A New Statistical Approach to Predicting Aromatic Hydroxylation Sites. Comparison with Model-Based Approaches

Yu. Borodina,[‡] A. Rudik,[‡] D. Filimonov,[‡] N. Kharchevnikova,[§] A. Dmitriev,[‡] V. Blinova,[†] and V. Poroikov*[,‡]

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

Received May 19, 2004

A new approach is described that is able to predict the most probable metabolic sites on the basis of a statistical analysis of various metabolic transformations reported in the literature. The approach is applied to the prediction of aromatic hydroxylation sites for diverse sets of substrates. Training is performed using the aromatic hydroxylation reactions from the Metabolism database (Accelrys). Validation is carried out on heterogeneous sets of aromatic compounds reported in the Metabolite database (MDL). The average accuracy of prediction of experimentally observed hydroxylation sites estimated for 1552 substrates from Metabolite is 84.5%. The proposed approach is compared with two electronic models for P450 mediated aromatic hydroxylation: the oxenoid model using the atomic oxygen and the model using the methoxy radical as a model for the heme active oxygen species. For benzene derivatives, the proposed method is inferior to the oxenoid model and as accurate as the methoxy-radical model. For hetero- and polycyclic compounds, the oxenoid model is not applicable, and the statistical method is the most accurate. Broad applicability and high speed of calculations provide the basis for using the proposed statistical approach for high-throughput metabolism prediction in the early stages of drug discovery.

INTRODUCTION

Prediction of site-specific metabolism is a challenging task for modern computational chemistry. Mammals express a plethora of metabolic enzymes that can specifically transform various substrates.[1,2] It is currently accepted that the site of metabolism depends on both electronic substrate-enzyme interactions and orientation of a substrate within the enzyme active center.[3] Therefore, the majority of methods developed for prediction of site-specific metabolism try to model a mechanism of catalysis[4−9] or an enzyme structure.[10−15] Usually these methods are time-consuming and strongly depend on the quality and restrictions of a particular model.

At the same time, the increasing amount of information on metabolic transformations available in the literature and commercial databases[16,17] creates the prerequisite for development of predictive tools based on a statistical approach. Anticipated advantages of such an approach are fast data processing, the possibility to use heterogeneous and noisy data for training, and no dependence on the initial hypothesis concerning a mechanism of catalysis or an enzyme model.

In a previous work, we described a statistical approach capable of predicting many classes of biotransformation for chemical compounds.[18] 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. Examples of predicted classes are "Aromatic Hydroxylation", "Aromatic Hydroxylation (Cytochrome P450)", "Aromatic Hydroxylation (Cytochrome P450, CYP2D6)", "Aliphatic Hydroxylation (Cytochrome P450, CYP3A4)", "N-Dealkylation (Cytochrome P450)", "Hydrolysis (Aminopeptidase)", "Oxidative Deamination (Monoamine Oxidase)", etc. The structural formula of a chemical compound is used as input information. The result of the prediction is a list of more-probable classes of biotransformation arranged in descending order of their probability.

Herein we present a further development of this approach for predicting regioselectivity for individual classes of biotransformation. The method is based on the assumption that different potentially possible metabolic sites have different probability within the set of all possible substrate molecules. Using statistical analysis of known transformations, it is possible to estimate the probability of a given transformation occurring at each of the potential metabolic sites for a new molecule. The approach is designed to be universally applicable for any class of biotransformation. In this study, we analyzed the applicability of the approach to the prediction of aromatic hydroxylation sites. We compare our method with two model-based approaches that simulate the mechanism of aromatic hydroxylation by cytochromes P450. The point for comparison was the ability of each method to predict experimentally observed transformations.

THEORETICAL BASIS

General Approach. We assume that some positions of the aromatic system are preferable for hydroxylation in many

* Corresponding author phone: +7-095 245-2753; e-mail: vvp@ibmh.msk.su.
[†] Russian Institute for Scientific and Technical Information.
[‡] Institute of Biomedical Chemistry of the Russian Academy of Medical Sciences.
[§] Institute of Human Ecology and Environmental Health.

10.1021/ci049834h CCC: $27.50
Published on Web 09/25/2004

different substrates, whereas other potential positions are hydroxylated very rarely or not hydroxylated at all in naturally occurring biotransformations. Therefore, structural schemes of "real" (experimentally observed) transformations must have some similarity with each other and differ from "unreal" (not experimentally observed) transformations. Then, the goal is to construct a decision rule for distinguishing real transformations from potentially possible but not really existing ones. For this, we created a training set consisting of both experimentally observed transformations and transformations generated artificially for the same set of substrates, taking into consideration all potential sites of aromatic hydroxylation. For the computer representation of a transformation, 2D topological descriptors were introduced that code a transformation as a single entity including both the substrate and the product. Any particular descriptor can be found with a certain frequency among "real" and "unreal" transformations in the training set. For any new compound, its potential transformations are generated and coded by the applicable descriptors. The contribution of all descriptors is estimated, and every transformation is classified as "real" or "unreal" with a certain probability.

Coding a Transformation by Descriptors. The structural formula of a particular transformation includes a substrate formula on the left of the arrow and a product formula on the right. The transformation is coded by 2D topological descriptors that we refer to as RMNA descriptors (Reacting Multilevel Neighborhood of Atom). In the development of the RMNA descriptors, we started from MNA descriptors[19] developed earlier for biological activity prediction. In addition to standard MNAs that describe atoms and bonds in one molecule, RMNAs take into account atoms and bonds involved in a transformation. Descriptors are generated according to the following algorithm.

The first-level RMNA descriptor for an atom *A* has the form

$$\mathrm{RNMA}_1(A) = ([-]A[T]D_1[B]D_2[B]..D_i[B]...)$$

where *A* is the atom type; [-] is the label added to nonring atoms; [*T*] is a label for transformed atoms, which can take the values "<" for attached atom and ">" for removed atom, respectively; D_1, .., D_i... are the immediate neighbors of the atom in a lexicographical order; [*B*] is a label for bonds changed during the transformation, which can take the values "/" if a bond between *A* and one of the neighboring atoms is broken, "[\]" if a new bond is formed, "]" if a bond changes bond type (indicated in a substrate), and "[" if a bond changes bond type (indicated in a product).

The *n*th-level RMNA descriptor for an atom *A* is created by an iterative procedure and has the following substructure notation

$$\mathrm{RNMA}_n(A) = (\mathrm{RNMA}_{n-1}(A)\ \mathrm{RNMA}_{n-1}(D_1)\ \mathrm{RNMA}_{n-1}(D_2)...\ \mathrm{RNMA}_{n-1}(D_i)...)$$

where $\mathrm{RNMA}_{n-1}(A)$ is the (*n*-1)-level RMNA descriptor for the atom *A* and $\mathrm{RNMA}_{n-1}(D_i)$ is the *n*-1-level RMNA descriptor for its *i*th immediate neighbor.

