Modeling of Human Cytochrome P450-Mediated Drug Metabolism Using Unsupervised Machine Learning Approach

Dmitry Korolev,[†,#] Konstantin V. Balakin,[*,‡,#] Yuri Nikolsky,[‡] Eugene Kirillov,[†] Yan A. Ivanenkov,[‡] Nikolay P. Savchuk,[‡] Andrey A. Ivashchenko,[‡] and Tatiana Nikolskaya[†]

GeneGo, Inc., 227 South Berrien Street, New Buffalo, Michigan 49117, and Chemical Diversity Labs, Inc., 11558 Sorrento Valley Road, San Diego, California 92121

Received March 5, 2003

We developed a computational algorithm for evaluating the possibility of cytochrome P450-mediated metabolic transformations that xenobiotics molecules undergo in the human body. First, we compiled a database of known human cytochrome P-450 substrates, products, and nonsubstrates for 38 enzyme-specific groups (total of 2200 compounds). Second, we determined the cytochrome-mediated metabolic reactions most typical for each group and examined the substrates and products of these reactions. To assess the probability of P450 transformations of novel compounds, we built a nonlinear quantitative structure−metabolism relationships (QSMR) model based on Kohonen self-organizing maps (SOM). This neural network QSMR model incorporated a predefined set of physicochemical descriptors encoding the key molecular properties that define the metabolic fate of individual molecules. Isozyme-specific groups of substrate molecules were visualized, thus facilitating prediction of tissue-specific metabolism. The developed algorithm can be used in early stages of drug discovery as an efficient tool for the assessment of human metabolism and toxicity of novel compounds in designing discovery libraries and in lead optimization.

Introduction

With the growing use of chemicals as therapeutic agents, food additives, cosmetics, agricultural fertilizers, and pest management agents, people are increasingly exposed to exogenous compounds (xenobiotics). Both the parent agent and the products of its metabolism in the liver and other organs may contribute to the composite toxic effect of an agent on the human organism. The metabolic transformations may profoundly affect the initial bioavailability, the desired activity, the tissue distribution, the toxic action, and the eventual elimination of a compound. The understanding of the possible toxicity and the metabolic fate of xenobiotics in the human body is particularly important in drug discovery, where such early assessment may eliminate the potentially toxic candidates from further development prior to expensive clinical trials. With the obvious importance of the early assessment problem, there is a need for in silico methodologies for uncovering the relations between the structure and the biological activity of novel molecules. For in silico models to be relevant, it is vital to understand the pathways of xenobiotics biotransformation in the body which define their activity, toxicity, and interaction with normal body metabolism.

In an attempt to systematize the overall complexity of the metabolism of xenobitics molecules, several groups tried to identify the general rules of metabolic biotransformations. The first general rules on the metabolic behavior of esters, O- and N-alkyl derivatives, and aromatic fragments were established almost 30 years ago.[1] Some of these rules were extensively used in rational drug development.[2] Later, with the steady accumulation of drug metabolism data, the data storage and management systems, such as MetabolExpert[3] and META[4], were built. By using computer-driven queries across these databases, one can identify the sites on a query molecule most prone to metabolic transformations. However, these databases contained indiscriminate data from studies on a variety of mammalian species. Consequently, the programs using these data sets tend to predict all the metabolic possibilities for an exogenous molecule, as it was placed in a theoretical "average" mammal.[5] Metabolite[6] is a newer database that represents a broad collection of metabolic data with its main strength in size (more than 25 000 compounds, 85 000 metabolic reactions), rather than in quality of data entries.

There are several concerns to be mentioned regarding computerized prediction of the metabolic fate of novel compounds. First, as mentioned above, the indiscriminate pooling of metabolic data from different species distorts substantially any attempt at generalization.[5] The metabolic pathways and corresponding networks can be very different even in close mammalian species, so any use of "pooled" data seems to be problematic.[7] Second, in vitro versus in vivo data may vary substantially even for the same species. The metabolic fate of a drug delivered to a human liver after intravenous administration is often quite different than in in vitro experiments using liver microsomal fractions. Third, in the last several years, the individual enzymatic profile (pharmacogenomics profile) became an important concern in drug toxicity and metabolism. The metabolism of the same drug may vary substantially between individuals depending on the expression level of particular enzymes, polymorphisms in enzyme-encoding and regulatory genes and the presence of particular

* To whom correspondence should be addressed. Tel.: (858) 794-4860. Fax: (858) 794-4931. E-mail: kvb@chemdiv.com.
† GeneGo, Inc.
‡ Chemical Diversity Labs, Inc.
# Equal contribution authors.

10.1021/jm030102a CCC: $25.00
Published on Web 07/23/2003

isoenzymes in normal[8] and disease states.[9] Fourth, in determining structure−metabolism relationships, one important question is whether it is the complete molecular structure or its structural components that actually undergo metabolism. The answer is vital for choosing the descriptors for a robust QSMR model capable of predicting the metabolic fate of novel compounds. Finally, the QSMR algorithms should be highly effective for real-time handling of the very large virtual and real discovery compound databases built during the last years.

These concerns make the accurate prediction of the metabolic outcome for a new compound a very complicated task, even for an individual structural query in one species. To facilitate the solution, we offer a multistep approach. In the first step, one should build and analyze a broad database of species-specific metabolic transformations of all possible metabolic functional groups. Using this database, the specific rules should be established to characterize the probability of transformations of the key "metabophore" elements. To further reduce the complexity of the problem, one should consider the metabolic pathways and subsystems responsible for xenobiotics catabolism separately. For these pathways and enzymatic families, we need to define the enzyme-specific substrate groups for which an effective QSMR model can be designed. Such a multistep approach, with a comprehensive initial database and working algorithms applied by experts, would be an extremely useful tool for early assessment of human metabolic transformations of lead compounds and drug candidates at the preclinical stage of drug discovery. The automated version of the program can be also applied for "filtering" the large virtual compound collections designed for initial bioscreening programs.

In this work, we assembled a comprehensive set of over 2200 substrate−product reactions for 38 human cytochromes. On the basis of this database, we developed a neural network computational algorithm for in silico assessment of the probability of cytochrome P450-mediated transformation for any novel drug-like compound.

Methods

Databases. A database of about 2200 compounds representing substrates, nonsubstrates, and products of human cytochrome P450-mediated metabolic reactions was compiled from the experimental literature.[11] The main descriptive statistics for this database is shown in Table 1. All the compounds were assigned to at least one enzyme-specific group. Prior to the neural network experiments, the molecules were filtered based on molecular weight (range 200−700) and atom type content (only C, N, O, H, S, P, F, Cl, Br, and I were permitted). Some specific chemical classes of compounds that typically are not related to drug-like agents, such as polyaromatic compounds, long-chain linear molecules (e.g., leukotrienes, fatty acids) were excluded.

Descriptors. Sixty molecular descriptors describing the important molecular properties, such as lipophilicity, charge distribution, topological features, steric and surface parameters were explored. These descriptors were calculated for the entire dataset using Cerius[2,12] and ChemoSoft [13] software tools. The SLIPPER program[14] (included in ChemoSoft) was used for $\log D_{7.4}$ and $\log S_w$ calculations. The number of descriptors was reduced to 26 by the omission of the low-variable and highly correlated ($R > 0.9$) descriptors. To further reduce the descriptor space, a principal component analysis was performed using ChemoSoft. Eventually, seven descriptors were selected as the most relevant and further used as input parameters in all neural network experiments. For each compound, the neural network scores for the back-propagated neural networks or the map coordinates for the Kohonen networks were calculated as the outcome descriptors.

Table 1. Descriptive Statistics for the Initial Database

CYP enzyme	reactions	substrates	products
CYP19	19	15	20
CYP11A1	3	3	4
CYP11B1	2	2	2
CYP11B2	6	5	6
CYP17	7	6	7
CYP1A	4	3	4
CYP1A1	205	135	211
CYP1A2	350	244	346
CYP1B1	76	53	75
CYP21B	2	2	2
CYP24	5	3	4
CYP26A1	3	2	3
CYP27A1	17	11	17
CYP2A13	10	8	12
CYP2A6	122	98	130
CYP2B6	170	139	178
CYP2C	2	2	3
CYP2C18	42	38	43
CYP2C19	229	162	239
CYP2C8	137	102	150
CYP2C9	283	207	293
CYP2D6	277	217	289
CYP2E1	220	174	218
CYP2F1	5	5	5
CYP2J2	8	3	8
CYP3A4	589	423	624
CYP3A5	78	61	87
CYP3A7	23	20	25
CYP4A11	17	16	18
CYP4B1	17	13	17
CYP4F12	5	5	5
CYP4F2	9	9	10
CYP4F3	14	14	14
CYP4F8	9	6	9
CYP7A1	8	8	8
CYP7B1	4	4	4
CYP8A1	4	3	5
CYP3A43	4	3	5

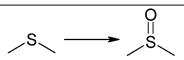
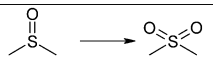
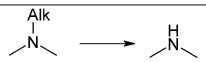
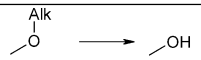
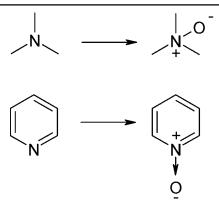
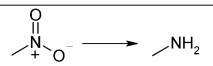
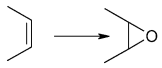
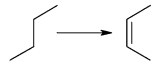
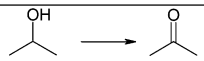
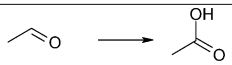
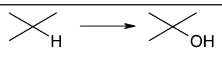
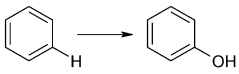
Neural Network (NN) Modeling. For the unsupervised learning procedure and generation of the Kohonen map, we used an internally developed program, part of ChemoSoft software suite.[13] The training parameters were as follows: the number of interactions for the training runs 2000; the starting adjustment radius for the training runs 0.1; the decay factor 0.001.

