Predicting Biotransformation Potential from Molecular Structure

Yu. Borodina,*,‡ A. Sadym,‡ D. Filimonov,‡ V. Blinova,† A. Dmitriev,‡ and V. Poroikov‡

Laboratory of Structure−Function Based Drug Design, Institute of Biomedical Chemistry of Russian Academy of Medical Sciences, 10 Pogodinskaya Str., Moscow 119121, Russia, and Russian Institute for Scientific and Technical Information, 20 Usievicha Str., Moscow 125190, Russia

Received April 17, 2003

The program PASS-BioTransfo is presented, which is capable of predicting many classes of biotransformation for chemical compounds. A particular class of biotransformation is defined by the chemical transformation type and may additionally include the name of the enzyme involved in a transformation. An evaluation of the approach is presented, using biotransformations taken from the databases Metabolite (MDL) and Metabolism (Accelrys), respectively. When trained with biotransformations from Metabolite, PASS-BioTransfo predicts 1927 classes of biotransformation; the average accuracy estimated in LOO cross-validation is about 88%. After training with the biotransformations from the Metabolism database, 178 classes of biotransformation are predicted with an average accuracy of about 85%. The results of cross-prediction with several training and evaluation sets are presented and discussed.

INTRODUCTION

According to the IUPAC vocabulary of terms used in medicinal chemistry, *biotransformation* is the modification of chemical compounds by living organisms or enzyme preparations.[1] Generally this process is applicable to both xenobiotics and endogenous compounds, but medicinal chemistry usually concentrates on metabolic biotransformation of pharmaceutical agents.

A plethora of enzymes is involved in drug metabolism. For the cytochromes P450 superfamily alone (major enzymes of the phase I metabolism), about 60 human isoforms are known,[2] and it is assumed that the total number may yet increase. The major enzyme systems responsible for drug biotransformation are the following:[3]

- cytochromes P450
- esterases
- epoxide hydrolase
- dihydropyrimidine dehydrogenase
- glutathione S-transferases
- N-acetyltransferases
- sulfotransferases
- thiopurine methyltransferase
- glucuronosyltransferases

In most cases more than one enzyme is involved in a particular drug metabolism. On the other hand, a particular enzyme is capable of catalyzing many biotransformations in terms of different individual chemical reactions. The ability of a particular drug to interact with certain enzymes and to undergo certain biotransformations constitutes its biotransformation (or metabolic) potential.

Early determination of the metabolic potential is a critical issue in the drug development cycle. Experimental studies of the metabolic potential in the human body are labor-intensive and represent a biomedical as well as an ethics problem. Existing in vitro systems for high-throughput metabolism screening do not cover whole enzymatic profiles and do not enable one to recognize all possible ways of a drug's biotransformation. The role of in silico technologies lies in decreasing the number of in vivo and in vitro experiments and in enabling the study of the metabolic potential at the presynthesis stage of drug design.

Most of the modern in silico approaches to metabolism prediction are concentrated on the analysis of drug biotransformation by various isoforms of cytochrome P450.[4−7] These studies are usually based on modeling of substrate−enzyme interaction and are not suitable for high-throughput application. Existing computer programs for rapid biotransformation prediction, META,[8,9] MetabolExpert,[10] and METEOR[11] are knowledge-based expert systems. To predict biotransformations they use special rules provided by experts in the field of xenobiotic metabolism. Therefore, the predictions are based on "human knowledge" rather than on robust and objective computational estimates.

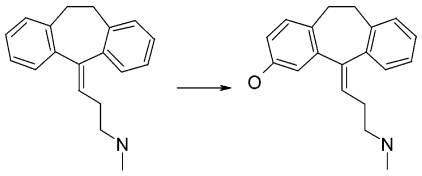
In this paper we describe a different approach for predicting biotransformation potentials. The existing published (or "in house" available) experimental data are used for automatic generation of "structure-biotransformation" relationships, which are then applied in the prediction of the biotransformation potential for a new molecule. The mathematical method underlying this approach was adopted from the program PASS[12−19] developed earlier for predicting many kinds of biological activity for chemical substances. Since the PASS approach was shown to be successful for biological activity prediction, it was used "as is" in the current version of the biotransformation prediction program, which we call "PASS-BioTransfo". Here, it is employed for the prediction of possible *classes of biotransformation*. To evaluate this approach, PASS-BioTransfo was applied to biotransformations taken from the MDL Metabolite[20] and Accelrys Metabolism[21] databases, generally seen as the two best commercially available biotransformation databases. These experiments address the following issues:

* Corresponding author phone: +7-095 247-3029; e-mail: borodina@ibmh.msk.su.

‡ Laboratory of Structure−Function Based Drug Design.

† Russian Institute for Scientific and Technical Information.

10.1021/ci034078l CCC: $25.00
Published on Web 08/20/2003



Aromatic Hydroxylation
Aromatic Hydroxylation (Cytochrome P450**)**
Aromatic Hydroxylation (Cytochrome P450, CYP2D6**)**

Figure 1. Example of notations of biotransformation classes.

1. Use of a uniform computational approach for predicting different classes of biotransformation from molecular structure.

2. Application of data from animal experiments to human biotransformations prediction.

CLASSES OF BIOTRANSFORMATION

IUPAC defines an organic chemical transformation as a modification of a substrate into a product: "A *transformation* is distinct from a *reaction* in that it describes only those changes that are involved in converting the structure of a substrate into that of a product, regardless of the reagent or the precise nature of the substrate, or (with some exception) the mechanism. Therefore all processes in which X-H is converted in X-NO2 are examples of the single transformation called 'nitration', whatever the nature of X, and irrespective of whether the reaction entails the replacement of H^+ by NO_2, of $H^\bullet$ by $NO_2^\bullet$, or of H^- by NO_2^-."[22]

The scheme of a particular transformation includes a substrate structure on the left of the arrow and a product structure on the right. The name of the transformation reflects the main structural modification in the *substrate* molecule.

Since a *biotransformation* is a transformation of enzymatic nature, we suggest that inclusion of the specification of the involved enzymes in the definition of a specific biotransformation is reasonable. We therefore use the following general form for designating *classes of biotransformation*:

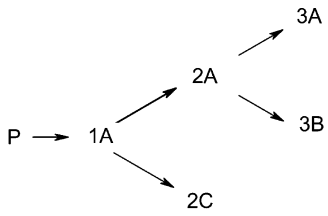
chemical transformation (enzyme system, isoenzyme)

The "chemical transformation" is the basic part of the notation, while "enzyme system" and "isoenzyme" are optional parts, which may be missing, especially when enzyme or isoenzyme are unknown. Figure 1 shows an example of the above notation for the aromatic hydroxylation by cytochrome P450 2D6.

The first class of biotransformation indicates only the class of chemical modification of the substrate, while the second and third ones specify which enzyme and isoenzyme is involved.