This iterative process can be continued including second, third, etc. neighborhoods of each atom. Currently we use the 4th level descriptors. Descriptors generated for a substrate and for a product are combined in the same set, and every unique descriptor is saved. Therefore, an entire transformation is represented by the set of descriptors derived from a substrate and a product. An example of descriptors of the 4th level is given in Figure 1.

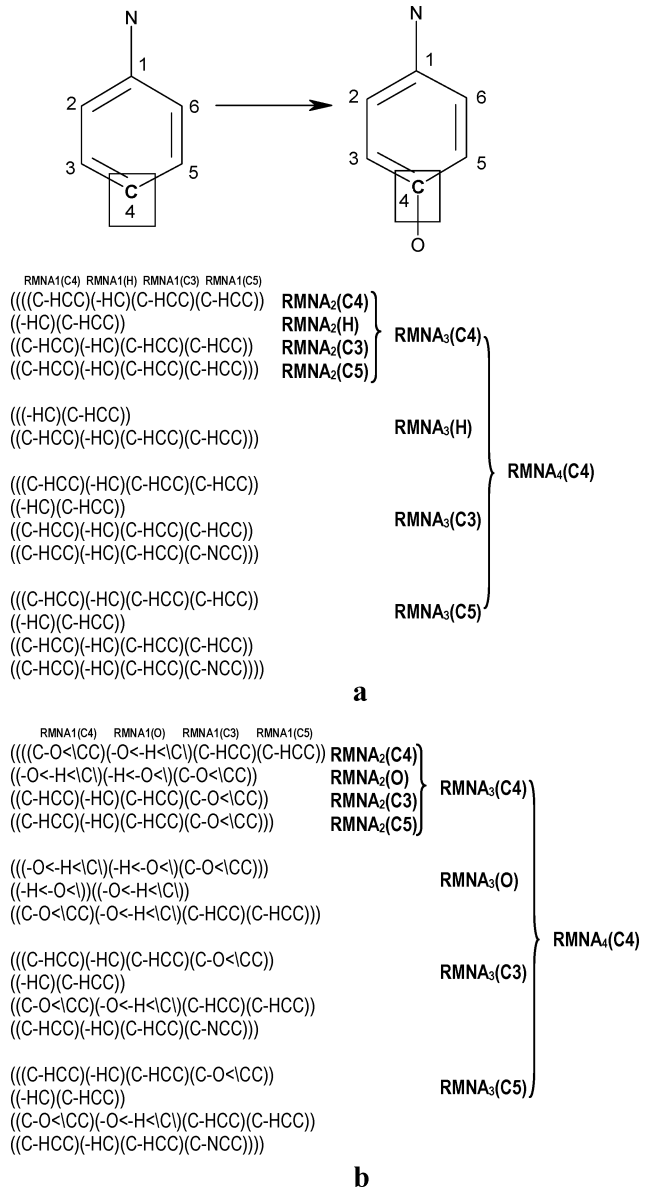


Figure 1. Two 4th level descriptors for a selected carbon atom (C4) in the substrate (a) and in the product (b). The iterative formation of descriptors from descriptors of previous levels is shown in bold type.

Generation of Potential Transformations. For any given substrate, all potential transformations of aromatic hydroxylation are generated using all potential sites within the molecule. The generation is based on the vocabulary of transformation patterns and a graph-searching algorithm. A transformation pattern is a transformation formula that describes the structural changes from a substrate to a product. When the substrate part of a pattern is found in a new molecule, it is replaced by the product part of the pattern. The vocabulary of patterns was prepared by automatic processing of transformations included in the Metabolism database (Accelrys)[16] followed by refinement by a human expert. Therefore, only patterns found in the database at least

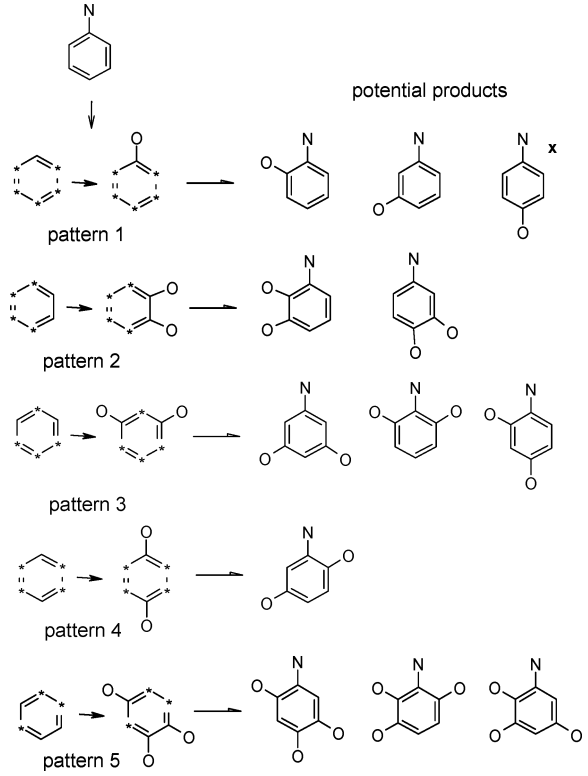


Figure 2. Transformation patterns and potential products generated for aniline. Asterisks indicate a carbon for which different numbers of neighbor atoms are allowed. The "x" indicates the experimentally observed product.

once are included in the vocabulary. Examples of transformation patterns and generated transformations are given in Figure 2.

Prediction Algorithm. The method of prediction is the same as the one used in the program PASS.[20−25] In the training set, containing n transformations, n_r transformations are real (experimentally reported), while others are not. Any descriptor i is found in n_i transformations, n_{ir} of those are real. For any given transformation of the training set its descriptors $\{d_1,...,d_i,...,d_m\}$ are generated, and the following value is calculated

$$t = (s - s_0)/(1 - s*s_0) \quad (1)$$

where

$$s = Sin\left(\sum_i ArcSin(2*n_{ir}/n_i - 1)/m\right), \quad s_0 = 2*n_r/n - 1$$

The empirical distributions of t-values for real transformations as well as for unreal transformations are estimated and saved in a file of a special format.

For a new transformation, RMNA descriptors are generated, and the t-statistics is calculated. Comparison of the t-value for a new transformation with the distributions of t-values for real and unreal, i.e., experimentally unobserved (see Materials and Methods section), transformations in the training set yields the probabilities P_r and P_u of assigning a transformation to the classes "real" and "unreal". The difference $\Delta P = P_r - P_u$ is the result of the prediction for one transformation. The higher the ΔP value, the more probable a transformation is. Note that if a transformation is already included in the training set it is automatically excluded from the training set before the prediction procedure starts for this transformation.

