Discriminating between Drugs and Nondrugs by Prediction of Activity Spectra for Substances (PASS)

Soheila Anzali,*,[†] Gerhard Barnickel,[†] Bertram Cezanne,[†] Michael Krug,[†] Dmitrii Filimonov,[‡] and Vladimir Poroikov[‡]

Bio- and Chemoinformatics Department, Merck KGaA, Darmstadt D-64271, Germany, and Institute of Biomedical Chemistry of Russian Academy of Medical Sciences, Pogodinskaya Street, 10, Moscow 119832, Russia

Received August 31, 2000

Using the computer system PASS (prediction of activity spectra for substances), which predicts simultaneously several hundreds of biological activities, a training set for discriminating between drugs and nondrugs is created. For the training set, two subsets of databases of drugs and nondrugs (a subset of the World Drug Index, WDI, vs the Available Chemicals Directory, ACD) are used. The high value of prediction accuracy shows that the chemical descriptors and algorithms used in PASS provide highly robust structure−activity relationships and reliable predictions. Compared to other methods applied in this field, the direct benchmark undertaken with this paper showed that the results obtained with PASS are in good accordance with these approaches. In addition, it has been shown that the more specific drug information used in the training set of PASS, the more specific discrimination between drug and nondrug can be obtained.

Introduction

In the past decade the drug discovery process has changed dramatically. The challenge to identify novel leads has driven the need for automated systems that can rapidly perform selection of compounds at the beginning of the drug discovery process, namely in the analysis and the extension of the high throughput screening (HTS) pool. The number of discovered hits depends on the cutoff level, e.g., 10 mM. First of all, the activity needs have to be confirmed and then followed by selectivity and functional assays.

An important task is the rejection of false hits and focus on the promising molecules. The lead molecule plays the pivotal role for the initiation of a lead optimization project. A promising lead compound with a desired pharmacological activity may have undesirable side effects, characteristics that limit its bioavailability, or structural features which adversely influence its metabolism and excretion from the body.

Therefore biological activity has to be balanced with "drug-like" properties, and the closer we get to a candidate compound, the more important drug-likeness becomes. Despite the many attempts[1−11] to classify compounds into the "drug" and "nondrug" categories, there is no unambiguous definition for drug and nondrug. Especially, it may vary depending the indications or diseases considered.[12] Reagent databases such as ACD,[13] as an example, is often used as a model database for nondrug compounds, while CMC,[14] WDI,[15] and MDDR[16] could be seen as databases for drugs. Certainly, if one could consider the fate of some compounds in the ACD database they may become drugs in the future, whereas a few compounds from MDDR and WDI will never be seen as drugs.

Because of the lack of discrimination among structural features for drug and nondrug compounds, different approaches have to be applied to compensate. As concluded by Walters et al.,[17] "future work is likely to include additional approaches and more robust attempts at validation of these methods."

The PASS program,[18−22] which is based on a regression approach applied to noncongeneric chemical series, provides highly robust predictions for more than 500 biological activities. Since PASS is trained to recognize drugs with activities on various targets, the approach may have potential use to discriminate drugs from nondrugs. The purpose of this work is to evaluate the ability of the PASS approach in discriminating between drug-like compounds and nondrugs.

Materials and Methods

PASS Approach. The computer system PASS (prediction of activity spectra for substances)[18−21] predicts several hundreds of biological activities (pharmacological main and side effects, mechanisms of action, mutagenicity, carcinogenicity, teratogenicity, and embryotoxicity).

Biological activity results from the interaction of chemical compounds with biological entities. In clinical studies, the biological entity is the whole human organism. In preclinical testing they are the experimental animal (in vivo) and/or the experimental model (in vitro). Biological activity depends on peculiarities of compound (structure and physicochemical properties), biological entity (species, gender, age, etc.), and mode of treatment (dose, route of administration, etc.).

The majority of biologically active compounds reveal often a wide spectrum of different effects. Some of them are useful in treatment of definite diseases; others cause various side and toxic effects. The whole complex of activities caused by the compound in biological entities is called the "biological activity spectrum of the substance".

