SAR and QSAR in Environmental Research, 2013
Vol. 24, No. 4, 585–597, http://dx.doi.org/10.1080/1062936X.2013.773373

Drugs, non-drugs, and disease category specificity: organ effects by ligand pharmacology

A.T. García-Sosa* and U. Maran

Institute of Chemistry, University of Tartu, Tartu, Estonia

(*Received 12 June 2012; in final form 6 September 2012*)

Important understanding can be gained from using molecular biology-based and chemistry-based techniques together. Bayesian classifiers have thus been developed in the present work using several statistically significant molecular properties of compiled datasets of drugs and non-drugs, including their disease category or organ. The results show they provide a useful classification and simplicity of several different ligand efficiencies and molecular properties. Early recall of drugs among non-drugs using the classifiers as a ranking tool is also provided. As the chemical space of compounds is addressed together with their anatomical characterization, chemical libraries can be improved to select for specific organ or disease. Eventually, by including even finer detail, the method may help in designing libraries with specific pharmacological or toxicological target chemical space. Alternatively, a lack of statistically significant differences in property density distributions may help in further describing compounds with possibility of activity on several organs or disease groups, and given their very similar or considerably overlapping chemical space, therefore wanted or unwanted side-effects. The overlaps between densities for several properties of organs or disease categories were calculated by integrating the area under the curves where they intersect. The naïve Bayesian classifiers are readily built, fast to score, and easily interpretable.

Keywords: Bayesian; drug design; organ; specificity; toxicity; chemical library

1. Introduction

Lately, there have been indications that important conclusions on selectivity and affinity to disease categories (DC) or organs can be gained from describing chemical compounds based on their physicochemical space, in addition to their target or biomolecular space [1]. The ligand space available for specific disease categories or organs, as described in the highest level of Anatomical Therapeutic Chemical Classification (ATC, provided by the World Health Organization [2]), may provide clues as to whether specific ligand molecular properties can imply DC or organ activity and/or specificity.

Molecular biology-based techniques look at the similarity of sequence or structure between protein targets, and ascribe function. Chemistry-based techniques see the target space from the ligand point of view, considering structure as well as molecular physicochemical properties. Combining both world views can achieve important benefits to decipher the relationships between chemical compounds, biomolecules, and organs and systems (finally

*Corresponding author. Email: alfonsog@ut.ee

affecting an organism). Examples are the different localization of receptors in different tissues and organs such as the different histamine receptors: H_2 acts on the gastric system while H_3 acts in the central nervous system.

Introducing specificity for each subtype of receptor is crucial in achieving the desired therapeutic effect without side-effects. Another factor to consider is the activation of very different and unrelated proteins by similar ligands, i.e. a small modification in ligand structure can imply a large biological effect; as well as the effect that small modification in ligand, protein, or associated water structure can have even on the same protein [3]. Another factor to consider is either the shared or different mechanisms of action and application routes of drugs [4].

A recent important development is the realization that breast cancer alone consists of 10 different diseases according to their different gene expression [5]. This shows how organ compartimentalization according to ligand chemical space also needs to account for individual target considerations (i.e. a disease in the same organ can have different molecular pathways and molecular targets).

Recent work using drug and non-drug compiled datasets has provided information on the molecular chemical space available to different groups of compounds, as well as their drug ratio and sensitivity using probability density functions [6,7]. Their classification into drugs and non-drugs has also been achieved using logistic functions [8]. In addition, compounds with similar side-effects have been linked to similar activities against a variety of biological targets [9]. Therefore, use of biological and chemical data can provide a means of charting and understanding chemical and biological systems in unison.

Ligand efficiencies are gaining usefulness in describing simultaneously, among others, the molecular properties of molecular binding, pharmacokinetics and size. Their utility is encompassed by the different molecular properties they can describe using molecular weight, number of heavy atoms, logP, surface areas, and others. Naïve Bayesian classifiers are useful in representing classes of elements by their underlying distributions, without assuming relationships between the properties of each element [10,11]. Therefore, in the present work, drugs and non-drugs as well as the different disease category or organ of drugs, have been studied using naïve Bayesian classifiers and the chemical space defined by the molecular properties of the different compounds according to different disease categories or organs has been used to define separate and statistically significant classifiers that can help in assigning and predicting specificity or multiple effects between drugs and non-drugs, and between organs or disease categories.

2. Methods

2.1 *Datasets*

The compounds used for compiling the training and validation datasets are from the same sources as used previously in our work [6–8].

A collection of drug compounds for the training set were compiled and curated from the PDBBind version 2005 [12], the K_i Bank [13] and the SCORPIO dataset [14]. Non-drug compounds were also collected for the training set from these sources and the DrugBank database [15] was used to verify their nondrug status. The data include experimentally determined binding constants for each compound with its related target. The number of molecules and the distribution of binding energies are similar for drugs and non-drugs.