MATERIALS AND METHODS

Training Set. Real transformations were taken from the Metabolism v.2002.1 database (Accelrys).[16] A total of 789 structurally unique first step mammalian transformations referred in the database as belonging to the "Aromatic Hydroxylation" class were used. A substrate's structure was extracted from a particular transformation, and its potential transformations were generated and added to the training set. The resulting training set included 5747 transformations, 789 of them experimentally observed (reported in the database) while others were not. Both real and unreal transformations were coded by RMNA descriptors and saved in a binary file. Every real transformation has a specific label indicating its belonging to the "Aromatic Hydroxylation" class. Specific enzymes participating in each transformation were not taken into account since the Metabolism database does not contain such information.

Evaluation Sets. We validated the approach against transformations reported in the Metabolite v.2001.1 database (MDL Information Systems Inc.).[17] Four evaluation sets were prepared. To prepare the first set, we retrieved from the database 1552 compounds undergoing aromatic hydroxylation. For every molecule, all potential reactions of aromatic hydroxylation were generated. Experimentally observed transformations were labeled "Aromatic Hydroxylation". The total evaluation set included 2124 observed and 16361 unobserved transformations.

The second, third, and fourth evaluation sets were used for comparison with model-based quantum chemical calculations. We selected structurally diverse compounds with multiple potential sites of aromatic hydroxylation. Only compounds that had both observed and unobserved sites of hydroxylation were included in each of the evaluation sets. The resulting evaluation sets included 15 substituted benzenes, 17 compounds containing poly- and heterocycles, and 15 substrates of cytochrome 2D6. For every molecule, we generated all possible reactions of *one substituted* aromatic hydroxylation. Experimentally observed transformations were labeled "Aromatic Hydroxylation". In the fourth set only reactions catalyzed by CYP2D6 were labeled.

Note that all experimental data were taken from the Metabolite database "as is". Other information sources were not used. Expert validation and correction of the data were not carried out. If the database contained at least one reference for a transformation, the transformation was considered to have been experimentally observed. Test system, route of administration, dose, metabolites' amount, and stereochemistry were not taken into account.

Model-Based Approaches Used for Comparison. Most often aromatic hydroxylation is mediated by the cytochrome P450 enzymes. Two different models of active heme-iron oxygen species of cytochrome P450 were used for the prediction of the aromatic hydroxylation site. Both models successfully reproduce the aromatic hydroxylation site in the series of substituted benzenes with simple substituents.[5,7] The first model was the so-called oxenoid model.[4−6] According to this model, P450 breaks a dioxygen molecule and generates the active atomic oxygen species (oxens), which

readily react with substrates. The stability of an intermediate with one tetrahedral carbon atom relative to the substrate molecule is considered to be a factor that determines the preferable hydroxylation position. The lower the difference ΔH between the heat of formation of a "tetrahedral" intermediate and that of the substrate, the more probable a transformation is. In our calculations of benzene derivatives, we optimized the geometries of substrates using the MNDO Hamiltonian method with the MOPAC 6.0 software. The structure of the benzene ring for any intermediate was that of the "tetrahedral" intermediate of unsubstituted benzene and was taken from the publication of K. Korzekwa and coworkers.[4] The substituents' geometry for intermediates was assumed to be the same as in the substrate molecule. The heats of formation of substrate and intermediates were calculated for all possible positions of aromatic hydroxylation in a substrate molecule.

The second model was that proposed by J. Jones and coworkers.[7] According to this model, a methoxy radical instead of oxene serves as a model system for oxygenating species. The difference ΔH between the heat of formation of a substrate and that of an intermediate describes approximately the activation energy of oxidation.[7] The structures of both the substrate and the intermediate were minimized using the default procedure in SYBYL 6.9 and then optimized with the AM1 semiempirical Hamiltonian using MOPAC 6.0. The heat of formation of the tetrahedral intermediate resulting from the addition of a methoxy radical to a carbon atom of the aromatic ring was calculated for all possible positions of aromatic hydroxylation. Open-shell systems were treated with an unrestricted Hartree−Fock method.

Validation Method. The goal was to estimate the ability of our approach and each of the model-based approaches to recognize experimentally observed transformations in any given evaluation set. To do this, we ranked all potential transformations related to a particular substrate in accordance to (1) the ΔP value and (2) the ΔH value. Ideally, real (experimentally observed) transformations should have higher ranks than unreal transformations. The accuracy of prediction, which we call IAP (Independent Accuracy of Prediction), was estimated as

$$\mathrm{IAP} = \frac{N(rank_r < rank_u)}{N_r \cdot N_u} \cdot 100\%$$

where $N(rank_r < rank_u)$ is the number of cases when a real transformation has higher rank then an unreal transformation (all pairs real-unreal are compared); N_r and N_u are the number of real and unreal transformations of the substrate, respectively.[26] The IAP statistics was calculated for every particular substrate and averaged over all substrates in the evaluation set.

RESULTS AND DISCUSSION

Table 1 contains IAP values for all evaluation sets calculated for the results obtained by our method as well as the quantum chemical approaches. Table 2 compares the average time required for processing of one transformation by each one of the methods.

In general, the accuracy of the statistical prediction is comparable with, or higher than, that of model-based predictions, while the time required per molecule is much less.

Table 1. Accuracy of Prediction (IAP) Estimated for the Statistical Approach (s), Oxenoid Model (o), and Methoxy Radical Model (m)

	all substrates[a]	benzene derivatives[b]	heteropolycyclic compounds[b]	CYP2D6 substrates[b]
no. of substrates	1552	15	17	15
IAPs, %	84.5	85.0	83.1	89.6
IAPo, %		95.0	not applicable	not applicable
IAPm, %		84.2	44.9	70.1

[a] All possible reactions of aromatic hydroxylation were used. [b] Only reactions of one substituted aromatic hydroxylation were used.

Table 2. Average CPU Time Required for Computation of One Transformation by the Statistical Approach, Oxenoid Model, and Methoxy Radical Model

method	CPU time, s
statistical[a]	0.03
oxenoid model[b]	2
methoxy-radical model[b]	60

[a] PC Pentium 4, 2.4GHz, 768RAM, OS MS Windows 2000. [b] Origin200, 2 × 180 MHz/1 MB cache R10000, OS IRIX 6.5 (minimization time is not included).

For the first training set we did not use the model-based approaches because these would be too time-consuming taking into account that the total number of potential transformations exceeds 18000. The accuracy of the statistical prediction is reasonable despite the evaluation set exceeding three times the size of the training set in terms of number of substrates.

For the second evaluation set (benzene derivatives), the accuracy of prediction was compared with two quantum chemical approaches. The oxenoid model provides the best accuracy of prediction. However, the model can be applied to substituted benzenes only since it uses a fixed benzene ring geometry taken from a published intermediate of a benzene molecule oxidation.[4] The statistical approach and the methoxy-radical model are almost equivalent in their predictive accuracy. In Table 3, ranking of transformations by ΔP values is compared with ranking by ΔH values. Appropriate IAP values are given in Table 4. As one can see, the majority of experimentally observed transformations have highest ranks by both the model-based and statistical approaches. The statistical predictions fail for substrates 3, 6, and 9. The oxenoid model correctly predicts all substrates except substrate 14. The methoxy-radical model fails for substrate 10. For substrates 3, 6, 14, and 15, the lowest energy also does not correspond with the experimentally observed site of hydroxylation. However, the difference between the lowest and the next lowest energy is insignificant. It is interesting that for substrates 3 and 6, the same hydroxylation sites are predicted with the highest ranks by both the methoxy-radical model and the statistical method.

