Robustness of Biological Activity Spectra Predicting by Computer Program PASS for Noncongeneric Sets of Chemical Compounds

V. V. Poroikov,*,† D. A. Filimonov,† Yu. V. Borodina,† A. A. Lagunin,† and A. Kos‡

Institute of Biomedical Chemistry RAMS, Pogodinskaya Str., 10, Moscow 119832, Russia, and AKos Consulting & Solutions GmbH, Rossligasse 2, CH-4125 Riehen, Switzerland

Received March 1, 2000

The computer system PASS provides simultaneous prediction of several hundreds of biological activity types for any drug-like compound. The prediction is based on the analysis of structure−activity relationships of the training set including more than 30000 known biologically active compounds. In this paper we investigate the influence on the accuracy of predicting the types of activity with PASS by (a) reduction of the number of structures in the training set and (b) reduction of the number of known activities in the training set. The compounds from the MDDR database are used to create heterogeneous training and evaluation sets. We demonstrate that predictions are robust despite the exclusion of up to 60% of information.

INTRODUCTION

Traditional QSAR and 3D molecular modeling are successful at predicting the biological activities for chemical structures, provided they work with small number of types of activity and usually stay in the same chemical series.[1−5] Similarity searching[6,7] and clustering methods[7,8] can be used to separate compounds into structural groups[9] and for the prediction of biological activities and compound selection.[10] In reality many biologically active compounds possess several types of activity. The computer system PASS *(Prediction of Activity Spectra for Substances)*[11−14] predicts simultaneously several hundred various biological activities. These are pharmacological effects, mechanisms of action, mutagenicity, carcinogenicity, teratogenicity, and embryotoxicity. PASS prediction is based on the analysis of structure−activity relationships of the training set including a great number of noncongeneric compounds with different biological activities. PASS once trained is able to predict many types of activity for a new substance. The example of prediction for known cerebrotonic drug Cavinton (Vinpocetin) is shown in Table 1. Many types of activity known for this drug are predicted. Some new ones (Multiple sclerosis treatment, Antineoplastic enhancer, etc.) display the directions for further study of Cavinton.

We had a long-term experience with PASS applications to select probable biologically active substances from databases of available samples and to arrange the experimental testing of compounds under study. It was shown that the mean accuracy of prediction with PASS is about 86% in leave-one-out cross-validation.[14] PASS prediction accuracy exceeds more than 3 times the expert's guess-work for an independent set of 33 different compounds studied as pharmacological agents.[15] Recently PASS was tested in blind mode by 9 scientists from 8 countries. The mean accuracy of prediction was shown to be 82.6%.[16]

Table 1. Some Predicted Biological Activities for Cavinton[a]

no.	Pa	Pi	activity	expt
1	0.929	0.004	peripheral vasodilator	
2	0.900	0.000	multiple sclerosis treatment	
3	0.855	0.005	vasodilator	+
4	0.844	0.003	abortion inducer	+
5	0.812	0.001	antineoplastic enhancer	
6	0.760	0.006	coronary vasodilator	+
7	0.732	0.007	spasmogenic	
8	0.700	0.036	antihypoxic	+
9	0.650	0.004	lipid peroxidase inhibitor	+
10	0.648	0.008	cognition disorders treatment	+
11	0.656	0.021	antiischemic	+
12	0.577	0.013	acute neurologic disorders treatment	+
13	0.540	0.039	spasmolytic	+
14	0.519	0.026	antianginal agent	
15	0.486	0.037	antihypertensive	+
16	0.449	0.035	antiarrhythmic	+
17	0.432	0.063	sympatholytic	
18	0.438	0.077	sedative	+
19	0.500	0.152	antiinflammatory, pancreatic	
20	0.328	0.020	antidepressant, imipramin-like	
21	0.300	0.010	thrombolytic	+
22	0.342	0.075	psychotropic	+
23	0.276	0.023	alpha 2 adrenoreceptor antagonist	+

[a] Pa and Pi are the probabilities of belonging to the classes of active and inactive compounds, respectively.

The accuracy of PASS prediction depends on several factors:[12] 1. description of the chemical structure, 2. description of the biological activity, 3. mathematical methods, 4. quality of the training set, 4.1. activity data, 4.2. structure data, and 5. errors in the data.

Quality of the training set seems to be the most critical factor in PASS approach. Really, the training set includes various compounds, which are investigated on various types of activity. Information about *each* compound is taken into account to predict *each* type of activity. If a compound from the training set was not investigated on a given type of activity, it is considered as inactive. However, we cannot be sure that all these compounds are really inactive. Therefore, there is the incompleteness of activity data in the training set. On the other hand, only part of known compounds is

* Corresponding author phone: (7-095) 245-2753; e-mail: vvp@ibmh.msk.su.
† Institute of Biomedical Chemistry RAMS.
‡ AKos Consulting & Solutions GmbH.