For the validation dataset, different compounds from a newer collection of the PDBBind database version 2009 [16], and PDSP database [17] that were not available at the time of compiling the training set were used. Thus, the validation set compounds are a completely independent and external validation dataset. The total number of drugs was 417 compounds. The total number of compounds was thus 823. The drug compounds were further grouped according to their Anatomical Therapeutic Chemical Classification (ATC) disease category (DC, 14 groups), as established by the World Health Organization [2] and as provided by the DrugBank [15]. These are: DC1 = Alimentary tract and metabolism; DC2 = Blood and blood forming; DC3 = Cardiovascular system; DC4 = Dermatological; DC5 = Genito-urinary system and sex hormones; DC6 = Systemic hormonal drugs excluding sex hormones and insulins; DC7 = Anti-infectives; DC8 = Anti-neoplastic and immunomodulating agents; DC9 = Musculo-skeletal system; DC10 = Nervous system; DC11 = Antiparasitics, insecticides and repellants; DC12 = Respiratory system; DC13 = Sensory organs; and DC14 = Various drugs. The list of all compounds used, as well as the DC each drug belongs to, are provided in Table S1 in the Supplementary Material which is available via the multimedia link on the online article webpage.

2.2 *Physicochemical properties and ligand efficiencies*

The logarithm of the octanol–water partition coefficient (logP) was calculated with XLOGP (an atom-additive method) [18]. Marvin Beans version 5.6.0.1 [19] was used for calculating aliphatic ring count, apolar surface area (APSA), aromatic atom count, aromatic ring count, atom count, Balaban index, bond count, exact mass (MW), Harary index, hydrogen bond acceptor count, hydrogen bond donor count, hydrogen count, hyper-Wiener index, molecular polarizability (molpol), molecular surface area (MSA), number of carbons (NoC), number of heavy atoms (NHA), Platt index, polar surface area (PSA), Randic index, ring count, rotatable bond count, Szeged index, Wiener index (Wiener), and Wiener polarity.

$\triangle G_{bind}$, the binding energy of compounds to their partner proteins, was calculated as before [6–8], using the experimental equilibrium inhibition or dissociation constants (K_i or K_d), and temperature of 300K.

Ligand efficiency indices (LE) were calculated by dividing $\triangle G_{bind}$ by a normalization factor: NHA [20], MW [21,22], NoC [23], PSA [24], MSA [8], APSA [8] and Wiener index [25].

2.3 *Statistics*

The statistical computing package R [26] was used for descriptive statistics, kernel density computations, *t*-tests, probability calculations, density overlap coefficient calculations, as well as some plots.

2.4 *Bayesian classifiers*

Bayesian classifiers were calculated according to Equation. 1, where μ is the mean, σ is the standard variation, and x is an independent variable [27]:

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1}$$

Since the population of drugs and non-drugs are similar in size, the *a priori* probability is approximated to 1. Accuracy (as a percentage) was measured according to Equation (S1) in the Supplementary Material available via the multimedia link on the online article webpage. Sensitivity, and specificity (as a percentage), were calculated according to Equation (S2) and Equation (S3), respectively, in the Supplementary Material. Mathew's correlation coefficients (MCC) were calculated as [28]:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TN + FN) \cdot (TN + FP) \cdot (TP + FN) \cdot (TP + FP)}} \tag{2}$$

where TP = true positives, TN = true negatives, FP = false positives and FN = false negatives. Overlap coefficients were calculated as [29]:

$$OC = \int_{-\infty}^{\infty} \min[f(x), g(x)]dx \tag{3}$$

where $f(x)$ and $g(x)$ are the function curves for each different group.

3. Results and discussion

3.1 *Drugs vs. non-drugs*

Descriptive statistics and kernel densities were calculated and compared. Pairwise comparisons of properties between drugs and non-drugs showed separation for several properties. An immediate assessment is provided by box plot figures that show views comparable to histograms viewed from atop. These box plots are shown in Figure 1.

From Figure 1 it can be seen that the binding affinity, ΔG_{bind}, is virtually identical for drugs and non-drugs. This is a feature design of our datasets in order to have non-drugs of similar potency of binding as drugs to provide a challenging background to distinguish both groups [6–8,25]. However, the properties of number of hydrogen bond acceptors, number of hydrogen bond donors, logP, MW, number of heavy atoms (NHA) and PSA show shifts between the distributions of properties between both groups. The ligand efficiencies of ΔG_{bind}/MW, ΔG_{bind}/NHA, ΔG_{bind}/PSA and ΔG_{bind}/MSA also show shifts in means, medians, and first and third quartiles between their distributions. To precisely quantify the difference between all distributions, pairwise Student *t*-tests were conducted using Welch, two-sample, unequal variance, two-sided statistical tests of the null hypothesis that the difference between distributions can be due to random variation. Those properties that were statistically significant at the 95% confidence level (i.e. p <0.05) are shown in Table 1, together with their means (μ) and standard deviations (σ), as well as those for ΔG_{bind} as a comparison.

