Available online at www.sciencedirect.com
SCIENCE DIRECT®
Mutation Research 586 (2005) 138–146
Genetic Toxicology and Environmental Mutagenesis
www.elsevier.com/locate/gentox
Community address: www.elsevier.com/locate/mutres

Computer-aided rodent carcinogenicity prediction

Alexey A. Lagunin [a,*], John C. Dearden [b], Dmitri A. Filimonov [a], Vladimir V. Poroikov [a]

[a] *Institute of Biomedical Chemistry RAMS, Pogodinskaya Str. 10, Moscow 119121, Russia*
[b] *School of Pharmacy and Chemistry, Liverpool John Moores University, Byrom Street, Liverpool L3 3AF, UK*

Received 7 December 2004; received in revised form 19 May 2005; accepted 18 June 2005

Abstract

The potential of the computer program PASS (Prediction Activity Spectra for Substances) to predict rodent carcinogenicity for chemical compounds was studied. PASS predicts carcinogenicity of chemical compounds on the basis of their structural formula and of structure–activity relationship analysis of known carcinogens and non-carcinogens. The data on structures and experimental results of 2-year carcinogenicity assays for 412 chemicals from the NTP (National Toxicological Program) and 1190 chemicals from the CPDB (Carcinogenic Potency Database) were used in our study. The predictions take into consideration information about species and sex of animals. For evaluation of the predictive accuracy we used two procedures: leave-one-out cross-validation (LOO CV) and leave-20%-out cross-validation. In the last case we randomly divided the studied data set 20 times into two subsets. The data from the first subset, containing 80% of the compounds, were added to the PASS training set (which includes about 46,000 compounds with about 1500 biological activity types collected during the last 20 years to predict biological activity spectra), the second subset with 20% of the compounds was used as an evaluation set. The mean accuracy of prediction calculated by LOO CV is about 73% for NTP compounds in the 'equivocal' category of carcinogenic activity and 80% for NTP compounds in the 'evidence' category of carcinogenicity. The mean accuracy of prediction for the CPDB database is 89.9% calculated by LOO CV and 63.4% calculated by leave-20%-out cross-validation. Influence of incorporation of species and sex data on the accuracy of carcinogenicity prediction was also investigated. It was shown that the accuracy was increased only for data on male animals.

Keywords: Carcinogenicity; Prediction; PASS; SAR; Rodents; CPDB; NTP

1. Introduction

Progress in combinatorial chemistry and high-throughput screening is resulting in significant increases in the number of known chemical compounds. The study of toxic effects of chemicals, includ-

* Corresponding author. Tel.: +7 95 247 3029; fax: +7 95 245 0857.
E-mail address: alexey.lagunin@ibmc.msk.ru (A.A. Lagunin).

1383-5718/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.mrgentox.2005.06.005

ing carcinogenicity, is an important part of both drug R&D and protection of the environment. Clearly, such toxicological investigations increase concomitantly with the increase in the number of chemical compounds. It is not possible to study carcinogenicity of chemicals in humans, and therefore several species of animals are used for carcinogenicity assessment. One of the most general methods of carcinogenicity assessment is 2-year carcinogenicity assay for rodents (mainly rats and mice), which is time-consuming and highly expensive. A study made by Gold et al. [1] shows that if a chemical is a carcinogen for rats and mice there is a 50% probability that it will be a human carcinogen. Our comparison of chemicals classified by the IARC human carcinogenic hazard classification (International Agency for Research on Cancer) [2] with rodent carcinogenicity data from CPDB (Carcinogenic Potency Database) shows that about 90% of evident, probable and possible human carcinogens (Groups 1, 2A and 2B by IARC classification) are rodent carcinogens. The possibility of extrapolating rodent bioassay data to human carcinogens was shown by other investigators [3]. According to the strength-of-evidence criteria used by IARC, chemicals that induce cancer in two species should be considered most likely to pose carcinogenic hazards to humans [4] and need more detailed and extensive evaluation of the available data and information. Many approaches have been developed to predict rodent carcinogenicity. These vary from short-term biological assays [5,6] and expert systems [7,8] to computer-aided carcinogenicity prediction on the basis of structure–activity relationships [9–11]. Unfortunately, by comparison of prediction results with experimental data it has been shown that the average prediction accuracy of these approaches is less than 65% [12,13]. The reasons for this poor accuracy are the limited number of chemical compounds from some chemical classes used as training sets, imperfection of applied approaches and complex pathogenesis of carcinogenicity. Nevertheless, whether or not a compound is a carcinogen depends (as do all its properties) on its molecular structure. Analysis of the second comparative exercise (Predictive Toxicology Challenge of US National Toxicology Program) for carcinogenicity prediction methods shows that the most difficult problem is a separation of true carcinogens from non-carcinogens that contain structural fragments of carcinogens [13]. As a consequence, too many false positives are predicted. Moreover, this experiment shows that testing in cell lines and using approaches based on expert evaluations provide more accurate results than do available computer-aided approaches.