BIOTRANSFORMATIONS USED IN THE STUDY

We used two well-known commercial databases (DB), Metabolite (MDL Information Systems Inc.) and Metabolism (Accelrys), for the evaluation of our approach. These databases contain biotransformation reactions collected from different in vitro and in vivo studies and assemble them in metabolic schemes. Structural information includes the RXN structure for the entire reaction, which may be augmented by structures of substrates and products in mol-file format. In addition to the structural information, the databases include



	Metabolite	Metabolism
1.	P → 1A	P → 1A
2.	**1A → 2A**	**P → 2A**
3.	1A → 2C	P → 2C
4.	2A → 3A	P → 3A
5.	2A → 3B	P →3B

Figure 2. Representation of biotransformations in Metabolite and Metabolism databases.

experimental parameters, biological activity and toxicity data, pharmacokinetic data, physicochemical properties of compounds, etc. Here we briefly describe those data from the databases that are of importance for our study. These data include the following:

- structural information
- description of transformation classes
- enzyme information
- species information

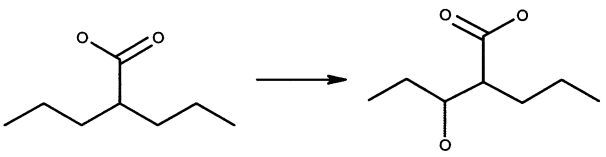
MDL Metabolite Database. The database Metabolite 2001.1 includes 55546 biotransformations of more than 9000 parent compounds. About 98% of the annotated parent compounds are pharmaceuticals, and about 2% are food additives, industrial chemicals, and agrochemicals.

Every biotransformation represents one of the reactions of a metabolic scheme. Within one metabolic scheme, every biotransformation has a particular substrate and product and is related to a parent compound. Figure 2 shows how biotransformations are represented in Metabolite. Every transformation is attributed to one or several classes indicated in the text field "reaction class". The complete vocabulary of reaction classes present in Metabolite includes 203 records. Enzyme information is contained in the fields "enzyme name" and "isoenzyme" and includes the common names of the enzyme system and isoenzymes that were reported in the experiment. In Metabolite 2001.1, 11 969 biotransformations have enzyme information, with 94 enzyme systems being present. Species information is held in the field "species". It is essentially a database of biotransformations found in mammals, 17 289 of which are human.

Accelrys Metabolism Database. The database Metabolism, version 2002, includes 25 000 biotransformations of 3164 parent compounds. By our estimates, about 52% of the annotated parent compounds are agrochemicals, 25% are pharmaceuticals, 10% are industrial and environmental chemicals, 8% are model compounds, and 5% are natural products and food additives.

Unlike MDL Metabolite, in Metabolism every biotransformation is a parent compound modification. Thus, a substrate is always a parent compound. Comparison of a metabolic sheme representation in Metabolite and Metabolism is shown in Figure 2.

Structural information includes the RXN structure and parent and product structures. Transformation classes are indicated in the field "key phrases". The total list of key phrases consists of 238 items. Enzyme information is not given. The field "test system" indicates the experimental



Metabolite: Aliphatic Hydroxylation. C-Hydroxylation

Metabolism: Aliphatic hydroxylation. Aliphatic oxidation. Aliphatic oxygenation. C-Hydroxylation. C-Oxidation. C-Oxygenation.

Figure 3. Example of the classification of the same biotransformation in Metabolite and Metabolism.

organism in which a biotransformation was found. The database contains biotransformations in vertebrates, invertebrates, and plants: 11 968 biotransformations are mammalian, 2635 are human.

Data Preparation. The biotransformations from the Metabolism and Metabolite databases were used for both training and evaluating the method. We prepared six data sets from the two databases:

1. Substrates of all biotransformations of Metabolite (15 044 unique substrates).
2. Substrates of the first step mammalian biotransformations of Metabolism (2598 unique substrates).
3. Substrates of animal biotransformations of Metabolite (11 599 unique substrates).
4. Substrates of human biotransformations of Metabolite (6376 unique substrates).
5. Substrates of animal first step biotransformations of Metabolism (2186 unique substrates).
6. Substrates of human first step biotransformations of Metabolism (708 unique substrates).

To prepare the set 2 we retrieved all mammalian biotransformations from Metabolism. As was mentioned above, in Metabolism a substrate is always a parent and intermediate substrates are not listed. Since we needed the real substrates for our study, we retrieved from Metabolism only the first step biotransformations for which parent compounds were real substrates. For the preparation of data sets 3 and 4, we divided all biotransformations of Metabolite into two subsets. The first subset contained biotransformations found in in vivo or in vitro human experiments or in in vitro experiments with human enzymes expressed in animal cells. The second one contained biotransformations found only in animal experiments. For the preparation of the 5th and 6th data sets, we repeated the same procedure with the first step biotransformations of the Metabolism DB.

For every biotransformation from Metabolite, the substrate structure, reaction classes, and names of enzyme and isoenzyme were used. From Metabolism, we used parent structure and key phrases. All data were converted to a unified format in which every structure was saved as a list of substructural descriptors (see the Methods Section), while reaction classes and enzyme information (for Metabolite) were saved in biotransformation classes.

It should be mentioned that Metabolite and Metabolism often classify biotransformations differently. Although the lists of terms used in Metabolite and Metabolism overlap significantly, the descriptions of particular transformations often differ. Classifications in Metabolism tend to be more, if not overly, complicated compared with the typically short descriptions in Metabolite. An example is given in Figure 3.

To allow for the best possible comparison between two databases, biotransformations of Metabolite and Metabolism should ideally be standardized on the basis of a uniform classification. However, we did not do that for the following reasons. (1) There is currently no such uniform classification of chemical transformations, and it would be beyond the scope of the current study to introduce one. (2) Standardization would lead to significant modification of the original information. Since the abovementioned databases are widely applied as reference sources, we used their information "as it is" for the *evaluation* of our approach.

METHODS

PASS-BioTransfo predicts possible classes of biotransformation from molecular structure using a method that has already been published elsewhere.[13] Here we present only a brief description of the method.

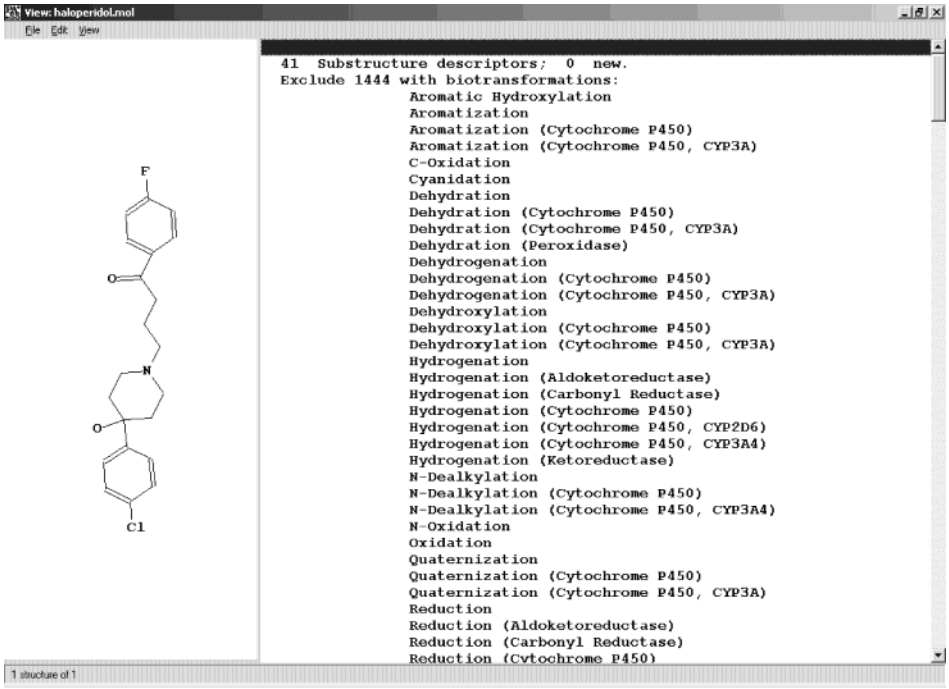
Chemical Structure. Chemical structure is represented by original descriptors called Multilevel Neighborhoods of Atoms (MNA). These descriptors are generated from the compound's structural formulas. A detailed definition of these descriptors can be found in a previous publication.[23] It has been shown that the MNA descriptors are rather universal and are capable of representing various structure−property relationships, including many types of biological activity,[13,14,17−19] mutagenicity and carcinogenicity,[19] drug-likeness,[15] etc. MNA descriptors describe surroundings of each atom in a molecule. Building on these successful applications of MNAs and assuming that such a description reflects the influence of neighboring atoms on the biotransformation center, we use them as the basis for predicting biotransformations for compounds.