The commercially available NeuroSolution 4.0 program[15] was used for the supervised learning. Feed-forward nets consisting of input neurons, one hidden layer, and two output neurons were constructed. The final score was calculated by subtracting the "substrates" score from the "products" score. The back-propagated nets were trained by the momentum learning rule as implemented in NeuroSolution. The training was performed with over 1000 iterations.

Results

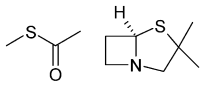
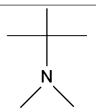
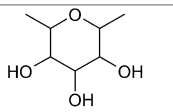
Cytochrome P450-Mediated Metabolism. Several groups of metabolizing enzymes are involved in Phase I processing of xenobiotics, including cytochrome P450 isozymes, hydrolases (esterases, amidases, epoxide hydrolases, glycosidases, glucuronidases), specific carboxylases, reductases (such as alcohol dehydrogenases and aldo-keto reductases), and some non-CYP450-related oxidases. Cytochrome P450 (CYP) enzyme superfamily plays a central role in Phase I metabolism of xenobiotics.[10] CYP enzymes represent mixed function monooxygenases capable of either inactivating or activating xeno- and endobiotics molecules for further processing by Phase II bioconjugation enzymes. The major compo-

Table 2. General Rules for Cytochrome P450-Mediated Metabolic Transformations

Reaction	Structural Representation	Number of enzymatic transformations of this type in the database
Sulfur(II) oxidation		35
Sulfur(IV) oxidation		19
N-dealkylation		226
O-dealkylation		134
N-oxide formation		60
Nitro-group reduction		9
Double bond epoxidation		75
Double bond formation (desaturation)		41
Alcohol oxidation		49
Aldehyde oxidation		9
Aliphatic hydroxylation		342
Aromatic hydroxylation		222

nents of Phase I enzymatic complex are a phospholipid, a flavoprotein, a NADPH-cytochrome P450 oxidoreductase, and the hemoprotein cytochrome P450.[16] CYPs are the terminal binding proteins of monooxygenase electron transport chain, important for catalyzing the oxidation of such endobiotics as fatty acids, steroids, ketones, polycyclic aromatic hydrocarbons, nitrosamines, hydrazines, and arylamines.[16] CYPs have been characterized as the most powerful in vivo oxidizing agents.[17] The recent reviews on CYPs detail their chemistry, regulation, membrane topology, molecular biology, and provide the models for substrate binding sites.[10] Out of about 40 human CYP genes cloned and described,[18] only three CYP families, a half-dozen subfamilies, and fewer than a dozen isoenzimes have been shown to play any significant role in hepatic processing of drugs. There may be a survival benefit associated with the use of such selected number of CYP isoforms.[19] The active CYP

Table 3. Examples of Specific Rules Governing Some Metabolic Transformations

reaction	initial molecular fragment	resulting molecular fragment	exclusions
Sulfur(II) oxidation			
N-dealkylation			
O-dealkylation			
alcohol oxidation			

enzymes have broad and overlapping substrate specificity, which poses a serious challenge to prediction of therapeutic or toxic outcomes of xenobiotic metabolism.

We analyzed the cytochrome P450 substrate/product base and deduced the general rules for the CYP-mediated metabolism. Table 2 summarizes all structural fragments that undergo a metabolic transformation. We also established a set of specific rules governing the transformations of molecular fragments that do not react according to the general scheme. Examples of such exclusions are shown in Table 3.

For most of the CYP-mediated enzymatic reactions, such as aldehyde or sulfide oxidation, the recognition of the reaction site is relatively straightforward. However, in some aromatic molecules, several sites may be sensitive to CYPs as several delocalized double bonds may be oxidized. In such cases, definition of relative sensitivity of molecular sites and the scope of P450 metabolites is not trivial and the metabolites can be observed at positions with relatively low sensitivity. Sometimes, the quantum mechanics calculations (such as evaluation of highest occupied molecular orbital (HOMO), the lowest unoccupied molecular orbital (LUMO), potential of ionization, etc.) can help to solve the problem of correct choice. Our data indicate that only a thorough estimation of a large number of input parameters, including steric, electronic and quantum, as well as an application of efficient substructure similarity evaluation tools leads to the effective solution.

For early evaluation of the metabolic fate of a compound, the aromatic hydroxylation itself is usually more important than the exact positioning of the hydroxyl group. The most typical consequence of hydroxylation is further glucuronidation of the hydroxyl group and rapid elimination of the metabolized compound from the organism. The same is true, in general, for aliphatic compounds.

To build a general model for metabolic fate of novel compounds, it is not sufficient to know the structural fragments metabolized by cytochromes and the empirical rules governing the particular metabolic reactions.

Such rules relate to the local molecular fragments, but do not take into account the properties of the whole molecule. Thus, prediction of even the most typical metabolic conversions, such as sulfur (II) oxidation or N-dealkylation, usually requires the whole-molecule approach.

Several predictive models for the metabolism of organic compounds by particular cytochromes have been reported, those based on pharmacophores, protein sequences,[20] and the assessment of reaction energies.[21] In this work, we designed a more general model for the prediction of human cytochrome P450-mediated drug metabolism. With our model, a substrate/nonsubstrate potential can be assessed for each compound based on its 2D molecular representation.

Unsupervised Kohonen Learning Approach. Neural network (NN) classification methodology has been used for multiple applications in rational drug design.[22] For instance, the ligands for certain protein classes, like GPCRs, were accurately differentiated based on some specific physicochemical features.[23] In most of the reported applications of NN in drug discovery, a supervised learning strategy was used. The alternative unsupervised learning method also becomes popular for comparative analysis and visualization of data sets.[24] Recently, a study on comparison of a benzodiazepine and dopamine data sets was performed with an implementation of a Kohonen network.[25] In another study, a dataset of 31 steroids binding to the corticosteroid binding globulin (CBG) receptor was modeled.[26] Kohonen self-organizing maps were used for distinguishing between drugs and nondrugs with a set of descriptors derived from semiempirical molecular orbital calculations.[27] It was emphasized that Kohonen map-based classification does not depend on the definition of a nondrug data set, and, therefore, the virtual screening of drug candidates can be conducted more objectively. This property of unsupervised Kohonen learning strategy is particularly important in cases when it is hard to define correctly the negative training set.

The choice between the supervised or the unsupervised approach depends on the problem and the available data. In both cases, the objects with known answers are needed. In supervised learning, the answers are directly used to influence the learning system; in unsupervised learning, the answers are needed to identify and label the output neurons. Whereas with supervised learning, the system adapts itself to a selected representation of classes, an unsupervised neural network method is more flexible due to its many possible outputs. Using the supervised learning, the multivariate objects should be split into three sets (the training, the control, and the test set). In unsupervised learning, the control set is not required, since the learning continues until the network stabilization. The number of available objects is critical, since the supervised learning procedure may take hundreds of thousands of epochs, and, for each object, the corrections of thousands of weights might be required even for medium-sized net. In unsupervised learning, a time for training with the same number of objects is much shorter since the single layer neural networks have much fewer epochs and weights.

In this work, we used the unsupervised learning methodology for in silico evaluation of a compound ability to be a cytochrome substrate. In this application, the available data do not describe all the possibilities, as only one distinct category of molecules can be unambiguously identified, namely, the cytochrome P450 substrates. The available data for nonsubstrates are related to specific isozymes, and a nonsubstrate for a particular isozyme is often described as a well-metabolized substrate of another isozyme. The products of individual metabolic reactions cannot be used as separate compound categories because these reactions often proceed in consecutive steps and the products of metabolic conversions can be metabolized further. This limitation excludes multilayer neural networks with a supervised learning procedure as an error back-propagation learning algorithm, and requires an unsupervised approach. Cytochrome substrates are extremely diverse structurally and, therefore, several different clusters are likely to appear. Therefore, a neural network should be capable of classifying the objects into none, one, or more classes, and not only into one out of several predefined existing classes. For classification of a large number of objects, the unsupervised strategy seems to be more efficient than the supervised one. We have chosen Kohonen neural network as the one with the most appropriate architecture and learning strategy.

Datasets. To effectively predict the metabolic processing of xenobiotics by CYP enzymes and to develop an effective QSMR model, one needs to compile the representative group of substrates of these isozymes. For the purposes of comparative analysis, it is also very desirable to consider the opposite category of compounds that manifests the properties of nonsubstrates.