The number of hydrogen bond donors, number of hydrogen bond acceptors, molecular weight, polar surface area and number of heavy atoms were lower for drug than non-drug compounds. However, logP was slightly higher, but in a better defined distribution of values than for non-drug compounds. The ligand efficiency indices were all deeper for drugs than for non-drug compounds, which arises from the fact that the former are optimized for binding and for their physicochemical properties [3,6,8,24].

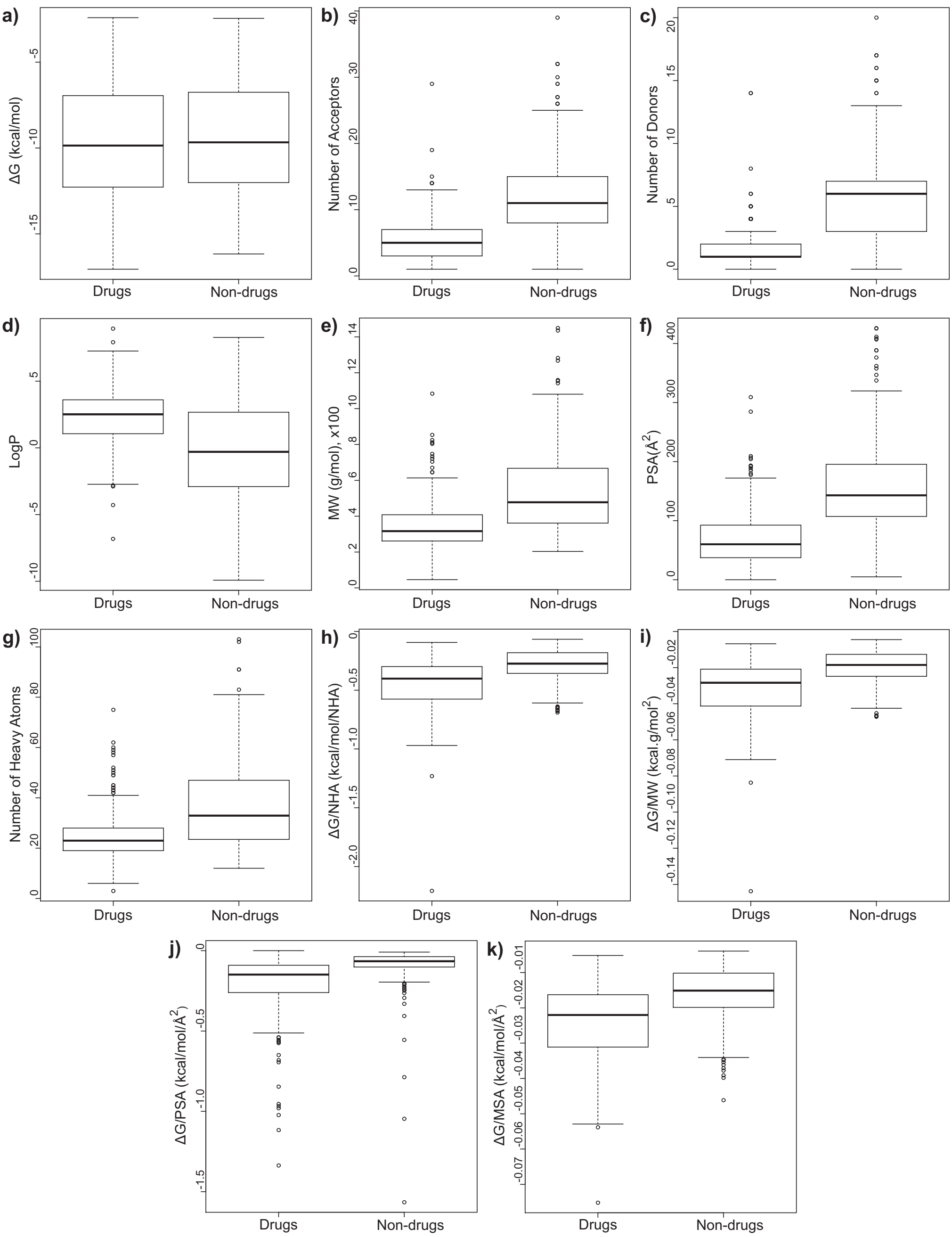

Figure 1. Box plots of physicochemical distributions of properties for drugs and non-drugs. Thick lines show means, while edges of boxes show lower and upper quartiles.

3.2 *Bayesian classifiers for drugs and non-drugs*

The comparison between distributions of physicochemical data of drug and non-drug compounds can be achieved using the densities of their distributions. Figure 2 shows comparisons

Table 1. Means (μ) and standard deviations (σ) for the statistically significant properties of drugs and non-drugs.