The biological activity spectrum of a compound presents all its activities despite the difference in essential conditions of

* Correspondence: Soheila Anzali, Ph.D., Merck KGaA, Bio- and Chemoinformatics Department, Frankfurter Str. 250, D-64271 Darmstadt, Germany. Tel: +49-6151-724863. Fax: +49-6151-7233299. E-mail: soheila.anzali@merck.de.

[†] Merck KGaA.

[‡] Institute of Biomedical Chemistry of Russian Academy of Medical Sciences.

10.1021/jm0010670 CCC: $20.00
Published on Web 06/14/2001

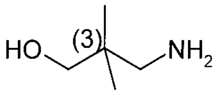

Figure 1.

their experimental determination. If the difference in species, gender, age, dose, route, etc., is neglected, the biological activity can be identified only qualitatively. Thus, "the biological activity spectrum" is defined as the "intrinsic" property of a compound depending only on its structure and physicochemical characteristics.

The prediction of this spectrum by PASS is based on SAR analysis of a training set containing more than 30 000 compounds which reveal more than 500 kinds of biological activities. Therefore, PASS once trained is able to predict for a test compound all likely biological activities, which are included in the training set.

It was shown that the mean accuracy of prediction with PASS is about 86% in leave-one-out cross-validation.[21] PASS prediction accuracy exceeds more than three times the expert's guess-work for an independent set of 33 different compounds studied as pharmacological agents, which are not included in the PASS training set.[22] Recently PASS was tested in a blind mode by nine scientists from eight countries on the heterogeneous set of 118 compounds having 138 activities, and the mean accuracy of prediction was shown to be 82.6%.[23] The PASS prediction is relatively successful even in the case of rather new compounds which have nontraditional structures and/or belong to new chemical classes. Like any other ligand-based design approach, PASS cannot predict the affinity for new targets, but even in that case PASS points to possible side effects which may also prevent the application of a drug candidate.

Besides this SAR-base available in PASS, it is also possible to create other SAR-bases or to enlarge it.

Activities Description. In this work, the investigated activity is "drug", so the compounds from WDI and the Cipsline DB were described as drugs and the compounds from the other sets were described as nondrugs.

Chemical Structure Description. We described in detail the substructure descriptors called "multilevel neighborhoods of atoms" (MNA) in a paper published recently.[24] MNA descriptors of a molecule are based on the 2D representation of its structure. According to the valences and partial charges of the atoms, hydrogens are included, whereas bond types are not explicitly specified. An MNA descriptors set is subdivided on levels and generated recursively. A zero-level MNA descriptor describes the atom itself. Any next level MNA descriptor is the substructure notation $\boldsymbol{A(D_1D_2...)}$, where $\boldsymbol{A}$ is the atom A descriptor, and $\boldsymbol{D_i}$ is the previous level MNA descriptor of *i*th neighbor atom for atom A. For example, for carbon(3) in Figure 1, the MNA descriptors are as follows: first, "C"; second, "C(CCCC)"; third, "C(C(HHHC)C(HHHC)C(HHCN)C(HH-CO))".

Different stereoisomers of a molecule have identical MNA descriptors and are considered as equivalent molecules in PASS. The use of MNA descriptors in PASS for prediction is described in the Appendix. In the present version of PASS, up to second level MNA descriptors are used.

Databases Used for the Training and Evaluation of PASS. To compare the PASS ability in discriminating drug-like compounds and nondrugs with the recently published results of Sadowski and Kubinyi,[3] we used the same subsets of WDI and ACD compounds for the training of PASS. These subsets include 5000 compounds from WDI ("drugs") and 5000 compounds from ACD ("nondrugs"). This data set was also used by Ajay et al.[2]

To evaluate the method we prepared several test sets. As a sample of drug compounds we extracted two data sets from the Cipsline database,[25] which is a subset of MDDR.[16] The first subset includes all launched, registered, and investigated compounds (LRID). At the second stage, in order to focus on real drug compounds, we extracted the subset of Cipsline with just launched and registered compounds (LRD). If the PASS approach could provide a reasonable discrimination between drugs/nondrugs, the expected results should be better for the second subset.