The main purpose of carcinogenicity prediction is a determination of carcinogenicity for chemical compounds without using the long-term biological assays. This conforms to the biological activity spectra concept, on which the PASS approach (Prediction of Activity Spectra for Substances) is based: all potential activities caused by the compound in biological entities are presented as a biological activity spectrum of the substance [14]. Carcinogenicity is one of many potential activities of chemical compounds that is predicted and it is described qualitatively. A qualitative description of biological activities provides the basis for including the data collected from many different sources in the PASS training set. The PASS approach uses a heterogeneous training set and universal descriptors of molecules that provide high accuracy and robustness of biological activity prediction [14–18]. Features of carcinogenic manifestations which may depend on species, sex and tissues may be taken into account as particular types of biological activity. On the basis of such statements we studied the prediction accuracy of the PASS approach for rodent carcinogenicity data from on-line representation of NTP [19,20] and CPDB [21,22] databases. This paper describes an investigation of PASS prediction accuracy for different categories of carcinogenicity (evident, equivocal and non-carcinogens) from the NTP database according to species and sex of animals. We analyzed the influence of combination of chemicals by species and sex on the accuracy of carcinogenicity prediction. The accuracy of carcinogenic prediction for chemicals from the CPDB database was studied and compared with that for compounds from the NTP database. Concordance, sensitivity and specificity values of carcinogenicity prediction were found to depend on the threshold of estimated probability.

2. Data and methods

2.1. PASS technology

The PASS approach [14–16] is described in detail on the web-site (http://www.ibmh.msk.su/PASS) and in a

book [17]. Here, we present only a brief description of the PASS method, necessary for a general understanding of how it works. PASS uses MNA (Multilevel Neighborhoods of Atoms) descriptors for presentation of a compound's structure [18]. They are based on a molecular structure description according to the valences and partial charges of connected atoms (including hydrogen atoms) although the bond types are not specified. MNA descriptors are generated as a recursively defined sequence:

- zero-level MNA descriptor for each atom is the label $\boldsymbol{A}$ of the atom itself;
- any next-level MNA descriptor for the atom is the sub-structure notation $\boldsymbol{A(D_1D_2\cdots D_i\cdots)}$;

where $\boldsymbol{D_i}$ is the previous-level MNA descriptor for ith immediate neighbor of the atom $\boldsymbol{A}$. The neighbor descriptors $\boldsymbol{D_1D_2\cdots D_i\cdots}$ are arranged in a unique manner, e.g., in lexicographic order. The atom label $\boldsymbol{A}$ may include not only the atomic type but also any additional information about the atom. In particular, if the atom is not included in any ring, it is marked by "-". The iterative procedure for MNA descriptor generation can be continued to cover the first, second, etc., neighborhoods of each atom. In this way the structure of any molecule is represented as a set of MNA descriptors. Fig. 1 shows the structure and MNA descriptors of Amiben (2,5-dichloro-3-aminobenzoic acid) which is a carcinogen in female mice.

In PASS the biological activity is described qualitatively ('active' or 'inactive'). The algorithm of prediction is based on $\boldsymbol{B}$-statistics specially designed for the PASS approach. For each type of activity $\boldsymbol{A}_k$, the following $\boldsymbol{B}_k$ values are calculated based on the molecule's structure represented by the set of $\boldsymbol{m}$ MNA descriptors $\{\boldsymbol{D_1}, \ldots, \boldsymbol{D_m}\}$:

$$\boldsymbol{B}_k = \frac{\boldsymbol{S}_k - \boldsymbol{S}_{0k}}{1 - \boldsymbol{S}_k\boldsymbol{S}_{0k}},$$

$$\boldsymbol{S}_k = \sin\left[\frac{\sum_i \arcsin(2\boldsymbol{P}(\boldsymbol{A}_k|\boldsymbol{D}_i) - 1)}{\boldsymbol{m}}\right],$$

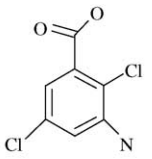
$$\boldsymbol{S}_{0k} = 2\boldsymbol{P}(\boldsymbol{A}_k) - 1$$

where $\boldsymbol{P}(\boldsymbol{A}_k|\boldsymbol{D_i})$ is a conditional probability of activity type $\boldsymbol{A}_k$ if the descriptor $\boldsymbol{D_i}$ present in the set of a molecule's descriptors and $\boldsymbol{P}(\boldsymbol{A}_k)$ is the a priori probability to find a compound with activity type $\boldsymbol{A}_k$. The estimates of the probabilities $\boldsymbol{P}(\boldsymbol{A}_k)$, $\boldsymbol{P}(\boldsymbol{A}_k|\boldsymbol{D_i})$ are given by