	Training set				Validation set			
Property	*μ Drugs*	*σ Drugs*	*μ Non-drugs*	*σ Non-drugs*	*μ Drugs*	*σ Drugs*	*μ Non-drugs*	*σ Non-drugs*
Acceptor count	5.5	3.3	12.2	6.1	5.4	4.6	9.8	0
Donor count	1.7	1.6	5.9	3.5	1.8	2.6	4.4	3
log P	2.29	2.01	−0.28	3.69	2.32	1.87	−0.10	4.1
MW	346.9	138.7	533.3	232.6	348.2	191.4	453	179.8
NHA	24.5	9.8	36.8	17	24.7	13.4	30.4	12.8
PSA	70	44.8	159.7	82.1	66.8	54.7	143.2	80.8
ΔG_{bind}	−9.64	2.92	−9.28	3.18	−10.22	2.31	−8.89	2.31
ΔG_{bind}/MSA	−0.024	0.010	−0.016	7.6e−3	−0.025	0.009	−0.02	0.010
ΔG_{bind}/MW	−0.032	0.015	−0.02	8.7e−3	−0.034	0.012	−0.023	0.011
ΔG_{bind}/NHA	−0.446	0.214	−0.288	0.133	−0.479	0.173	−0.353	0.176
ΔG_{bind}/PSA	−0.208	0.184	−0.088	0.123	−0.239	0.196	−0.081	0.047

(including means) between the density distributions of the statistically significant properties of number of hydrogen bond donors, logP and PSA of drugs and non-drugs, together with $\triangle G_{bind}$ as control.

Figure 2 shows that these differences between the distributions are clear. The means and standard deviations of the statistically significant properties, as well as their reasonably continuous distribution, allow use of a naïve Bayesian classifier to separate both groups into classes. Using Equation (1), where x is the value for a given property for a compound, it is possible to calculate the probability (P) of a compound belonging to one of the two classes by using μ and σ for each group (drugs and non-drugs in the training set). If the value calculated for P_{drug} is larger than that of $P_{\mathrm{non\text{-}drug}}$, then the likelihood of a compound belonging to the drug class is larger than belonging to the class of non-drugs. Using this classifier, the probabilities of belonging to each class by the compounds of the validation set were calculated. The results are shown in Table 2, which gives accuracies, sensitivities, specificities and Mathew's correlation coefficients.

Table 2 also shows the results obtained for the combination (multiplication) of all the probabilities of the properties and how they perform to classify the compounds, as well as the combination of all the probabilities of the ligand efficiencies. Some of the properties perform better than others, but those at the top of the table have good values for accuracy, sensitivity and specificity (approaching 100%), as well as MCC values (an MCC value of 1.0 would imply a perfect distinction between classes). Those with accuracies of 70% or higher, and/or MCC values higher than 0.5, can be considered the best classifiers among these properties. It is important to note that these classifiers may be used sequentially or in combination, in order to daisy chain different properties.

Another measure of assessing the ability to place into two classes the compounds was provided by using the naïve Bayesian classifier probability as a rank and computing Receiver–Operator characteristic (ROC) curves. These are presented in Figure S1 in the Supplementary Material which is available via the multimedia link on the online article webpage. From the ROC curves, the best ranking of drugs over non-drugs is provided, in descending order, by the combined probability of all properties, the probability based on number of hydrogen bond donors, the probability based on the number of hydrogen bond

