JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

Automated Information Extraction and Structure−Activity Relationship Analysis of Cytochrome P450 Substrates

Fumiyoshi Yamashita,*[,†] Chunlai Feng,[†] Shuya Yoshida,[†] Takayuki Itoh,[§] and Mitsuru Hashida[†,‡]

[†]Graduate School of Pharmaceutical Sciences

[‡]Institute for Integrated Cell-Material Sciences

Kyoto University, Yoshidaushinomiya-cho, Sakyo-ku, Kyoto 606-8501, Japan

[§]Department of Information Sciences, Faculty of Science, Ochanomizu University, 2-1-1 Otsuka, Bunkyo-ku, Tokyo 112-8610, Japan

Supporting Information

ABSTRACT: Information on CYP−chemical interactions was comprehensively explored by a text-mining technique, to confirm our previous structure−activity relationship model for CYP substrates (Yamashita et al. *J. Chem. Inf. Model.* **2008**, *48*, 364−369). The text-mining technique is based on natural language processing and can extract chemical names and their interaction patterns according to sentence context. After chemicals were automatically extracted and classified into CYP substrates, inhibitors, and inducers, 709 substrates were retrieved from the PubChem database and categorized as 216, 145, 136, 217, 156, and 379 substrates for CYP1A2, CYP2C9, CYP2C19, CYP2D6, CYP2E1, and CYP3A4, respectively. Although the previous classification model was developed using data from only 161 compounds, the model classified the substrates found by text-mining analysis with reasonable accuracy. This confirmed the validity of both the multi-objective classification model for CYP substrates and the text-mining procedure.

INTRODUCTION

The cytochrome P450s (CYPs) are a superfamily of heme-containing mixed function oxygenases that catalyze the regio- and stereoselective oxidation of a wide variety of xenobiotics, including drugs. The broad substrate specificity of CYPs often leads to unexpected drug−drug interactions. Competitive or noncompetitive inhibition of CYPs by coadministered drugs retards the body clearance of the drugs and results in unexpected rises in their blood concentrations. On the other hand, induced expression of CYPs reduces or shortens the duration of pharmacological activity of the drugs by accelerating the body clearance. Prediction of drug−drug interactions associated with CYPs is an important issue in drug discovery and development as well as in their clinical applications.

The human CYP superfamily consists of more than 50 functional genes. Over 90% of human drug oxidation can be attributed to the following CYPs: 1A2 (4%), 2A6 (2%), 2C9 (10%), 2C19 (2%), 2E1 (2%), 2D6 (30%), and 3A4 (50%).[1] Each individual CYP isoform has a unique substrate specificity, often to a particular region of a drug molecule, to a particular stereoisomer, or to both, although considerable overlap in substrate specificity also exists.[2] We recently proposed a multi-objective structure−activity relationship (SAR) analysis method that consists of hierarchical classification and large-scale visualization techniques, and we adopted it to obtain a bird's-eye view of the substrate specificity of 6 CYP isoforms.[3] The analysis revealed several SARs for CYP-mediated metabolism: (1) CYP2C9 substrates are mostly anionizable compounds, (2) in contrast, many CYP2D6 substrates are cationic compounds, (3) CYP2E1 preferentially metabolizes smaller compounds, and (4) there is a positive correlation between metabolic susceptibility toward CYP3A4 and molecular volume. These findings are essentially the same as those reported by Smith et al.[4] and Lewis.[5]

In the previous analysis, we used the Bonnabry et al.[6] data set of 161 clinically relevant drugs. Although the chemical structures were diverse, the data might be too limited to generalize SARs. In the present study, we aim to follow up the previous findings with comprehensive literature searches to increase the data available for analysis. We have recently developed a text-mining system for automatically extracting information from the literature on chemical−CYP3A4 interactions (substrate, induction, and inhibition).[7] The system is based on a natural language processing technique. It identifies chemicals and CYP3A4 name variants in the text according to a combination of name dictionaries and context features and extracts information on chemical−CYP3A4 interactions based on the order of three keywords: chemicals, CYP3A4 name variants, and key verbs. Although it is a simple pattern matching method, the system achieved 87.4% recall and 92.3% precision for identification of chemical names and 85.2% recall and 92.0% precision for the extraction of chemical−CYP3A4 interactions. We have now collected information on CYP isoforms other than CYP3A4 by the text-mining system and used it to investigate the validity of the previous decision tree model.

MATERIALS AND METHODS

Text Mining of Chemical−CYP Interactions. The automated information extraction of chemical−CYP interactions

Received: August 30, 2010
Published: January 19, 2011

Table 1. A Typical Example of the Interaction Extraction Processes[a]

step no.	process	result
0		timolol is used topically for the treatment of glaucoma and metabolized by cytochrome P450 (CYP) 2D6 in the liver.
1	part of speech and morphological processing	timolol$_{/NN}$ is$_{/VBZ(be)}$ used$_{/VBN(use)}$ topically$_{/RB}$ for$_{/IN}$ the$_{/AT}$ treatment$_{/NN}$ of$_{/IN}$ glaucoma$_{/NN}$ and$_{/CC}$ metabolized$_{/VBN(metabolize)}$ by$_{/IN}$ cytochrome$_{/NN}$ P$_{/NNP}$450$_{/CD}$ (CYP$_{/NNP}$) 2$_{/CD}$D$_{/NNP}$6$_{/CD}$ in$_{/IN}$ the$_{/AT}$ liver$_{/NN./}$
2	Gazetter lookup and name identification	**timolol** is used topically for the treatment of glaucoma and metabolized by **cytochrome P450 (CYP) 2D6** in the liver.
3	noun group creation	timolol *is used* topically for the treatment of glaucoma and *metabolized* by cytochrome P450 (CYP) 2D6 in the liver
4	noun phrase creation	[timolol] (is used topically for) [the treatment of glaucoma] and (metabolized by) [cytochrome P450 (CYP) 2D6 in the liver]
5	noun phrase group creation	**timolol** {is used topically for} **the treatment of glaucoma** and {metabolized by} **cytochrome P450 (CYP) 2D6 in the liver**
6	clause reconstruction	timolol metabolized by cytochrome P450 (CYP) 2D6 in the liver
7	interaction extraction	timolol metabolized by 2D6 (substrate); rule: (Chem) (keyVerb, passive) (CYP)