Mathematical Approach. PASS-BioTransfo discriminates between compounds that undergo or do not undergo a particular biotransformation by analysis of the multivariate space of MNA descriptors. The contribution of every descriptor d_i to a particular biotransformation B_j is estimated as a conditional probability value $p(B_j|d_i) = n_{ij}/n_i$, where n_{ij} is the total number of compounds in the training set that contain the descriptor d_i and belong to the class of biotransformation B_j and n_i is the total number of compounds that contain the descriptor d_i. The contributions of all the descriptors of the molecule to the biotransformation B_j are summarized in the specially designed statistic

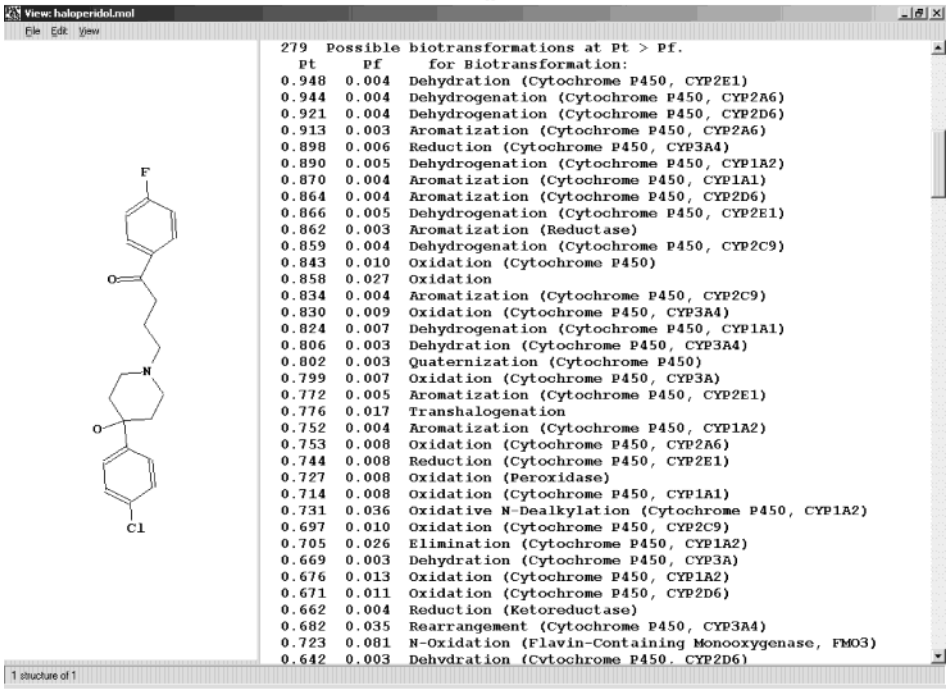
$$t = (1+(s - s_0)/(1 - s^*s_0))/2 \qquad (1)$$

where $s = \mathrm{Sin}(\sum_i ArcSin(2^*p(B_j|d_i) - 1)/m)$, $s_0 = 2^*p(B_j) - 1$, $p(B_j)$ is the *a priory* probability of biotransformation B_j, and m is the number of molecule's descriptors.

A smooth estimate of empirical distributions of *t*-values for compounds from the training set, which undergo a particular biotransformation (t_t) as well as for those that do not undergo this biotransformation (t_f) are estimated and stored (see Appendix 1 of the Supporting Information). When PASS-BioTransfo encounters a new compound, MNA descriptors are generated and the *t*-statistic is calculated. Comparison of the *t*-value for the new compound with the distributions of t_t and t_f of the training set yields the probabilities P_t and P_f of assigning a compound to the classes of "compounds which undergo biotransformation B_j" and "compounds which do not undergo biotransformation B_j".



a



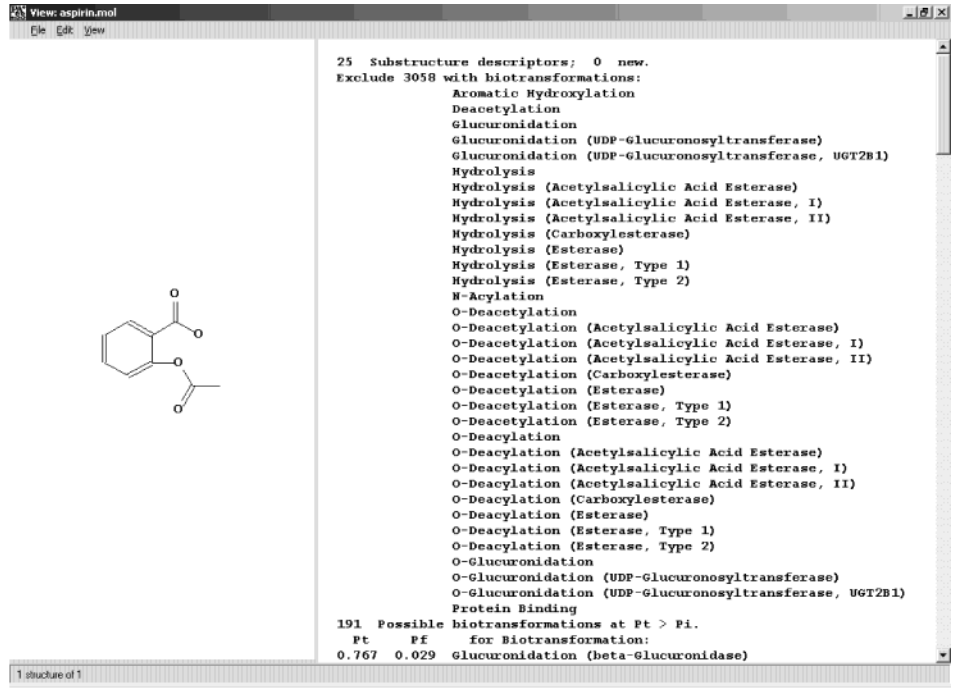
b

Figure 4. Prediction of biotransformation potential for haloperidol. The Metabolite database was used for training. Biotransformations that are known for the compound (a) may be compared with predicted biotransformations (b).