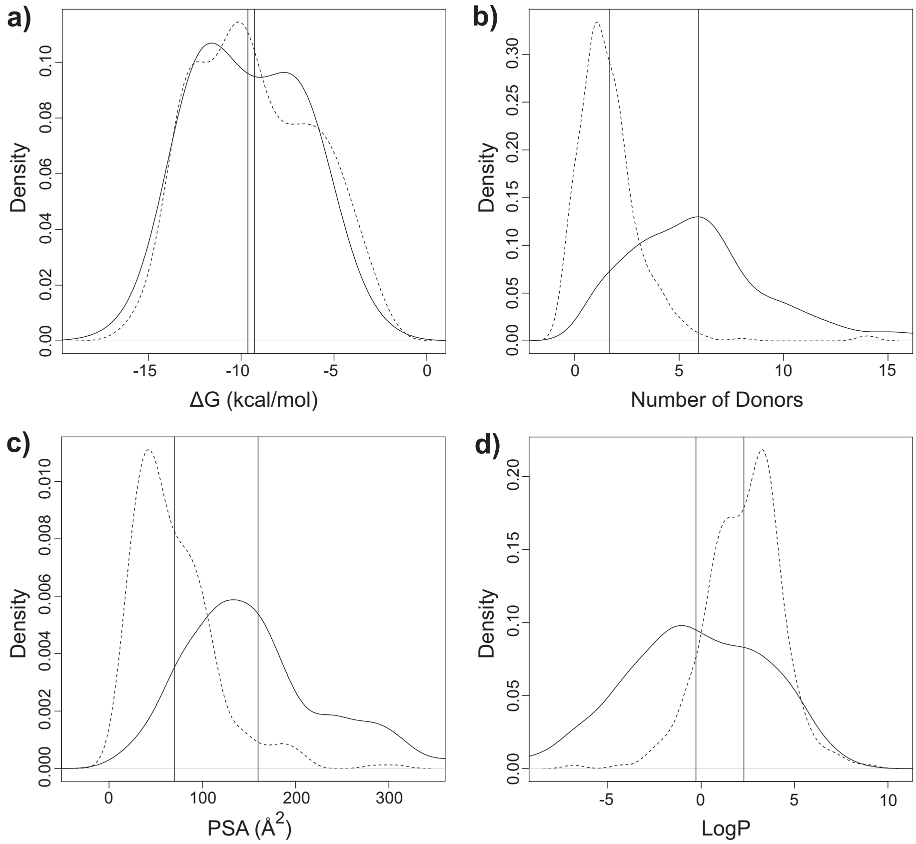

Figure 2. Kernel densities. Drugs are in dotted lines, non-drugs in full line curves; means as vertical lines.

Table 2. Accuracy, selectivity, specificity, and Mathew's correlation coefficients (MCC) for the statistically significant properties of drugs and non-drugs.

	Training set				*Validation set*			
Property	*Accuracy (%)*	*Sensitivity (%)*	*Specificity (%)*	*MCC*	*Accuracy (%)*	*Sensitivity (%)*	*Specificity (%)*	*MCC*
Donor count	82.4	90.7	74.4	0.658	73.0	89.6	57.4	0.496
All together (*P*)	81.5	89.4	73.8	0.638	70.2	91.5	50.0	0.455
PSA	76.5	90.7	62.8	0.556	71.2	91.5	51.9	0.471
Acceptor count	76.1	90.7	61.9	0.547	66.5	93.4	40.7	0.401
log P	71.3	87.5	55.6	0.453	70.2	90.6	51.0	0.451
MW	68.5	88.4	49.1	0.407	62.8	89.6	37.1	0.313
ΔG/MW	68.2	61.7	75.3	0.374	73.5	82.1	65.7	0.484
All EI together (*P*)	67.5	50.5	84.1	0.367	70.2	74.5	66.7	0.413
ΔG/NHA	66.7	47.6	85.3	0.356	68.8	66.1	72.2	0.383
NHA	66.6	88.4	45.3	0.373	63.7	88.7	39.8	0.326
ΔG/MSA	65.8	41.8	89.1	0.351	65.6	55.7	75.9	0.323
ΔG	–	–	–	–	–	–	–	–

Table 3. Area under the curve (AUC) for the Receiver–Operator characteristic curves for statistically significant properties.

	AUC	
Property	*Training set*	*Validation set*
All together	0.851	0.742
Donor count	0.847	0.714
Acceptor count	0.819	0.688
PSA	0.778	0.713
logP	0.750	0.744
All EI together	0.662	0.543
MW	0.657	0.644
NHA	0.649	0.645
ΔG/MW	0.606	0.655
ΔG/NHA	0.602	0.652
ΔG/MSA	0.528	0.547

acceptors, as well as the probability based on PSA. This is shown by the early recall of true positives (in this case, drug compounds) and much better performance than a random choice, which is represented by the diagonal black line. This is shown quantitatively by the areas under the curve, the integration of the ROC curves of Figure S1 over the *x* axis, which are presented in Table 3.

The values of the area under the curve (AUC) in Table 3 for the properties at the top of the table show a high recall of drug compounds, where an AUC of 1.0 would show a complete and perfect separation, and an AUC of 0.50 represents the area that would be covered by a random choice represented by the black diagonal line. Here, the best classifiers for ranking were all the probabilities combined, followed by hydrogen bond donors, hydrogen bond acceptors, PSA and logP. The rest are less successful at this ranking.

It is interesting to note that some of the top properties correspond to some of the components of Lipinski's rule-of-five [30]. These properties such as small values for MW, logP, number of hydrogen bond acceptors and donors have a strong role to play in the bioavailability of compounds, specifically their oral bioavailability, describing physicochemical properties of molecules that may make them more easily distributed across membranes. The rule-of-five has been very useful in profiling compounds and libraries of compounds. However, coded as a test, it provided an accuracy of only 60%, as well as a low MCC of 0.292 for classifying the drugs and non-drug compounds in the validation set. Many of the Bayesian classifiers in this work show better accuracy and MCC values. In addition, the naïve Bayesian classifiers can provide a gradual and continuous ranking of compounds, instead of a harsh limit as in the rule-of-five, which can be useful when profiling compounds and chemical libraries.

The naïve Bayesian classifiers constructed seem to be of help in scoring the probability of a compound to belong to a class of drug or non-drug, based on their molecular properties, and being applicable to new datasets. As they are naïve, they do not assume any structure or relationship between the variables, but are of use to classify compounds.