$$\boldsymbol{P}(\boldsymbol{A}_k) = \frac{\boldsymbol{n}_k}{\boldsymbol{n}}, \qquad \boldsymbol{P}(\boldsymbol{A}_k|\boldsymbol{D_i}) = \frac{\boldsymbol{n}_{ik}}{\boldsymbol{n}_i},$$

where $\boldsymbol{n}$ is the total number of compounds in the training set, $\boldsymbol{n}_i$ the number of compounds containing descriptor $\boldsymbol{D_i}$ in the structure description, $\boldsymbol{n}_k$ the number of compounds containing the activity type $\boldsymbol{A}_k$ in the training set and $\boldsymbol{n}_{ik}$ is the number of compounds containing both the activity type $\boldsymbol{A}_k$ and descriptor $\boldsymbol{D_i}$.

The PASS training set used for this work contained 45,660 substances represented by 43,000 MNA descriptors and possessing about 1500 different types of biological activity which are molecular mechanisms of action, pharmacological activity, side and toxic effects. The training set was collected during 25 years from the literature and databases that describe the real experimental data. The structural information input of PASS is a MOL- or SD-file (MDL Information Systems Inc.) [23]. The results of prediction (output) can be represented as CSV, TXT or SD-files containing a



MNA/1	MNA/2
HC	C(C(CC-H)C(CC-H)-Cl(C))
HN	C(C(CC-H)C(CC-Cl)-C(C-O-O))
HO	C(C(CC-H)C(CC-Cl)-N(C-H-H))
CHCC	C(C(CC-C)C(CC-N)-Cl(C))
CCCC	C(C(CC-C)C(CC-Cl)-H(C))
CCCN	C(C(CC-N)C(CC-Cl)-H(C))
CCCCl	-H(C(CC-H))
CCOO	-H(-N(C-H-H))
NHHC	-H(-O(-H-C))
OHC	-C(C(CC-C)-O(-H-C)-O(-C))
OC	-N(C(CC-N)-H(-N)-H(-N))
ClC	-O(-H(-O)-C(C-O-O))
	-O(-C(C-O-O))
	-Cl(C(CC-Cl))

Fig. 1. List of the MNA descriptors for Amiben (2,5-dichloro-3-aminobenzoic acid). MNA/1 and MNA/2 are descriptors of the first- and second-level, respectively.

list of activities with the estimates of probability of being active (P_a) and inactive (P_i), respectively. Only activities for which $P_a > P_i$ can be considered as probable.

We used carcinogenicity data from NTP and CPDB databases as training sets separately or in combination with the existing PASS training set. In the last case all compounds from the PASS training set different from those of NTP and CPDB databases were considered as non-carcinogens. Actually some of these compounds may be rodent carcinogens. Analysis of drugs included in the open part of the FDA database on rodent carcinogens (http://www.predictive-toxicology.org/data/fda/) shows that only 17% of 223 drugs reveal carcinogenicity in two species and 43% at least in one. In spite of lack of knowledge whether all of these compounds are non-carcinogens or not, our experiments show that such a procedure improves the accuracy of prediction (see Section 3). In that way we increase the training set. The positive influence of the database size on the informational content was confirmed by Takihi et al. [24]. Increasing the number of inactive compounds in the training set also improves the accuracy of prediction that was shown in the study on mutagenicity prediction by Liu et al. [25].

2.2. Databases

Quality of experimental data plays a key role in computer-aided carcinogenicity prediction. We used two on-line resources of rodent carcinogenicity data: the US National Toxicology Program (NTP) database [19,20] and the Carcinogenic Potency Data Base (CPDB) [21,22]. Both databases contain results of the standard 2-year rodent carcinogenicity bioassay for chemical compounds, which takes into consideration species and sex of animals (male or female, mice or rats). The NTP database (http://ntp-server.niehs.nih.gov/htdocs/pub.html) includes more than 480 diverse chemical compounds and their mixtures, which were evaluated under a strict experimental protocol. Five categories of carcinogenic activity are used in the NTP database to summarize the strength of the evidence observed in each experiment: two categories for positive results (clear evidence and some evidence); one category for uncertain findings (equivocal evidence); one category for no observable effects (no evidence); and one category for experiments that because of major flaws cannot be evaluated (inadequate study). We used a set of 412 chemical compounds from the NTP database for training in PASS. Small inorganic compounds (e.g., NO_2), oils, paraffins and mixtures of compounds were excluded from the set. The previously existing PASS training set already included 337 structures of carcinogens, and thus only 75 new structures were added. The PASS training set includes compounds with carcinogenic activity, but this activity is not specified by species. From 337 compounds which are in the NTP database and in the PASS training set 152 compounds are carcinogens, 119 are mutagens, 107 compounds have embryotoxic effects, 105 compounds have toxic effects and 68 compounds are teratogens. These compounds reveal also about 200 other types of biological activity, but with a frequency of less than 40 compounds. We combined two categories of positive results (clear evidence and some evidence) into one, which was described according to the species and sex of animals as ‘carcinogen, female mice’; ‘carcinogen, male mice’; ‘carcinogen, female rats’ and ‘carcinogen, male rats’. For the ‘equivocal evidence’ category we used the following descriptions: ‘carcinogen, equivocal, female mice’; ‘carcinogen, equivocal, male mice’; ‘carcinogen, equivocal, female rats’ and ‘carcinogen, equivocal, male rats’. All compounds that did not fall into any category were considered as non-carcinogens. The CPDB database (http://potency.berkeley.edu/cpdb.html) includes information from the NTP database and data of carcinogenicity studies reported in the literature with minimum protocol requirements for 1190 chemical compounds [26]. The results of carcinogenicity bioassays are represented by TD_{50} values (dose at which tumorogenesis was found in 50% of tested animals) for each species (mice, rats) and by tissue targets for tumor formation for each species and sex. The CPDB database contains data about tumor manifestation in 36 tissues. The PASS training set already included 753 structures from the CPDB database, and hence 437 structures were added as new ones. From 753 compounds which are in the CPDB database and in the PASS training set 282 compounds are carcinogens, 249 are mutagens, 223 have embryotoxic effect, 221 compounds have toxic effect and 136 compounds are teratogens. These compounds reveal also about 400 biological activity types, but with a frequency of less than 60 compounds.