For hetero- and polycyclic compounds, the statistical approach was compared with the methoxy-radical model only. As one can see from Tables 5 and 6, the methoxy-radical model fails for 11 of 17 substrates (2, 5, 6, 7, 10, 11, 12, 13, 14, 16, 17). The average accuracy of 44.9% means that the result of the prediction does not differ significantly

Table 3. Benzene Derivatives (Ranking of Potential Hydroxylation Sites According to ΔP and ΔH Values)

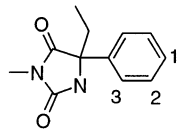
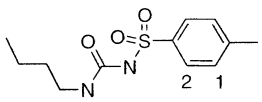
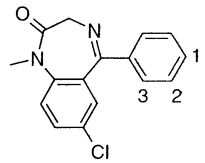
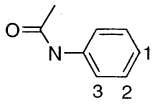
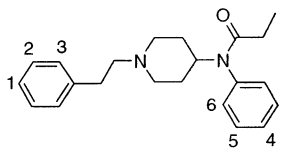
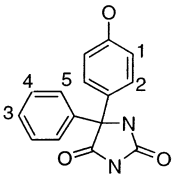
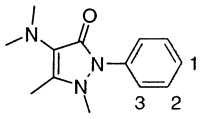
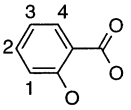
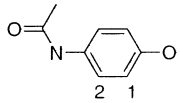
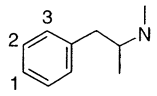
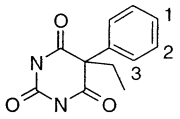
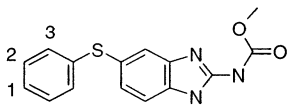
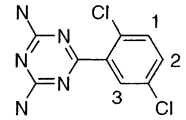
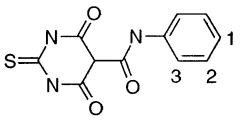
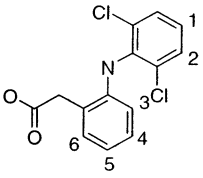
	Structure	Hydroxylation site	Experiment [a]	ΔP	Rank ΔP	ΔH^o, Kcal/M	Rank ΔH^o	ΔH^m, Kcal/M	Rank ΔH^m
1	 Mephenytoin	1	+	0.274	1	73.05	1	-27.89	1
		2	-	0.026	3	73.38	2	-26.76	2
		3	-	0.106	2	80.07	3	-26.08	3
2	 Tolbutamide	1	+	0.233	1	74.01	1	-28.83	1
		2	-	0.068	2	80.51	2	-21.92	2
3	 Diazepam	1	+	0.171	2	70.91	1	-29.19	2
		2	-	0.500	1	71.35	2	-29.38	1
		3	-	-0.488	3	76.34	3	-28.70	3
4	 Acetanilide	1	+	0.740	1	64.05	2	-33.33	2
		2	-	-0.321	3	73.76	3	-28.83	3
		3	+	0.215	2	57.20	1	-33.44	1
5	 Fentanyl	1	+	0.561	2	70.08	2	-38.94	1
		2	-	-0.102	3	71.10	3	-37.30	5
		3	-	-0.568	6	76.42	5	-37.40	4
		4	+	0.618	1	68.33	1	-38.57	3
		5	-	-0.349	5	78.28	6	-31.94	6
		6	-	-0.324	4	73.98	4	-38.76	2
6	 5-(p-Hydroxyphenyl)-5-phenylhydantoin	1	+	0.216	3	68.24	1	-28.94	2
		2	-	-0.535	5	78.19	5	-23.72	5
		3	-	0.639	2	70.84	2	-28.10	3
		4	-	0.720	1	71.13	3	-28.96	1
		5	-	-0.462	4	78.09	4	-26.33	4
7	 Aminophenazone	1	+	0.425	1	67.45	1	-31.67	1
		2	-	-0.683	2	68.87	2	-27.42	3
		3	-	-0.854	3	71.92	3	-28.71	2
8	 Salicylic acid	1	+	0.355	2	58.96	1	-29.87	1
		2	+	0.022	3	88.07	3	-25.45	3
		3	+	0.617	1	66.22	2	-28.94	2
		4	-	-0.635	4	93.99	4	-24.87	4
9	 Paracetamol	1	+	-0.331	2	62.98	1	-33.17	1
		2	-	0.528	1	68.09	2	-29.70	2

Table 3 (Continued)

	Structure	Hydroxylation site	Experiment [a]	ΔP	Rank ΔP	ΔH^o, Kcal/M	Rank ΔH^o	ΔH^m, Kcal/M	Rank ΔH^m
10		1	+	0.734	1	67.95	1	-27.68	2
		2	-	0.048	2	70.36	2	-29.54	1
		3	-	-0.315	3	70.96	3	-27.53	3
	1-Phenyl-2-methylaminopropane								
11		1	+	0.683	1	73.17	2	-28.05	1
		2	+	-0.218	2	72.79	1	-27.33	2
		3	-	-0.515	3	74.75	3	-25.84	3
	Phenobarbital								
12		1	+	-0.061	1	68.35	1	-38.39	1
		2	-	-0.423	2	69.41	2	-37.06	2
		3	-	-0.631	3	71.40	3	-35.61	3
	Fenbendazole								
13		1	+	0.855	2	83.39	1	-29.49	1
		2	+	0.920	1	84.10	2	-27.44	2
		3	-	0.422	3	93.33	3	-26.95	3
	Irsogladine								
14		1	+	0.847	1	72.59	2	-34.01	2
		2	-	-0.246	3	70.14	1	-33.00	3
		3	-	0.567	2	73.91	3	-34.30	1
	Merbarone								
15		1	+	0.858	1	67.05	1	-30.09	2
		2	+	0.128	6	71.35	5	-28.71	4
		3	-	0.278	4	68.40	3	-30.24	1
		4	+	0.363	3	69.99	4	-30.09	2
		5	+	0.815	2	68.33	2	-30.03	3
		6	-	0.173	5	74.03	6	-27.46	5
	Diclofenac								

[a] + the transformation is reported in the Metabolite database. − the transformation is not reported in the Metabolite database. [o] Oxenoid model. [m] Methoxy radical model.