Table 1. Functional Groups Describing Nondrug Compounds[a]

aldehyde
α halogenated
α-β unsat. (>2)
amide (>2)
aryl-Br
aryl-Cl (>3)
aryl-F (>3)
aryl-I
azide
azo compds
benzylhalogenide
boron
carbamate (>3)
carbamoyl chloride
carbonate
carbox. acid ester (>2)
carbox. acid anhydride
carbox. acid hal.
carbox. acid (>2)
crown ether
diazonium salt
disulfide
dithio acid ester
dithio carbamate
double bond (>1)
enolate (>1)
epoxide
hydrazine (>1)
imines (>1)
iso(thio)cyanante
isothiourea
ketone (>2)
nitro (>2)
oxime
P=S bond
arom. hydroxy (>2)
phosphine
phosphoric acid ester
phosphorus (>1)
prim. alcohol (>2)
prim. amine (>1)
prim. arom. amine (>1)
sec. alcohol (>2)
sec. amine (>2)
sec. aromatic amine (>2)
silicon
S−N bond
sulfone (>1)
sulfonic acid ester (>1)
sulfonic acid (>1)
sulfonamide (>2)
sulfonylhalogenide
sulfur (>3)
sulfuric acid ester
tert. aliphat. amines (>2)
tert. arom. amines (>2)
thioester (>1)
thioureas (>1)
1,1,1-trichloroethyl-2,2-diamino
trifluoromethyl (>2)

[a] Minimum frequency of a certain functional group is indicated in parentheses; in all other cases it is 1. Compounds with MW < 150 were also classified as nondrugs.

As an example for a nondrug data set, we prepared 9737 compounds (ND) from a supplier database of approximately 57 000 commercially available compounds. A compound was identified as nondrug by the analysis of 60 different functional groups/fragments. Most of them are reactive groups, which are unfavorable for drugs. Some examples of such groups are shown in Table 1.

In addition, all compounds with a molecular weight less than 150 Da were classified as nondrugs.

As an independent evaluation set of drugs (TOP-100), we use a list of top-100 prescription pharmaceuticals[26] (Table 2). Twelve of these entries are biopolymers and were not included in the evaluation (Table 3).

Computation Time. The calculation time on a PC (Pentium 2; 300 MHz; 128 Mb RAM) for the prediction of one compound is 4 ms, which demonstrates the ability of PASS to handle huge data sets, as they are used, for example, in the analysis of virtual libraries or supplier databases.

Results and Discussion

Training of PASS. The results of a leave-one-out cross-validation (LOO), which characterizes the quality of obtained structure−property relationships, are shown in Table 4, no. 1. The quality of the prediction is described by the percentage of false classification. During model building (including LOO cross-validation), the quality is expressed as the mean error of prediction (MEP). The mean accuracy for prediction in the LOO cross-validation is about 80%, which is slightly less than in the current version of PASS applied for the prediction of the biological activity spectra, but which is still satisfactory to discriminate between drugs and nondrugs. Such accuracy of prediction is comparable to the results obtained by Sadowsky and Kubinyi.[3]