3.3 *Disease (organ) category vs. disease (organ) category*

The drugs in the training and in the validation set were then grouped into their disease category (DC). For each DC, pairwise comparisons were made for each molecular property calcu-

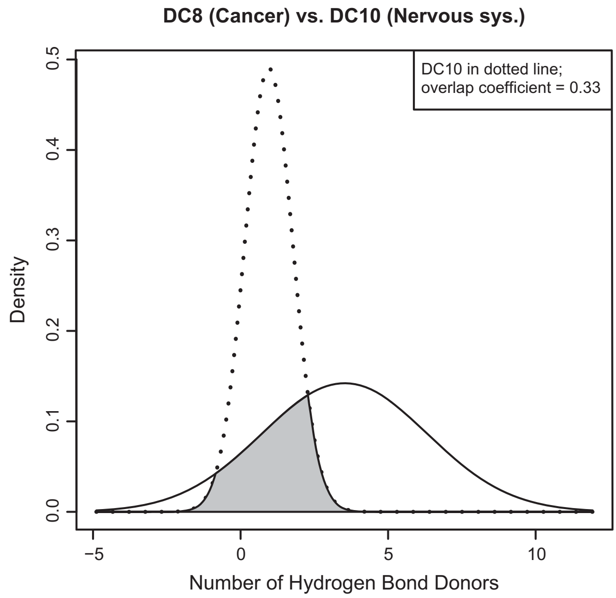

Figure 3. Overlap coefficient between density Gaussians of number of hydrogen bond donors for DC8 anti-neoplastic agents and DC10 nervous system drugs.

lated in a similar way as in section 3.1. The statistically significant differences for all comparisons *per* DC and *per* molecular property are shown in Table S2 in the Supplementary Material which is available via the multimedia link on the online article webpage. A further characterization was provided by calculating the overlap coefficient between the density Gaussians as calculated in Equation (3) and shown in Figure 3.

Figure 3 shows a broader distribution for DC8 (antineoplastic) than for DC10 (nervous system), with separate peaks as well as an overlap coefficient of 0.33 between cancer (full line) and nervous system (dotted line) drugs, shaded in grey. An overlap coefficient of 1.0 would show a complete overlap between distributions, which would render impossible any separation between the groups. Thus, the property of number of hydrogen bond donors serves as a good separator between the distributions for drugs used in cancer therapy and those acting on the nervous system. The best properties were thus selected and a similar comparison between the groups as in section 3.2 was computed.

3.4 *Bayesian classifiers for DC (organ) vs. DC (organ)*

The results for the best separations between drugs belonging to different DC or organ according to molecular properties are shown in Table 4, with overlap coefficients shown in Table S3 in the Supplementary Material which is available via the multimedia link on the online article webpage.

A variety of molecular properties were useful for building naïve Bayesian classifiers between drugs of different DC or organ. Some perform better than others, and in a couple of cases, a combination (multiplication) of several probabilities derived from molecular properties provided the best results. For example, for the comparison of DC5 *vs*. DC7, the combination of the probability of classification provided by logP and number of hydrogen bond donors was best; whilst for DC5 *vs*. DC10, the probability provided by the number of bonds,

Table 4. Accuracy for prediction of disease category in pair-wise comparisons according to statistically significant properties.

Comparison	*Property*	*Training set (%)*	*Validation set (%)*	*Sensitivity (%) Training; Validation*	*Specificity (%) Training; Validation*	*MCC*
DC3 ↔ DC5 Cardio ↔ Genito	Number of hydrogens	65.5	75	82.6; 100	26.9; 42.9	0.111; 0.832
DC3 ↔ DC7 Cardio ↔ Antiinf.	Number of aliphatic rings	69.4	75	65.5; 75	77.8; -	0.404; -
DC3 ↔ DC8 Cardio ↔ Cancer	Molecular Surface Area	78.8	78.6	94.8; 100	36.4; 40	0.404; 0.548
DC5 ↔ DC7 Genito ↔ Antiinf.	PlogP* P#HBacceptors[1]	83.3	57.1	82.1; 57.1	84.6; -	0.667; -
DC5 ↔ DC8 Genito ↔ Cancer	Wiener polarity	76.1	66.7	96.4; 66.7	44.4; -	0.503; -
DC5 ↔ DC10 Genito ↔ Nerv.	P#bonds*P#aliphatic rings*P#hydrogens[1]	82.5	78.8	60.7; 57.1	89.1; 84.6	0.505; 0.400
DC8 ↔ DC10 Cancer ↔ Nerv.	Number of HB donors	88.1	90.3	50; 60	96.9; 96.2	0.565; 0.616

[1]P property refers to the Bayesian probability calculated for that property.

number of aliphatic rings and number of hydrogens together gave the best results as defined by accuracy and MCC. The DCs presented are for cardiovascular drugs (DC3), genito-urinary system and sex hormones (DC5), anti-infectives (DC7), anti-neoplastic agents (DC8) and nervous system agents (DC10).