10.1021/ci000383k CCC: $19.00 © 2000 American Chemical Society
Published on Web 09/30/2000

included into training set. This is incompleteness of structural data. Is able PASS to cope with such incomplete data in the training set and to give a reasonable prediction for a new compound without retraining? Should the complete *spectrum of activity* for each compound in the training set be known for providing accurate prediction, or a partial knowledge is quite enough?

The purpose of the present work is to determine how robust are the results of prediction depending on the incompleteness of the training set. We investigate the influence on the accuracy of predicting types of activity with PASS by (a) reduction of the number of structures in the training set and (b) reduction of the number of known activities in the training set.

GENERAL DESCRIPTION OF PASS METHOD

Basic elements of PASS include the following: presentation of biological activity, description of chemical structure, training set of compounds, training procedure, and prediction procedure. The current version of PASS differs essentially from the previous.[11]

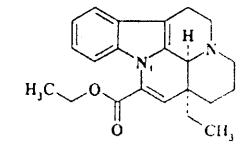
Biological Activity. Biological activities in PASS are described qualitatively: presence or absence. List of activity types that have been found for each compound represents the biological activity data in the training set. This list for current version of PASS is available via Internet.[14]

Chemical Structure Description. In our paper published recently[17] we described the substructure descriptors called "Multilevel Neighborhoods of Atoms" (MNA). MNA descriptors are based on structure representation, which does not specify the bond types and includes hydrogens according to valence and partial charge of atoms. MNA descriptors are generated as a recursively defined sequence: 1. zero-level MNA descriptor for each atom is the mark A of the atom itself and 2. any next-level MNA descriptor for each atom is the substructure notation $A(D_1D_2 \ldots D_i \ldots)$, where Di is the previous-level MNA descriptor for the i-th immediate neighbors of the atom.

This iterative process can be continued enclosing second, third, etc. neighborhoods of each atom. It is important to emphasize that the atom mark may include not only the atom type but also any additional information about the atom, for example, its belonging to a cycle or a chain. A structure of molecule is represented in PASS as a set of the first- and second-level MNA descriptors. In second-level MNA descriptors we use the mark "-" of belonging to a chain. Figure 1 shows the structure and MNA descriptors of Cavinton.

Structure equivalence is the important feature of PASS concept. The structures are considered as equivalent if they have the same molecular formulas and the same MNA descriptors set. Only unique structures are included in the training set. Since MNA descriptors do not represent the stereochemical peculiarities of a molecule, the compounds, which have only stereochemical differences in the structure, are formally considered as the equivalent.

Training Set. The prediction is based on the analysis of the training set of biologically active compounds. For each compound from the training set we store MNA descriptors and a list of activity types. Every unique MNA descriptor is included into the descriptors dictionary.



MNA/1	MNA/2
HC	C(C(CC-H-H)C(CN-H-H)-H(C)-H(C))
CHHHC	C(C(CC-H-H)N(CCC)-H(C)-H(C))
CHHCC	C(C(CC-H)C(CC-H)-H(C))
CHHCN	C(C(CCN-H)C(CC-H)C(CC-H-H)-C(C-H-H-C))
CHHCO	C(C(CC-H)N(CCC)-C(C-O-O))
CHCC	C(C(CCC-C)C(CC-H-H)-H(C)-H(C))
CHCCN	C(C(CCC-C)C(CN-C)-H(C))
CCCC	C(C(CCC-C)C(CCN)N(CCC)-H(C))
CCCCC	C(C(CCC)C(CN-H-H)-H(C)-H(C))
CCCN	C(C(CCC)C(CC-H)-H(C))
CCOO	C(C(CCN)C(CC-H)-H(C))
NCCC	C(C(CCC)C(CC-H)N(CCC))
OC	C(C(CCC)C(CCN-H)N(CCC))
OCC	C(C(CCC)C(CCN)C(CC-H))
	C(C(CCC)C(CCN)C(CC-H-H))
	N(C(CCN-H)C(CN-H-H)C(CN-H-H))
	N(C(CCN)C(CCN)C(CN-C))
	-H(C(CC-H))
	-H(C(CCN-H))
	-H(C(CC-H-H))
	-H(C(CN-H-H))
	-H(-C(-H-H-H-C))
	-H(-C(C-H-H-C))
	-H(-C(-H-H-C-O))
	-C(-H(-C)-H(-C)-H(-C)-C(C-H-H-C))
	-C(-H(-C)-H(-C)-H(-C)-C(-H-H-C-O))
	-C(-H(-C)-H(-C)-C(-H-H-H-C)-O(-C-C))
	-C(C(CCC-C)-H(-C)-H(-C)-C(-H-H-H-C))
	-C(C(CN-C)-O(-C)-O(-C-C))
	-O(-C(C-O-O))
	-O(-C(C-O-O)-C(-H-H-C-O))

Figure 1. List of the MNA descriptors for Cavinton. MNA/1 and MNA/2 are descriptors of the first- and second-level, respectively.