We also compared carcinogenic activities for the chemicals that are present in both NTP and CPDB databases. It appeared that the difference between these is about 15%. We did not correct these data and used them as it is.

2.3. Validation of the prediction accuracy

We used the independent accuracy of prediction (IAP) values to estimate the prediction accuracy. IAP was calculated for each type of activity in PASS predictions:

$$\mathrm{IAP} = \frac{N(p_1 > p_0)}{n_1 n_0}$$

where $N(p_1 > p_0)$ is the number of cases when the predicted probability of an active compound (p_1) to be active is greater than the predicted probability to be active for an inactive compound (p_0); n_1 and n_0 is the number of active and inactive compounds, respectively in the evaluation subset.

For evaluation of the accuracy of prediction we used two procedures: leave-one-out cross-validation (LOO CV) and leave-20%-out cross-validation. In the last case we randomly divided the studied set 20 times into two subsets. The data from the first subset containing 80% of the compounds were added to the PASS training set, the second subset with 20% compounds was used as an evaluation set. Thus, we prepared 20 pairs of training and evaluation sets. Each training set was added to the PASS training set or used separately, then PASS was retrained and a prediction of carcinogenicity was made for the appropriate evaluation set. The reason for partition of the set 20 times is the following: if we have 20 values of variate then 18 of them or 90% are between minimal and maximal values. Therefore, minimal and maximal values from 20 values of variate reflect the 90% confidence interval: $90\% = (100 \times (20-2)/20)$.

To compare the quality of PASS prediction with other methods, we used characteristics generally applied in evaluation of carcinogenicity: concordance, sensitivity and specificity [27]. Concordance is the ratio of total correct predictions to total number of predictions, sensitivity is the percentage of correct predictions for carcinogens and specificity is the percentage of correct predictions for non-carcinogens.

3. Results and discussion

3.1. Prediction of carcinogenicity on the basis of NTP data

Initially, we studied the predictive accuracy of PASS for data from the NTP database, which were used together with the PASS training set, and separately. The number of compounds and the accuracy of carcinogenicity prediction calculated by LOO CV for the NTP set ($\mathrm{IAP_{NTP}}$) and in combination with the PASS training set ($\mathrm{IAP_{PASS}}$) also as leave-20%-out cross-validation ($\mathrm{IAP_{20PASS}}$) procedures for each animal group are displayed in Table 1.

The prediction accuracy calculated by LOO CV for combined NTP and PASS training sets ($\mathrm{IAP_{PASS}}$) shows that the prediction accuracy for compounds from the category ‘equivocal’ (nos. 1–4) was worse than for positive carcinogens (nos. 10–14), varying from 75.6 (carcinogen, equivocal, male rats) to 82.1% (carcinogen, equivocal, female mice). The mean accuracy for ‘equivocal’ activities is 78.9%. At the same time the mean accuracy for positive carcinogens is 86.7%. It varies from 85.7 (carcinogen, female rats) to 87.6% (carcinogen, female mice). Combining activities by sex (nos. 5, 6, 14, 15) or by species (nos. 7–9, 16–18) did not lead to any significant increase of the accuracy of carcinogenic prediction. Analysis of the prediction accuracy activity types show that the combination of animal groups makes sense only for male data (carcinogen, equivocal, male mice—77.3%, carcinogen, equivocal, male rats—75.6%, carcinogen, equivocal, male rodent—80.6%; carcinogen, male mice—86.2, carcinogen, male rats—87.3, carcinogen, male rodent—87.9). When we used only the NTP set as a training set for carcinogenicity prediction then the accuracy of prediction ($\mathrm{IAP_{NTP}}$) was considerably reduced by 30–40% and some activity types could not be predicted. The discussion about the reasons for such a distinction is given in Section 3.2. At the same time, also for the joint NTP and PASS training sets, positive carcinogens were predicted better than compounds from the ‘equivocal’ category. The accuracy of carcinogenicity prediction for NTP data calculated by the leave-20%-out cross-validation ($\mathrm{IAP_{20PASS}}$) also shows that positive carcinogens were predicted better then compounds from the ‘equivocal’ category. The mean accuracy prediction for positive carcinogens was