Two overlapping data sets were distinguished within the database, filtered as described in Methods, and used in the neural network experiments. The first data set consisted of 485 compounds described as substrates for the cytochromes listed in Table 1. The second data set comprised 523 products of the cytochrome-mediated biotransformations for which no data on their further cytochrome-mediated metabolism were found. It was assumed that this dataset models the properties of the nonsubstrates to the whole cytochrome P450 family. This assumption probably results in a certain number of false negatives among the classified compounds. Because of incompleteness of the data, this selection cannot be considered as true negative training set with respect to the cytochrome substrates. In this work, we used it for illustrative purposes only. The complete 1008-compound database was used in further modeling experiments.

Molecular Descriptors. The principal component (PC) analysis[28] for 26 descriptors was performed for the 1008-compound dataset (Table 4). About 90% of the variance can be explained by the first 11 PCs. We conducted two tests to evaluate the significance of PCs for our set of descriptors and data, based on the plot of eigenvalues of the PCs against the number of the PC (not shown). The first one, Kaiser-Guttmann criterion,[29] establishes all PCs with eigenvalues larger than one as significant. In our experiment, the first six PCs appeared to be significant in this test. The second test, Scree test,[30] proposes that the eigenvalue plots should have a kink between the significant and the less significant PCs. Our plot has a kink between PCs six

Table 4. Descriptive Statistics for the Principal Component Analysis

p.c.	eigenvalue	variance	accum variance
1	6.4663	24.87	24.87
2	6.1199	23.54	48.41
3	3.3335	12.82	61.23
4	1.6702	6.42	67.65
5	1.3071	5.03	72.68
6	1.0821	4.16	76.84
7	0.8805	3.39	80.23
8	0.8521	3.28	83.51
9	0.7041	2.71	86.21
10	0.5758	2.21	88.43
11	0.5531	2.13	90.56
12	0.4565	1.76	92.31
13	0.4430	1.7	94.02
14	0.3874	1.49	95.51
15	0.3147	1.21	96.72
16	0.2293	0.88	97.6
17	0.1857	0.71	98.31
18	0.1644	0.63	98.94
19	0.1160	0.45	99.39
20	0.0674	0.26	99.65
21	0.0462	0.18	99.83
22	0.0281	0.11	99.94
23	0.0089	0.03	99.97
24	0.0056	0.02	99.99
25	0.0023	0.01	100
26	0.0000	0	100

and seven. As the latter agrees with the Kaiser-Guttmann criterion, we conclude that the space described by our 26 descriptors for the studied database is 6−7-dimensional.

The coefficients for these six PCs are shown in Table 5. The first PC, which explains 25% of the total variance, consists of size descriptors such as the molecular volume, partial negative and solvent accessible molecular surface area. The second PC, which accounts for about 24% of the total variance, consists mainly of the total polar surface area and the number of H-bond acceptors. The third PC, explaining about 13% of the total variance, is dominated by molecular lipophilicity parameters. PC4 (6% of the total variance) can be attributed to molecular polarity as it includes the contributions from molecular lipophilicity and the surface charge distribution parameters. PC5 (5% of the total variance) is dominated by the factors describing positive and negative surface charges. The remaining PC6, accounting for 4% of the variance, has strong contributions from the quantum descriptors that can contribute to the reactivity of the active sites of a molecule.

To reduce the calculation time and to make the parameters intuitively more understandable, it is appropriate to use particular descriptors instead of the principal components. Therefore, for all neural network experiments performed in this work, we used seven molecular descriptors italicized in Table 5. We selected the descriptors, which maximally contribute to the first six PCs, using the contribution coefficients and the descriptor loadings plot (not shown). An exception was made for the descriptor "molecular weight", included into the final descriptor set instead of V_m. The former is a classical molecular property correlating well with the molecular volume, and the contributions of these coefficients are comparable. The chosen descriptors are readily computable and, in combination, provide a reasonable basis for assessment of the cytochrome P450 substrate potential. In terms of relative importance, the physical properties of descriptors in descending order are as follows: molecular surface area and volume, H-bonding potential, surface charge properties contributing to all PCs, molecular lipophilicity and reactivity.

Kohonen Neural Networks. We used a Kohonen net with a 2D organization of the network nodes (neurons). To prevent the border effects, the neurons were organized toroidally, so that every neuron is

Table 5. Six Most Significant PCs for 26 Descriptors of the 1008-Compound Database[a,b]

descriptor	definition	PC1	PC2	PC3	PC4	PC5	PC6
logD$_{7.4}$	log of 1-octanol/water partition coeff. at pH 7.4	−0.1375	0.0155	**−0.4002**	0.2971	−0.1163	0.0363
logS$_{w\ 7.4}$	log of water solubility at pH 7.4	0.0861	0.0373	**0.3948**	−0.2077	0.0821	−0.0495
FA	fractional absorption	0.0661	0.1910	−0.2389	0.2343	0.1160	−0.1247
DipM	dipole moment	−0.0527	−0.1661	−0.0630	0.2832	0.1016	**−0.4786**
HOMO	highest occupied molecular orbital	−0.0129	−0.0049	−0.0518	−0.2076	**−0.4556**	**−0.4537**
LUMO	lowest unoccupied molecular orbital	−0.0774	0.0932	0.1271	0.2430	−0.0022	**0.4967**
Jurs-PPSA-1	partial negative surface area	**−0.3762**	−0.0529	0.0758	0.0618	−0.0254	−0.0088
Jurs-PNSA-1	partial negative surface area	0.0816	**−0.3178**	−0.2483	−0.1671	−0.0401	0.0765
Jurs-PNSA-3	atomic charge weighted negative surface area	−0.0759	**0.3738**	0.0643	0.0038	0.0485	−0.1140
Jurs-FNSA-1	partial negative surface area divided by TPSA	0.2875	−0.1860	−0.2299	−0.1087	−0.0139	0.0809
Jurs-FPSA-3	atomic charge weighted positive surface area divided by TPSA	−0.1357	−0.0509	**0.3840**	**0.3612**	−0.1331	−0.1566
Jurs-FNSA-3	atomic charge weighted negative surface area divided by TPSA	−0.2595	0.2722	0.0631	−0.0366	0.0255	−0.1197
Jurs-RPCG	relative positive charge	0.2656	−0.0177	−0.0544	0.1042	0.1913	−0.1786
Jurs-RNCG	relative negative charge	0.2453	0.1433	−0.0280	0.1395	**−0.3032**	−0.0550
Jurs-RPCS	relative positive charge surface area	0.1083	0.0837	0.0896	0.1607	**−0.5933**	0.0634
Jurs-RNCS	relative negative charge surface area	0.1838	0.0418	−0.1300	−0.0698	−0.2670	**0.3331**
Jurs-TPSA	total polar surface area	0.0399	**−0.3772**	0.0869	0.1977	−0.0248	−0.0693
Jurs-TASA	total sovent-accessible surface area	**−0.3448**	0.0854	−0.1195	−0.1807	−0.0239	0.0842
Jurs-RPSA	relative positive surface area	0.2165	−0.2854	0.0969	0.2252	−0.0060	−0.0622
Vm	molecular volume	**−0.3245**	−0.2062	−0.0549	0.0000	−0.0522	−0.0025
logP	log of 1-octanol/water partition coeff	−0.1701	−0.0196	**-0.3482**	**0.3129**	−0.0761	0.0857
Balaban	Balaban index	0.2017	0.0661	0.0486	**0.3110**	0.2841	0.0702
MW	molecular weight	−0.2519	−0.2805	−0.0994	−0.1035	−0.0885	−0.0372
B-rot	no. of rotatable bonds	−0.2363	−0.1974	0.1120	0.1609	0.0755	0.1820
HBA	no. of H-bond acceptors	0.0096	**−0.3447**	0.0532	−0.1878	0.1222	0.0379
HBD	no. of H-bond donors	0.0140	−0.1840	**0.3414**	0.0772	−0.2335	0.1447
eigenvalue		6.466	6.120	3.334	1.670	1.307	1.082
% variance explained		24.87	23.54	12.82	6.42	5.03	4.16
total % variance explained		24.87	48.41	61.23	67.65	72.68	76.84

[a] Coefficients larger than 0.30 are shown in boldface. [b] Descriptors, used in furher neural network experiments, are italicized.