In the current version of PASS the training set consists of about 35000 biologically active compounds compiled from scientific literature, in-house and commercial databases. The descriptor's dictionary contains about 36000 MNA descriptors. In different published sources biological activities are named by different terms. In PASS this information is represented in a standard form that combines all biological activity data about equivalent compounds collected from many sources. The number of different types of activity exceeds 800, but many of them are represented by less than 3 compounds. Total "activity spectrum", i.e., the list of predictable types of biological activity, includes more than 500 items.

In this work we use different subsets of compounds from MDDR database as *training sets*. A more detailed description of the training sets is given below.

Training Procedure. For every type of activity we generate the structure−activity relationships in the following way: n is the total amount of compounds in the training set; n_i is the amount of compounds, containing MNA descriptor i; n_j is the amount of compounds, containing the type of activity j in activity spectrum; and n_{ij} is the amount of compounds, containing MNA descriptor i and the type of activity j. For j-th type of activity we calculate the initial estimates t_j for each compound in the training set.

Each compound is excluded from the training set once; values n, n_i, n_j, and n_{ij} are recalculated from the remaining compounds, and the following values are calculated

$$s_j = \mathrm{Sin}\left(\sum_i \mathrm{ArcSin}(r_{i*}(2_*p_{ij} - 1))/m\right)$$

$$s_{0j} = \mathrm{Sin}\left(\sum_i \mathrm{ArcSin}(r_{i*}(2_*p_j - 1))/m\right)$$

$$t_j = (1 + (s_j - s_{0j})/(1 - s_{j*}s_{0j}))/2$$

where the summation is taken over all MNA descriptors of a given compound and m is the total number of descriptors in it, $r_i = n_i/(n_i + 0.5/m)$ is the regulating factor, $p_j = n_j/n$ is the estimation of the a priori probability of the type of activity j, $p_{ij} = n_{ij}/n_i$ is the estimation of conditional probability of the type of activity j for the MNA descriptor i. A priory probability p_j estimates the chance to find a compound with type of activity j by random search. Conditional probability p_{ij} estimates the same chance under the condition that the search is done among the compounds containing the descriptor i.

Estimates t_j for active compounds are sorted in ascending order; the estimates t_j for inactive compounds are sorted in descending order. The conditional expectations A_j and I_j are calculated as

$$A_j(F) = \sum_p Pr(p-1, n_j-1, F)t_{jp}$$

$$I_j(F) = \sum_q Pr(q-1, n-n_j-1, F)t_{jq}$$

where $Pr(m, n, F) = C_n^m F^m(1-F)^{n-m}$ is the binomial distribution, $C_n^m = n!/m!(n-m)!$ is the binomial coefficient, p is an active compound and q is an inactive compound, and F is in the range [0, 1]. It is clear that $A_j(F)$ and $I_j(F)$ are the calculated quantiles of the probability distributions of the initial estimates. Functions $A_j(F)$ and $I_j(F)$ together with values n, n_i, n_j, and n_{ij} represent the SAR data for j-th type of activity.

Prediction Procedure. To estimate the activity spectrum for a new compound (C) its MNA descriptors are generated. For each type of activity (j) the value of t_j^C is calculated. The probabilities of presence Pa_j and absence Pi_j of j-th activity type in the compound are calculated according to the next equations:

$$A_j(\mathrm{Pa}) = t_j^C; \quad I_j(\mathrm{Pi}) = t_j^C$$

In other words, Pa and Pi are the probabilities of belonging to the classes of active and inactive compounds, respectively.

The result of prediction for a new compound is the *activity spectrum,* which is the ranked list of activity types with estimated Pa and Pi values. The ranking is executed on descending order of Pa−Pi; thus, more probable activity types are at the top of predicted spectrum. Compound is considered as active if Pa−Pi exceeds the cutoff value. By default we use cutoff of Pa − Pi = 0, but any user may accept his own cutoff value, for example 0.5. Table 1 shows the top part of predicted activity spectrum for Cavinton.

Validation of Prediction Accuracy. To estimate the accuracy of prediction for *evaluation set* of compounds (i.e. set of compounds with known biological activity, not included into the training set) we use the next procedure.

MNA descriptors are generated for each compound in the *evaluation set.* For jth type of activity t_j value is calculated. To estimate the quality of prediction of jth type of activity we use the expression called the Independent Accuracy of Prediction

$$\mathrm{IAP}_j = \mathrm{N}\{t_j^{act} > t_j^{inact}\}/(n_{act} * n_{inact})$$

where $\mathrm{N}\{t_j^{act} > t_j^{inact}\}$ is the number of cases when t_j for an active compound is greater than t_j for an inactive compound, when all pairs of active and inactive compounds in the evaluation set are compared; n_{act} and n_{inact} are the numbers of active and inactive compounds in the evaluation set.