Table 4. Accuracy of Prediction (IAP) Estimated for Substrates Given in Table 3 by the Statistical Approach (s), Oxenoid Model (o) and Methoxy-Radical Model (m), Respectively

	substrate name	IAPs, %	IAPo, %	IAPm, %
1	mephenytoin	100.0	100.0	100.0
2	tolbutamide	100.0	100.0	100.0
3	diazepam	50.0	100.0	50.0
4	acetanilide	100.0	100.0	100.0
5	fentanyl	100.0	100.0	87.5
6	5-(p-hydroxyphenyl)-5-phenyl-hydantoin	50.0	100.0	75.0
7	aminophenazone	100.0	100.0	100.0
8	salicylic acid	100.0	100.0	100.0
9	paracetamol	0.0	100.0	100.0
10	1-phenyl-2-methylaminopropane	100.0	100.0	50.0
11	phenobarbital	100.0	100.0	100.0
12	fenbendazole	100.0	100.0	100.0
13	irsogladine	100.0	100.0	100.0
14	merbarone	100.0	50.0	50.0
15	diclofenac	75.0	75.0	50.0

from random selection of potential hydroxylation sites. In contrast to this model, the statistical approach provides a reasonable accuracy of prediction (83.1%). However, for substrates 12 and 16, the first rank does not correspond to an experimentally observed (i.e. reported in the Metabolite database) hydroxylation site, and for substrate 6, two experimentally observed sites are predicted with lowest ranks. The failure of the methoxy-radical model might be related to not taking into account a possible epoxidation that can precede the phenol formation stage. If an epoxidation mechanism is involved, additional calculation of the electronic effect of epoxide ring opening[9] would be necessary to predict the hydroxylation site. This calculation was not done in this study. The relative success of the statistical approach may be explained by its independence of any hypothesis about the hydroxylation mechanism and its use of the structures of the final products for training.

Table 5. Hetero- and Polycyclic Compounds (Ranking of Potential Hydroxylation Sites According to ΔP and ΔH Values)

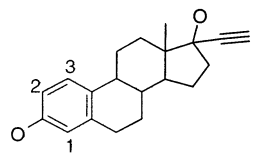
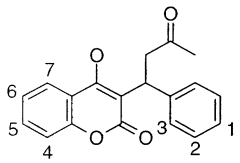
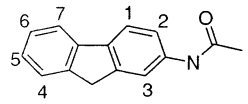
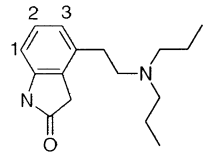
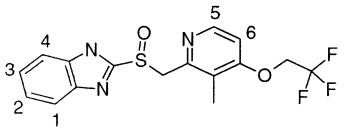
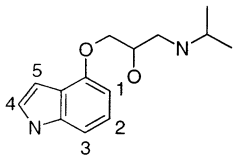
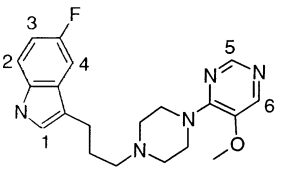
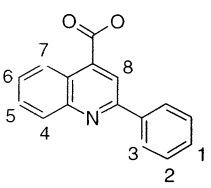
	Structure	Hydroxylation site	Experiment[a]	ΔP	Rank ΔP	ΔH, Kcal/M	Rank ΔH
1	 Ethinylestradiol	1	+	0.924	1	-29.96	2
		2	+	0.924	1	-31.45	1
		3	-	0.646	2	-29.07	3
2	 Warfarin	1	+	0.685	4	-28.70	5
		2	+	0.480	5	-27.67	6
		3	-	0.035	7	-28.70	5
		4	+	0.743	3	-31.19	2
		5	+	0.774	2	-29.39	4
		6	+	0.785	1	-30.61	3
		7	-	0.090	6	-34.75	1
3	 N-2-Fluorenylacetamide	1	-	-0.002	6	-29.60	6
		2	+	-0.056	7	-32.15	4
		3	+	0.078	5	-34.33	1
		4	+	0.179	4	-28.79	7
		5	+	0.794	1	-32.52	2
		6	-	0.464	3	-32.39	3
		7	+	0.557	2	-30.03	5
4	 Ropinirole	1	+	0.921	1	-32.49	1
		2	-	0.799	3	-28.39	3
		3	-	0.800	2	-30.71	2
5	 Lansoprazole	1	-	0.268	5	-33.85	1
		2	+	0.775	1	-28.16	5
		3	+	0.540	2	-29.73	4
		4	-	-0.160	6	-33.37	2
		5	-	0.303	4	-22.57	6
		6	-	0.417	3	-31.09	3
6	 Pindolol	1	+	0.773	1	-27.23	5
		2	+	0.643	4	-32.00	4
		3	-	0.699	3	-35.19	3
		4	-	0.736	2	-43.61	1
		5	+	0.615	5	-39.31	2
7	 BMS 181101	1	-	-0.325	5	-42.49	1
		2	-	0.276	3	-33.28	2
		3	+	0.534	1	-30.50	4
		4	-	-0.598	6	-31.73	3
		5	-	0.312	2	-19.96	6
		6	-	-0.108	4	-21.64	5
8	 Cinchophen	1	+	0.690	2	-31.87	5
		2	-	0.598	3	-29.04	8
		3	-	-0.203	8	-30.18	7
		4	+	0.282	5	-37.49	1
		5	-	0.413	4	-31.25	6
		6	+	0.710	1	-34.97	2
		7	-	0.175	6	-33.75	3
		8	-	0.134	7	-33.18	4

Table 5 (Continued)