Table 2. Evaluation Set Based on the List of Top-100 Drugs

		WDI/ACD		LR/ND				WDI/ACD		LR/ND	
drug	indication	drug	nondrug	drug	nondrug	drug	indication	drug	nondrug	drug	nondrug
Accupril	hypertension	0.427[a]		0.802		Parlodel	parkinsonism	0.937		0.916	
Adalat	hypertension/angina		0.335	0.246		Pepcid	antiulcer		0.204	0.154	
Amoxil	antiinfective	0.723		0.902		Premarin	oestrogen replacement	0.580			0.163
Ativan	antianxiety	0.310		0.319		Prepulsid/Propulsid	nocturnal heartburn	0.325		0.457	
Augmentin	antiinfective	0.723		0.902		Prinivil	hypertension	0.516		0.833	
Axid	antiulcer	0.287		0.469		Procardia	hypertension/angina		0.335	0.246	
Becotide/Beclovent	asthma	0.835		0.938		Proscar	benign prostatic hypertrophy	0.615		0.613	
Buspar	antianxiety	0.265		0.431		Proventil	asthma	0.691		0.567	
Calan	hypertension/angina	0.448		0.859		Provera	hormone therapy/contraceptive	0.677		0.857	
Capoten	hypertension	0.565		0.712		Prozac	antidepressant	0.245		0.335	
Cardizem	hypertension/angina	0.43		0.506		Relifex/Relafen	arthritis	0.308		0.131	
Cardura	hypertension/BPH	0.443		0.538		Renitec/Vasotec	hypertension	0.512		0.825	
Ceclor	antiinfective	0.622		0.820		Retrovir	AIDS	0.524		0.182	
Ciproxin	antiinfective	0.622		0.846		Risperdal	antipsychotic	0.300		0.207	
Claforan	antiinfective	0.580		0.968		Roaccutane Accutane	acne	0.832		0.731	
Claritin	allergy		0.238	0.643		Rocephin	antiinfective	0.303		0.919	
Dalacin/Cleocin	antiinfective	0.654		0.293		Sandimmun/Sandimmune	immunosuppressive	0.944		0.821	
Depakote	epilepsy	0.405		0.551		Seldane	allergy	0.532		0.735	
Diflucan	antifungal		0.215	0.120		Serevent	asthma	0.662		0.530	
Diprivan	anasthesia inducer	0.436			0.115	Seroxat/Paxil	antidepressant	0.528		0.566	
Dormicum/Versed	anaesthesia inducer		0.228	0.644		Sporanox	fungal	0.208		0.787	
Duricef	antiinfective	0.694		0.894		Tagamet	antiulcer	0.558		0.939	
Estraderm	oestrogen replacement	0.686		0.655		Taxol	anticancer	0.806		0.907	
Eulexin	anticancer		0.749		0.347	Tegretol	epilepsy	0.258		0.605	
Feldene	arthritis	0.224		0.607		Tenormin	hypertension	0.535		0.583	
Fortum/Fortaz	antiinfective	0.554		0.981		Timoptic	glaucoma	0.365		0.516	
Hytrin	hypertension/BPH	0.431		0.646		Toradol	analgesic	0.475		0.759	
Imigran/Imitrex	migraine	0.252		0.720		Transderm-Nitro	angina	0.967		0.791	
Intal	asthma	0.529		0.548		Trental	haemorheologic	0.367		0.289	
Istin/Norvasc	hypertension/angina	0.276		0.426		Unasyn	antiinfective	0.668		0.976	
Klacid/Klaricid/Biaxin	antiinfective	0.994		0.994		Vancenase/Vanceril	asthma/antiallergy	0.776		0.922	
Klonopin	epilepsy		0.323	0.186		Vancocin	antiinfective	0.863		0.807	
Lamisil	antifungal	0.477		0.239		Ventolin	asthma	0.691		0.567	
Lasix	diuretic		0.369		0.362	Voltaren	arthritis		0.382		0.102
Leponex/Clozaril	antipsychotic		0.251	0.377		Xanax/Alprazolam	antianxiety		0.207	0.651	
Lipostat/Pravachol	cholesterol reducer	0.916		0.956		Zantac	antiulcer	0.239		0.619	
Lodine	arthritis	0.624		0.179		Zestril	hypertension	0.516		0.833	
Losec/Prilosec	antiulcer	0.373		0.564		Zinnat/Ceftin	antiinfective	0.493		0.840	
Lotensin	hypertension	0.452		0.626		Zithromax	antiinfective	0.993		0.993	
Lupron	anticancer/antiendometriosis	0.506		0.919		Zocor	cholesterol reducer	0.960		0.957	
Lustral/Zoloft	antidepressant	0.309		0.656		Zofran	antiemetic	0.361		0.413	
Mevacor	cholesterol reducer	0.979		0.982		Zoladex	anticancer	0.393		0.837	
Nizoral	antifungal	0.414		0.573		Zovirax	herpes	0.749		0.751	
Nolvadex	anticancer	0.572		0.400							
Paraplatin	anticancer	0.314		0.902							

[a] Pa scores representing probability belong to this therapeutic class.

Evaluation of PASS vs "Drugs". Formally the first test set (LRID) includes 7468 presumed drug compounds. Their structures were checked for being present in the training set yielding 632 compounds. These compounds were eliminated from the test set, as were 688 compounds which had no connection table fields or had errors in structural formulas (invalid compounds). After filtering, the final test set contained 6148 compounds. A total of 4514 (73.4%) compounds were predicted as drugs and 1634 (26.6%) compounds as nondrugs (Table 4; no. 2).