[a] Step 1: Subscript text indicates the part of speech, and parentheses indicate the base forms. Definitions: AT, article; CC, coordinating conjunction; CD, cardinal numeral; IN, preposition; NN, noun; NNP, proper noun; RB, adverb; VBN, verb past participle; VBZ, verb, third singular present. Step 2: Bold indicates chemical or CYP names listed in dictionaries. Step 3: Underlining indicates noun groups, and italic indicates verb groups. Step 4: Square brackets [ ] indicate noun phrases, and parentheses () indicate verb phrases. Step 5: Bold text indicates noun phrase groups, and curly brackets { } indicate verb phrase groups. Step 6: Clauses involving relationships between chemicals and CYPs are reconstructed. Step 7: Chemical−CYP interactions are extracted by pattern matching based on the order of keywords.

consists of three major steps. First, chemicals and CYP names are identified in the text. Second, sentences containing any names of chemicals and CYP are transformed into simple clauses, each of which contains a single event. Finally, information on chemical−CYP interactions is extracted from the clauses by a pattern matching method. Since details of each step were reported in our previous article,[7] only a brief summary is provided herein.

The CYP name dictionaries containing name variants of each isoform and the chemical name dictionary comprising approximately 100 000 entries were created by extracting the names of chemicals and proteins from the MeSH of the U.S. National Library of Medicine (NLM). Because simple dictionary-based approaches miss a number of chemical names or cause partial matches of chemical names, a context-based approach comprising a set of pattern matching rules was adopted together with a dictionary-based one. Specific term dictionaries containing words that are related to chemical names (concentration, effect, mg/mL, etc.) were used for the purpose. To create a chemical name dictionary, the first scan of an entire text was performed by pattern matching.

After the entire text was rescanned to identify chemical and CYP names, sentences containing both of them were subjected to sentence processing. Each sentence was divided into noun and verb phrases based on part-of-speech (POS) tags. Simple clauses expressing a single event involving chemicals and CYPs were then reconstructed by considering the voice of each verb phrase and syntactic structure. When a sentence was complex, noun phrases paired with each verb phrase were identified by a set of pattern matching rules. Additional modification to the previous system[7] was only that, if a sentence had a noun phrase group containing plural chemicals and another containing plural CYPs, it was excluded from the information extraction analysis of chemical−CYP interactions.

The mode of action of compounds with CYP was specified by semantics of verbs. A key verb list was created. The list includes verbs indicating chemical−CYP interaction (e.g., metabolize, transform, inhibit) along with the verbal nouns (e.g., metabolism, oxidation, inhibition) and adjectives (e.g., metabolic, inhibitory) derived from the verbs. Chemical-denoting nouns (e.g., substrate) were also added to the list. The pattern matching analysis was performed based on the order of appearance of three types of keywords (chemical names, CYP names, and key verbs) in a clause and by taking prepositions and coordinate conjunctions into account. Several patterns were created in each case depending on whether chemical and CYP names existed in the same phrase or in different phrases of a simple clause. In addition, negative elements (e.g., “not”, “no”, “n’t”, “unable”, and “unlikely”) were identified to check if the clause implied a negative context.

The overall architecture of our system was implemented on the general architecture for text engineering (GATE) platform developed at the University of Sheffield, U.K.[8] For tokenization, POS tagging, and morphological processing, we used the tools provided by GATE without any change. A typical example of the chemical−CYP interaction extraction process is shown in Table 1.

Multi-Objective Recursive Partitioning Analysis and its Validation. Recursive partitioning is an exploratory data mining technique, which successively splits a data set into increasingly homogeneous subsets. To deal with multi-objective problems, we have developed an extension of the recursive partitioning technique.[3] In the previous article,[3] we used the Bonnabry et al. data set[6] of 161 clinically relevant drugs to analyzed CYP-mediated metabolism. Similar to conventional techniques, our technique performed a brute-force search to find the best splitting rule on the basis of a quality-of-split criterion ($O(S)$).

