided Mol. Design 16, 325–334 (2002)
10. Veber, D.F., Johnson, S.R., Cheng, H.-Y., Smith, B.R., Ward, K.W., Kopple, K.D.: Molecular properties that influence the oral bioavailability of drug candidates. J. Med. Chem. 45, 2615–2623 (2002)
11. Baurin, N., Baker, R., Richardson, C.M., Chen, I.-J., Foloppe, N., Potter, A., Jordan, A., Roughley, S., Parratt, M.J., Greaney, P., Morley, D., Hubbard, R.E.: Drug-like Annotation and Duplicate Analysis of a 23-Supplier Chemical Database Totalling 2.7 Million Compounds. Journal of Chemical Information and Modeling 44(2), 643–651 (2004)

12. King, R.D., Muggleton, S.H., Srinivasan, A., Sternberg, M.J.E.: Structure activity relationships derived by machine learning: The use of atoms and their bond connectivities to predict mutagenicity using inductive logic programming. Proceedings of the National Academy of Sciences, USA 93, 438–442 (1996)
13. Buttingsrud, B., Ryeng, E., King, R.D., Alsberg, B.K.: Representation of molecular structure using quantum topology with inductive logic programming in structure-activity relationships. Journal of Computer-Aided Molecular Design 20(6), 361–373 (2006)
14. Bader, R.F.W.: Atoms in Molecules - A Quantum Theory. Oxford University Press, Oxford (1990)
15. Liu, K., Feng, J., Young, S.S.: PowerMV: A Software Environment for Molecular Viewing, Descriptor Generation, Data Analysis and Hit Evaluation. J. Chem. Inf. Model. 45, 515–522 (2005)
16. Guha, R., Howard, M.T., Hutchison, G.R., Murray-Rust, P., Rzepa, H., Steinbeck, C., Wegner, J.K., Willighagen, E.: The Blue Obelisk – Interoperability in Chemical Informatics. J. Chem. Inf. Model. 46(3), 991–998 (2006)
17. Codd, E.F.: Recent Investigations into Relational Data Base Systems. IBM Research Report RJ1385 (April 23, 1974); republished in Proc. 1974 Congress, Stockholm, Sweden. North-Holland, New York (1974)
18. Blockeel, H., De Raedt, L.: Top-down induction of first order logical decision trees. Artificial Intelligence 101(1-2), 285–297 (1998)
19. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann series in Machine Learning. Morgan Kaufmann, San Francisco (1993)