This criterion is defined as "independent" because it does not depend on any additional assumptions concerning the parent population and risk function.

DESIGN OF THE EXPERIMENT

Database Used in This Study. We use the compounds from MDDR[18] (MDL Drug Data Report) as it is one of the largest collections of structures, which include information about biological activity. MDDR 97.2 from MDL Information Systems, Inc.[18] contains the information about 87486 pharmacological agents compiled mainly from the patent literature. About 92% of them are under biological testing, 7% are drug candidates, and about 1% of the compounds are registered drugs. Every compound in MDDR has one or several records in the field "activity class", indicating that compound is related to certain therapeutic area. However, not every one was really tested in experiments. Those substances, for which biological activity was studied in detail, have records in the field "Action", such as experimental data on activity, LD50, IC50, Ki, etc.

We considered only those compounds, which have some records in the field "Action". These are called the *principal compounds.* For example, compound A-83094A is described in the field "activity class" as *"Antibiotic"* and in the field "Action" as *"Pyrrole-ether antibiotic produced by Streptomyces setonii, active in vitro against Gram-positive bacteria as well as coccidia. LD50 =196.4 mg/kg i.p. and 630 mg/kg p.o. in mice".* So it was included into our study. Compound MUREIDOMYCIN A contains the word *"Antibiotic"* in the field "activity class", and nothing in the field "Action". This compound was not used in our study.

Following this rule, we have prepared a subset from MDDR that includes 20561 *principal compounds.*

Activities Considered in This Study. The types of activity were selected which represent specific pharmacological effects or molecular mechanisms of actions. Some unspecified terms, such as diagnostic agent, chemical delivery system, pharmacological tool, etc., were not considered. When synonyms encountered, the common term was chosen. Table 2 shows the examples of how the types of activity were constructed from terms used in MDDR.

In this way a list of 517 types of activity was obtained. Since we planned to exclude a significant part of information from the training sets in the frame of our experiment, only those types of activity were chosen for which more than 80 *principal compounds* were found in MDDR. Based on this criterion 124 types of activity were selected. The majority of them is represented by compounds of various chemical classes, but there are some activity categories in which the diversity is limited by compounds of the same chemical series (e.g. "Antibiotic Carbapenem-like", "Antibiotic Quinolone-like").

Descriptors Database. We exported the set of *principal compounds* as an SDFile containing only data on structures and activities. We excluded the entries, containing undetermined structures (monoclonal antibodies, vaccines, etc.), undefined R, X-groups, atoms with incorrect valencies or polypeptides (insulin, regulatory peptide, etc.). For each

Table 2. Examples of Activities Used in This Study

activity	MDDR terms
5 hydroxytryptamine 1D agonist	5 HT1D agonist
alpha 2 adrenoreceptor antagonist	adrenergic (alpha2) blocker adrenoceptor (alpha2) antagonist
antibacterial	antibacterial, topical antibacterial
antibiotic beta lactam-like	monocyclic beta-lactam lactam (beta) enhancer lactam (beta) antibiotic
benzodiazepine agonist	benzodiazepine benzodiazepine agonist
choleretic	cholagogue choleretic
corneal wound healing stimulator	wound healing agent corneal wound healing stimulator
male reproductive disfunction treatment	male sexual disorders, agent for antiinfertility, male
psychostimulant	stimulant, central centrally acting agent
renal disease treatment	crf antagonist renal failure, agent for
spasmolytic	spasmolytic antispastic
thyroid hormone agonist	thyromimetic thyroid hormone

structure in the SDFile we build the MNA descriptors, which can also be called keys, and store them in a database called *SARBase*. In this way we generate about 30,000 descriptors and arrange them as a binary file in *SARBase*. The *SARBase* contains 18977 unique compounds with their activities.

Creation of the Training and Evaluation Sets. The set of compounds in *SARBase* was 50 times divided at random into two equal subsets. The first subset was used as the training set, the second one as the evaluation set and vice versa. So we prepared 100 pairs of the training and evaluation sets.

Cross-Validation. We carried out the leave-one-out (LOO) procedure for each of 100 training sets of compounds. Every compound was consequently excluded from the set, and its types of activity were predicted by PASS trained on the other compounds. Then the IAP value for each type of activity was calculated.

Simulation of Incompleteness of Activity Data. The crucial question was how robust are the prediction results depending on the quality of the training set. In particular we wanted to evaluate how the accuracy of prediction with PASS is influenced by leaving out activity data for a number of compounds. The result is that some compounds have no activity data any more. The other ones, which had originally several types of activity, still have some activity data. We proposed the following experimental procedure.