dx.doi.org/10.1021/ci100334z |J. Chem. Inf. Model. 2011, 51, 378–385

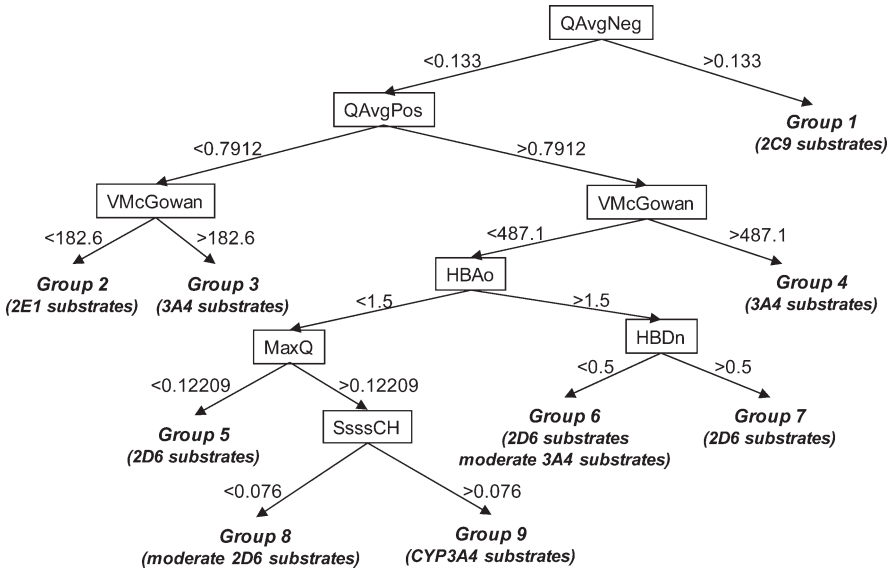

Figure 1. A multi-objective decision tree model for CYP substrates. The model was developed by a multi-objective recursive partitioning method which we reported previously.[3] The Bonnabry et al. data set[6] comprising 161 clinically relevant drugs was used for construction of the model. Keys: QAvgNeg, absolute value of the population average across all ionized species of the net formal negative charge at pH 7.4; QAvgPos, absolute value of the population average across all ionized species of the net formal positive charge at pH 7.4; VMcGowan, McGowan molecular volume; HBAo, number of oxygen-based hydrogen-bond acceptors; HBDn, number of nitrogen-based hydrogen-bond donors; MaxQ, maximal partial equalization of orbital electronegativity (PEOE) partial atomic charge; SsssCH, atom-type E-state index for −CH-group.

The quality-of-split criterion that we used is as follows:

$$O(S) = \sum_i \mathrm{IG}_i(S) \quad (1)$$

$$\mathrm{IG}_i(S) = \mathrm{IE}_i(S) - \left(\frac{N_{S_1}}{N_S}\mathrm{IE}_i(S_1) + \frac{N_{S_2}}{N_S}\mathrm{IE}_i(S_2)\right) \quad (2)$$

$$\mathrm{IE}_i(S) = -\sum_j \frac{N_j^i}{N_S}\log\frac{N_j^i}{N_S} \quad (3)$$

where N_j^i is the number of elements of which the i-th attribute belongs to category j; N_S is the total number of elements in group S (that is, $N_S = \sum_j N_j^i$); and S_1 and S_2 indicate subgroups obtained by binary splitting of S. The parameters IE and IG are generally referred to as information entropy and information gain, respectively. As shown in eq 1, to obtain the best compromise of multiple objective variables, the quality-of-split criterion was defined as the sum of the information gains for each objective attribute. After the binary tree was fully grown, pruning of the tree was performed with reference to the misclassification rate determined by the "leave-some-out" cross-validation procedure. The decision tree model that was finally obtained is shown in Figure 1.[3]

In the present study, we classified the CYP substrates that were extracted from the literature by text mining according to the following model: First, molecular descriptors of the CYP substrates were calculated using ADMET Predictor, ver. 4.0 (Simulations Plus, Inc., Lancaster, CA). The descriptors include constitutional descriptors and topological and electrotopological descriptors and descriptors for hydrophobicity, electronic properties, hydrogen bonding, and molecular ionization. The CYP substrates were then classified based on the decision tree model.

Visualization of Hierarchically Structured Data. Hierarchical data classified by the decision tree model were displayed using a data visualization technique, HeiankyoView.[9,10] The technique represents leaf nodes of hierarchical data as square icons and nonleaf nodes as nested rectangular borders. The technique first places the leaf nodes at the lowest level of the hierarchy onto a display space and represents a nonleaf node by enclosing the leaf nodes. The technique then places leaf nodes and nonleaf nodes at a higher level and again encloses them by another rectangle. The technique represents the entire hierarchical data by repeating the process until it reaches the top of the hierarchy. How to solve the rectangle packing problem of finding the optimal display layout of leaf and nonleaf nodes has been described elsewhere.[9,10] Basically, the technique attempts to satisfy the following conditions: (1) rectangles never overlap one another; (2) the area of the rectangular region enclosing the placed rectangles should be minimized; and (3) the aspect ratio of the rectangular region enclosing the placed rectangles should be optimized.

RESULTS

Extraction of CYP Substrates, Inhibitors, And Inducers from PubMed Abstracts. Full-text PubMed abstracts relating to each CYP isoform were searched for by entering the key phrase comprising CYP name variants (e.g., CYP1A2 or CYP450 1A2 or CYTOCHROME P450 1A2 or P450 1A2) in the search box. The numbers of abstracts retrieved were 4108, 2670, 2275, 4423, 4301, and 5278 for CYP1A2, CYP2C9, CYP2C19, CYP2D6, CYP2E1, and CYP3A4, respectively. Each set of abstracts was subjected to the text-mining algorithm. The texts were analyzed in a sentence by sentence manner, and the sentences that include both chemical names and CYP names were further subjected to chemical−CYP interaction extraction analysis. Table 2 summarizes key verbs or verbal nouns used for the classification of retrieved chemicals into substrates, inhibitors, or inducers. In the case that a verb or verbal noun detected

Table 2. Verbs and Verbal Nouns Used for Classification into Substrates, Inhibitors, and Inducers

Substrate	demethylenation	nitroreduction	induce
activate	denitrosation	substrate	inducer
bioactivation	depropargylation	sulfoxidation	induction
biotransformation	depropylation	thiono-oxidation	stimulate
catalyze	desalkylation	transform	
clear	despropylation	transformation	**Chemicals**[a]
clearance	desulfuration		decrease
conversion	detoxication	**Inhibitor**	enhance
dealkylation	detoxification	inactivate	increase
dearylation		inactivation	incubate
debenzylation	eliminate	inhibit	incubation
debutylation	elimination	inhibition	interact
dechloroethylation	epoxidation	inhibitor	interaction
de-esterification	form	interfere	modulate
de-ethylation	formation	suppress	modulation
defluorination	hydrogenation	suppression	probe
dehalogenation	hydroxylate		reduce
deisopropylation	inducible	**Inducer**	regulate
demethylate	metabolism	activator	regulation
demethylation	metabolize	elevate	response

[a] This category of words was used to detect chemical names.