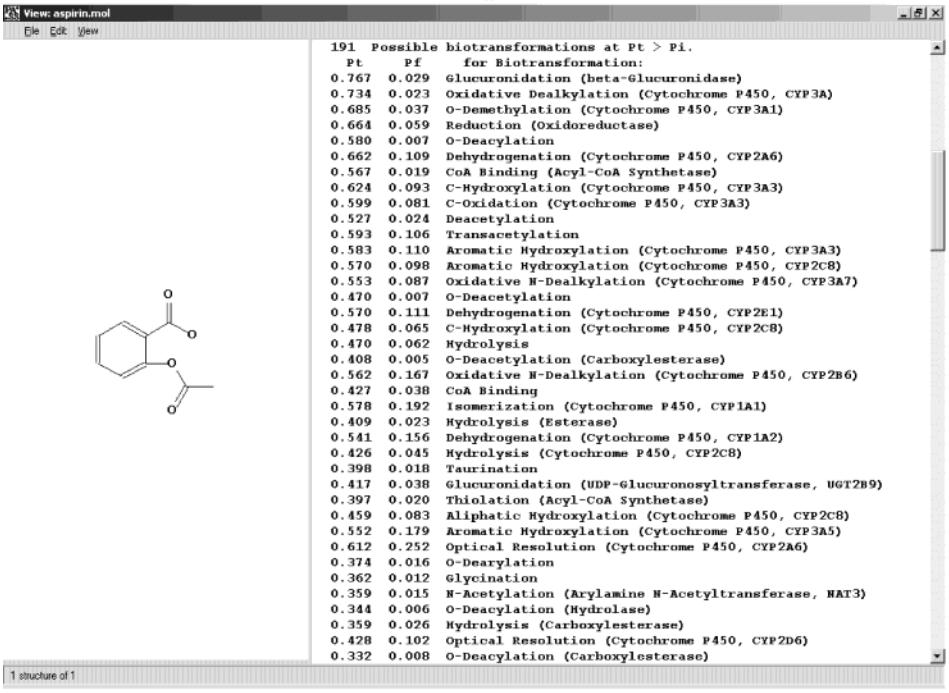
This training procedure and analysis is carried out for several hundred classes of biotransformation automatically.

Application. A prediction is possible for those classes of biotransformation that are represented by at least 3 substrates in the training set. For example, if we use all biotransformations from the Metabolite database, the training set includes 15 044 compounds with a unique structure and represents 7728 classes of biotransformation, but prediction is possible for only 1927 classes of biotransformation. When the program finds a new compound in its input, the list of possible biotransformation classes is predicted. If the structure of the compound under study coincides with that of one of the training set compounds, it is left out of the training set before the program calculates the prediction.

Figure 4 shows as an example of predicted biotransformation classes, the results for the haloperidol molecule, using all biotransformations from the Metabolite database as a training set. Since this structure was found in the training set, it was left out before the prediction was made. The upper list represents biotransformations found for the compound in the training set. The bottom list shows a part of the predicted biotransformations arranged in descending order of $P_t - P_f$ values. Only those biotransformaton classes are included in the list for which the $P_t - P_f$ value is positive.



a



b

Figure 5. Prediction of biotransformation potential for acetylsalicylic acid. The Metabolite database was used for training. Biotransformations that are known for the compound (a) may be compared with predicted biotransformations (b).

One can see that such transformations as *dehydration*, *dehydrogenation*, *aromatization*, *reduction*, *oxidation*, *quaternization*, and *oxidative N-dealkylation* are predicted correctly. The type of enzyme is predicted correctly, e.g., for *aromatization (cytochrome P450, CYP3A4)*, *reduction (cytochrome P450, CYP3A4)*, and *N-dealkylation (cytochrome P450)*. One can see that, for some transformations, additional enzymes and isoenzymes are predicted by PASS-BioTransfo.

Figure 5 shows another example, the predictions for acetylcalicylic acid.

Among the transformations included in the list for acetylsalicylic acid *hydrolysis*, *O-deacylation*, *O-deacetylation*, *glucuronidation*, and *aromatic hydroxylation* are predicted correctly. For some of them additional suggestions about involved enzymes have been made. Such reaction as *O-demethylation*, *dehydrogenation*, *oxidative N-dealkylation*, *aliphatic hydroxylation*, and *N-acetylation* are metabolically meaningless for this compound. It is interesting that transformations involving a nitrogen atom are present in both the known (*N-acylation*) and the predicted (*oxidative N-dealkylation*, *N-acetylation*) lists although acetylsalicylic acid does not possess a nitrogen atom. These errors are probably related to the data presented in the training set. For example, a more detailed analysis of the Metabolism database revealed

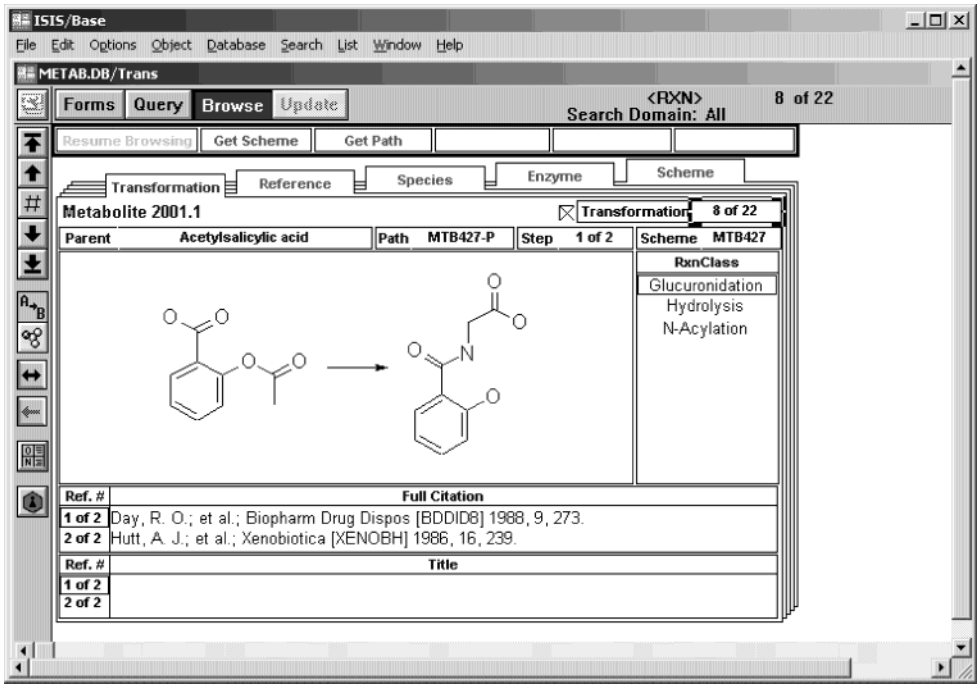


Figure 6. N-Acylation of acetylsalicylic acid reported in the Metabolite database.

presence the reaction of N-acylation of acetylsalicylic acid (Figure 6).

It is obvious that the more correct name of this reaction would be *glycination*. As one can see in Figure 5(b), *glycination* is included in the prediction list.

EVALUATION OF THE APPROACH

The purpose of our analysis was to assess the applicability of the approach to the prediction of different biotransformation classes. For this, we carried out several experiments in silico. From the six sets of biotrsansformations described above we prepared several pairs of training and evaluation sets. For every such pair, we estimated the accuracy of prediction of a particular biotransformation class. We used the following algorithm for the evaluation:

1. Leave-one-out (LOO) cross-validation with a training set.
2. Validation with an independent evaluation set.
3. Comparison of the results obtained in steps 1 and 2.

The accuracy of predicting a particular biotransformation class was estimated as

$$\mathrm{IAP} = 100{*}N\{t_t > t_f\}/(N_t{*}N_f),\ \%$$

where $N\{t_t > t_f\}$ is the number of cases when the *t*-value estimated for a compound that undergoes the biotransformation of that class exceeds the *t*-value for a compound that does not undergo this biotransformation, all pairs of transformed/untransformed compounds of the evaluation set being compared; N_t and N_f are the number of compounds that undergo and do not undergo the biotransformation of that class, respectively. *IAP* ranges from 0 to 100% and is interpreted as follows.

- *IAP* > 50% − Statistically significant prediction. In *IAP* the percent of cases the program correctly recognizes if a compound undergoes or does not undergo a particular class of biotransformation.
- *IAP* = 50% − No prediction. The probability of correct prediction is equal to random.
- *IAP* < 50% − False prediction. The prediction is positive for compounds that do not really undergo a particular class of biotransformation. This event is possible in very rare cases when (1) the data describe a particular class of biotransformation incorrectly or (2) the number of compounds in the evaluation/training set is not large enough for the evaluation.