Table 1
The number of compounds, accuracy of prediction and categories of carcinogenic activity for NTP data

No.	Number	IAP_{PASS} (%)	IAP_{NTP} (%)	IAP_{20PASS} (%)	Activity type
1	23	82.1	46.5	54.2	Carcinogen, equivocal, female mice
2	36	80.7	41.8	50.9	Carcinogen, equivocal, female rats
3	39	77.3	–	46.0	Carcinogen, equivocal, male mice
4	41	75.6	46.8	38.0	Carcinogen, equivocal, male rats
5	55	81.1	–	47.6	Carcinogen, equivocal, female rodent
6	75	80.6	37.1	43.4	Carcinogen, equivocal, male rodent
7	55	77.1	–	41.6	Carcinogen, equivocal, mice
8	67	82.8	42.4	49.3	Carcinogen, equivocal, rats
9	110	81.1	–	42.4	Carcinogen, equivocal, rodent
10	147	87.6	48.4	57.5	Carcinogen, female mice
11	120	85.7	48.2	57.0	Carcinogen, female rats
12	125	86.2	47.9	60.8	Carcinogen, male mice
13	152	87.3	47.3	57.0	Carcinogen, male rats
14	190	86.6	48.2	56.0	Carcinogen, female rodent
15	201	87.9	48.3	60.0	Carcinogen, male rodent
16	165	87.4	47.4	59.0	Carcinogen, mice
17	172	86.9	49.6	57.0	Carcinogen, rats
18	228	87.4	46.9	57.2	Carcinogen, rodent

IAP_{PASS}: invariant accuracy of prediction calculated by the leave-one-out cross-validation procedure for joint PASS and NTP data; IAP_{NTP}: invariant accuracy of prediction calculated by the leave-one-out cross-validation procedure for NTP data; IAP_{20PASS}: invariant accuracy of prediction calculated by the leave-20%-out cross-validation procedure on the basis of joint PASS and NTP data; –: activity types that could not be predicted by PASS.

58.1%, which is 10% better than for compounds from the 'equivocal' category (47.3%).

3.2. Prediction of rodent carcinogenicity on the basis of CPDB data

We estimated the accuracy of rodent carcinogenicity prediction for 1190 compounds from the CPDB set. In this investigation we studied the influence of data on possible non-carcinogens contained in the PASS training set on the accuracy of prediction. For this we compared the prediction results for the CPDB set using the PASS training set with drug-like compounds and the CPDB set together and separately as training sets for prediction of rodent carcinogenicity. We calculated IAP by LOO CV for all compounds from the training set (IAP_{PASS} and IAP_{CPDB}) and IAP by leave-20%-out cross-validation for the CPDB set (IAP_{20PASS} and IAP_{20CPDB}). Table 2 shows the result of these calculations for each group of animals.

The mean accuracy calculated by LOO CV for the combination of the PASS training set and compounds from the CPDB database (IAP_{PASS}) was 89.9%. This value is 3% higher than the accuracy of carcinogenicity

Table 2
The number of compounds, accuracy of carcinogenicity prediction calculated by the leave-one-out cross-validation (LOO CV) procedure (IAP) for CPDB data

Activity type	Number	IAP_{PASS} (%)	IAP_{CPDB} (%)	IAP_{20PASS} (%)	IAP_{20CPDB} (%)
Carcinogen, female mice	294	90.8	74.0	65.3	58.9
Carcinogen, female rats	319	89.7	70.5	65.5	59.1
Carcinogen, male mice	277	90.1	65.3	62.8	57.8
Carcinogen, male rats	357	88.9	71.8	59.8	51.9
Mean values	312	89.9	70.4	63.4	56.9

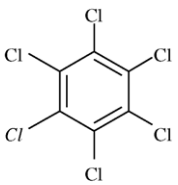
IAP_{PASS}: invariant accuracy of prediction calculated by the leave-one-out cross-validation procedure for joint PASS and CPDB data; IAP_{CPDB}: invariant accuracy of prediction calculated by the leave-one-out cross-validation procedure for CPDB data; IAP_{20PASS}: invariant accuracy of prediction calculated by the leave-20%-out cross-validation procedure on the basis of joint PASS and CPDB data; IAP_{20CPDB}: invariant accuracy of prediction calculated by the leave-20%-out cross-validation procedure for CPDB data.

prediction for the NTP data in spite of the fact that the NTP data were obtained under the more strict protocols and are less diverse chemically in comparison with the CPDB set. This may be explained by the higher number of compounds in the CPDB set than in the NTP database.