There exists no independent criteria to be sure that some compounds predicted as nondrug will not become drugs in the future; therefore we eliminated all the investigated compounds from the LRID set. The remaining 1184 compounds were launched and registered

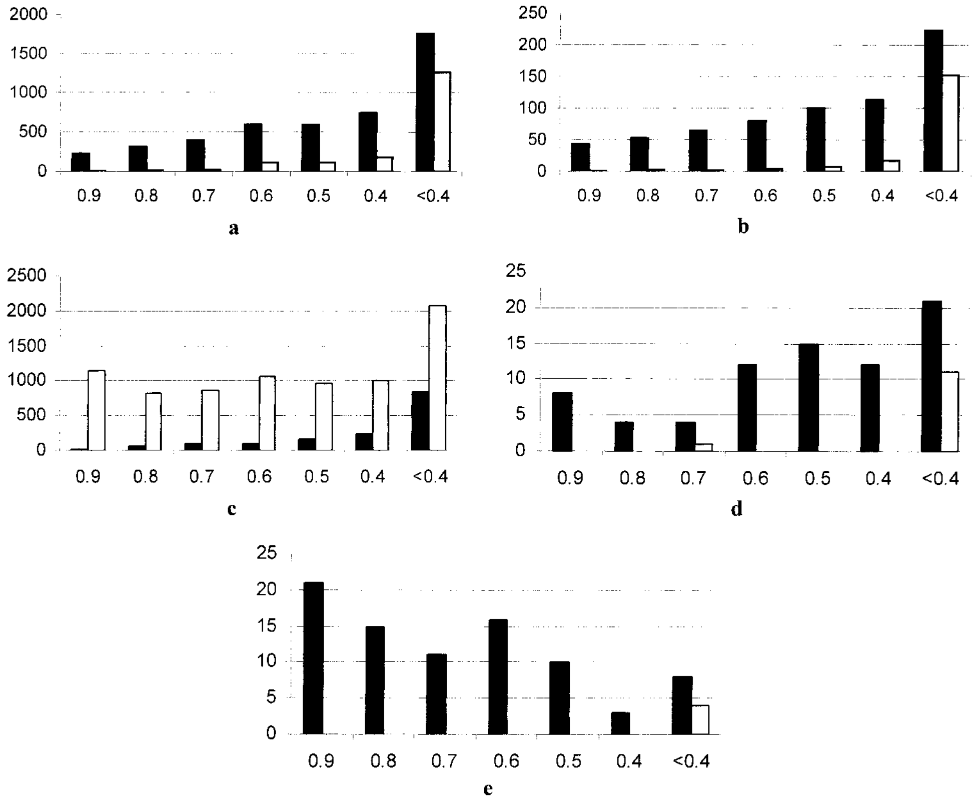

Figure 2. Distribution of predicted scores Pa for drugs (black) and nondrugs (white): a, WDI/ACD training set and LRID test set (Table 4, no. 2); b, WDI/ACD training set and LR test set (Table 4, no. 3); c, WDI/ACD training set and ND test set (Table 4, no. 4); d, WDI/ACD training set and TOP-100 test set (Table 4, no. 5); e, LR/ND training set and TOP-100 test set (Table 4, no. 11).

Table 3. Entries Excluded from Evaluation (Biologicals)

drug	indication	comment
Activase	clot dissolver	tissue plasminogen activator
Clexane/Lovenox	anticoagulant	low MW heparin
Engerix-B	immunostimulant	hepatitis B vaccine
Epogen	erythropoiesis	glycoprotein
Eprex/Procit	erythropoiesis	glycoprotein
Genotropin	growth hormone	protein
Humatrope	growth hormone	
Humulin	diabetes	insulin
Intron_A	biological response modifier	low MW glycoprotein
Neupogen	biological response modifier	protein
Ortho-Novum	oral contraceptive	
Primaxin	antiinfective	