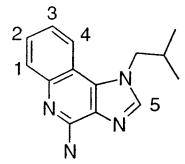
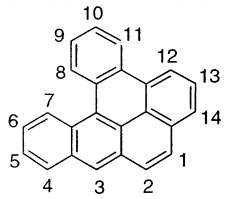
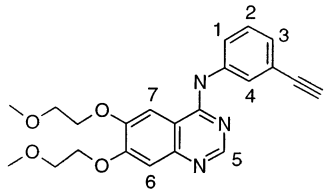
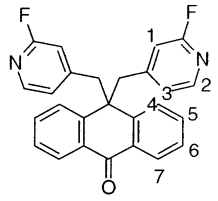
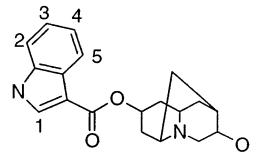
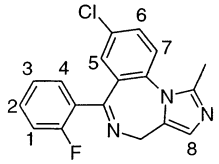
	Structure	Hydroxylation site	Experiment[a]	ΔP	Rank ΔP	ΔH, Kcal/M	Rank ΔH
9	 Imiquimod	1	+	0.442	2	-34.27	1
		2	+	0.428	3	-30.28	5
		3	+	0.671	1	-33.47	3
		4	-	0.286	4	-33.82	2
		5	-	0.035	5	-32.11	4
10	 Dibenzo(a,l)pyrene	1	-	0.320	13	-41.65	10
		2	-	0.359	12	-44.41	7
		3	-	0.782	2	-52.51	1
		4	-	0.722	3	-44.99	4
		5	-	0.635	7	-44.45	6
		6	-	0.605	9	-46.27	2
		7	-	0.674	5	-36.95	11
		8	-	0.481	10	-30.67	13
		9	-	0.658	6	-44.45	6
		10	-	0.622	8	-45.84	3
		11	-	0.460	11	-44.85	5
		12	-	0.708	4	-36.04	12
		13	-	0.196	14	-41.97	9
		14	+	0.818	1	-43.90	8
11	 CP 358774	1	-	0.309	3	-35.67	2
		2	+	-0.681	7	-31.83	5
		3	+	0.460	2	-32.58	4
		4	-	-0.331	5	-34.11	3
		5	+	0.532	1	-20.51	6
		6	-	-0.358	6	-35.67	2
		7	-	0.220	4	-38.03	1
12	 DMP 543	1	-	0.480	6	-32.86	1
		2	+	0.641	3	-22.45	7
		3	-	0.702	2	-31.76	2
		4	-	0.490	5	-24.77	6
		5	-	0.616	4	-28.30	4
		6	-	0.719	1	-29.47	3
		7	-	0.367	7	-26.02	5
13	 Hydrodolasetron	1	-	0.267	5	-49.52	1
		2	+	0.879	4	-43.50	2
		3	+	0.907	2	-40.66	5
		4	+	0.916	1	-41.75	3
		5	-	0.883	3	-40.85	4
14	 Midazolam	1	-	0.109	4	-31.77	2
		2	+	0.629	1	-27.93	6
		3	-	0.209	3	-31.66	3
		4	-	0.083	5	-26.91	7
		5	-	-0.468	8	-26.59	8
		6	-	-0.076	6	-28.37	5
		7	-	-0.077	7	-30.75	4
		8	-	0.258	2	-33.44	1

Table 5 (Continued)

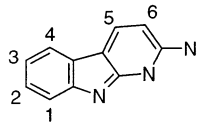
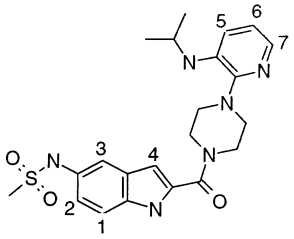
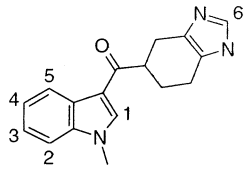
	Structure	Hydroxylation site	Experiment[a]	ΔP	Rank ΔP	ΔH, Kcal/M	Rank ΔH
15	 2-Amino-alpha-carboline	1	-	0.431	4	-37.79	2
		2	-	0.551	3	-33.01	5
		3	+	0.681	2	-35.57	3
		4	-	0.331	5	-33.29	4
		5	-	0.161	6	-31.59	6
		6	+	0.792	1	-38.77	1
16	 Delavirdine	1	-	0.558	6	-36.67	3
		2	-	0.608	4	-35.02	6
		3	+	0.664	3	-38.90	2
		4	-	0.749	2	-46.78	1
		5	-	0.194	7	-36.47	4
		6	-	0.939	1	-36.19	5
		7	+	0.582	5	-31.37	7
17	 Ramosetron	1	-	0.364	4	-35.07	2
		2	+	0.310	5	-18.94	6
		3	+	0.684	3	-29.35	4
		4	+	0.771	1	-30.09	3
		5	-	0.131	6	-28.22	5
		6	-	0.749	2	-35.45	1

Table 6. Accuracy of Prediction (IAP) Estimated for Substrates Given in Table 5 by the Statistical Approach (s) and Methoxy-Radical-Model (m), Respectively

	substrate name	IAPs, %	IAPm, %
1	ethinylestradiol	100.0	100.0
2	warfarin	100.0	35.0
3	N-2-fluorenylacetamide	60.0	60.0
4	ropinirole	100.0	100.0
5	lansoprazole	100.0	25.0
6	pindolol	33.3	16.7
7	BMS 181101	100.0	40.0
8	cinchophen	86.7	86.7
9	imiquimod	100.0	50.0
10	dibenzo(a,l)pyrene	100.0	38.5
11	CP 358774	66.7	0.0
12	DMP 543	66.7	0.0
13	hydrodolasetron	83.3	33.3
14	midazolam	100.0	28.6
15	2-amino-alpha-carboline	100.0	87.5
16	delavirdine	50.0	40.0
17	ramosetron	66.7	22.2

In the above-mentioned evaluation sets, we did not take into account isoforms of P450 involved in any transformation. To prepare the fourth evaluation set, we selected only substrates of CYP2D6. It is known from literature that the metabolic site of CYP2D6 substrates depends not only on electronic factors but also on specific orientation of a substrate within the enzyme molecule.[10−12] In light of this knowledge, we initially assumed that the prediction based on the training set that does not contain any enzymatic information should fail in the case of CYP2D6 substrates. However, our results contradicted this assumption. The accuracy of 89.6% achieved by the statistical approach is comparable or even higher than for other evaluation sets. It shows that the site of aromatic hydroxylation by CYP2D6 can be predicted with the statistical approach without the enzyme model. The methoxy-radical model yields 70% accuracy, which is also higher than the result obtained by the same model for hetero- and polycyclic compounds. As one can see in Table 7, for substrate 5 (carvedilol) the methoxy-radical model better predicts hydroxylation sites reported for different isoforms of P450, while the statistical prediction places the sites reported for CYP2D6 at the top of the list. The same holds for substrate 14 (tamoxifen). However, for the majority of molecules (see Tables 7 and 8) highest ranks obtained by both the quantum chemical and statistical calculations coincide with the experimentally observed sites of CYP2D6 hydroxylation rather than with sites reported for other enzymes. Therefore, it might be assumed that aromatic hydroxylation by CYP2D6 occurs at the most energetically favorable position of the aromatic ring and does not depend on the orientation of the substrate within the enzyme active site. We emphasize that this assumption concerns only sites of aromatic hydroxylation. The influence of orientation on other potential reactions of CYP2D6, e.g. N-demethylation, was not investigated in this study.

CONCLUSIONS

We have shown that formal statistical analysis of reaction schemes of known transformations can be used for prediction of possible aromatic hydroxylation sites for new substrates. The accuracy of prediction of experimentally observed hydroxylation sites is not significantly affected by the presence of hetero- and polycyclic moieties in a substrate's structure.