Using only CPDB set for training, we obtained a mean accuracy of 70.4% (IAP_{CPDB}), which is 20% less than for the joint PASS and CPDB data. Since the PASS training set contains mostly non-carcinogenic drug-like compounds, its addition to the data from CPDB increases the structural diversity of non-carcinogens. As a consequence, the discrimination between carcinogens and non-carcinogens is also increased, thus improving the accuracy of prediction. This is clear from the leave-20%-out cross-validation. When 80% of the CPDB set was used together with the PASS training set (IAP_{20PASS}) for the training of PASS, then the mean accuracy of prediction for 20% of the CPDB set was 63.4%. When 80% of the CPDB set was used separately from the PASS training set (IAP_{20CPDB}) for the training of PASS, then the mean accuracy of prediction for 20% of the CPDB set decreased to 56.9%, a difference of 6.5%. Therefore, in further experiments we combined the PASS training set and the CPDB data.

Fig. 2 shows an example of rodent carcinogenicity prediction for 1,2,3,4,5,6-hexachloro-benzene.

Carcinogenic effects for three animal groups with probability $P_a > P_i$ were predicted for this compound.



> <CAS>
58-89-9

> <ACTIVITY_PREDICTION>
6 Substructure descriptors; 0 new.
3 Possible activities.
Pa Pi for Activity:
0.811 0.004 Carcinogen, male mice
0.761 0.005 Carcinogen, female mice
0.183 0.167 Carcinogen, male rats

Fig. 2. Example of carcinogenicity prediction for 1,2,3,4,5,6-hexachlorobenzene. Known activities are marked in bold.

Predictions for two groups of animals (carcinogen, male mice and carcinogen, female mice) coincided with the experimental results. The probability of a carcinogenic effect for these groups exceeds 75%. Carcinogenicity was predicted also for male rats, whereas it was not found experimentally. However, the low value of the probability of a carcinogenic effect in this group of animals and the small difference between P_a (0.183) and P_i (0.167) values of probability shows that this prediction is not significant.

As mentioned above, one of the main conclusions from the NTP comparative exercises was the difficulty in separating true carcinogens from non-carcinogens that contain structural fragments of carcinogens. To reproduce the same condition with methods that were applied in the NTP comparative exercises (1996), we used as a training set only those data from NTP that were available at that time, i.e. from 1976 to 1995. Our computer experiment with prediction of carcinogenicity for 30 compounds used in the second NTP comparative exercise shows that concordance of carcinogenicity prediction by PASS was 62.5%, sensitivity 86.7% and specificity 22.2% at $P_a > P_i$ threshold. That is, despite the high number of correctly predicted carcinogens, most non-carcinogens were predicted as carcinogens. A similar result was obtained for most of the best approaches applied in the NTP comparative exercises [13]. Prediction results of carcinogenicity expressed by P_a and P_i values solve the problem of separating carcinogens and non-carcinogens by means of selection of a particular difference between P_a and P_i as a threshold (Fig. 3).

The curves in Fig. 3 were calculated by LOO CV for CPDB data. Fig. 3 shows that changing the '$P_a - P_i$' threshold from 0 to 95% leads to an increase in the prediction accuracy of non-carcinogens and decreases the number of false positives. Specificity increases from 36.4 ($P_a - P_i > 0$) to 99.5% ($P_a - P_i > 95\%$). At the same time, however, such change leads to a decrease in the prediction accuracy of carcinogens. Sensitivity decreases from 84.9 ($P_a - P_i > 0$) to 4.5% ($P_a - P_i > 95\%$). Taking into consideration all three curves presented in Fig. 3, it was suggested that the optimal threshold is at $P_a - P_i > 30\%$. In this case concordance has a maximal value of 64.4%, specificity 62.4% and sensitivity 66.2%.

In this investigation we can see an inverse negative relationship between sensitivity and specificity.