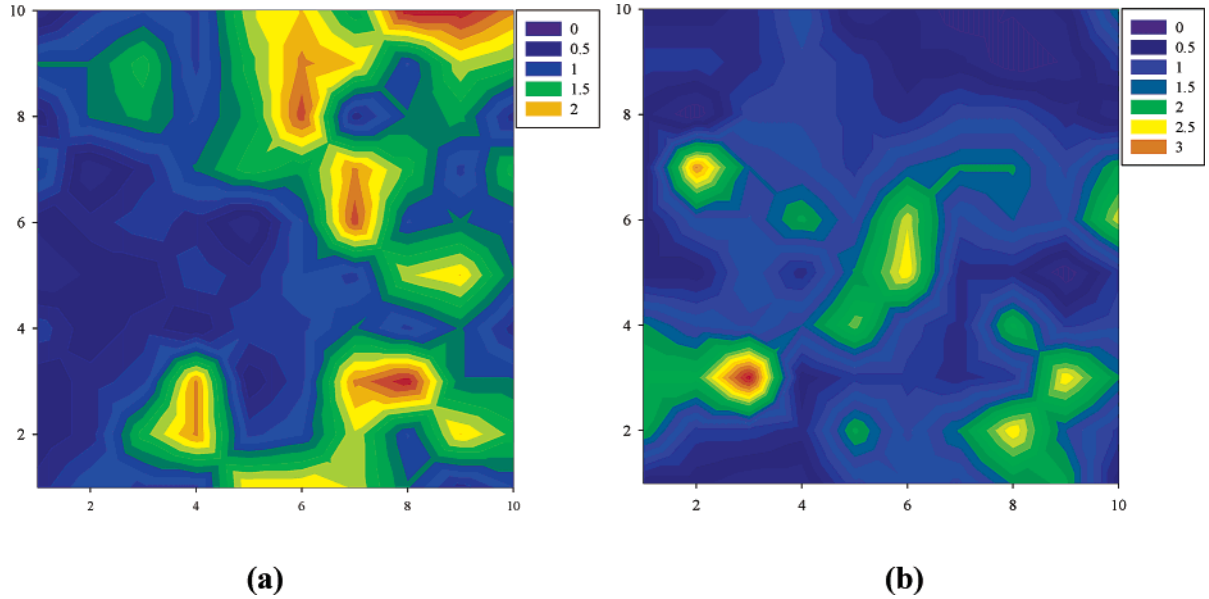


Figure 1. (a) 10 × 10 Kohonen network trained with seven selected descriptors for cytochrome substrates (485 compounds). (b) Final cytochrome reaction products (523 compounds) processed within the same map. The data have been smoothed.

equivalent to the others. For all Kohonen net calculations reported here, we used the internally developed program, part of ChemoSoft software environment.[13] A 10 × 10 node architecture was chosen to provide the studied molecules (485 CYP substrates) with sufficient distribution space. The nodes were arranged in a rectangular grid. The "bubble" function was used as the radial adjustment function.

The smoothed projection of the combined data set of cytochrome substrates onto the 10 × 10 Kohonen map was conducted using the seven descriptors selected by principal component analysis (Figure 1a). The cytochrome substrates are distributed throughout the map as the irregularly shaped islands, with a clearly defined trend toward the right side of the map. The area occupied by the cytochrome substrates is relatively large, which reflects the broad substrate specificity of the studied set of cytochromes. We suggest that the physicochemical properties of a molecule falling into the positive regions of the Kohonen map are consistent with the molecule's ability to be a cytochrome substrate.

For the comparison, we also processed the additional data set of 523 products of cytochrome-mediated biotransformations, on the same Kohonen map (Figure 1b). This data set occupies distinct areas on the map substantially different from the regions of the substrates localization. The "product" compound category is unified by a combination of physicochemical properties distinctly different from the cytochrome substrates. Therefore, the sites of "products" localization on the Kohonen map can be used for the enhancement of prediction quality.

On the basis of these distributions, we built the smoothed contour plots of the occurrences of these two compound categories within the Kohonen map (Figure 2). The area of "substrates" is marked in green, the area of "products" is in blue, and the low-populated area is in brown. The contours correspond to at least 1.5% of compounds per node, belonging to the particular category. Therefore, these areas have a higher concentration of compounds compared to random distribution. Some overlap (5% of the total surface occupied by both

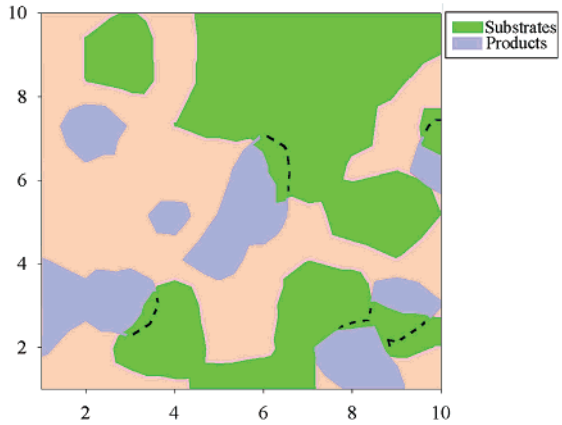


Figure 2. Smoothed contour plots of the occurrences of substrates and final products within the Kohonen map. The area of substrates is depicted in green, the area of products is in blue, and the low-populated area is in brown. The contours correspond to at least 1.5% of compounds, from a particular category, per node.

substrates and products) is observed between these two distributions, which can be explained by the incompleteness of data for the set of "products". We suggest that a fraction of compounds assigned to the category of "products" from the overlapping regions, indeed, represents the cytochrome substrates. For this reason, the overlapping areas were assigned to "substrates".

The model correctly classified 76.7% of substrates and 62.7% of products, as defined by their localization in the corresponding areas of the Kohonen map (Table 6). A certain number of compounds (12.6 and 22.4% for substrates and products, correspondingly) falls into the area for which no specific assignment could be made. Additional criteria are needed for assessing the cytochrome substrate potential for these compounds. The observed limited classification power of the model can be explained by several factors. First, the cytochrome substrates are widely diverse that leads to high variability of molecular properties and high heterogeneity of the input data. More statistical data are usually required for accurate prediction in such case. Second,

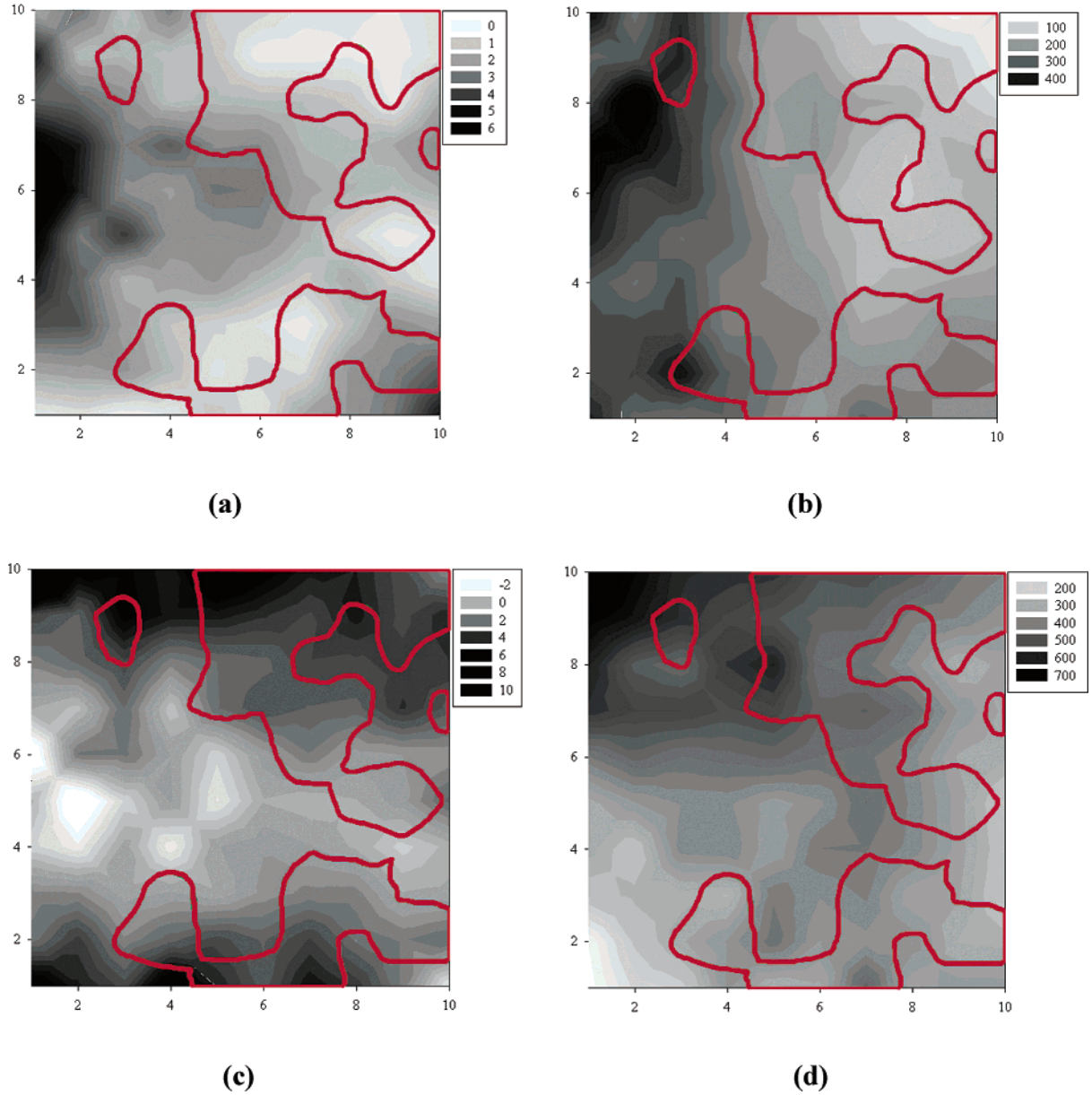


Figure 3. Gradient Kohonen maps of the values for four descriptors used for neural network generation: (a) HBD, (b) Jurs TPSA, (c) $logD_{74}$, and (d) Jurs PPSA-1. The red contour corresponds to the contour plot of cytochrome substrates given in Figure 2.