The detailed explanation of IAP statistics is given in Appendix 2 of the Supporting Information.

For the first step of the evaluation, every compound was sequentially left out of the training set and the *t*-value calculated. For the second step, the *t*-value was calculated for every compound from the evaluation set.

MDL Metabolite vs Accelrys Metabolism Databases. To evaluate the accuracy of prediction for different biotransformation classes we used two data sets: (1) substrates from the Metabolite database and (2) substrates from the Metabolism database.

We used the first set as a training set and the second one as an evaluation set and vice versa. Therefore, we have two pairs of the training-evaluation sets. First we trained the program and estimated the accuracy of predicting a particular biotransformation by a LOO procedure with the training set (IAP-LOO). Then, we estimated the accuracy of prediction for the evaluation set (IAP-ES). For this comparison, we selected only those classes of biotransformation for which there were at least 10 substrates in both the training and evaluation sets.

Human vs Animal Biotransformations. The next task was to estimate if it is possible to use the data of animal experiments for predicting biotransformations in human. For this, we trained the program on animal biotransformations and evaluated it on human biotransformations.

The first experiment was carried out with the biotransformations from the MDL Metabolite database. We used substrates of animal transformations as the training set and

Table 1. Examples of Biotransformation Classes Predicted after Training with Metabolite

biotransformation	no. of compds	IAP-LOO,[a] %
aliphatic hydroxylation	1477	79.3
aliphatic hydroxylation (cytochrome P450)	443	81.4
aliphatic hydroxylation (cytochrome P450, CYP2D6)	52	82.5
aromatic hydroxylation	1575	84.0
aromatic hydroxylation (cytochrome P450)	403	85.5
aromatic hydroxylation (cytochrome P450, CYP3A4)	78	85.0
reduction	1075	84.9
reduction (aldehyde reductase)	12	87.1
reduction (aldehyde reductase (NADPH), AR-H)	16	97.9
hydrolysis	3364	87.4
hydrolysis (aminopeptidase)	53	98.9
hydrolysis (carboxylesterase)	148	93.3
hydrolysis (epoxide hydrolase)	105	99.0
conjugation	860	82.1
conjugation (glutathione transferase)	13	86.3
conjugation (UDP-glucuronosyltransferase)	72	83.5
conjugation (sulfotransferase)	35	92.9

[a] IAP-LOO is the accuracy of prediction estimated in a LOO procedure.

Table 2. Examples of Biotransformation Classes Predicted after Training with Metabolism

biotransformation	no. of compds	IAP-LOO,[a] %
aliphatic hydroxylation	260	86.4
allylic hydroxylation	99	95.1
aromatic hydroxylation	567	82.1
reduction	298	84.0
hydrolysis	544	85.0
N-deoxygenation	110	95.8
S-methylation	32	94.3
O-deacylation	123	95.1
epoxide cleavage	30	98.4
conjugate formation	677	73.4

[a] IAP-LOO is the accuracy of prediction estimated in a LOO procedure.

substrates of human transformations as the evaluation set. IAP-LOO and IAP-ES values were calculated for those classes of biotransformations for which there were at least 10 substrates in both the training and the evaluation sets.

The equivalent experiment was then carried out with the Accelrys Metabolism database.

RESULTS AND DISCUSSION

MDL Metabolite vs Accelrys Metabolism Databases. When we used the substrates of Metabolite for the training of the program, 1927 classes of biotransformation were predicted with an average accuracy of prediction of 88.5%. When training with the substrates from Metabolism database, 178 classes of biotransformation were predicted with an average accuracy of 85.2%. Examples of biotransformations predicted for both databases are given in Tables 1 and 2. The complete list of predicted classes of biotransformation is available as Supporting Information.

In Table 3, the accuracy of prediction estimated by LOO cross-validation for the training sets is compared with that estimated for the evaluation sets. As one can see, the average

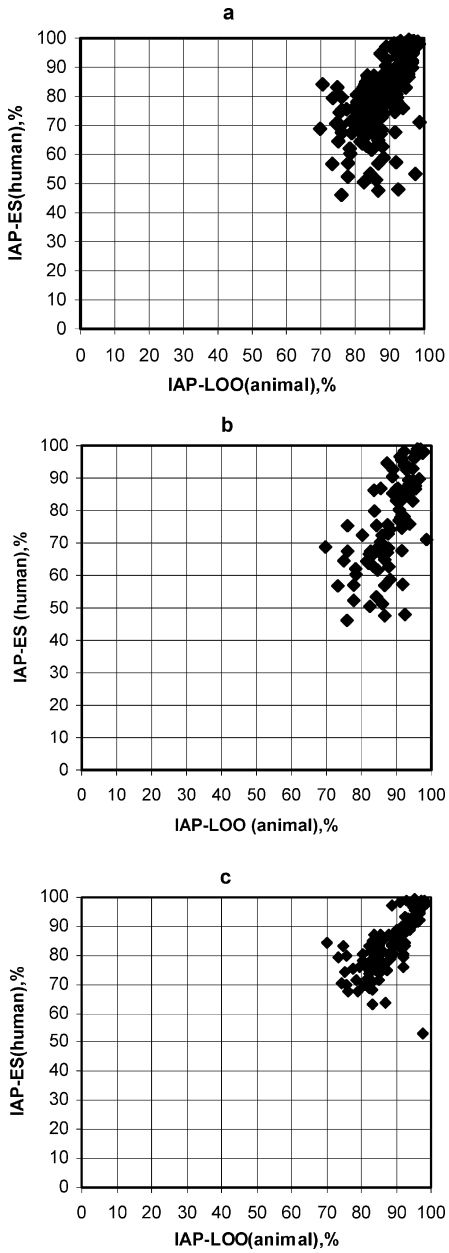


Figure 7. Human vs animal biotransformations (Metabolite). IAP-LOO is the accuracy of prediction estimated in LOO procedure with the training set based on animal biotransformations. IAP-ES is the accuracy of prediction estimated for the evaluation set based on human biotransformations. These values are given for all classes of biotransformation (**a**); for classes of biotransformation that include (**b**) and do not include (**c**) enzyme information, respectively.

accuracy of prediction estimated for both evaluation sets is about 77%, which is lower than that estimated for each one of the training sets. Classes of biotransformation such as *O-deacetylation, S-dealkylation, N-demethylation, O-demethylation, S-oxidation, N-deacetylation, S-methylation*, and *N-debenzylation* are cross-predicted with very high accuracy for all sets of data. Several classes are predicted with significantly lower accuracy in all experiments, for example, *oxidation, hydroxylation*, and *cleavage*. These are less specific biotransformations that may involve many different reaction centers and many different enzymes. Some classes of biotransformation are predicted with very low accuracy (or not predicted at all) for the evaluation sets despite the accuracy estimated for both the Accelrys and MDL training