Among the approaches compared in this study, the oxenoid model is the most accurate in predicting hydroxylation sites in benzene derivatives. However, it is not applicable to

Table 7. CYP2D6 Substrates (Ranking of Potential Hydroxylation Sites According to ΔP and ΔH Values)[b]

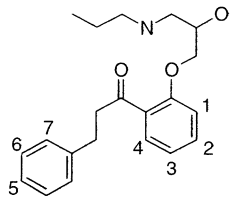
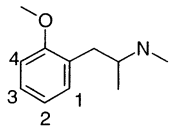
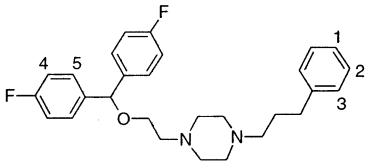
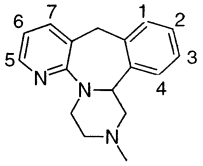
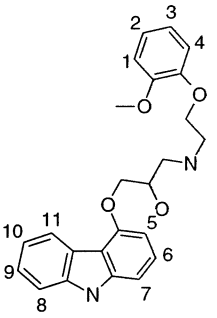
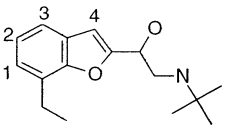
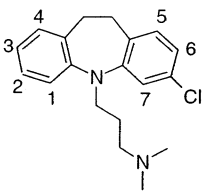
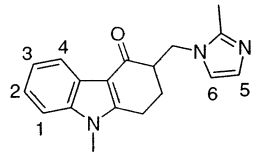
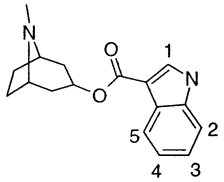
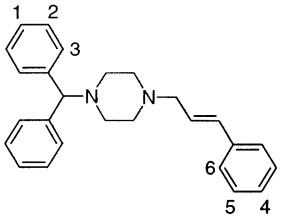
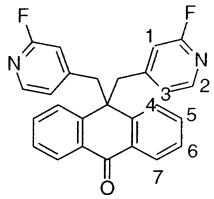
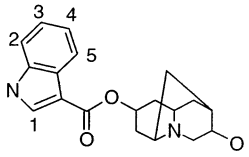
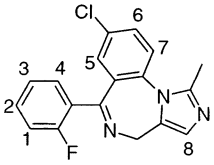
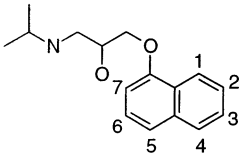
	Structure	Hydroxylation site	Experiment[a]	ΔP	Rank ΔP	ΔH, Kcal/M	Rank ΔH
1	 Propafenone	1	-	0.087	6	-26.14	7
		2	-*	0.229	4	-27.97	6
		3	+	0.552	2	-30.88	1
		4	-	0.099	5	-30.00	2
		5	-*	0.676	1	-29.06	3
		6	-	0.276	3	-28.01	5
		7	-	0.017	7	-28.65	4
2	 Methoxyphenamine	1	-	-0.192	4	-27.07	3
		2	+	0.598	1	-30.64	1
		3	+	0.330	2	-28.20	2
		4	+	0.307	3	-20.69	4
3	 GBR 12909	1	+	0.747	1	-28.74	3
		2	-	-0.168	3	-29.81	2
		3	-	-0.530	5	-27.38	4
		4	-	-0.367	4	-37.18	1
		5	-	0.210	2	-26.63	5
4	 Mirtazapine	1	-	0.493	3	-30.47	3
		2	-	0.651	2	-28.79	5
		3	-	0.329	5	-31.18	2
		4	-	0.126	7	-26.14	7
		5	-	0.448	4	-27.11	6
		6	+	0.780	1	-33.31	1
		7	-	0.147	6	-29.61	4
5	 Carvedilol	1	-	0.396	7	-32.40	4
		2	+	0.633	2	-28.60	11
		3	+	0.652	1	-30.87	8
		4	-	0.375	8	-31.59	7
		5	-*	0.436	6	-31.79	6
		6	-	0.267	11	-30.84	9
		7	-x	0.270	10	-35.52	1
		8	-x	0.356	9	-33.13	2
		9	-	0.566	4	-28.87	10
		10	-*	0.617	3	-33.04	3
		11	-*	0.482	5	-32.02	5
6	 Bufuralol	1	-*	0.320	4	-32.03	3
		2	-	0.822	1	-29.05	4
		3	+	0.343	3	-33.94	2
		4	-*	0.784	2	-37.55	1
7	 Clomipramine	1	-	-0.110	6	-27.97	5
		2	-	0.095	3	-28.12	4
		3	+	0.827	1	-31.58	1
		4	-	-0.284	7	-29.66	2
		5	-	0.043	4	-27.14	6
		6	+	0.451	2	-29.02	3
		7	-	-0.033	5	-24.07	7

Table 7 (Continued)

	Structure	Hydroxylation site	Experiment[a]	ΔP	Rank ΔP	ΔH, Kcal/M	Rank ΔH
8		1	+	0.222	5	-32.60	3
		2	+	0.642	2	-29.34	6
		3	+	0.715	1	-34.02	2
		4	-	0.084	6	-29.54	5
		5	-	0.625	3	-31.49	4
		6	-	0.337	4	-38.95	1
	Ondansetron						
9		1	-	0.765	1	-35.16	1
		2	+	0.526	4	-34.01	2
		3	+	0.718	3	-31.18	3
		4	+	0.747	2	-30.47	4
		5	-	0.500	5	-29.63	5
	Tropisetron						
10		1	+	0.140	2	-28.71	4
		2	-	-0.243	3	-27.87	6
		3	-	-0.542	5	-27.99	5
		4	+	0.561	1	-31.67	1
		5	-	-0.247	4	-30.44	3
		6	-	-0.817	6	-31.38	2
	Cinnarizine						
11		1	-	-0.094	3	-27.68	3
		2	+	0.149	2	-28.07	2
		3	+	0.714	1	-31.61	1
		7	-	0.367	7	-26.02	5
	DMP 543						
13		1	-	0.267	5	-49.52	1
		2	+	0.879	4	-43.50	2
		3	+	0.907	2	-40.66	5
		4	+	0.916	1	-41.75	3
		5	-	0.883	3	-40.85	4
	Hydrodolasetron						
14		1	-	0.109	4	-31.77	2
		2	+	0.629	1	-27.93	6
		3	-	0.209	3	-31.66	3
		4	-	0.083	5	-26.91	7
		5	-	-0.468	8	-26.59	8
		6	-	-0.076	6	-28.37	5
	Midazolam	7	-	-0.077	7	-30.75	4
		8	-	0.258	2	-33.44	1
15		1	-*	0.708	3	-36.26	2
		2	-[x]	0.625	5	-34.23	3
		3	-*	0.575	6	-32.07	7
		4	+	0.851	2	-33.46	6
		5	+	0.873	1	-37.06	1
		6	-*	0.647	4	-33.92	4
	Propranolol	7	-*	0.464	7	-33.55	5

[a] + a transformation is reported in the Metabolite database for CYP2D6. − a transformation is not reported in the Metabolite database. [b] -[x] a transformation is reported for different isoforms of P450. -* a transformation is reported without indication of a P450 isoform.