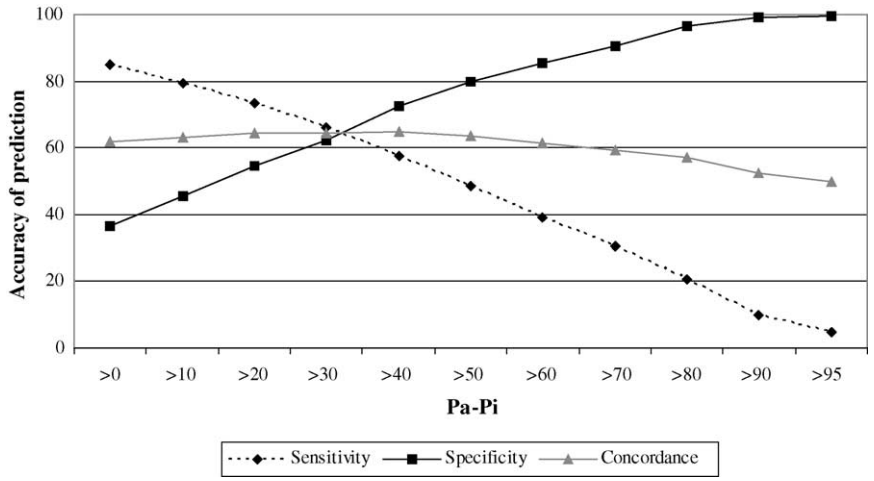


Fig. 3. Accuracy of carcinogenicity prediction depending on threshold of difference between P_a and P_i values. Concordance = total correct predictions/total number of predictions, sensitivity = percentage of correct predictions of carcinogens, specificity = percentage of correct predictions of non-carcinogens.

The same observation was made on the study by Matthews and Contrera, which reported a modified FDA-MCASE model for carcinogenicity prediction based on an expanded version of the CPDB enriched with pharmaceuticals and with an FDA weight-of-evidence-graded activity assignment [28]. They also had discordance between sensitivity and specificity. In that investigation the authors reported high sensitivity and low specificity in rodent carcinogenicity prediction but did not provide a possibility to find an optimum between two. They supposed that the relationships between sensitivity and specificity depend on the quality of a training set. We think that it also depends on the algorithm of the method and an output of prediction results. In our opinion it is very important to have a possibility to change one to improve another and our approach provides this.

Some of the rodent carcinogenicity prediction approaches used in the NTP comparative exercises, such as MULTICASE, provide useful information (rules or structural features, which may be reason of this effect) for rationalization and justification of a prediction. We consider that the representation of relationships between the PASS descriptors and carcinogenicity is possible but it has not been realized yet. We studied this possibility, but at present it may be done only in statistical terms.

4. Conclusions

Studying the prediction of rodent carcinogenicity by the PASS software has shown that the PASS algorithm can be successfully applied for this purpose. Analysis of prediction results of rodent carcinogenicity shows that use of data on carcinogenicity together with data for drug-like compounds from the PASS training set, which are represented as possible non-carcinogens, increases the accuracy of carcinogenicity prediction. Changing of the '$P_a - P_i$' threshold leads to revision of sensitivity and specificity of carcinogenicity prediction. It may be used to increase the number of correctly predicted carcinogens or non-carcinogens. Our study shows that the mean prediction accuracy calculated by LOO CV was 78.9% for 'equivocal' and 86.7% for 'evident' carcinogens. The study also shows that combining NTP data on species and sex did not increase the accuracy of carcinogenicity prediction with the exception of data for male animals. The mean accuracy for combined CPDB data and the PASS training set was 89.9% calculated by LOO CV and 63.4% calculated by leave-20%-out cross-validation for 'evident' carcinogens. This accuracy was achieved without expert evaluation of prediction results and is comparable with the best currently available methods of carcinogenicity prediction. Carcinogenicity prediction for rats and

mice of both sexes may be useful for extrapolation of rodent carcinogenicity to humans.

Acknowledgement

One author (Alexey A. Lagunin) is grateful to the Royal Society for support of this work (study grant No. 15515).