Table 6. Classification Quality for the Kohonen Neural Network Algorithm

	predicted substrates	predicted products	unassigned area	total map
substrates	363 (76.7%)	52 (10.7%)	70 (12.6%)	485 (100%)
products	78 (14.9%)	328 (62.7%)	117 (22.4%)	523 (100%)

the substrates and products for all the studied CYP enzymes were considered as a group, without taking into account the isozyme-specific differences in substrate's properties. Separation of substrates/nonsubstrates into the groups according to particular CYP isozymes should enhance the classification power. Third, some compounds may be misclassified because of the incompleteness of the reliable data for the "products" and lack of comparative quantitative data on CYP-mediated metabolic reactions.

Although the general classification power of our model is moderate, it reasonably descriminates between CYP450 substrates and nonsubstrates when the studied compounds fall into meaningful regions of the map. The enhancement factors for the areas of substrates and nonsubstrates are equal to 7.17 and 4.21, correspondingly. The enhancement factor is a ratio between the fractions of correctly and incorrectly classified compounds within the corresponding areas on the map. Effectively, it shows how many folds the number of CYP450 substrates or nonsubstrates found in the corresponding areas on the map exceeds the random distribution expectation. We believe that the developed algorithm is useful in assessment of compound's ability to be a cytochrome P450 substrate.

Distribution of the Descriptor Values. On Figure 3, we showed the contour plot of the occurrence of different CYP substrates and the distribution of four selected descriptor values within the Kohonen map. The distributions of the number of H-bond donors and the total polar surface area values (Figure 3a,b) are similar

Table 7. Type and Number of Metabolic Reactions Catalyzed by Seven Selected Cytochromes

CYP enzyme	N-dealkylation	O-dealkylation	sulfide oxidation	sulfoxide oxidation	aromatic hydroxylation	aliphatic hydroxylation	N-oxide formation	nitro-group reduction	double bond peroxidation	hydroxyl-carbonyl oxidation	double bond formation (desaturation)	aldehyde oxidation	all reactions
CYP1A1	39	20	4	3	53	28	14	1	29	6	5	3	205
CYP1A2	85	35	9	3	88	46	25	4	34	8	10	2	349
CYP2C19	54	27	8	1	49	64	6	2	11	3	3	1	229
CYP2C9	56	25	10	1	67	68	14	1	20	12	9	0	283
CYP2D6	66	63	13	5	68	33	15	0	2	7	4	1	277
CYP2E1	33	18	4	1	46	54	13	1	32	10	3	2	217
CYP3A4	155	56	24	10	71	144	38	6	20	24	33	5	586
total set	488	244	72	24	442	437	125	15	148	70	67	14	2146

to that of the cytochrome substrates. The distribution of substrates correlates well with $logD_{7.4}$ values: both distributions display a positive trend toward the upper and the lower parts of the map (Figure 3c). These correlations reflect an important pattern observed for the cytochrome substrates: the reduction of compound's lipophilicity positively correlates with the decreased probability of its CYP metabolism, as CYPs' binding sites are lipophilic. The deviations from this rule may be related to the other molecular properties affecting metabolism, such as the nonspecific protein binding. For instance, the level of CYP metabolism prediction is poor when plasma protein binding is incorporated into clearance extrapolations, particularly for drugs with high plasma protein binding.[31] A study that specifically examined this phenomenon, showed different binding characteristics for basic and acidic drugs.[32] Both the serum-free fraction and the free microsomal fraction contained a relatively high fraction of the basic drugs propranolol and imipramine: 12 and 6% for serum-free and 38 and 16% for microsomes of a total drug content in the body, respectively. In contrast, the acidic drug warfarin showed a much lower concentration of the serum-free fraction (0.8%), but high microsomal binding (free fraction of 73%). Probably, this aspect can partly explain a relative significance of another molecular parameter, Jurs PPSA-1, i.e., the sum of solvent-accessible surface areas of all positively charged atoms (Figure 3d), describing what part of the molecule is positively charged.

Kohonen Maps for Individual CYP-specific Groups of Substrates. To evaluate the partitioning within the combined substrate dataset, we selected the seven largest enzyme-specific substrate groups. These CYPs (with the exception of CYP1A1) are responsible for the transformation of >95% of metabolized drugs.[33] The number of reactions for each CYP type from the database is shown in Table 7. All these cytochromes possess the broad and overlapping substrate specificity. The prevailing types of metabolic conversions are the aromatic and aliphatic hydroxylation and *N*- and *O*-dealkylation. The distribution of each group within the Kohonen map is shown in Figure 4. The substrates for some cytochromes, such as CYP2C9 and CYP2E1, are distributed widely within the map, which indicates the high substrate specificity for these particular isozymes. On the other hand, the substrates for the remaining isozymes are typically mapped in different distinct areas. This is particularly evident for CYP2D6, CYP2C19, and CYP1A1 substrates, with 2−3 distinct sites of localization different from the areas occupied by other substrate groups. These differences can be explained by the fact that the binding sites of CYP enzymes are not conservative, with binding site architecture and amino acid composition varying greatly for each isozyme.[34] Therefore, different micro environmental conditions define the possibility of specific enzyme−substrate interactions. Importantly, despite a significant diversity in each enzyme-specific group, the number of distinct clusters is relatively low. Probably, it indicates that the substrates can bind to the active sites of CYP enzymes only if they fit a rather narrow range of variability of several molecular properties.

The combined contour plot with localization of particular isozyme-specific substrate groups is shown in Figure 4h. Using these fingerprints, one can address the cytochrome-mediated metabolism in different tissues, most importantly in liver, muscles and lungs. For example, CYP1A1, largely undetectable in uninduced human liver, is found in lungs and placentas of cigarette smokers.[16] CYP2A6, CYP2B7, CYP2F1, and 4B1 have been identified in the human lung.[35] CYP2C9, CYP2D6, CYP2E1, and CYP3A4 are expressed in the liver and intestine, and CYP2D6 in liver and kidney.[36] A combination of such tissue-specific models with the knowledge of tissue distribution of organic compounds (literature data is available in our data set) is important for early evaluation of pharmacokinetic parameters.

Comparison of Supervised and Unsupervised Learning Approaches. Our observations indicate that the difference between human cytochrome substrates and the products of cytochrome-mediated reactions can be described by a combination of specific physicochemical features. Correspondingly, these categories of compounds have distinctly different localizations on the Kohonen map. We chose an unsupervised learning method because we cannot consider the subset of 523 CYP "products" as a true negative training set with respect to the "substrates". As a result, only the cytochrome P450 "substrates" can be unambiguously identified. However, we believe it was useful to evaluate the alternative classification algorithm, the supervised learning, for its ability to discriminate between these compound categories using the same set of molecular descriptors. Such comparison is particularly interesting as the back-propagation method was applied in almost 90% of all publications on neural networks in chemistry.

The NeuroSolution 4.0 program[15] was used for generation of the neural networks. We constructed feed-forward nets consisting of seven input neurons, one hidden layer, and two output neurons (italicized descriptors in Table 5). The networks were trained with the molecular descriptors as input values and the scores as output values. The back-propagated nets were trained following the momentum learning rule as implemented in NeuroSolution 4.0 over 1000 iterations. All scores were scaled between 0 and 1. The complete training set