(1) Train PASS using the *initial* training set and run the prediction for the evaluation set.

(2) Exclude from the initial training set at random *20% of total number of activities*.

(3) Retrain PASS and run the prediction for the evaluation set again.

(4) Repeat step 3 excluding 40, 60, and 80% of total number of activities from the initial training set.

(5) Compare the results of predictions, based on the training set with different degree of incompleteness of activity data.

Simulation of Incompleteness of Structure Data. The purpose of this test was to evaluate how structural incompleteness of the training set influences the accuracy of predicting with PASS. For each of 100 pairs of the training and evaluation sets we carried out the experiment similar to the previous one but leaving out structures instead of activities.

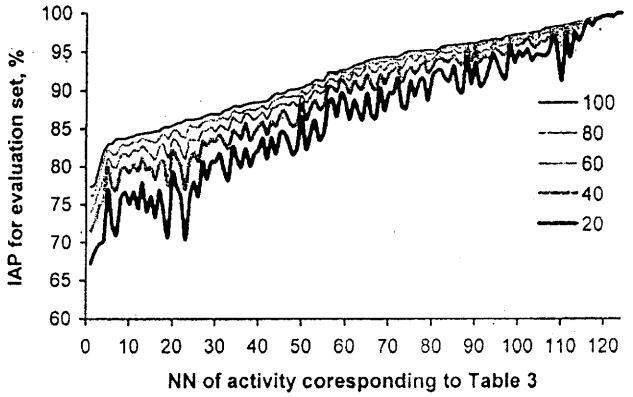


Figure 2. Influence of incompleteness of activity data in the training set on the accuracy of prediction. The legend shows the percentage of activity data in the training set.

RESULTS AND DISCUSSION

Table 3 shows for each type of activity, the number of compounds in the whole set, average results of prediction for 100 evaluation sets, obtained by PASS trained on respective training sets, and the average IAP_{LOO} calculated by LOO procedure for each training set. The last line of the table shows the mean value for IAP over all types of activity, IAP_m.

The data in Table 3 are sorted in ascending order of IAP. The best results are obtained for the compounds with the following actions: antibiotic carbapenem-like (99.96%), antibiotic quinolone-like (99.94%), and antibiotic macrolide-like (99.75%). The worst but still satisfactory accuracy of prediction is observed for anticerebroischemic (77.39%), antiarthritic (77.74%), and septic shock treatment (79.89%) actions.

In general, Table 3 demonstrates that mean IAP and IAP_{LOO} values are very close to one another (91.95 and 91.70%, respectively). This means that the leave-one-out approach can be used to estimate the accuracy of prediction.

Influence of Incompleteness of Activity Data on the Quality of Prediction with PASS. Figure 2 shows how IAP values for each type of activity change depending on incompleteness of activity data in the training set. The *x*-axis plots the numbers of types of activity corresponding to Table 3.

As one can see from Figure 2, IAP values are decreased depending on incompleteness of activity data for the majority of activity types. In general, the decrease of IAP value is greater for those types of activity, which have a smaller initial value of IAP. For example, IAP value for activity "diuretic" changes from 86.18 to 70.44 when the activity data in the training set are reduced from 100 to 20%, while IAP value for activity "antibiotic beta lactam-like" changes from 99.58 to 99.54.

The minima in the graph are caused by removing data from the types of activity, which are originally represented