by pattern matching was in the key verb list (Table 2), the record comprising a set of chemical/CYP/verb was stored in the database. The total number of records assigned to each of the three categories were 2391, 1477, 1246, 2482, 2430, and 4466 for CYP1A2, CYP2C9, CYP2C19, CYP2D6, CYP2E1, and CYP3A4, respectively. After records that contained duplicate chemical names were deleted, the chemical structures were searched for in the PubChem database. For the compounds that were retrieved from PubChem, molecular descriptors were calculated from the two-dimensional (2D) structure. When all of the molecular descriptors of any two chemicals coincided with each other, the two chemicals were assumed to be the same. Finally, a list of chemicals interacting with CYP isoforms was obtained (Supporting Information, Tables 1−3: It should be noted that the list in the table was created based only on the text-mining procedure. The accuracy of all information has not been confirmed.) When there were contradictory records for a chemical, the following rule was applied: if "positive" records exceeded "negative" ones by three-fold or more, then the chemicals were regarded as an interactant (that is, substrate, inhibitor, or inducer). Here, a "positive" means that the chemical is a substrate/inhibitor/inducer, while a "negative" means that the chemical is explicitly a nonsubstrate/noninhibitor/noninducer. If "positive" and "negative" records existed but in a ratio less than 3:1 or if there are no related records, then we deferred the conclusion that the chemical is an interactant and practically regarded it as a noninteractant. Finally, the numbers of substrates were 216, 145, 136, 217, 156, and 379, inhibitors were 167, 130, 111, 154, 121, and 274; and inducers were 101, 31, 16, 22, 81, and 125 for CYP1A2, CYP2C9, CYP2C19, CYP2D6, CYP2E1, and CYP3A4, respectively.

To confirm the validity of the text-mining system, 200 abstracts were randomly selected from the entire abstract corpus, and the recall and the precision were calculated for both chemical name identification and interaction extraction procedures. Here, the recall was the ratio of the number of relevant items retrieved to the number of relevant items in collection, while the precision was the ratio of the number of relevant items retrieved to the number of items retrieved. The system achieved 86.3% recall and 86.7% precision for the identification of chemical names and 72.2% recall and 83.9% precision for the extraction of chemical−CYP interactions.

Analysis of Commonality in Substrates, Inhibitors, And Inducers between CYP Isoforms. For a given pair of CYP isoforms, the Tanimoto similarity index (TI) was then calculated

$$\mathrm{TI} = \frac{c}{a+b-c}$$

where a and b are the numbers of substrates of each CYP isoform, and c is the number of substrates in common between the two. The same calculation was done with inhibitors and inducers. Table 3 summarizes the Tanimoto similarity indices for substrates, inhibitors, and inducers in common between any pair of CYP isoforms. Here, 95% bootstrap confidence intervals were also calculated by producing 1000 bootstrap samples for each. As shown in Table 3, combination of CYP2C9 and CYP2C19 exhibited the highest Tanimoto similarity indices in commonality of substrates, inhibitors, and inducers. The Tanimoto similarity indices regarding CYP2E1 were generally lower than those regarding CYP3A4.

When the amino acid sequences of human CYP isoforms were compared using ClustalW (http://www.clustal.org/), CYP2C9 and CYP2C19, with a score of 91%, were found to share the highest homology of the CYP isoforms studied here. The homology of CYP2E1 with CYP2C9 and CYP2C19 was 57% and with CYP2D6 it was 38%. Although the chemical interaction profiles of CYP2E1 were quite different from those of the other CYP isoforms, its homology with the other isoforms is higher than the homology between CYP3A4 and all other isoforms (19−22%). It means that commonality in substrates/inhibitors/inducers among the CYP isoforms cannot be explained only by their homologies of amino acid sequence.

Validation of the Multi-Objective Decision Tree Model for CYP Substrates. The CYP substrates were classified by the multi-objective decision tree model reported previously[3] (Figure 1). The decision tree model classifies compounds hierarchically according to the molecular descriptors obtained by the ADMET Predictor. The compounds are first divided into two groups depending on whether the population average across all ionized species of the net formal negative charge at pH 7.4 (QAvgNeg) is greater or less than 0.133. If QAvgNeg is less than the specified value, then the compounds are subdivided depending on whether the net formal positive charge (QAvgPos) is greater or less than 0.7912. By repeating this procedure and applying the other descriptors in turn as shown in Figure 1, all compounds are finally classified into one of nine terminal node groups.

The classification result for the 709 CYP substrates obtained by the text-mining analysis is shown in Figure 2A. The bar length represents the proportion of each of the CYP substrates within the groups. The ID number of the groups corresponds to that in Figure 1. Figure 2B represents the classification of the 161 drugs that were used for construction of the model. The population profiles in all groups, except Group 9, are well correlated with the original profiles. The poor correlation in Group 9 might be due to too few members in the group (23 of 709 for the present data and 5 of 161 for the training data). In addition, the population of CYP2C9 substrates in Group 1 is lower than expected from the previously reported decision tree model.

Table 3. Tanimoto Similarity Indices Representing Commonality in Substrates, Inhibitors, And Inducers between Any Pair of CYP Isoforms[a]