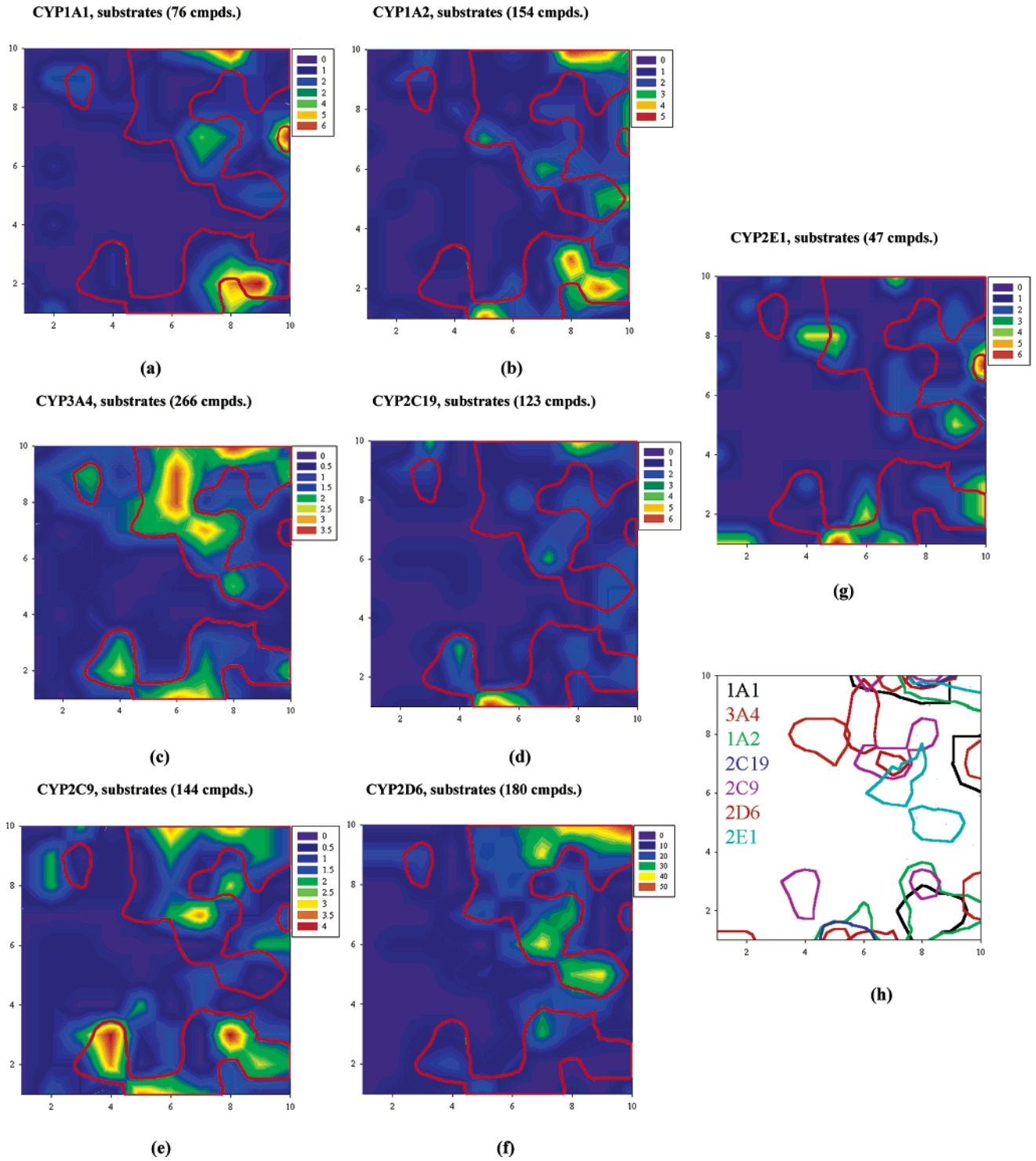


Figure 4. (a−g) Distributions of seven large isozyme-specific substrate groups within the Kohonen map. (h) The combined contour map of the occurrences of substrates within the Kohonen map. The contours restrict the areas that contain at least 70% of compounds belonging to each isozyme-specific group.

of 1008 compounds (substrates and products) was randomized and subdivided into three categories: (1) the training set (60% of the total number of compounds), (2) the cross-validation set (20%), and (3) the test set (20%). The cross-validation set was used to avoid overtraining while building the models. Control sets are essential for supervised learning, but are not required for unsupervised learning as in the latter case learning continues until the network stabilization. We conducted three independent training-testing experiments with the seven-descriptor set. The discriminative power of the trained network was moderate, as it demonstrated by the distribution of the substrates and products of CYP-mediated reactions (Figure 5). The classification quality was approximately the same in each of these three independent cycles: on average, 60% of CYP substrates and 65% of CYP reaction products were correctly classified in the corresponding test sets (Table 8). We assigned the score value 0.5 as a reasonable threshold (Figure 5).

We conclude that the unsupervised learning procedure provides with more accurate discrimination be-

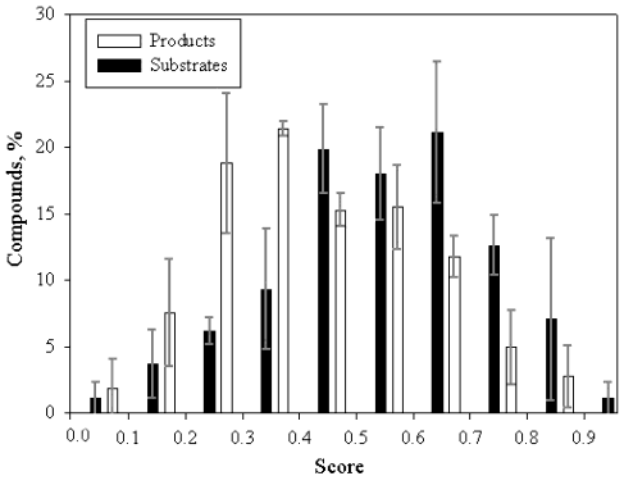


Figure 5. Compound distributions on the scale of calculated neural network scores for three independent test sets.

Table 8. Fraction of Correctly Classified Compounds Using Supervised Neural Network Approach

randomization	compound category	training set (60%)	cross-validation set (20%)	test set (20%)
rand 1	substrates	67.4	64.6	62.4
	products	68.4	66.3	63.5
rand 2	substrates	57.4	50.5	57.3
	products	73.4	67.3	65.2
rand 3	substrates	60.3	60	60
	products	70.4	78	65.7
average	substrates	61.7	58.4	59.9
	products	70.8	70.5	64.8

tween the studied compound categories than the supervised learning (see Figures 1 and 2 and Table 6). The moderate level of discrimination can be explained by the extreme diversity of cytochrome substrates and products. Such diversity resulted in several distinct clusters in the investigated property space corresponding to separate islands on the Kohonen map. The neural network classified objects into more than two classes, and not only into one out of several predefined ones. Overall, the unsupervised strategy was more efficient than the supervised method for solving the CYP-mediated metabolism problem.

Discussion

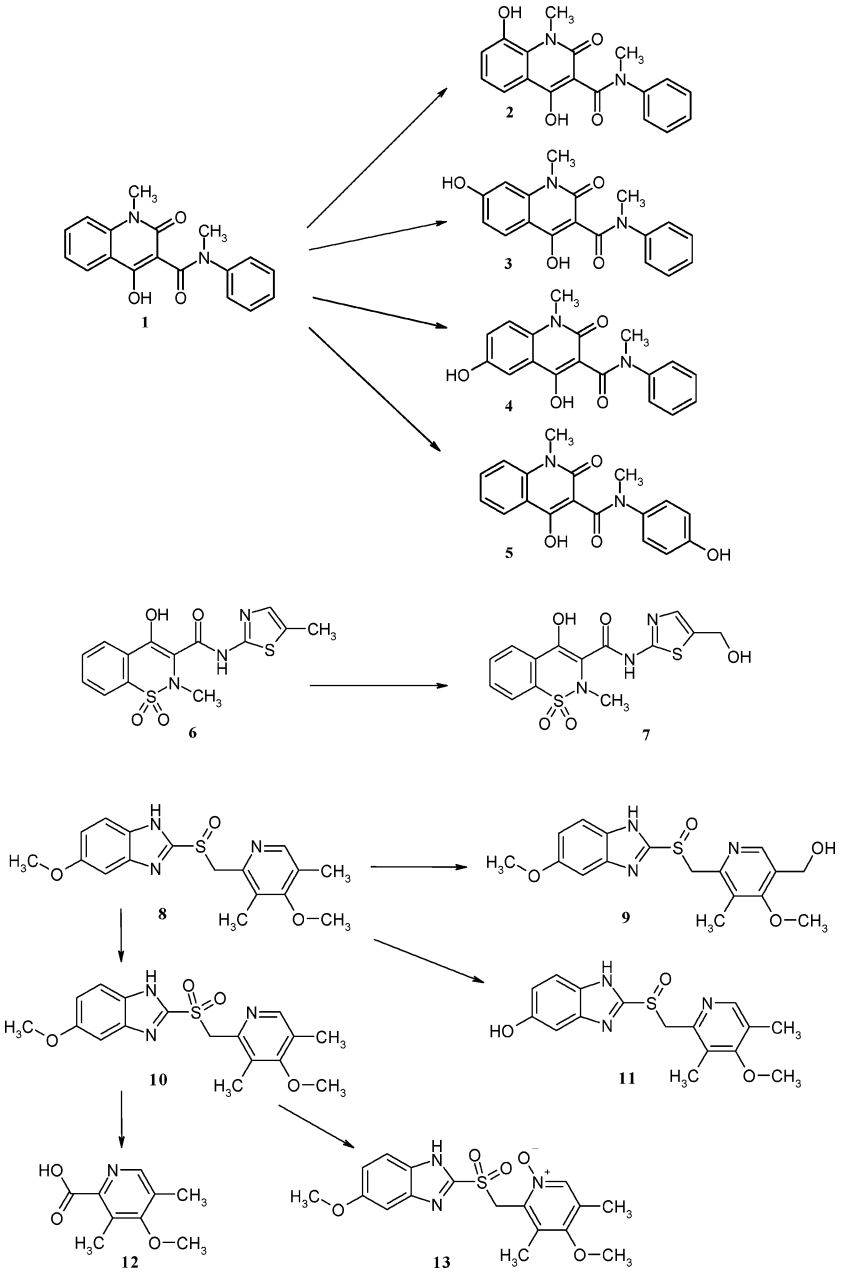
Early Consideration of Cytochrome-Mediated Metabolism in Drug Discovery. Currently, drug metabolism and toxicity in the human body is primarily a subject of clinical trials. Its outcome can be extrapolated based on preclinical experiments, both in vitro (hepatocytes, organ slices, etc.) and in vivo (several animal models). The acute human toxicity is predicted fairly well; chronic toxicity is reasonable on average, and nothing can be done for idiosyncratic toxicity. We believe that experiment-based in silico models allow us to consider human metabolism and toxicity at earlier stages of drug discovery. The most important application is at the level of advanced drug leads. Typically, a drug development program has several such leads with similar potency and selectivity but unknown human toxicity, and only one drug candidate is chosen for expensive clinical studies. The expert use of metabolic models could help to eliminate the ones in which metabolites are likely to appear toxic to humans. Moreover, our algorithm is applicable at the level of synthesis SAR hit-to-lead libraries. The compounds with the highest risk of formation of toxic intermediates can be excluded from synthesis, for example, by choosing "metabolism-friendly" building blocks. Prediction of metabolism can be useful for retro-metabolic drug design, including chemical delivery systems and soft drug approaches.[2] As liver metabolism is the predominant drug clearance mechanism, its early consideration may help in establishing the therapeutic dose of a novel agent. It is also useful to select metabolically "stable" compounds within a chemical series in order to incorporate stability into the candidate selection.[37] Finally, prediction of human metabolism can open new opportunities of evaluating synthetic combinatorial libraries for bioscreening.[38]