Table 8. Accuracy of Prediction (IAP) Estimated for Substrates Given in Table 7 by the Statistical Approach (s) and Methoxy-Radical-Model (m), Respectively

	substrate name	IAPs, %	IAPm, %
1	propafenone	83.3	100.0
2	methoxyphenamine	100.0	66.7
3	GBR 12909	100.0	50.0
4	mirtazapine	100.0	100.0
5	carvedilol	100.0	11.1
6	bufuralol	33.3	66.7
7	clomipramine	100.0	90.0
8	ondansetron	77.8	44.4
9	tropisetron	50.0	50.0
10	cinnarizine	100.0	75.0
11	desipramine	100.0	100.0
12	fluperlapine	100.0	100.0
13	imipramine	100.0	66.7
14	tamoxifen	100.0	71.4
15	propranolol	100.0	60.0

hetero- or polycyclic compounds. The methoxy-radical model and statistical approach provide approximately the same accuracy for benzene derivatives. For more complex compounds, the methoxy-radical model fails, while the statistical approach demonstrates relatively good prediction.

The statistical approach can be used for prediction of aromatic hydroxylation sites for CYP2D6 substrates without modeling of the enzyme structure. Both quantum chemical and statistical calculations used in this study show no specific influence of CYP2D6 on the site of aromatic hydroxylation.

The proposed statistical approach can be used for high-throughput metabolism prediction due to its broad applicability and high speed of calculation. The applicability of the approach to other biotransformation classes and the influence of other P450′ isoforms on the prediction accuracy will be evaluated in the future.

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. through providing the Institute of Biomedical Chemistry of the Russian Academy of Medical Sciences with a license for ISIS and the Metabolite database used in this study.

REFERENCES AND NOTES

(1) Wilkinson, G. R. Pharmacokinetics: The Dynamics of Drug Absorption, Distribution, and Elimination. In *Goodman & Gilman's The Pharmacological Basis of Therapeutics*,10th ed.; Hardman, J. G., Limbird L. E., Gilman, A. G., Eds.; McGraw-Hill: 2001; pp 3−29.
(2) Rendic, S. Summary of information on human CYP enzymes: human P450 metabolism data. *Drug Met. Rev.* **2002**, *34*, 83−448.
(3) Higgins, L.; Korzekwa, K. R.; Rao, S.; Shou, M.; Jones, J. P. An Assessment of the Reaction Energetics for Cytochrome P450-Mediated Reactions. *Arch. Biochem. Biophys.* **2001**, *385*, 220−230.
(4) Korzekwa, K.; Trager, W.; Gouterman, M.; Spangler, D.; Loew, G. H. Cytochrome P450 Mediated Aromatic Oxidation: A Theoretical Study. *J. Am. Chem. Soc.* **1985**, *107*, 4273−4279.
(5) Kuznetsov A. V. To the Oxenoid Model of Cytochrome P450 Mediated Activation of Molecular Oxygen: the Role of Substrate Structure (rus). *Molecularnaya Biologya* **1990**, *24*, 1373−1380.
(6) Dyachkov P. N. Quantum-chemical calculations in the study of action mechanism and toxicity of xenobiotics (rus). Itogi nauki I techniki. VINITI. *Ser. Toksikologya* **1990**, *16*, 1−280.
(7) Jones, J. P.; Mysinger, M.; Korzekwa, K. R. Computational Models for Cytochrome P450: A Predictive Electronic Model for Aromatic Oxidation and Hydrogen Atom Abstraction. *Drug Metab. Dispos.* **2002**, *30*, 7−12.
(8) Singh, S. B.; Shen, L. Q.; Walker, M. J.; Sheridan, R. P. A Model for Predicting Likely Sites of CYP3A4-mediated Metabolism on Drug-like Molecules. *J. Med. Chem.* **2003**, *46*, 1330−1336.
(9) Dowers, T. S.; Rock D. A.; Rock D. A.; Oerkins B. N. S.; Jones J. P. An analysis of the regioselectivity of aromatic hydroxylation and N-oxygenation by cytochrome P450 enzymes. *Drug Metab. Dispos.* **2004**, *32*, 328−332.
(10) Ekins, S.; de Groot, M. J.; Jones, J. P. Pharmacophore and three-dimensional quantitative structure activity relationship methods for modeling cytochrome P450 active sites. *Drug Metab. Dispos.* **2001**, *29*, 936−944.
(11) De Groot, M. J.; Ackland, M. J.; Horne, V. A.; Alex, A. A.; Jones, B. C. Novel Approach to Predicting P450-Mediated Drug Metabolism: Development of a Combined Protein and Pharmacophore Model for CYP2D6. *J. Med. Chem.* **1999**, *42*, 1515−1524.
(12) De Groot, M. J.; Ackland, M. J.; Horne, V. A.; Alex, A. A.; Jones, B. C. A Novel Approach to Predicting P450-Mediated Drug Metabolism. CYP2D6 Catalyzed N-Dealkylation Reactions and Qualitative Metabolite Predictions Using a Combined Protein and Pharmacophore Model for CYP2D6. *J. Med. Chem.* **1999**, *42*, 4062−4070.
(13) De Groot, M. J.; Alex, A. A.; Jones, B. C. Development of a Combined Protein and Pharmacophore Model for Cytochrome P450 2C9. *J. Med. Chem.* **2002**, *45*, 1983−1993.
(14) Park, J.-Y.; Harris, D. Construction and Assessment of Models of CYP2E1: Predictions of Metabolism from Docking, Molecular Dynamics, and Density Functional Theoretical Calculations. *J. Med. Chem.* **2003**, *46*, 1645−1660.
(15) Zamora, I.; Afzelius, L.; Cruciani, G. Predicting Drug Metabolism: A Site of Metabolism Prediction Tool Applied to the Cytochrome P450 2C9. *J. Med. Chem.* **2003**, *46*, 2313−2324.
(16) Accelrys Inc., 9685 Scranton Road, San Diego, CA 92121-3752, U.S.A. (http://www.accelrys.com).
(17) MDL Information Systems, Inc., 14600 Catalina Street, San Leandro, CA, U.S.A. (http://www.mdli.com).
(18) Borodina, Yu.; Sadym, A.; Filimonov, D.; Blinova, V.; Dmitriev, A.; Poroikov, V. Predicting Biotransformation Potential from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1636−1646.
(19) 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.
(20) 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.
(21) 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.
(22) Poroikov, V. V.; Filimonov, D. A. How to acquire new biological activities in old compounds by computer prediction. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 819−824.
(23) 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.
(24) 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.
(25) Lagunin, A.; Gomazkov, O.; Filimonov, D.; Gureeva, T.; Dilakyan, E.; Kugaevskaya, E.; Elisseeva, Yu.; Solovyeva, N.; Poroikov, V. Computer-Aided Selection of Potential Antihypertensive Compounds with Dual Mechanism of Action. *J. Med. Chem.* **2003**, *46*, 3326−3332.
(26) If ranks of real and unreal transformations coincide, the number in the numerator increases by 0.5.

CI049834H