Table 3. Independent Accuracy of Prediction

no.	activity	amt	IAP, %	IAP_{LOO}, %	no.	activity	amt	IAP, %	IAP_{LOO}, %
1	anticerebroischemic	154	77.39	77.21	64	peristaltic stimulant	127	93.58	93.82
2	antiarthritic	563	77.74	77.24	65	acetylcholine agonist	184	93.84	93.59
3	septic shock treatment	157	79.89	79.91	66	antiemphysemic	117	93.99	93.74
4	irritable bowel syndrome therapy	107	82.01	81.32	67	alpha adrenoreceptor antagonist	150	94.10	93.85
5	immunomodulator	626	82.99	82.71	68	lipoxygenase inhibitor	490	94.21	94.08
6	vasodilator	203	83.13	82.98	69	adrenalin antagonist	200	94.32	94.11
7	urinary incontinence treatment	82	83.72	82.68	70	5 hydroxytryptamine 1 antagonist	88	94.33	94.18
8	antipsoriatic	276	83.72	83.66	71	endothelin antagonist	134	94.34	94.40
9	antiinflammatory	962	83.76	83.40	72	calcium channel antagonist	331	94.34	94.13
10	antiulcerative	376	83.95	83.57	73	NMDA antagonist	247	94.62	93.82
11	antianginal	410	84.13	83.45	74	elastase inhibitor	127	95.03	94.98
12	sedative	85	84.38	83.72	75	5 hydroxytryptamine 2 antagonist	133	95.06	94.97
13	dermatologic	449	84.43	84.38	76	alpha 1 adrenoreceptor antagonist	87	95.08	94.81
14	mediator release inhibitor	112	84.52	84.25	77	antihistaminic	137	95.11	94.94
15	acute neurologic disorders treatment	610	84.59	83.60	78	thromboxane synthase inhibitor	114	95.16	94.93
16	spasmolytic	106	84.73	82.93	79	dopamine D2 aihtagonist	99	95.20	95.08
17	analgesic, nonopioid	407	84.82	84.39	80	H+/K+-transporting ATPase inhibitor	117	95.20	94.31
18	antiosteoporotic	109	85.27	85.25	81	leukotriene antagonist	372	95.21	95.26
19	antineoplastic enhancer	81	85.27	84.33	82	antiemetic	212	95.59	95.33
20	cognition disorders treatment	930	85.49	85.21	83	anticoagulant	169	95.73	95.53
21	antiobesity	114	85.74	85.28	84	antiacne	186	95.85	95.78
22	reverse transcriptase inhibitor	83	85.86	85.38	85	thromboxane antagonist	238	95.86	95.70
23	diuretic	125	86.18	85.15	86	aldose reductase inhibitor	161	95.88	95.98
24	antiprotozoal	166	86.23	85.58	87	androgen antagonist	87	95.94	95.79
25	lipid peroxidase inhibitor	117	86.26	86.17	88	antibacterial	1473	96.03	95.89
26	anticonvulsant	380	86.37	85.74	89	phosphodiesterase inhibitor	216	96.04	95.87
27	immunostimulant	109	86.52	86.29	90	5 hydroxytryptamine antagonist	473	96.07	95.97
28	ophthalmic drug	229	86.84	86.80	91	platelet activating factor antagonist	272	96.23	95.97
29	antineoplastic	2410	86.94	86.73	92	acetylcholinesterase inhibitor	102	96.39	96.36
30	immunosuppressant	276	86.95	86.62	93	phosphodiesterase IV inhibitor	128	96.43	96.08
31	antiallergic	1164	86.98	86.73	94	thrombin inhibitor	123	96.57	96.49
32	cardiotonic	779	87.39	87.01	95	acetyl CoA transferase inhibitor	232	96.60	96.42
33	antiparkinsonian	171	87.60	87.04	96	dopamine antagonist	204	96.60	96.49
34	antiviral	598	87.93	87.94	97	analgesic, opioid	169	96.89	96.67
35	analgesic	577	87.95	87.59	98	antimitotic	88	96.94	96.98
36	bronchodilator	320	87.99	87.21	99	5 hydroxytryptamine agonist	290	97.17	97.12
37	calcium regulator	94	88.18	88.04	100	antimetabolite	137	97.19	97.00
38	tumor necrosis factor antagonist	92	88.34	88.13	101	acetylcholine muscarinic agonist	138	97.23	96.89
39	antidiabetic	319	88.42	88.10	102	5 alpha reductase inhibitor	141	97.27	97.37
40	antidepressant	549	88.47	88.24	103	5 hydroxytryptamine 1A agonist	159	97.37	97.35
41	platelet aggregation inhibitor	783	88.57	88.31	104	adrenalin agonist	86	97.49	96.58
42	anti-HIV	693	88.63	88.46	105	5 hydroxytryptamine 1 agonist	250	97.55	97.51
43	antihelmintic	108	89.15	89.12	106	substance P antagonist	174	97.93	97.93
44	antiglaucomic	195	89.56	89.55	107	cholecystokinin antagonist	156	97.96	97.95
45	antihypertensive	1894	89.57	89.29	108	antibiotic	1301	98.07	98.05
46	gastric antisecretory	311	89.73	89.50	109	HIV-1 protease inhibitor	152	98.15	98.05
47	phospholipase inhibitor	118	90.07	88.98	110	squalene synthetase inhibitor	86	98.22	98.11
48	protein kinase C inhibitor	84	90.12	89.88	111	5 hydroxytryptamine 3 antagonist	203	98.36	98.33
49	antiviral (AIDS)	638	90.23	89.99	112	aromatase inhibitor	89	98.41	98.20
50	psychotropic	1492	90.39	90.27	113	GP IIb/IIIa antagonist	209	98.54	98.52
51	phospholipase A2 inhibitor	113	90.63	89.52	114	potassium channel activator	156	98.69	98.56
52	antiarrhythmic	373	90.86	90.37	115	angiotensin converting enzyme inhibitor	124	98.72	98.75
53	anxiolytic	710	91.33	91.09	116	prostaglandin agonist	94	99.19	99.19
54	antidiabetic symptomatic	200	91.33	91.36	117	HMG CoA reductase inhibitor	184	99.25	99.16
55	cyclooxygenase inhibitor	125	91.34	91.50	118	angiotensin II antagonist	465	99.44	99.41
56	chemoprotective	236	92.31	92.38	119	renininhibitor	218	99.58	99.56
57	antipsychotic	597	92.32	92.11	120	antibiotic beta lactam-like	655	99.58	99.57
58	prostate disorders treatment	194	92.38	92.20	121	antibiotic cephalosporin-like	315	99.65	99.65
59	protease inhibitor	127	92.65	92.83	122	antibiotic macrolide-like	109	99.75	99.73
60	antifungal	469	92.72	92.46	123	antibiotic quinolone-like	254	99.94	99.94
61	leukotriene synthesis inhibitor	115	92.90	93.10	124	antibiotic carbapenem-like	162	99.96	99.97
62	hypolipemic	812	93.22	93.11		IAP_m		91.95	91.70
63	antimigraine	187	93.39	93.12					