To summarize, reliable computational tools for early consideration of metabolic transformations of drug-like compounds are seriously needed. In this work, we developed a computational algorithm capable of recognizing substrates to human cytochromes P450. This algorithm allows early in silico evaluation of many metabolism-related effects, especially those associated with pharmacokinetics and toxicology.

Validation of CYP Metabolism Prediction. To validate the effectiveness of the developed model for prediction of metabolic transformation of drugs, we analyzed three different series of cytochrome-mediated reactions from our database (Scheme 1). These compounds were not used in the stage of building the Kohonen map, and, therefore, could be used as an independent validation set. Each compound was assigned to "substrates" or "products" within the reaction schemes; the descriptors were calculated for each compound and plotted on the Kohonen map used in all described experiments (Figure 6). The model correctly classified all the compounds with substrate potential, and most of all final products with no further metabolic degradation. Two final products were misclassified (**11** and **13**), but their localization in close proximity with the area of products allows us to classify them as likely final products.

We consider our model as a step in the development of a comprehensive algorithm for the assessment of all possible cytochrome-mediated and Phase II metabolic transformations that any "foreign" compound can undergo in the human body. For such algorithms to be relevant, additional groundwork is needed. First, a complete set of all possible metabolites for both Phase I and Phase II should be compiled. MetaDrug[11] developed by GeneGo, Inc. has the most complete dataset available today, with 12 000 "natural" human metabolites, 4000 xenobiotics metabolites. Second, one needs to establish and store in logically linked dictionaries a complete set of rules governing metabolic reactions based on the dataset (work in progress at GenGo). It is important to develop reliable criteria for metabolic "termination", i.e., to identify that which cannot be further metabolized. This is a nontrivial problem as many of Phase II xenobiotics metabolites become the substrates of normal metabolic reactions (over 8000 human reactions in MetaCore[11], a commercial product available from GeneGo, www.genego.com). The algorithm recognizes and applies these rules and then

Scheme 1. Metabolic reactions used for evaluation of the developed model[a]



[a] Positioning of substrates and products of these reactions on the Kohonen map is shown in Figure 6.

displays the potential metabolites of a xenobiotic. To assess the possibility of transformations for which no unambiguous rule-based prediction can be made, the nonlinear quantitative structure−metabolism relationships based on Kohonen self-organizing maps should be implemented. The probability of the transformation is defined by the position of a compound on the map. The final set of predicted metabolites is generated from the corresponding substrates localized within the positive areas on the map. Third, the established rules should be applied in the framework of a special computational platform able to combine all parts of the system into one mechanism. This platform should integrate the chemical database management program, the algorithm for generation of potential metabolite structures from the queried compounds, the programs for descriptor calculation and Kohonen network testing. Currently, we are working on a module for CYP450 metabolism prediction based on the CDL proprietary chemo informatics tool, ChemoSoft.

Isozyme-Specific Substrate Distribution. It is known that the CYT P450s responsible for the metabolism of most drugs has broad and overlapping substrate specificity.[33] This is well in line with our data on individual distributions of seven large isozyme-specific groups of substrates on the Kohonen map (Figure 4). At the same time, the distributions of several particular groups of substrates are substantially different. To explain this observation, we should emphasize that, unlike most enzyme families, the CYP superfamily does not have any highly conserved catalytic motif.[34] While CYPs do share a common heme unit for the delivery of the active oxygen atom to substrates, the substrate binding site per se is in one of the most variable regions throughout the family. This variability manifests at the levels of the gene sequences, the amino acid composition

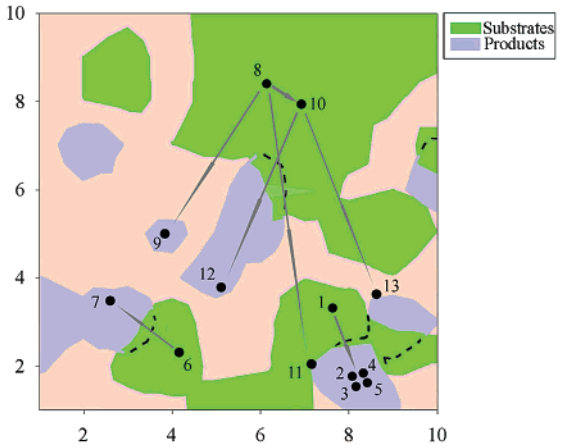


Figure 6. Application of the developed classification algorithm. Three groups of substrates and the final products correspond to the cytochrome-mediated metabolic reactions depicted in Scheme 1. The internal identification numbers are used for individual compounds. Arrows indicate the direction of each reaction.

and in the three-dimensional architecture of the binding sites. Presumably, such variability is necessary for the accommodation of such chemically different substrates.

Identification of the isozyme-specific areas on the Kohonen map is important for assessing the tissue-specificity of drug metabolism. Using the Kohonen maps for individual isozyme-specific groups of substrates, one can make a prognosis about cytochrome-mediated metabolism in different tissues. This tool could be particularly useful in combination with special algorithms capable of predicting the distribution of compounds within specific tissues.

Conclusions

In this work, we developed a neural network model for early evaluation of human cytochrome P450-mediated metabolism of drug-like compounds. The model is based on an unsupervised Kohonen learning approach and a preselected set of molecular descriptors. We have chosen the unsupervised learning method over the more popular supervised techniques due to the nature of CYP P450 metabolism as application and particular features of the available data set. Namely, these are (i) the lack of reliable data on CYP P450 nonsubstrates, and (ii) the extreme diversity of cytochrome substrates. The direct comparison of the efficiency of substrate/nonsubstrate discrimination conducted with unsupervised and supervised learning strategies, demonstrates the superiority of the former approach. The developed neural network model represents an effective tool for classification and visualization of drug-like compounds based on their ability to be cytochrome P450 substrates.

The model allows for the development of an automated computational algorithm for early assessment of possible cytochrome-mediated metabolic transformations that any compound can undergo in the human body. Using this algorithm, the position of a compound on the Kohonen map will determine the probability of its cytochrome-mediated metabolic transformation even in cases when no unambiguous rule-based prediction can be made.

Another useful extension of the proposed methodologyconsists of an early assessment of tissue-specificity of cytochrome-mediated metabolism. This is based on the map of tissue-specific fingerprints defining the localization of potential isozyme-specific substrate groups for the seven most significant cytochromes.

References

(1) Smith, R. V.; Erhardt, P. W.; Leslie, S. W. Microsomal O-Demethylation, N-Demethylation and Aromatic Hydroxylation in the Presence of Bisulfite and Dithiothreitol. *Res. Commun. Chem. Path. Pharmacol.* **1975**, *12*, 181−184.

(2) Bodor, N. Retrometabolic Approaches for Drug Design and Targeting. *Pharmazie* **1997**, *52*, 491−499.

(3) (a) Darvas, F. MetabolExpert: an Expert System for Predicting the Metabolism of Substances. In*QSAR in Environmental Toxicology II*; Kaiser, K. L., Ed.; Reidel Co.: Dordrecht, 1987; pp 71−81. (b) Darvas, F.; Marokházi, S.; Kormos, P.; Kulkarni, G.; Kalász, H.; Papp, Á. MetabolExpert: Its Use in Metabolism Research and in Combinatorial Chemistry. In *Drug Metabolism. Databases and High-Throughput Testing During Drug Design and Development*; Erhardt, P. W., Ed.; Blackwell Science Ltd., 1999; pp 237−270.