		CYP1A2	CYP2C9	CYP2C19	CYP2D6	CYP2E1	CYP3A4
substrate	CYP1A2	—	0.168 (0.126−0.212)	0.189 (0.144−0.235)	0.213 (0.173−0.256)	0.163 (0.123−0.202)	0.207 (0.171−0.242)
	CYP2C9		—	0.301 (0.235−0.364)	0.175 (0.132−0.219)	0.132 (0.093−0.169)	0.196 (0.162−0.237)
	CYP2C19			—	0.201 (0.155−0.250)	0.066 (0.038−0.097)	0.192 (0.157−0.230)
	CYP2D6				—	0.104 (0.073−0.138)	0.221 (0.183−0.257)
	CYP2E1					—	0.117 (0.090−0.144)
	CYP3A4						—
inhibitor	CYP1A2	—	0.222 (0.171−0.279)	0.219 (0.169−0.274)	0.180 (0.138−0.224)	0.176 (0.126−0.224)	0.195 (0.156−0.231)
	CYP2C9		—	0.317 (0.255−0.385)	0.235 (0.183−0.291)	0.096 (0.060−0.133)	0.213 (0.170−0.257)
	CYP2C19			—	0.274 (0.213−0.335)	0.143 (0.098−0.194)	0.215 (0.170−0.261)
	CYP2D6				—	0.091 (0.058−0.128)	0.206 (0.164−0.249)
	CYP2E1					—	0.125 (0.091−0.161)
	CYP3A4						—
inducer	CYP1A2	—	0.073 (0.032−0.124)	0.045 (0.009−0.088)	0.060 (0.018−0.104)	0.124 (0.075−0.175)	0.136 (0.086−0.185)
	CYP2C9		—	0.175 (0.053−0.308)	0.060 (0−0.134)	0.062 (0.020−0.112)	0.147 (0.090−0.205)
	CYP2C19			—	0.056 (0−0.139)	0.029 (0−0.067)	0.093 (0.044−0.144)
	CYP2D6				—	0.037 (0.009−0.279)	0.050 (0.015−0.279)
	CYP2E1					—	0.092 (0.052−0.279)
	CYP3A4						—

[a] The values in parentheses represent a 95% bootstrap confidence interval.

All classified data were visualized by the large-scale data visualization technique called HeiankyoView. In Figure 3, for each of the CYP isoform, the red and blue icons indicate substrates and nonsubstrates, respectively, and the nested rectangular borders indicate class groups that were divided hierarchically according to the rules in the decision tree. It should be noted that the data are arranged symmetrically in each diagram and the icons located in the same position indicate the same compounds. The CYP2C9 substrates are most dense in the upper rectangular group that depicts a QAvgNeg of greater than 0.133. The CYP2D6 substrates are densest in the bottom-right region that indicates a QAvgNeg of less than 0.133 and a QAvgPos of greater than 0.7912. Most of the CYP2E1 substrates are in the bottom-left region that indicates a QAvgNeg of less than 0.133, a QAvgPos of less than 0.7912, and a McGowan molecular volume (VMcGowan) of less than 182.6 cm^3/mol. While the substrates of CYP3A4 substrates are many, they are mainly in the region that is just above the CYP2E1 region representing the group with a VMcGowan of greater than 182.6 cm^3/mol. The substrates of CYP1A2 and CYP2C19 are distributed equally in all regions of the diagram, indicating that these two isoforms are the least specific in their binding. Thus, the HeiankyoView provides intuitively understandable visual images for the classification of CYP substrates by the decision tree model.

DISCUSSION

The recursive partitioning method has been used as one of the SAR modeling approaches. The recursive partitioning model creates a decision tree that, in addition to its nonparametric and nonlinear characteristics, is easily understood. We previously proposed a multi-objective recursive partitioning method, to assist in simultaneous multi-objective optimization of chemical