The combination of terms, as in multiplication of variables in multivariate analysis, can lead to the augmentation or dampening of the effect of dependent variables on the outcome variable, or even spurious results if one is not careful when choosing the combination terms. However, this is not an issue for the present work since the combination terms are the multiplication of the probability of a compound belonging to the drug distribution *vs*. nondrug distributions, and so the multiplication is comparing the same effect, that is, how different in nature is the observed physicochemical value for a chemical compound in drugs as opposed to non-drugs. That is, at no point are the molecular properties multiplied or mixed, only the predicted naïve Bayesian probabilities of belonging to the drug distribution as opposed to the non-drug distribution are combined (i.e. effect of several probabilities).

The ability to score a probability to class a chemical compound based on its molecular properties to have a higher likelihood of belonging to a particular DC as opposed to another may help in introducing specificity to chemical libraries and thus avoid unwanted side-effects. Chemical libraries that could be targeted to specific organs or diseases would be a welcome improvement for drug design or, conversely, for avoiding toxicity at a specific organ. In addition, when the same sort of biomolecular targets are present in several organs, the naïve Bayesian classifiers may help in sorting compounds to tailor them to the organ in which they would be effective or for a specific application. They also possess the advantage of being relatively simple to use, fast (since they can be used on the fly) and are able to be rebuilt according to specific information. For those cases where there is no statistically significant difference between properties for two specific DC comparisons, this is also valuable information in the sense that it may hint at a possible effect on multiple DCs or organs of a specific compound or library.

4. Conclusions

The statistically significant physicochemical properties of ligands have been used to build naïve Bayesian classifiers using the density of the property distributions. These have been used to classify drugs *vs*. non-drugs, as well as drugs of one disease category or organ against other disease categories or organs. These classifiers perform well for some molecular properties, as shown by different tests on compound sets. They can also have utility by recalling true positives (such as drugs) from among non-drugs as evidenced by early receiver–operator characteristics curves and areas under the curve which approach 1.0. Naïve Bayesian classifiers were also able to be built for some disease categories or organs, and the overlap coefficients between their Gaussian densities indicated favourable (small) overlaps in some cases. In addition, the naïve Bayesian classifiers remain easily interpretable in their physicochemical properties.

These functions could be of help in designing chemical libraries or compounds that could target a specific organ or disease category. The advantages could be in conferring specificity to compounds according to organ or disease category, and thus helping to limit side-effects and target-specific compartmentalization of compounds to an organ or disease. This would be of use in the cases of a very similar (or identical) receptor present in different tissues or organs and only one is indicated for therapeutic intervention. Conversely, discovering non-statistically significant chemical property distributions between libraries of compounds may provide a gross indication that there may be possible activity at more than one organ or disease category, given that their chemical space is very similar or overlaps considerably. It is envisaged that this method can be employed on biomolecule- and organ specific-characterization of chemical libraries, so that even finer detail may be incorporated.

Supplementary material can be found on via the multimedia link on the online article webpage.

Acknowledgements

We thank Mare Oja for help with data collation and Dr Csaba Hetényi for helpful discussions and fruitful collaboration on similar topics. We thank the Estonian Science Foundation Grant 7709 and the Estonian Ministry for Education and Research Grant SF0140031Bs09.

References

[1] M.J. Keiser, J.J. Irwin, and B.K. Shoichet, *The chemical basis of pharmacology*, Biochemistry 49 (2010), pp. 10267–10276.

[2] World Health Organization, *Anatomical Therapeutic Chemical Classification* [online], 2012. Available at www.whocc.no/atc/structure_and_principles/

[3] A.T. Garcia-Sosa and R.L. Mancera, *Free energy calculations of mutations involving a tightly bound water molecule and ligand substitutions in a ligand-protein complex*, Mol. Inf. 29 (2010), pp. 589–600.

[4] G. Panagiotou and O. Taboureau, *The impact of network biology in pharmacology and toxicology*, SAR QSAR Environ. Res. 23 (2012), pp. 221–235.

[5] C. Curtis, S. Shah, S.-F. Chin, G. Turashvili, O.M. Rueda, M.J. Dunning, D. Speed, A.G. Lynch, S. Samarajiwa, Y. Yuan, S. Graf, G. Ha, G. Haffari, A. Bashashati, R. Russell, S. McKinney, METABRIC Group, A. Langerød, A. Green, E. Provenzano, G. Wishart, S. Pinder, P. Watson, F. Markowetz, L. Murphy, I. Ellis, A. Purushotham, A.-L. Børresen-Dale, J.D. Brenton, S. Tavare1, C. Caldas, and S. Aparicio, *The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups*, Nature 486 (2012), pp. 346–352.

[6] A.T. Garcia-Sosa, U. Maran, and C. Hetenyi, *Molecular property filters describing pharmacokinetics and drug binding*, Curr. Med. Chem. 19 (2012), pp. 1646–1662.

[7] A.T. Garcia-Sosa, M. Oja, C. Hetenyi, and U. Maran, *Disease-specific differentiation between drugs and non-drugs using principal component analysis of their molecular descriptor space*, Mol. Inf. 31 (2012), pp. 369–383.