Table 3. Cross-Prediction with MDL Metabolite (MT) and Accelrys Metabolism (MM)

biotransformation	no. of compds (MT)	no. of compds (MM)	IAP-LOO[c] (MT), %	IAP-ES[c] (MT),[a] %	IAP-LOO[c] (MM), %	IAP-ES[c] (MM),[b] %
O-deacetylation	111	28	97.7	98.2	98.8	94.9
S-dealkylation	161	43	97.0	96.1	96.7	93.8
N-demethylation	759	205	96.5	95.8	94.8	94.7
O-demethylation	500	149	95.4	95.3	95.4	91.6
dehalogenation	473	42	96.1	95.0	91.3	85.0
S-oxidation	483	138	96.9	95.0	96.0	95.6
N-deacetylation	96	19	96.8	93.8	94.0	94.9
S-methylation	144	32	98.0	93.5	94.3	91.4
dehydration	304	13	88.0	93.1	76.4	71.4
deacetylation	32	21	84.7	92.1	90.5	87.5
N-debenzylation	15	34	97.1	91.7	97.1	96.9
ring contraction	30	17	94.3	90.7	80.4	80.2
N-formylation	27	10	85.3	89.5	82.7	80.6
N-methylation	96	14	87.6	88.8	72.8	71.7
N-dearylation	37	13	88.1	88.2	91.6	86.6
N-acetylation	559	88	91.6	87.7	89.2	87.5
N-hydroxylation	177	65	91.3	87.5	92.5	86.8
O-dealkylation	825	197	92.0	86.6	89.4	80.9
O-dearylation	68	14	95.0	86.6	80.6	76.2
decarboxylation	316	18	84.2	86.5	87.9	69.7
decarboxylation	316	18	84.2	86.5	87.9	69.7
N-dealkylation	1703	387	91.4	86.3	88.1	86.8
N-oxidation	487	210	91.4	84.9	86.9	84.5
O-deacylation	214	123	95.6	84.1	95.1	91.9
S-alkylation	112	25	98.2	84.0	90.3	87.0
inversion	62	14	80.6	83.8	74.6	68.8
N-deacylation	360	132	91.9	83.6	86.9	88.1
N-acylation	314	90	92.6	81.9	86.7	88.4
aromatic hydroxylation	1575	567	84.0	80.8	82.1	78.6
aromatic methoxylation	58	38	88.3	80.3	81.0	79.2
hydrolysis	3364	544	87.4	79.5	85.0	79.4
O-methylation	270	19	89.8	78.8	77.5	78.4
O-alkylation	191	18	89.8	77.6	66.7	79.2
hydrogenation	469	61	85.8	77.2	84.9	76.5
C-demethylation	29	13	84.7	76.8	70.3	66.1
reduction	1075	298	84.9	76.5	84.0	77.3
aliphatic hydroxylation	1477	260	79.3	76.5	86.4	66.1
acetylation	42	31	87.0	74.5	85.4	79.0
deamination	356	39	89.1	74.4	81.3	72.0
dehydrogenation	542	76	82.4	73.3	78.3	66.7
esterification	98	43	86.1	72.5	85.2	63.7
rearrangement	249	40	81.9	71.7	82.0	61.6
N-alkylation	76	20	89.1	70.9	70.2	73.0
C-deacylation	11	13	95.4	70.7	61.9	86.2
C-hydroxylation	2576	930	76.2	70.1	74.6	68.5
methylation	14	38	97.3	67.8	84.3	94.6
C-oxidation	2373	1284	77.9	67.6	72.4	65.1
elimination	274	55	81.7	66.6	81.9	53.9
O-acylation	45	26	90.8	65.5	71.1	78.5
C-dealkylation	75	60	82.4	64.8	70.9	70.2
dearylation	10	31	95.3	63.5	78.5	85.8
amination	38	35	86.2	63.1	79.3	64.3
hydration	172	20	88.2	61.8	81.8	64.3
C-alkylation	18	15	89.1	61.4	77.3	68.5
alkylation	15	53	94.8	60.9	72.2	85.5
deacylation	29	167	84.7	55.9	84.4	68.5
deoxygenation	28	113	87.9	54.0	85.8	68.1
cleavage	66	582	79.0	53.1	79.9	54.4
hydroxylation	430	593	83.9	52.4	74.5	58.5
oxidation	1265	771	76.8	50.7	69.7	54.1
demethylation	36	175	85.6	50.5	86.8	55.7
dealkylation	92	359	85.4	48.2	81.8	64.9
average			88.9	77.2	83.1	77.4

[a] Accelrys Metabolism is used as the evaluation set. [b] MDL Metabolite is used as the evaluation set. [c] IAP-LOO is the accuracy of prediction estimated in LOO procedure; IAP-ES is the accuracy of prediction estimated for the evaluation set.

sets themselves being rather good. This is the case, for example, for *dealkylation*, *demethylation*, *deoxygenation*, *deacylation*, *rearrangement*, and *elimination*. These kinds of differences may be caused by differences in the classification of transformations in the Metabolite and Metabolism databases.

In general, reasonable prediction is possible for many different classes of biotransformation, both for rather un-