by a small number of compounds. This fact must have a larger influence on the result of prediction. For example, there are only 86 compounds in the SARBase for activity "squalene synthetase inhibitor", no. 110 in Table 3 and on the graph. The IAP value for "squalene synthetase inhibitor" changes from 98.21% to 91.32%.

Figure 3 shows the IAP_m, mean value for IAP over all types of activity, versus percentage of activity data in the training set. The *x*-axis plots the relative number of types of activity in the training set, while the *y*-axis plots IAP_m. The extreme right point on the graph shows the IAP_m calculated for the initial training set. Moving from right to left across

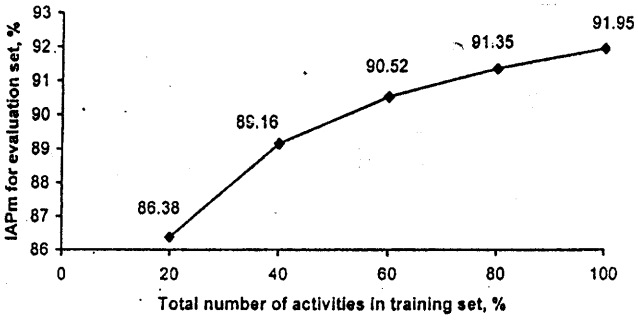


Figure 3. Mean accuracy of prediction versus percentage of activity data in the training set.

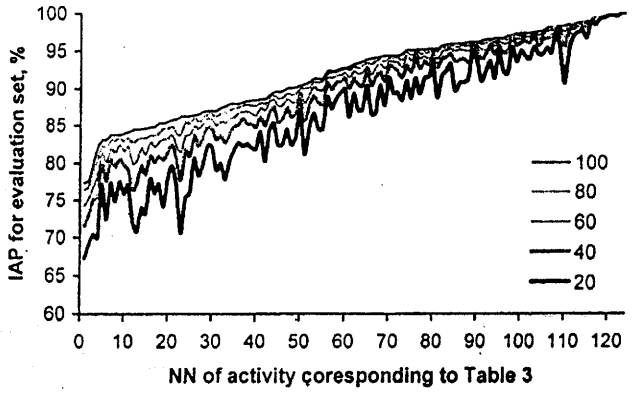


Figure 4. Influence of structural incompleteness of the training set on the accuracy of prediction. The legend shows the percentage of structures in the training set.

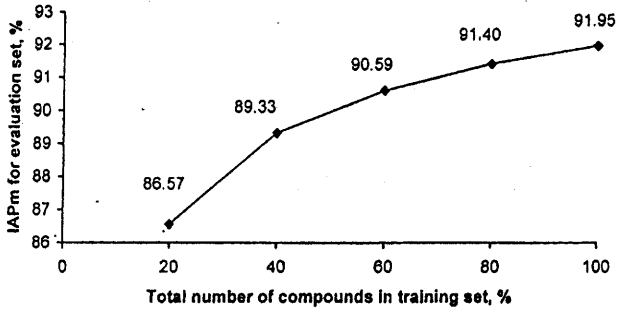


Figure 5. Mean accuracy of prediction versus percentage of compounds in the training set.

the graph corresponds to the reduction of the total number of activities in the training set and decreases in the mean prediction accuracy. However, even working with 40% of the available activities gives reasonably accurate predictions.

Influence of Structural Incompleteness on the Quality of Prediction with PASS. Figure 4 shows the change of IAP values for each type of activity depending on incompleteness of structure data in the training set. The *x*-axis plots the numbers of types of activity corresponding to Table 3.

The influence of structural incompleteness on the accuracy of prediction is similar to incompleteness of activity data. In general, the decrease of IAP value is greater if the initial value of IAP is smaller and the initial number of active compounds is less.

Figure 5 shows the IAP_m depending on the number of compounds in the training set. The *x*-axis plots the relative number of compounds in the training set, while the *y*-axis plots IAP_m.