References

[1] L.S. Gold, N.B. Manley, B.N. Ames, Extrapolation of carcinogenicity between species: qualitative and quantitative factors, Risk Anal. 12 (4) (1992) 579–588.
[2] International Agency for Research on Cancer (IARC), http://monographs.iarc.fr/.
[3] V.A. Fung, J.C. Barrett, J. Huff, The carcinogenesis bioassay in perspective: application in identifying human cancer hazards, Environ. Health Perspect. 103 (1995) 680–683.
[4] R. Tennant, Stratification of rodent carcinogenicity bioassay results to reflect relative human hazards, Mutat. Res. 286 (1993) 111–118.
[5] G.A. Kerckaert, R. Brauninger, R.A. LeBoeuf, R.J. Istorf, Use of the Syrian hamster embryo cell transformation assay for carcinogenicity prediction of chemical currently being tested by the National Toxicology Program in rodent bioassays, Environ. Health Perspect. 104 (1996) 1075–1084.
[6] V.J. Feron, M. Schwarz, K. Hemminki, D. Krewski, Long- and medium-term carcinogenicity studies in animals and short-term genotoxicity tests, IARC Sci. Publ. 131 (1999) 103–129.
[7] Y.T. Woo, D.Y. Lai, J.C. Arcos, M.F. Argus, M.C. Cimino, S. DeVito, L. Keifer, Mechanism-based structure–activity relationship (SAR) analysis of carcinogenic potential of 30 NTP test chemicals, Environ. Carcin. Eco. Rev. C 15 (1997) 139–160.
[8] C.A. Marchant, Prediction of rodent carcinogenicity using the DEREK system for 30 chemicals currently being tested by the National Toxicology Program, Environ. Health Perspect. 104 (1996) 1065–1074.
[9] Y.P. Zhang, N. Sussman, O.T. Macina, H.S. Rosenkranz, G. Klopman, Prediction of the carcinogenicity of a second group of organic chemicals undergoing carcinogenicity testing, Environ. Health Perspect. 104 (1996) 1045–1050.
[10] D.F.V. Lewis, C. Ioannides, D.V. Parke, COMPACT and molecular structure in toxicity assessment: a prospective evaluation of 30 chemicals currently being tested for rodent carcinogenicity by the NCI/NTP, Environ. Health Perspect. 104 (1996) 1011–1016.
[11] R.D. King, A. Srinivasan, Prediction of rodent carcinogenicity biossays from molecular structure using inductive logic programming, Environ. Health Perspect. 104 (1996) 1031–1040.
[12] A.M. Richard, R. Benigni, AI and SAR approaches for predicting chemical carcinogenicity: survey and status report, SAR QSAR Environ. Res. 13 (1) (2002) 1–20.
[13] R. Benigni, R. Zito, The second National Toxicology Program comparative exercise on the prediction of rodent carcinogenicity: definitive results, Mutat. Res. 566 (1) (2004) 49–63.
[14] V.V. Poroikov, D.A. Filimonov, Yu.V. Borodina, A.A. Lagunin, A. Kos, Robustness of biological activity spectra predicting by computer program pass for non-congeneric sets of chemical compounds, J. Chem. Inf. Comput. Sci. 40 (6) (2000) 1349–1355.
[15] PASS, http://www.ibmh.msk.su/PASS.
[16] A.V. Stepanchikova, A.A. Lagunin, D.A. Filimonov, V.V. Poroikov, Prediction of biological activity spectra for substances: evaluation on the diverse set of drug-like structures, Curr. Med. Chem. 10 (2003) 225–233.
[17] V. Poroikov, D. Filimonov, PASS: prediction of biological activity spectra for substances, in: Helma Christoph (Ed.), Predictive Toxicology, Taylor & Francis Group, LLC, Boca Raton, 2005, pp. 459–478.
[18] V.V. Poroikov, D.A. Filimonov, How to acquire new biological activities in old compounds by computer prediction, J. Comput. Aid. Mol. Des. 16 (11) (2003) 819–824.
[19] National Toxicology Program, Reports of NTP toxicology and carcinogenesis studies of chemicals in F344/N rats and B6C3F1 mice, National Toxicology Program, Research Triangle Park, NC, 1980–1998.
[20] National Toxicology Program, http://ntp-server.niehs.nih.gov/htdocs/pub.html.
[21] L.S. Gold, N.B. Manley, T.H. Slone, J.M. Ward, Compendium of chemical carcinogens by target organ: results of chronic bioassays in rats, mice, hamsters, dogs, and monkeys, Toxicol. Pathol. 29 (6) (2001) 639–652.
[22] Carcinogenic Potency Database, http://potency.berkeley.edu/cpdb.html.
[23] MDL Information Systems Inc., http://www.mdl.com.
[24] N. Takihi, Y.P. Zhang, G. Klopman, H.S. Rosenkranz, An approach for evaluating and increasing the informational content of mutagenicity and clastogenicity data base, Mutagenesis 8 (1993) 257–264.
[25] M. Liu, N. Sussman, G. Klopman, H.S. Rosenkranz, Estimation of the optimal data base size for structure-activity analyses: the Salmonella mutagenicity data base, Mut. Res. 358 (1996) 63–72.
[26] L.C. Gold, N.B. Manley, T.H. Slone, L. Rohrbach, Supplement to the Carcinogenic Potency Database (CPDB): results of animal bioassays published in the general literature in 1993 to 1994 and by the National Toxicology Program in 1995 to 1996, Environ. Health Perspect. Suppl. 107 (Suppl. 4) (1999) 527–602.
[27] C.A. Marchant, D.B. Combes, Artificial intelligence: the use of computer methods in the prediction of metabolism and toxicity, in: M.G. Ford, R. Greenwood, G.T. Brooks, R. Franke (Eds.), Bioactive Compound Design: Possibilities for Industrial Use, BIOS Scientific Publishers Ltd., 1996, pp. 153–162.
[28] E.J. Matthews, J.F. Contrera, A new highly specific method for predicting the carcinogenic potential of pharmaceuticals in rodents using enhanced MCASE QSAR-ES software, Regul. Toxicol. Pharm. 28 (1998) 242–264.