Table 4. Prediction for Human and Animal Sets of Compounds

biotransformation	no. of compds (human)	no. of compds (animal)	IAP-LOO[d] (animal)[a]	IAP-LOO[d] (human)[b]	IAP-ES[d] (human)[c]
hydrolysis (aminopeptidase)	21	30	96.1	97.6	99.0
dehalogenation (cytochrome P450, CYP2E1)	15	16	97.0	96.9	98.9
dehalogenation (cytochrome P450)	24	65	96.1	98.0	98.6
nucleophilic addition (cytochrome P450)	11	17	92.3	93.1	98.2
hydrolysis (epoxide hydrolase)	41	72	97.9	96.8	98.1
ring opening (epoxide hydrolase)	36	62	97.6	96.4	97.7
nucleophilic addition (epoxide hydrolase)	10	31	95.6	87.1	97.6
tautomerization (aldehyde oxidase)	10	13	91.1	86.8	96.7
dehalogenation (glutathione transferase)	26	44	95.5	95.6	96.5
tautomerization (xanthine oxidase)	15	13	95.1	97.7	96.3
ring opening (glutathione transferase)	12	25	91.6	91.3	96.2
S-oxidation (flavoprotein-linked monooxygenase)	11	21	91.5	90.0	95.6
C-hydroxylation (aldehyde oxidase)	12	16	87.4	83.9	94.6
O-alkylation (catechol O-methyltransferase)	23	29	92.2	91.7	93.9
O-methylation (catechol O-methyltransferase)	23	36	93.4	91.7	93.4
aromatization (cytochrome P450)	35	63	94.7	92.5	93.0
N-oxidation (flavoprotein-linked monooxygenase)	15	15	89.1	92.1	92.6
oxidative deamination (monoamine oxidase)	15	12	93.0	96.2	92.5
glutathionation (glutathione transferase)	61	175	88.9	90.6	90.4
S-oxidation (cytochrome P450)	44	49	96.6	89.1	89.8
O-demethylation (cytochrome P450)	97	53	93.6	92.6	89.4
N-demethylation (cytochrome P450)	134	86	94.9	93.4	89.2
C-oxidation (aldehyde dehydrogenase)	14	25	93.6	85.7	88.0
epoxidation (cytochrome P450, CYP1A1)	14	11	95.4	88.4	87.6
O-dealkylation (cytochrome P450)	119	85	92.9	90.8	87.0
epoxidation (cytochrome P450)	63	132	90.5	86.8	86.8
hydration (cytochrome P450)	10	17	85.6	77.2	86.8
N-hydroxylation (cytochrome P450)	23	51	95.5	84.2	86.6
ring opening (cytochrome P450)	42	70	90.0	89.8	86.5
O-sulfation (phenol sulfotransferase)	58	19	83.7	93.1	86.3
N-reduction (cytochrome P450)	11	32	93.0	77.5	85.7
N-acylation (n-Acetyltransferase)	30	30	93.8	94.0	85.3
hydrolysis (carboxylesterase)	76	58	89.0	92.1	85.3
dearomatization (cytochrome P450)	51	111	91.5	87.2	83.9
N-demethylation (cytochrome P450, CYP3A)	36	28	90.0	90.3	83.0
N-acetylation (N-acetyltransferase)	36	54	94.7	95.7	83.0
dearomatization (cytochrome P450, CYP1A1)	13	13	91.1	94.7	82.8
C-oxidation (alcohol dehydrogenase)	26	21	90.9	84.5	80.2
O-sulfation (sulfotransferase)	42	56	83.8	87.2	79.9
O-deacylation (carboxylesterase)	11	16	91.6	93.2	78.6
hydrolysis (esterase)	123	90	92.2	91.5	78.2
oxidative N-dealkylation (cytochrome P450)	60	42	92.2	87.9	78.0
epoxidation (cytochrome P450, CYP2E1)	23	10	91.3	88.5	77.9
N-dealkylation (cytochrome P450)	247	119	90.8	90.3	77.0
reduction (carbonyl reductase)	18	32	93.8	86.9	75.9
O-deacylation (esterase)	32	14	87.5	94.5	75.7
O-glucuronidation (UDP-glucuronosyltransferase)	307	167	84.4	88.1	75.3
ring closure (cytochrome P450)	12	16	76.1	81.3	75.3
aromatic hydroxylation (cytochrome P450, CYP1A1)	38	29	91.6	83.5	74.6
N-oxidation (cytochrome P450)	53	80	88.3	84.3	74.2
glucuronidation (UDP-glucuronosyltransferase)	182	98	87.7	85.1	72.8
oxidative deamination (cytochrome P450)	31	27	80.3	81.3	72.4
aromatic hydroxylation (cytochrome P450)	211	202	85.9	83.4	72.3
conjugation (sulfotransferase)	12	23	98.7	77.5	71.1
dehydration (cytochrome P450)	15	26	85.5	78.0	70.3
hydrolysis (cytochrome P450)	91	69	85.0	83.3	68.9
C-hydroxylation (cytochrome P450, CYP3A)	90	19	69.7	77.5	68.8
dehydrogenation (cytochrome P450)	52	56	87.7	80.7	68.5
tautomerization (cytochrome P450)	36	63	91.6	87.4	67.7
oxidative dealkylation (cytochrome P450)	21	21	76.1	74.0	67.5
C-oxidation (aldehyde oxidase)	17	11	82.7	88.0	67.4
isomerization (cytochrome P450)	14	23	87.4	80.8	67.1
deamination (cytochrome P450)	18	27	83.5	77.2	66.7
reduction (cytochrome P450)	34	41	85.7	64.1	66.7
oxidation (cytochrome P450, CYP2E1)	19	10	82.3	73.7	66.6
C-oxidation (cytochrome P450, CYP2E1)	66	40	86.7	79.7	64.9
oxidation (cytochrome P450)	76	70	75.0	74.1	64.5
C-hydroxylation (cytochrome P450)	456	459	81.5	79.1	64.4
aliphatic hydroxylation (cytochrome P450)	255	201	82.0	78.7	63.7
C-oxidation (cytochrome P450)	260	300	83.4	75.6	62.9
elimination (cytochrome P450)	15	24	87.9	80.4	62.7
rearrangement (cytochrome P450)	21	16	78.3	72.8	62.1

Table 4 (Continued)

biotransformation	no. of compds (human)	no. of compds (animal)	IAP-LOO[d] (animal)[a]	IAP-LOO[d] (human)[b]	IAP-ES[d] (human)[c]
C-hydroxylation (cytochrome P450, CYP1A1)	65	36	84.7	79.0	61.6
sulfation (sulfotransferase)	37	12	78.6	91.8	60.3
aromatic hydroxylation (cytochrome P450, CYP1A2)	60	18	88.2	85.3	58.8
aliphatic hydroxylation (cytochrome P450, CYP2E1)	48	11	91.8	81.8	57.3
hydrogenation (cytochrome P450)	23	19	77.9	69.4	57.0
aromatic hydroxylation (cytochrome P450, CYP2E1)	39	15	86.7	80.9	56.9
hydroxylation (cytochrome P450)	43	26	73.3	75.7	56.7
C-hydroxylation (cytochrome P450, CYP2E1)	87	28	84.3	80.1	53.4
C-hydroxylation (cytochrome P450, CYP1A2)	115	23	77.8	83.3	52.3
O-glucuronidation (UDP-glucuronosyltransferase, UGT2B1)	16	16	86.0	87.9	51.2
C-oxidation (cytochrome P450, CYP3A)	40	16	82.4	67.0	50.4
conjugation (UDP-glucuronosyltransferase)	42	29	92.5	77.1	47.9
hydrolysis (cytochrome P450, CYP2E1)	16	16	86.7	79.6	47.6
C-hydroxylation (cytochrome P450, CYP1A)	11	14	75.9	75.8	46.1
average			88.3	85.5	76.8

[a] Animal biotransformations are used for training. [b] Human biotransformations are used for training. [c] Animal biotransformations are used for training; human biotransformations are used as the evaluation set. [d] IAP-LOO is the accuracy of prediction estimated in LOO procedure; IAP-ES is the accuracy of prediction estimated for the evaluation set.

specific ones such as *C-oxidation* and for highly specific ones such as *C-oxidation (monoamine oxidase, MAO-A)*. Cross-prediction is satisfactory although Metabolite contains mostly biotransformations of pharmaceutical compounds, while Metabolism database focuses on biotransformations of agrochemicals, food additives, environmental, and industrial chemicals.

Human vs Animal Biotransformations. For the MDL Metabolite database, the training set comprised 11 599 unique substrates of animal biotransformations. The evaluation set included 6376 unique substrates of human biotransformations. In Figure 7(a), IAP values estimated for different classes of biotransformation by the LOO procedure using the animal training set are plotted against IAP values estimated for the same classes using human data. As in the previous experiment, we compare those classes of biotransformation that are represented by at least 10 substrates in both the training and the evaluation set. In total, the graph includes 207 points, each of which represents one biotransformation class. The average accuracy of prediction is 85.8% for the LOO procedure and 79.7% for the validation with human data.

It can be seen from Figure 7(a) that in the majority of cases, the accuracy of prediction evaluated on human data exceeds 70%. There is no clear relationship between IAP-LOO and IAP-ES values. However, IAP-ES tends to be higher when IAP-LOO is higher. We also separated classes of biotransformation into those that include and those that do not include enzyme information. Diagrams 7(b) and (c) represent the IAP-LOO vs IAP-ES for classes of biotransformation that include and do not include enzyme information, respectively. As one can see, classes of biotransformation that do not include enzyme information are predicted better from animal data. This probably indicates that a general type of chemical modification depends basically on the molecule structure, while an enzyme involved in a particular modification depends also on peculiarities of the whole enzymatic spectrum of a particular species. The results of the prediction for those classes of biotransformation which include enzyme and isoenzyme information are given in more detail in Table 4, including the IAP-LOO (animal) and IAP-ES (human) values shown in Figure 7b. IAP-LOO (human)

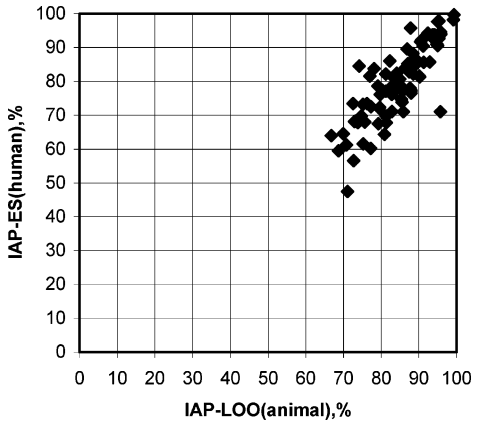


Figure 8. Human vs animal biotransformations (Metabolism). IAP-LOO is the accuracy of prediction estimated in LOO procedure with the training set based on animal biotransformations. IAP-ES is the accuracy of prediction estimated for the evaluation set based on human biotransformations.

is the accuracy estimated in LOO cross-validation for the human set. The data are arranged by descending IAP-ES values. As one can see, for some classes of biotransformation, IAP-ES is equal to or even exceeds IAP-LOO. At the same time, there are many classes of biotransformation that are predicted with low accuracy or not predicted at all from animal data. For example, *aromatic hydroxylation (cytochrome P450, CYP1A2)* is predicted well for both animal and human sets separately but is not predicted for human set when the program is trained with animal data. One might hypothesize that a compound undergoes the same chemical modification in the animal and human body, while the enzymes preferably catalyzing the modification are different.