The effect of reducing the number of structures on the accuracy of prediction is very similar to reducing the activities. In this particular case, a reduction of 60% of the data still gives reasonably accurate predictions. Such similarity is probably caused by the fact that the majority of compounds in MDDR have only one type of activity. So exclusion of the activity and exclusion of the compound with this activity causes similar change of the total number of activities in the training set.

CONCLUSIONS

We have shown that for a large set of compounds, like principal compounds from MDDR, the accuracy of prediction by PASS is still excellent for many types of activity, even when up to 60% of the information is left out. It means that chemical descriptors, biological activity representation, and mathematical methods used in PASS provide the robust approach to analyze SAR in large data sets.

The accuracy of prediction can be less if a new type of activity is encountered that is not well represented in the training set.

PASS, therefore, produces reasonably accurate results for many predictions without retraining the system for each special case.

ACKNOWLEDGMENT

We gratefully acknowledge the assistance of MDL Information Systems, Inc. for providing the Institute of Biomedical Chemistry RAMS with a license to ISIS and the database used in this study.

REFERENCES AND NOTES

(1) Hansch, C. Quantitative Structure–Activity Relationships and the Unnamed Science. *Acc. Chem. Res.* **1993**, *26*, 147–153.
(2) Apostolakis, J.; Caflisch, A. Computational Ligand Design. *Comb. Chem. High Throughput Screen* **1999**, *2*(2), 91–104.
(3) Lipnick, R. L. Correlative and Mechanistic QSAR Models in Toxicology. *SAR QSAR Environ. Res.* **1999**, *10*(2–3), 239–248.
(4) *3D QSAR in Drug Design: Theory, Methods and Application*; Kubinyi, H. Ed.; ESCOM: Leiden, 1993.
(5) *5. 3D QSAR in Drug Design: Recent Advances*; Kubinyi, H., Folkers, G., Martin, Y. C., Eds.; Kluwer/Escom: 1998.
(6) *Molecular Similarity in Drug Design*; Dean, P. M., Ed.; Chapman & Hall: London, 1995.
(7) Willet, P. *Similarity, and Clustering in Chemical Information Systems*; Research Studies Press Ltd: Letchworth, U.K., 1987.
(8) Barnard. J. M.; Downs, G. M. Clustering of Chemical Structures on the Basis of Two-Dimensional Similarity Measures. *J. Chem. Inf Comput. Sci.* **1992**, *32*, 644–649.
(9) Wild, D. J.; Blankey, C. J. Comparison of 2D Fingerprints Types and Hierarchy Level Selection Methods for Structural Grouping Using Ward's Clustering. *J. Chem. Inf Comnpul. Sci.* **2000**, *40*(1), 155–162.
(10) Brown, R. D.; Martin, Y. C. Use of Structure–Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572–584.
(11) Filimonov, D. A.; Poroikov, V. V.; Karaicheva, E. I.; Boudunova, A. P.; Shilova, E. V.; Rudnitskikh, A. V.; Seleznieva, T. M.; Goncharenko, L. V. Computer-Aided Prediction of Biological Activity Spectra of Chemical Substances on The Basis of Their Structural Formulae: Computerized System PASS. *Exptl. Clinical Pharmacol. (Russ.)* **1995**, *58* (2), 56–62.
(12) Filimonov, D. A.; Poroikov, V. V. In *Bioactive Compound Design: Possibilities For Industrial Use*; BIOS Scientific Publishers: Oxford, 1996; pp 47–56.
(13) Poroikov, V. V.; Filimonov, D. A.; Stepanchikova, A. V.; Kazarian, R. K.; Boudunova, A. P.; Mihailovskiy, E. M.; Rudnitskih, A. V.; Goncharenko, L. V.; Burov, Yu. V. Opimization of Synthesis and Pharmacological Testing of New Compounds Based on Computerized Prediction of Their Biological Activity Spectra. *Chim.-Pharm. J.*

(Russ.) **1996**, *30*(9), 20−23 (English translation by Consultants Bureau, New York: *Pharm. Chem. J.* **1996**, *30*(9), 570−573).
(14) http://www.ibmh.msk.su/PASS.
(15) Poroikov, V. V.; Filimonov, D. A.; Boudunova, A. P. In *Automatic Documentation and Mathematical Linguistics*; Allerton Press Inc.: New York, 1993; Vol. 27, No. 3, pp 40−43.
(16) http://wwwv.vei.co.uk/chemweb/library/lecturel17/slideroom_babaev/transcript.html.
(17) Filimonov, D. A.; Poroikov, V. V.; Borodina, Y.; Gloriozova, T. Chemical Similarity Assessment Trough Multilev.el Neighborhoods of Atoms: Definition and Comparison with The Other Descriptors. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 666−670.
(18) MDL Drug Data Report 97.2; MDL Information Systems, Inc.: 1997.

CI000383K