[8] A.T. Garcia-Sosa, M. Oja, C. Hetenyi, and U. Maran, *DrugLogit: Logistic discrimination between drugs and non-drugs including disease-specificity by assigning probabilities based on molecular properties*, J. Chem Inf. Model. 52 (2012), pp. 2165–2180.

[9] M. Campillos, M. Kuhn, A.-C. Gavin, L.J. Jensen, and P. Bork, *Drug target identification using side-effect similarity*, Science 321 (2008), pp. 263–266.

[10] Nidhi., M. Glick, J.W. Davies, and J.L. Jenkins, *Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases*, J. Chem. Inf. Model. 46 (2006), pp. 1124–1133.

[11] K. Azzaoui, J. Hamon, B. Faller, S. Whitebread, E. Jacoby, A. Bender, J.L. Jenkins, and L. Urban, *Modeling promiscuity based on in vitro safety pharmacology profiling data*, ChemMedChem 2 (2007), pp. 874–880.

[12] R. Wang, X. Fang, Y. Lu, and S. Wang, *The PDBbind database: Collection of binding affinities for protein-ligand complexes with known three-dimensional structures*, J. Med. Chem. 47 (2004), pp. 2977–2980.

[13] J.–W. Zhang, M. Aizawa, S. Amari, Y. Iwasawa, T. Nakano, and K. Nakata, *Development of KiBank, a database supporting structure- based drug design*, Comput. Biol. Chem. 28 (2004), pp. 401–407.

[14] A. Ababou and J.E. Ladbury, *Survey of the year 2005: Literature on applications of isothermal titration calorimetry*, J. Mol. Recognit. 20 (2007), pp. 4–14.

[15] D.S. Wishart, C. Knox, A.C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, and J. Woolsey, *DrugBank: A Knowledgebase for Drugs, Drug Actions and Drug Targets*, Nucleic Acids Res. 34 (2006), pp. D668–D672.

[16] Shanghai Institute of Organic Chemistry, *PDBbind database version 2009*; software available at http://www.pdbbind-cn.org/

[17] National Institute of Mental Health (NIMH) Psychoactive Drug Screening Program, *PDSP* K_i *Database*; available at http://pdsp.med.unc.edu/pdsp.php

[18] T. Cheng, Y. Zhao, X. Li, F. Lin, Y. Xu, X. Zhang, Y. Li, and R. Wang, *Computation of octanol-water partition coefficients by guiding an additive model with knowledge*, J. Chem. Inf. Model. 47 (2007), pp. 2140–2148.

[19] ChemAxon, *Marvin Beans version 5.6.0.1*; software available at http://www.chemaxon.com

[20] I.D. Kuntz, K. Chen, K.A. Sharp, and P.A. Kollman, *The maximal affinity of ligands*, Proc. Natl. Acad. Sci. U. S. A. 96 (1999), pp. 9997–10002.

[21] A.L. Hopkins and C.R. Groom, *Ligand efficiency: A useful metric for lead selection*, Drug Discovery Today 9 (2004), pp. 430–431.

[22] A.T. Garcia-Sosa, S. Sild, and U. Maran, *Docking and virtual screening using distributed grid technology*, QSAR Comb. Sci. 28 (2009), pp. 815–821.

[23] A.T. Garcia-Sosa, C. Hetényi, and U. Maran, *Drug efficiency indices for improvement of molecular docking scoring functions*, J. Comput. Chem. 31 (2010), pp. 174–184.

[24] A.T. Garcia-Sosa, S. Sild, K. Takkis, and U. Maran, *Combined approach using ligand efficiency, cross-docking, and antitarget hits for wild-type and drug-resistant Y181C HIV-1 reverse transcriptase*, J. Chem. Inf. Model. 51 (2011), pp. 2595–2611.

[25] C. Hetenyi, U. Maran, A.T. Garcia-Sosa, and M. Karelson, *Structure-based calculation of drug efficiency indices*, Bioinformatics 23 (2007), pp. 2678–2685.

[26] R Development Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna; software available at http://www.R-project.org

[27] J. Crawshaw and J. Chambers, *A Concise Course in Advanced Level Statistics*, Nelson Thornes, Cheltenham, U.K, 2001.

[28] B.W. Mathews, *Comparison of the predicted and observed secondary structure of T4 phage lysozyme*, Biochim. Biophys. Acta 405 (1975), pp. 442–451.

[29] S. Mizuno, T. Yamaguchi, A. Fukushima, Y. Matsuyama, and Y. Ohasi, *Overlap coefficient for assessing the similarity of pharmacokinetic data between ethnically different populations*, Clinical Trials 2 (2005), pp. 174–181.

[30] C. Lipinski, F. Lombardo, B.W. Dominy, and P.J. Feeney, *Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings*, Adv. Drug Delivery Rev. 23 (1997), pp. 3–25.