Table 4. Quality of Discriminating between Drugs and Nondrugs by Different Methods

no.	traing set/ref	procedure	test set	false classification (%)
1	WDI/ACD	LOO c-v[a]		20.1[b]
2	WDI/ACD		LRID	26.6
3	WDI/ACD		LR	21.5
4	WDI/ACD		ND	16.2
5	WDI/ACD		TOP-100	12.5
6	Sadowski and Kubinyi[3]		WDI	23.0
7	Sadowski and Kubinyi[3]		ACD	17.0
8	Ajay[2]		MDDR	20.0
9	Ajay[2]		CMC	10.0
10	LR/ND	LOO c-v[a]		10.1[b]
11	LR/ND		TOP-100	4.5

[a] LOO c-v: leave-one-out cross-validation. [b] MEP: maximal error of prediction in LOO cross validation.

compounds (LR) and represent real drugs. Their molecular structures were again checked for presence in the training set (111 compounds), and 208 compounds were removed as being invalid. A total of 864 structures were calculated, and 678 (78.5%) compounds were predicted as drugs and 186 (21.5%) as nondrugs (Table 4; no. 3). It is obvious that the fraction of compounds classified as drugs is higher in comparison with the first test set. This can be explained by a more objective definition of drug and nondrug for the second test set, which provides better recognition of real drugs from the suggested drugs of WDI.

Evaluation of PASS vs "Nondrugs". The third evaluation set (ND) included 9737 compounds from different sources carefully selected as nondrugs according to the criteria discussed above. After the same filtering procedure, 9484 compounds were left for prediction. A total of 7950 compounds (83.8%) were predicted as nondrugs and 1534 (16.2%) compounds as drugs (Table 4; no. 4). These results show that cleaning of the test set gave a higher prediction accuracy.

Evaluation of PASS vs Drugs from the Top-100 List. As we suggested that most of drugs from this list may be also included into the WDI set, all predictions were carried out under exclusion of the equivalent compounds from the training set. For 88 compounds remaining from the list of top-100 prescription pharmaceuticals, 77 compounds (87.5%) were predicted as drugs and 11 (12.5%) were predicted as nondrugs (Table 4; no. 5).

Evaluation of PASS with the Cleaned Training Set. It was interesting to see if the cleaning of the training set could also increase the accuracy of the PASS prediction. Therefore, we trained PASS with a new drug/nondrug SAR-base represented by the test sets LR and ND. The results of the LOO cross-validation are listed in the Table 4; no. 10. It is obvious that the accuracy of prediction is about 90%. That is significantly higher than in the WDI/ACD training procedure used in the first training set.

The results of prediction for the 88 compounds from the list of top-100 prescription pharmaceuticals were even better than in the LOO cross-validation. A total of 84 compounds (95.5%) were predicted as drugs, while only four compounds (4.5%) were predicted as nondrugs.

In Figure 2 the distributions of the numbers of drugs/nondrugs predicted with different training and test sets are presented versus the value of the PASS score Pa, which represents the estimated probability of compound belonging to the class of "drugs". It is clear that the discriminating ability of PASS is significantly higher in case of the cleaned training set, as it was obviously demonstrated for the test set of the top-100 prescription drugs.

Conclusions

The discrimination between drug and nondrug is facing three problems: (i) not well-defined databases, (ii) choice of a method to discriminate, and (iii) the selection of appropriate descriptors.

The widely used databases for the discrimination between drugs and nondrugs are relatively noisy: some compounds assigned as drugs are nondrugs in reality and vice versa. Since this problem lies in the nature of the complex term "drug-likeness", there seems no simple way to overcome the underlying problem.

Our experiments provide the evidence that information-guided selection of the data sets gives higher accuracy in discrimination between the classes of drug-like compounds and nondrugs. The high value of prediction accuracy shows that the chemical descriptors and algorithms used in PASS provide highly robust structure−activity relationships and reliable predictions on this basis. Compared to other methods applied in the field, the direct benchmark undertaken with this paper showed that the results obtained with PASS are in good accordance with these approaches.

Since no specific adaption of the prediction scheme implemented in the PASS program was required, the advantage of the PASS approach lies in the fact that only two annotated data pools for drug and nondrug cases are necessary to allow a reliable prediction of discrimination of given features. So the PASS methodology opens the door to include more specific drug information in order to get a more specific discrimination. This may also be extended to physical−chemical properties as well as the interplay of those properties with dedicated pharmacological properties.