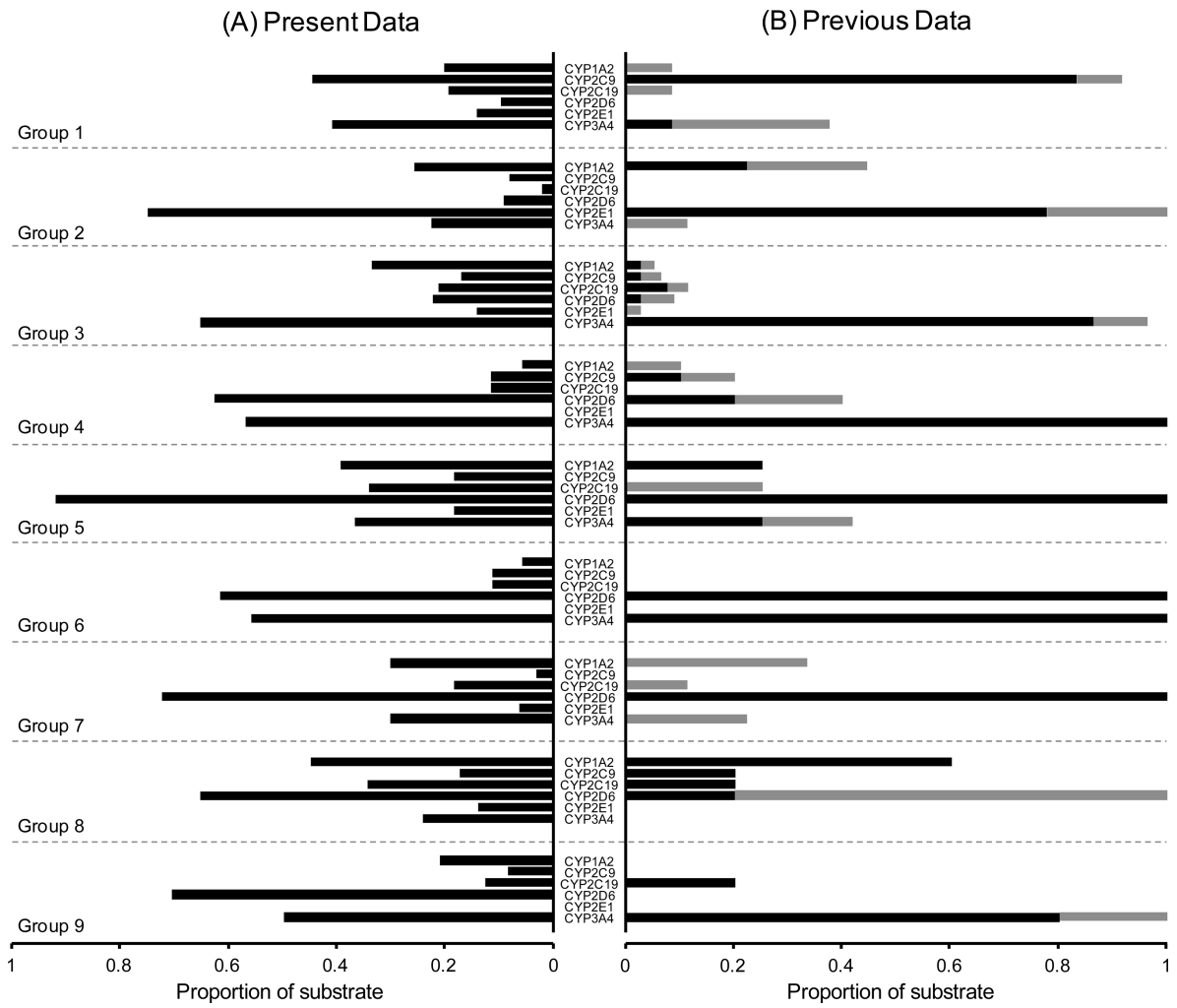

Figure 2. Proportions of each CYP substrate in the terminal node groups of the multi-objective decision tree model shown in Figure 1. (A) The 709 compounds obtained by the present text-mining procedure were classified according to their chemical structure. (B) The 161 compounds used in the construction of the decision tree model were similarly classified. Solid bars indicate extensively metabolized compounds, and shaded bars indicate moderately metabolized compounds.

properties.[3] A decision tree model obtained by this method would enable us to explore the properties and discover lead compounds of better quality. In the previous study, we applied our method to analyze the SAR relationship of CYP substrates.[3] The decision tree model that we obtained was able to discriminate, with reasonable accuracy (the misclassification rate was ~12%), between the substrates for six CYP isoforms based on their molecular descriptors. In addition, the large-scale visualization of the classification results by extended HeiankyoView was effective in understanding the SAR of the CYP substrates. While the effectiveness of multi-objective classification and visualization in SAR analysis was demonstrated, the model obtained for the classification of CYP isoforms was not fully examined. The external validation test carried out in the previous study only examined 16 drugs, corresponding to one-tenth of the data set of Bonnabry et al.[6]

In this context, the present text-mining analysis was performed to comprehensively collect CYP substrates in a bias-free manner. As a result, over 1000 compounds that interact with CYP isoforms were found in the PubMed domain. The text-mining system we developed has the unique feature that the dictionary- and context-based approaches were combined to effectively identify chemical names.[7] The recall of dictionary-based name identification is generally low due to the presence of too many chemical names or their variants. In addition, partial matches often occur in simple text matching procedures, lowering the precision of name identification. Therefore, we implemented rules in our system for the detection of chemical names from contexts that were based on verbs or verbal nouns relating to the action of chemicals as well as specific terms relating to quantity, quality, and usage of chemicals. Implementation of the context-based method can avoid partial matches or mismatches of chemical names and identify names which are not listed in a dictionary. Another feature identifies the pattern of chemical−CYP interactions based on the kind of verbs/verbal nouns and sequence of keywords (chemicals name, CYP3A4 name, and verbs) in a clause. In spite of its simple pattern matching, the system achieved 87.4% recall and 92.3% precision for the identification of chemical names and 85.2% recall and 92.0% precision for the extraction of chemical−CYP3A4 interactions.[7]

In a preliminary study of text-mining for other CYP isoforms than CYP3A4, we found a typical false-positive error which occurs when plural CYP names appear in one sentence. In the sentence "mephenytoin (MEPH), dextromethorphan, diclofenac, caffeine,

dx.doi.org/10.1021/ci100334z |J. Chem. Inf. Model. 2011, 51, 378–385

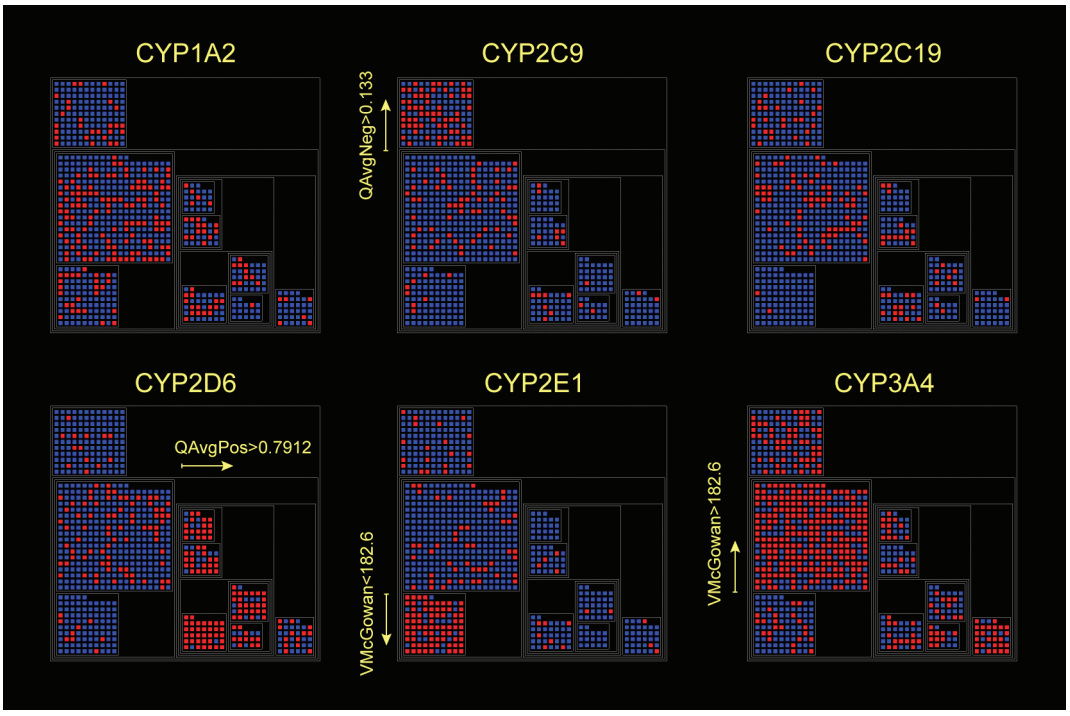

Figure 3. The HeiankyoView representation of CYP substrates classified by the multi-objective decision tree model. Nested rectangular borders indicate a hierarchical structure of groups according to the decision tree model. Red and blue icons represent CYP substrates and nonsubstrates, respectively. Note that compounds are positioned at the same position in each graph for all six isoforms.