(4) (a) Klopman, G.; Dimayuga, M.; Talafous, J. META. 1. A Program for the Evaluation of Metabolic Transformations of Chemicals. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1320−1325. (b) Klopman, G.; Tu, M. META: A Program for the Prediction of the Products of Mammal Metabolism of Xenobiotics. In *Drug Metabolism. Databases and High-Throughput Testing During Drug Design and Development*; Erhardt, P. W., Ed.; Blackwell Science Ltd., 1999; pp 271−276.

(5) Darvas, F. Predicting Metabolic Pathways by Logic Programming. *J. Mol. Graphics* **1988**, *6*, 80−86.

(6) Snyder, R. W.; Grethe, G. Metabolite. In *Drug Metabolism. Databases and High-Throughput Testing During Drug Design and Development*; Erhardt, P. W., Ed.; Blackwell Science Ltd., 1999; pp 277−280.

(7) Mulder, G. *Conjugation Reactions in Drug Metabolism*; Taylor & Francis, London, 1990.

(8) Hayashi, S.; Watanabe, J.; Kawajiri, K. Genetic Polymorphism in the 5′-Flanking Region Change Transcriptional Regulation of the Human Cytochrome P450IIE1 Gene. *J. Biochem. (Tokyo)* **1991**, *110*, 559−565.

(9) Kato, S.; Onda, M.; Matsukura, N.; Tokunaga, A.; Tajiri, T.; Kim, D. Y.; Tsuruta, H.; Matsuda, N.; Yamashita, K.; Shields, P. G. Cytochrome P4502E1 (CYP2E1) Genetic Polymorphism in a Case-Control Study of Gastric Cancer and Liver Disease. *Pharmacogenetics* **1995**, *5*, S141−S144.

(10) (a) Ioannides, C.; Parke, D. V. *Cytochromes P450: Metabolic and Toxicological Aspects*. CRC Press: New York, 1996. (b) Lechner, C. *Cytochrome P450: Biochemistry, Biophysics and Molecular Biology*; John Libbey Eurotext, Montrouge, France, 1995.

(11) Commercially available from GeneGo, Inc. (New Buffalo, USA). URL: http://www.genego.com/.

(12) Accelrys, Inc., 2000. URL: http://www.accelrys.com/.

(13) Chemical Diversity Labs, Inc., 2002. URL: http://www.chemdiv.com/.

(14) Raevsky, O. A.; Trepalin, S. V.; Trepalina, H. P.; Gerasimenko, V. A.; Raevskaya, O. E. SLIPPER-2001 − Software for Predicting Molecular Properties on the Basis of Physicochemical Descriptors and Structural Similarity. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 540−549.

(15) NeuroDimension, Inc., 2001. URL: http://www.nd.com/.

(16) Gonzales, F. The Molecular Biology of Cytochrome P450s. *Pharmacol. Rev.* **1989**, *40*, 243−288.

(17) Koymans, L.; Kelder, G.; Te, J.; Vermeulen, N. Cytochromes P450: Their Active-Site Structure and Mechanism of Oxidation. *Drug Metab. Rev.* **1993**, *25*, 325−327.

(18) Estabrook, R. W. Cytochrome P450: from a Single Protein to a Family of Proteins − with some personal reflections. In *Cytochromes P450: Metabolic and Toxicological Aspects*; Ionnides and Parke, Eds.; CRC Press: New York, 1996; pp 4−28.

(19) Nebert, D. Multiple Forms of Inducible Drug-Metabolizing Enzymes: a Reasonable Mechanism by Which Any Organism Can Cope with Adversity. *Mol. Cell. Biochem.* **1979**, *27*, 27−46.

(20) (a) de Groot, M. J.; Bijloo, G. J.; Martens, B. J.; van Acker, F. A. A.; Vermeulen, N. P. E. A Refined Substrate Model for Human Cytochrome P450 2D6. *Chem. Res. Toxicol.* **1997**, *10*, 41−48. (b) 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. (c) Ekins, S.; Bravi. G.; Binkley, S.; Gillespie, J. S.; Ring, B. J.; Wikel, J. H.; Wrighton, S. A. Three- and Four-dimensional Quantitative Structure−Activity Relationship Analyses of Cytochrome P-450 3A4 Iinhibitors. *J. Pharmacol. Exp. Ther.* **1999**, *290*, 429−438.

(21) 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.
(22) (a) Ajay, A.; Walters, W. P.; Murcko, M. A. Can We Learn to Distinguish between "Drug-Like" and "Nondrug-Like" Molecules? *J. Med. Chem.* **1998**, *41*, 3314−3324. (b) Ajay; Bemis, G. W.; Murcko, M. A. Designing Libraries with CNS Activity. *J. Med. Chem.* **1999**, *42*, 4942−4951. (c) Sadowski, J.; Kubinyi, H. A Scoring Scheme for Discriminating between Drugs and Nondrugs. *J. Med. Chem.* **1998**, *41*, 3325−3329.
(23) Balakin, K. V.; Tkachenko, S. E.; Lang, S. A.; Okun, I.; Ivashchenko, A. I.; Savchuk, N. P. Property-Based Design of GPCR−Targeted Library. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1332−1342.
(24) Anzali, S.; Gasteiger, J.; Holzgrabe, U.; Polanski, J.; Sadowski, J.; Teckentrup, A.; Wagener, M. The Use of Self-Organizing Neural Networks in Drug Design. In *3D QSAR im Drug Design − Volume 2;* Kubinyi, H.; Folkers, G.; Martin, Y. C., Eds.; Kluwer/ESCOM, Dordrecht, NL, 1998; pp 273−299.
(25) Bauknecht, H.; Zell, A.; Bayer, H.; Levi, P.; Wagener, M.; Sadowski, J.; Gasteiger, J. Locating Biologically Active Compounds in Medium-Sized Heterogeneous Datasets by Topological Autocorrelation Vectors: Dopamine and Benzodiazepine Agonists. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1205−1213.
(26) Anzali, S.; Barnickel, G.; Krug, M.; Sadowski, J.; Wagener, M.; Gasteiger, J.; Polanski, J. The Comparison of Geometric and Electronic Properties of Molecular Surfaces by Neural Networks: Application to the Analysis of Corticosteroid Binding Globulin Activity of Steroids. *J. Comput.-Aid. Mol. Des.* **1996**, *10*, 521−534.
(27) Brüstle, M.; Beck, B.; Schindler, T.; King, W.; Mitchell, T.; Clark, T. Descriptors, Physical Properties, and Drug-Likeness. *J. Med. Chem.* **2002**, *45*, 3345−3355.
(28) Jolliffe, I. T. *Principal Component Analysis*; Springer-Verlag: New York, 1986.
(29) (a) Guttmann, L. Some Necessary Conditions for Common Factor Analysis. *Psychometrika* **1954**, *19*, 149−162. (b) Kaiser, H. F.; Dickmann, K. Analytic Determination of Common Factors. *Am. Psychol.* **1959**, *14*, 425−439.
(30) (a) Catell, R. B.; Vogelmann, S. A Comprehensive Trial of the Scree and KG-Criteria for Determining the Number of Factors. *Multi. Behav. Res.* **1977**, *12*, 289−325. (b) Catell, R. B. The Scree Test for the Number of Factors. *Multi. Behav. Res.* **1966**, *1*, 245−276.
(31) Obach, R. S.; Baxter, J. G.; Liston, T. E.; Silber, B. M.; Jones, B. C.; Macintyre, F.; Rance, D. J.; Wastall, P. The Prediction of Human Pharmacokinetic Parameters from Preclinical and in vitro Metabolism Data. *J. Pharm. Exp. Ther.* **1997**, *283*, 46−58.
(32) Obach, R. S. Nonspecific Binding to Microsomes: Impact on Scale-up of in vitro Intrinsic Clearance to Hepatic Clearance as Assessed Through Examination of Warfarin, Imipramine and Propranolol. *Drug Metab. Dispos.* **1997**, *25*, 1359−1369.
(33) Spatzenegger, M.; Jaeger, W. Clinical Importance of Hepatic Cytochrome P450 in Drug Metabolism. *Drug Metab. Rev.* **1995**, *27*, 397−417.
(34) (a) Chang, Y.-T.; Stiffelman, O. B.; Vakser, I. A.; Loew, G. H.; Bridges, A.; Waskell, L. Construction of a 3D Model of Cytochrome P4502B4. *Prot. Engin.* **1997**, *10*, 119−129. (b) Chang, Y.-T.; Loew, G. H. Computer Modeling of 3D Structures of Cytochrome P450s. *Biochimie* **1996**, *78*, 771−779.
(35) Gonzales, F.; Gelboin, H. Human Cytochromes P450: Evolution and cDNA-directed Expression. *Environ. Health Persp.* **1992**, *98*, 81−85.
(36) Gonzales, F. Human Cytochromes P450: Problems and Prospects. *Trends Pharm. Sci.* **1992**, *13*, 346−352.
(37) Rodrigues, A. D. Use of in vitro Metabolism Studies in Drug Development: an Industrial Perspective. *Biochem. Pharmacol.* **1994**, *48*, 2147−2156.
(38) Darvas, F.; Dormán, G.; Papp, A. Diversity Measures for Enhancing ADME Admissibility of Combinatorial Libraries. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 314−322.

JM030102A