For the Accelrys Metabolism database, 2186 unique substrates found in animal biotransformations were used as a training set, and 708 substrates of human biotransformations were used as an evaluation set. The average accuracy of prediction was approximately 85% for the training set and 80% for the evaluation set. The comparative results for 86 biotransformation classes are given in Figure 8.

From Figure 8 one can see that the accuracy of a human biotransformation prediction is the higher, the higher the accuracy is that was determined by the LOO procedure with

the animal training set. For some biotransformations, the accuracies estimated for human data and for animal data are practically the same.

CONCLUSIONS

A series of in silico experiments was carried out with two well-known commercially available metabolic databases: Metabolite (MDL) and Metabolism (Accelrys). At the present stage of the work, we have not tried to verify or correct the classification of transformations as well as enzymatic information contained in these databases. We also did not change the algorithm and descriptors developed originally for biological activity prediction. No expert data were used to determine a priori impossible transformations. However, even given all these limitations we come to definite conclusions:

1. The method can be used for the prediction of many different classes of biotransformation from chemical structure for drug-like compounds.

2. The data of animal experiments (mammals), being used for training, provide reasonable accuracy of prediction for human biotransformation classes that do not include information about enzymes.

3. For robust prediction of human biotransformation classes that include enzyme and isoenzyme information, use of data of human experiments is necessary.

ACKNOWLEDGMENT

We gratefully acknowledge the support of this work by the Russian Ministry of Science and Technology (Grant # 43.071.1.1.2530) and the assistance of MDL Information Systems Inc. by providing the Institute of Biomedical Chemistry RAMS with a license to ISIS and the Metabolite database used in this study.

Supporting Information Available: Smooth estimation of the distribution function (Appendix 1), detailed explanation of IAP statistics (Appendix 2), and complete list of predicted classes of biotransformation (Tables S1 and S2). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Wermuth, C. G. et al. Glossary of terms used in medicinal chemistry. *Ann. Rep. Med. Chem.* **1998**, *33*, 385−395.

(2) Rendic, S. Summary of information on human CYP enzymes: human P450 metabolism data. *Drug Met. Rev.* **2002**, *34*, 83−448. In Special Issue on Human Cytochromes P450 (Human CYPs): Human Cytochrome P450 Enzymes, a Status Report Summarizing their Reactions, Substrates, Inducers, and Inhibitors − 1st Update. Guengerich, F. P., Rendic, S., Eds.; *Drug Met. Rev.* **2002**, *34* (parts 1 and 2).

(3) Wilkinson, G. R. In *Goodman & Gilman's The Pharmacological Basis of Therapeutics, 10/e*; Hardman, J. G., Limbird, L. E., Gilman, A. G., Eds.; McGraw-Hill: 2001; Chapter 1, pp 3−29.

(4) Langowski, J.; Long, A. Computer systems for the prediction of xenobiotic metabolism. *Adv. Drug Delivery Rev.* **2002**, *54*, 407−415.

(5) Van de Waterbeemd, H. High-throughput and *in silico* techniques in drug metabolism and pharmacokinetics. *Curr. Opin. Drug Discovery Dev.* **2002**, *5*, 33−43.

(6) Lewis, D. On the recognition of mammalian microsomal cytochrome P450 substrates and their characteristics. *Biochem. Pharmacol.* **2000**, *60*, 293−306.

(7) Lewis, D. COMPACT: a structural approach to the modeling of cytochromes P450 and their interactions with xenobiotics. *J. Chem. Technol. Biotechnol.* **2001**, *76*, 237−244.

(8) Klopman, G.; Dimayuga, M.; Talafous, J. META. 1. A program for the evaluation of metabolic transformation of chemicals. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1320−1325.

(9) Talafous, J.; Sayre, L.; Mieyal, J.; Klopman, G. META. 2. A dictionary model of mammalian xenobiotic metabolism. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1326−1333.

(10) Darvas, F. Predicting metabolic pathways by logic programming. *J. Mol. Graphics* **1998**, *6*, 80−86.

(11) Greene, N.; Judson, P.; Langowski, J.; Marchant, C. Knowledge-based expert systems for toxicity and metabolism prediction: DEREK, StAR, and METEOR. *SAR QSAR Environ. Res.* **1999**, *10*, 299−313.

(12) http://www.ibmh.msk.su/PASS.

(13) Poroikov, V.; Filimonov, D.; Borodina, Yu.; Lagunin, A.; Kos, A. Robustness of biological activity spectra predicting by computer program PASS for noncongeneric sets of chemical compounds. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1349−1355.

(14) Poroikov, V. V.; Filimonov, D. A. Computer-assisted prediction of biological activity in a search for and optimization of new drugs. In *Nitrogen-containing heterocycles and alcaloides*; Iridium Press: Moscow, 2001; Vol. 1, pp 149−154.

(15) Anzali, S.; Barnickel, G.; Cezanne, B.; Krug, M.; Filimonov, D.; Poroikov, V. Discriminating between Drugs and Nondrugs by Prediction of Activity Spectra for Substances (PASS). *J. Med. Chem.* **2001**, *44*, 2432−2437.

(16) Poroikov, V. V.; Filimonov D. A. How to acquire new biological activities in old compounds by computer prediction. *J. Comput. Aid. Mol. Des.* **2002**, *16*, 819−824.

(17) Poroikov, V.; Filimonov, D.; Ihlenfeldt, W.-D.; Gloriozova, T.; Lagunin, A. A.; Borodina, Yu.; Stepanchikova, A. V.; Nicklaus, M. C. PASS Biological Activity Spectrum Predictions in the Enhanced Open NCI Database Browser. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 228−236.

(18) Stepanchikova, A. V.; Lagunin, A. A.; Filimonov, D. A.; Poroikov, V. V. Prediction of biological activity spectra for substances: evaluation on the diverse set of drug-like structures. *Curr. Med. Chem.* **2003**, *10*, 225−233.

(19) Suchkov, A. P.; Filimonov, D. A.; Stepanchikova, A. V.; Poroikov, V. V. *Abstr. 11th European Symposium on Quantitative Structure−Activity Relationships: Computer-Assisted Lead Finding and Optimisation.* Lausanne, Switzerland, 1996, P-32C.

(20) MDL Information Systems, Inc., 14600 Catalina Street, San Leandro, CA, U.S.A. (http://www.mdli.com).

(21) Accelrys Inc., 9685 Scranton Road, San Diego, CA 92121-3752, U.S.A. (http://www.accelrys.com).

(22) Jones, R. A. Y.; Bunnett, J. F. Nomenclature for organic chemical transformations. *Pure Appl. Chem.* **1989**, *61*, 725−768.

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

CI034078L