and methadone (MET) were successfully applied as test substrates for CYP2C19, CYP2D6*1, CYP2C9*1, CYP1A2, and CYP3A4, respectively." (PubMed ID: 12900870), our old system[7] classified all combinations of five drugs and five isoforms as "substrate". In the previous analysis targeted to CYP3A4 alone,[7] the error was not so influential in total estimation. For other CYP isoforms, however, the error was relatively critical to the precision of information extraction. It would reflect that the probability of error is naturally higher due to much fewer substrates than CYP3A4. Therefore, we added a rule to exclude from the analysis the sentences in which plural chemical names and plural CYP names appear. As a result, the precision of information extraction remained high enough, although the recall was reduced.

While there are still some false-positive errors in the automatically collected information, it does not appear to be a major problem in the SAR analysis. As shown in Figures 2 and 3, the information on the metabolism of the compounds by the CYP isoforms that we collected correlates well with prediction results based on their chemical structures. It should be remembered that the structure−metabolism relationship model was created using the Bonnabry et al. data set,[3] which is only about one-fourth the size of ours. Thus, the good correlation in the data suggests that both the SAR model developed previously and the information collected here are reasonably reliable.

From our analysis of the Bonnabry et al. data set, we previously demonstrated that CYP2E1 and CYP3A4 metabolize small and large molecular compounds, respectively, and CYP2C9 and CYP2D6 metabolizes anionic and cationic compounds, respectively.[3] Clearly, such trends should still hold true in the present study. From the HeiankyoView image of the current data for these isoforms (Figure 3), we find only a slight difference for CYP2C9 substrates, where the proportion of anionic substrates of CYP2C9 was not as high as was been indicated by the previous SAR model or by the Bonnabry et al. data set. The 70 of the 145 CYP2C9 substrates (Supporting Information) that were not classified as anionic at neutral pH include alosetron, celecoxib, coumarin, fluoxetine, ketamine, nicardipine, and propofol. While, because of the possibility of false-negatives as mentioned above we need to be cautious, it would be interesting to examine how these neutral compounds might interact with CYP2C9.

Many investigators have shown that CYP2C9 exhibits selectivity for anionic compounds,[5,11−16] and site-directed mutagenesis[17,18] and X-ray structure analysis[19] have demonstrated the important role of Arg108 residue in interaction with anionic substrates. For celecoxib and its analogues, it has been demonstrated that when the methyl group that is a site of oxidation is oriented toward the heme, the oxygen atom in the sulfonamide moiety is within interaction distance of the Arg108 residue.[20] In contrast, the X-ray structure of CYP2C9 with warfarin cocrystallized revealed that Arg108 was oriented away from the active site cavity and was not involved in binding with warfarin.[21] The warfarin lies in a predominantly hydrophobic pocket with no interaction between its negatively charged group and any of the amino residues.[21] Thus, the two different CYP2C9 crystal structures, one with flurbiprofen[19] and one with warfarin,[21] lead to conflicting conclusions on the involvement of Arg108 and indicate that CYP2C9 may have conformational flexibility. Molecular dynamics simulations of the crystal structures with different ligands would help us to better understand the interaction modes of compounds with CYP2C9.[22,23]

As discussed in the previous paper,[3] the use of a multi-objective decision tree is beneficial in obtaining a bird's-eye view of CYP metabolism, and a large-scale visualization of classification provides us with intuitive understanding of the overall SARs. Therefore, this approach would help us in having general guides for drug design to avoid pharmacokinetic or toxicological

problems associated with CYP metabolism. However, it should also be noted that such a multi-objective model sacrifices the accuracy of prediction. In particular, SARs for CYP1A2 and CYP2C19, which are less significant in statistics, were overwhelmed during the process of developing the single multi-objective decision tree model. If better accuracy of prediction is required, individual models for each CYP isoform should be constructed.

In conclusion, we have successfully reconfirmed the multi-objective decision tree model for substrates of CYP isoforms by using information comprehensively retrieved from PubMed by a text-mining procedure. In spite of the small number of false-positives that we detected in a partial survey, the automatically retrieved information was sufficiently reliable to develop the SAR model. The name lists of CYP-interacting compounds developed in the present study will be useful for CYP-related metabolism and interaction studies in both preclinical and clinical investigations.

■ ASSOCIATED CONTENT

Supporting Information. Tables 1−3 provides the lists of substrates, inhibitors, and inducers of 6 CYP isoforms. The lists were created based only on the text-mining procedure. The accuracy of all information has not been confirmed. This information is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: yama@pharm.kyoto-u.ac.jp, telephone: 81-75-753-4535, fax: 81-75-753-4575.

■ ACKNOWLEDGMENT

This research was supported in part by a grant-in-aid for Scientific Research (B) (21390008) from the Japan Society for the Promotion of Science.

■ REFERENCES

(1) Rendic, S.; Di Carlo, F. Human cytochrome P450 enzymes: a status report summarizing their reactions, substrates, inducers, and inhibitors. *Drug Metab. Rev.* **1997**, *29*, 413–580.