Acknowledgment. We are sincerely grateful to Jens Sadowski (AstraZeneca) for providing us with subsets from WDI and ACD, which were used as an initial training set for PASS in this work.

Appendix: Mathematical Method

Abbreviations:

n is the total amount of compounds in the training set.

n_i is the amount of compounds, containing descriptor i.

n_j is the amount of compounds, revealing activity j.

n_{ij} is the amount of compounds, containing descriptor i and revealing activity j.

$p_j = n_j/n$ is the estimate of the a priori probability of activity j.

$p_{ij} = n_{ij}/n_i$ is the estimate of the conditional probability of the activity j for the descriptor i.

m is the number of descriptors for the compound under prediction.

$r_i = n_i/(n_i + 0.5/m)$ is the regulating factor.

Pr_j is the initial estimate of the probability of the activity j for the compound under prediction.

LOO is leave-one-out procedure: for each compound in the training set, the values n, n_i, n_j, n_{ij} are changed for $n - 1$, $n_i - 1$, and $n_j - 1$, $n_{ij} - 1$ when one is active, and the estimates Pr_j are calculated.

Algorithm of prediction:

For the compound under prediction, the structure descriptors are generated.

For each activity, the following values are calculated:

$u_j = \sum_i \arcsin\{r_i(2p_{ij} - 1)\}$

$u_{0j} = \sum_\mathrm{i} \arcsin\{r_i(2p_j - 1)\}$

$s_j = \sin(u_j/m)$

$s_{0j} = \sin(u_{0j}/m)$

$\mathrm{Pr}_j = (1 + (s_j - s_{0j})/(1 - s_j s_{0j}))/2$

Validation criterion:

For each compound in the training set, the LOO estimates of Pr_j are calculated:

$\mathrm{EF}_j(\mathrm{CP})$ is the estimate of the first kind of error probability.

$\mathrm{ES}_j(\mathrm{CP})$ is the estimate of the second kind of error probability.

CP is the cutting point.

The first kind of error is fixed when the compound under prediction actually is active but $\mathrm{Pr}_j < \mathrm{CP}$.

The second kind of error is fixed when the compound under prediction is inactive and $\mathrm{Pr}_j > \mathrm{CP}$.

For each activity, the estimates of $\mathrm{EF}_j(\mathrm{CP})$ and $\mathrm{ES}_j(\mathrm{CP})$ are calculated.

The cutting points $\mathrm{CP}_j{}^*$ which gives equality:

$\mathrm{EF}_j(\mathrm{CP}_j{}^*) = \mathrm{ES}_j(\mathrm{CP}_j{}^*)$ are calculated.

The maximal error of prediction MEP is as follows:

$\mathrm{MEP}_j = \mathrm{EF}_j(\mathrm{CP}_j{}^*) = \mathrm{ES}_j(\mathrm{CP}_j{}^*)$

Results of prediction:

The probability to be active is $\mathrm{Pa}_j = \mathrm{EF}_j(\mathrm{Pr}_j)$.

The probability to be inactive is $\mathrm{Pi}_j = \mathrm{ES}_j(\mathrm{Pr}_j)$.

Pa (Pi) can be considered as the probability of the first (second) kind of errors for the compound under prediction or as the probability of the compound belonging to classes of active (inactive) compounds, respectively.

References

(1) Cummins, D. J.; Andrews, C. W.; Bentley, J. A.; Cory, M. Molecular Diversity in Chemical Databases: Comparison of Medicinal Chemistry Knowledge Bases and Databases of Commercially Available Compounds. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 750−763.

(2) Ajay; Walters, W. P.; Murcko, M. A. Can We Learn To Distinguish between "Drug-like" and "Nondrug-like" Molecules? *J. Med. Chem.* **1998**, *41*, 3314−3324.

(3) Sadowski, J.; Kubinyi, H. A scoring Scheme for Discriminating between Drugs and Nondrugs. *J. Med. Chem.* **1998**, *41*, 3325−3329.

(4) Gillett, V. J.; Willett, P.; Bradshaw, J. Identification of Biological Activity Profiles Using Substructural Analysis and Genetic Algorithms. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 165−179.