(2) Wilkinson, G. Drug metabolism and variability among patients in drug response. *N. Engl. J. Med.* **2005**, *352*, 2211–21.

(3) Yamashita, F.; Hara, H.; Ito, T.; Hashida, M. Novel hierarchical classification and visualization method for multi-objective optimization of drug properties: Application to structure-activity relationship analysis of cytochrome P450 metabolism. *J. Chem. Inf. Model.* **2008**, *48*, 364–369.

(4) Smith, D. A.; Ackland, M. J.; Jones, B. C. Properties of cytochrome P450 isoenzymes and their substrates 0.2. properties of cytochrome P450 substrates. *Drug Discovery Today* **1997**, *2*, 479–486.

(5) Lewis, D. On the recognition of mammalian microsomal cytochrome P450 substrates and their characteristics: towards the prediction of human p450 substrate specificity and metabolism. *Biochem. Pharmacol.* **2000**, *60*, 293–306.

(6) Bonnabry, P.; Sievering, J.; Leemann, T.; Dayer, P. Quantitative drug interactions prediction system (Q-DIPS) - A dynamic computer-based method to assist in the choice of clinically relevant in vivo studies. *Clin. Pharmacokinet.* **2001**, *40*, 631–640.

(7) Feng, C. L.; Yamashita, F.; Hashida, M. Automated extraction of information from the literature on chemical−CYP3A4 interactions. *J. Chem. Inf. Model.* **2007**, *47*, 2449–2455.

(8) Cunningham, H.; Maynard, D.; Bontcheva, K.; Tablan, V. In *GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications*, The 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02), Philadelphia, PA, July 6−12, 2002; Association for Computational Linguistics: Stroudsburg, PA, 2002.

(9) Itoh, T.; Takakura, H.; Sawada, A.; Koyamada, K. Hierarchical visualization of network intrusion detection data. *IEEE Comp. Graphics Applic.* **2006**, *26*, 40–47.

(10) Yamashita, F.; Itoh, T.; Hara, H.; Hashida, M. Visualization of large-scale aqueous solubility data using a novel hierarchical data visualization technique. *J. Chem. Inf. Model.* **2006**, *46*, 1054–1059.

(11) Miners, J. O.; Birkett, D. J. Cytochrome P4502C9: an enzyme of major importance in human drug metabolism. *Br. J. Clin. Pharmacol.* **1998**, *45*, 525–538.

(12) Mo, S. L.; Zhou, Z. W.; Yang, L. P.; Wei, M. Q.; Zhou, S. F. New Insights into the Structural Features and Functional Relevance of Human Cytochrome P450 2C9. Part I. *Curr. Drug Metab.* **2009**, *10*, 1075–1126.

(13) Rettie, A. E.; Jones, J. P. Clinical and toxicological relevance of CYP2C9: Drug-drug interactions and pharmacogenetics. *Annu. Rev. Pharmacol. Toxicol.* **2005**, *45*, 477–494.

(14) Tai, G.; Dickmann, L. J.; Matovic, N.; DeVoss, J. J.; Gillam, E. M. J.; Rettie, A. E. Re-engineering of CYP2C9 to probe acid-base substrate selectivity. *Drug Metab. Dispos.* **2008**, *36*, 1992–1997.

(15) Jones, B. C.; Hawksworth, G.; Horne, V. A.; Newlands, A.; Morsman, J.; Tute, M. S.; Smith, D. A. Putative active site template model for cytochrome P4502C9 (tolbutamide hydroxylase). *Drug Metab. Dispos.* **1996**, *24*, 260–266.

(16) Mancy, A.; Broto, P.; Dijols, S.; Dansette, P. M.; Mansuy, D. The substrate-binding site of human liver cytochrome-p450 2c9 - an approach using designed tienilic acid-derivatives and molecular modeling. *Biochemistry* **1995**, *34*, 10365–10375.

(17) Dickmann, L. J.; Locuson, C. W.; Jones, J. P.; Rettie, A. E. Differential roles of Arg97, Asp293, and Arg108 in enzyme stability and substrate specificity of CYP2C9. *Mol. Pharmacol.* **2004**, *65*, 842–850.

(18) Ridderstrom, M.; Masimirembwa, C.; Trump-Kallmeyer, S.; Ahlefelt, M.; Otter, C.; Anderson, T. B. Arginines 97 and 108 in CYP2C9 are important determinants of the catalytic function. *Biochem. Biophys. Res. Commun.* **2000**, *270*, 983–987.

(19) Wester, M. R.; Yano, J. K.; Schoch, G. A.; Yang, C.; Griffin, K. J.; Stout, C. D.; Johnson, E. F. The structure of human cytochrome P4502C9 complexed with flurbiprofen at 2.0-angstrom resolution *J. Biol. Chem.* **2004**, *279*, 35630–35637.

(20) Ahlstrom, M. M.; Ridderstrom, M.; Zamora, I. CYP2C9 structure-metabolism relationships: Substrates, inhibitors, and metabolites. *J. Med. Chem.* **2007**, *50*, 5382–5391.

(21) Williams, P. A.; Cosme, J.; Ward, A.; Angova, H. C.; Vinkovic, D. M.; Jhoti, H. Crystal structure of human cytochrome P4502C9 with bound warfarin. *Nature* **2003**, *424*, 464–468.

(22) Bikadi, Z.; Hazai, E. In silico description of differential enantioselectivity in methoxychlor O-demethylation by CYP2C enzymes. *Biochim. Biophys. Acta, Gen. Subj.* **2008**, *1780*, 1070–1079.

(23) Yao, Y.; Han, W. W.; Zhou, Y. H.; Li, Z. S.; Li, Q.; Chen, X. Y.; Zhong, D. F. The metabolism of CYP2C9 and CYP2C19 for gliclazide by homology modeling and docking study. *Eur. J. Med. Chem.* **2009**, *44*, 854–861.