(5) Ghose, A. K.; Viswanadhan, V. N.; Wendolowski, J. J. A Knowledge-Based Approach in Designing Combinatorial and Medicinal Chemistry Libraries for Drug Discovery. 1. A Qualitative and Quantitative Characterization of Known Drug Databases. *J. Comb. Chem.* **1999**, *1*, 55−68.

(6) Blake, J. F. Chemoinformatics − predicting the physicochemical properties of "drug-like" molecules. *Curr. Opin. Biotechnol.* **2000**, *11*, 104−107.

(7) Teague, S. J.; Davis, A. M.; Leeson, P. D.; Oprea, T. The design of leadlike combinatorial libraries. *Angew. Chem., Int. Ed.* **1999**, *38*, 3743−3748.

(8) Oprea, T. I. Property distribution of drug-related chemical databases. *J. Comput.-Aided Mol. Des.* 2000, *14*, 251−264.

(9) Wagener, M.; van Geerestein, V. J. Potential drugs and nondrugs: Prediction and identification of important structural features. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 280−292.

(10) Clark, D. E.; Picket, S. D. Computational methods for the prediction of "drug-likeness". *Drug Discovery Today.* **2000**, *5*, 49−58.

(11) Frimurer, Th.; Bywater, R.; Naerum, L.; Lauritsen, L. N.; Brunak, S.; Improving the odds in discriminating "drug-like" from "non drug-like" compounds. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1315−1324.

(12) Ajay; Bemis, G. W.; Murcko, M. A. Designing Libraries with CNS Activity. *J. Med. Chem.* **1999**, *42*, 4942−4951.

(13) ACD: Available Chemicals Directory, Version 2/97, MDL Information Systems, 1997.

(14) CMC: Comprehensive Medicinal Chemistry, Version 1/97, MDL Information Systems, 1997.

(15) WDI: World Drug Index, Version 2/96; Derwent Information, 1996.

(16) MDDR: MDL Drug Report, Version 2/97; MDL Information Systems, 1997.

(17) Walters, W. P.; Ajay; Murcko, M. A. Recognizing Molecules with Drug-Like Properties. *Curr. Opin. Chem. Biol.* **1999**, *3*, 384−387.

(18) Filimonov, D. A.; Poroikov, V. V.; Karaicheva, E. I.; Kazaryan, R. K.; Boudunova, A. P.; Mikhailovsky, E. M.; Rudnitskih, A. V.; Goncharenko, L. V.; Burov, Yu. V. Computer-Aided Prediction of Biological Activity Spectra of Chemical Substances on the Basis of Their Structural Formulae: Computerized System PASS. *Exp. Clin. Pharmacol. (Rus)* **1995**, *58*, 56−62.

(19) Filimonov, D. A.; Poroikov, V. V. PASS: Computerized prediction of biological activity spectra for chemical substances. In *Bioactive Compound Design: Possibilities for Industrial Use*; BIOS Scientific Publishers: Oxford, 1996, 47−56.

(20) Poroikov, V. V.; Filimonov, D. A.; Stepanchikova, A. V.; Boudunova, A. P.; Shilova, E. V.; Rudnitskih, A. V.; Selezneva, T. M.; Goncharenko, L. V. Optimization of synthesis and pharmacological testing of new compounds based on computerized prediction of their biological activity spectra. *Chim.-Pharm. J. (Rus)* **1996**, *30*, 20−23.

(21) Web site: http://www.ibmh.msk.su/PASS.

(22) Poroikov, V. V.; Filimonov, D. A.; Boudunova, A. P. Comparison of the Results of Prediction of the Spectra of Biological Activity of Chemical Compounds by Experts and the PASS System. *Autom. Doc. Math. Linguist.* **1993**, *27*, 40−43.

(23) Website: http://www.vei.co.uk/chemweb/library/lecture17/slideroom_babaev/transcript.html.

(24) Filimonov, D. A.; Poroikov, V. V.; Borodina, Y.; Gloriozova, T. Chemical Similarity Assessment trough Multilevel Neighborhoods of Atoms: Definition and Comparison with the Other Descriptors. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 666−670.

(25) Cipsline, Correlates in Pharmacostructures Online, Version 2/2000; Prous Science: 2000.

(26) *Pharma Business* **1996**, July/August, 18−53.

JM0010670