Novel Hierarchical Classification and Visualization Method for Multiobjective Optimization of Drug Properties: Application to Structure−Activity Relationship Analysis of Cytochrome P450 Metabolism

Fumiyoshi Yamashita,*,[†] Hideto Hara,[†] Takayuki Ito,[‡] and Mitsuru Hashida[†]

Department of Drug Delivery Research, Graduate School of Pharmaceutical Sciences, Kyoto University,
46-29 Yoshidashimoadachi-cho, Sakyo-ku, Kyoto 606-8501, Japan, and Department of Information Sciences,
Faculty of Science, Ochanomizu University, 2-1-1 Otsuka, Bunkyo-ku, Tokyo 112-8610, Japan

Received July 21, 2007

In the lead optimization process, medicinal chemists must consider various chemical properties of active compounds, including ADME/Tox properties, and find the best compromise among these. This study presents a novel data mining method for multiobjective optimization of chemical properties, which consists of the hierarchical classification and visualization of multidimensional data. A hierarchical classification tree model is generated by an extension of recursive partitioning that utilizes averaged information gains for multiple objective variables as a quality-of-split criterion. All the hierarchically structured data objects are represented using a large-scale data visualization technique. The technique is an extension of HeiankyoView, which displays data objects as colored icons and group nodes as rectangular borders. Each icon is divided into subregions with different colors, so that it can present multidimensional data according to brightness of the colors. The proposed method was applied to the structure−activity relationship analysis for cytochrome P450 (CYP) substrates. The substrate specificity of six CYP isoforms was successfully delineated: e.g., CYP2C9 substrates are anionic compounds, while CYP2D6 substrates are cationic; and CYP2E1 substrates are smaller compounds, while CYP3A4 substrates are larger compounds.

INTRODUCTION

Lead optimization is the critical process that determines the chances of the eventual success of any drugs. During this process, medicinal chemists attempt to improve primary lead compounds in terms of pharmacological activity, pharmacokinetics, and safety. However, lead optimization is extremely difficult, since trade-offs between chemical properties often occur. Considering that the diversity of a group of compounds derived from a lead is not necessarily high, selecting a high-quality lead is a key issue in determining the success of drug discovery and development. In recent years, in vitro screening of ADME/Tox properties that are the major reason for the attrition of clinical development has been performed in a routine drug discovery setting.[1−3] A problem which has arisen in these situations is how to manage large-scale data produced by various screening procedures. There is a great need to develop data mining techniques that can extract hidden trends from data sets and identify promising leads or candidates.

Quantitative structure−activity relationship (QSAR) models have often been applied to find relationships between molecular properties and biological activity. These models suggest a direction of exploratory chemical synthesis to improve the chemical properties of the lead compound. However, to predict a particular chemical property is not sufficient in developing drug candidates in the lead optimization process. Medicinal chemists have to consider multiple properties relating to druglikeness and find the best compromise among these properties. An inconsistent collection of individual models will be unable to provide the right direction for drug design, even if each model has a reasonable accuracy of prediction. There is a need to develop a method that enables us to systematically examine how the transformation of chemical structure influences all properties to be considered.

This article presents a novel data mining method for multiobjective SAR analysis, which consists of two basic techniques: one is an extension of the recursive partitioning method to fit multiobjective analysis, and the other is a visualization technique to displaying all data objects that have been classified hierarchically by recursive partitioning analysis. The recursive partitioning method has been used as one of the SAR modeling approaches.[4−7] There are several advantages associated with the recursive partitioning model compared with statistical models (multiple linear regression, partial least-squares, etc.) or nonlinear pattern recognition models (neural network, support vector machine, etc.). The recursive partitioning model is in a decision tree structure that is easily understandable, in addition to nonparametric and nonlinear characteristics. In the present study, we propose a quality-of-split criterion enabling simultaneous optimization of multiple objective variables in recursive partitioning analysis. Moreover, we present a novel data visualization technique to display hierarchically structured data in an intuitively understandable manner. We have previously reported a novel data visualization technique, named "He-

* Corresponding author phone: 81-75-753-4535; fax: 81-75-753-4575; e-mail: yama@pharm.kyoto-u.ac.jp.
[†] Kyoto University.
[‡] Ochanomizu University.

10.1021/ci700262y CCC: $40.75 © 2008 American Chemical Society
Published on Web 01/23/2008

iankyoView", which displays hierarchical data as color icons surrounded by nested rectangular frames.[8,9] This novel data visualization technique was effective in visually extracting hidden trends in a data set involving solubility data for 908 compounds.[9]

The applicability of proposed multiobjective SAR analysis was investigated using metabolic stability data for clinically relevant drugs against cytochrome P450 (CYP) isoforms. CYP is a superfamily of heme-containing mixed function oxygenases that catalyzes the regio- and stereoselective oxidation of a wide variety of drugs.[10] Such broad substrate specificity of CYP isoforms often leads to undesirable drug–drug interactions.[11,12] Competitive or noncompetitive inhibition of CYP by coadministered drugs retards the clearance of drugs from the body, leading to an unexpected rise in their blood concentrations. On the other hand, induced expression of CYP reduces or shortens the duration of pharmacological activity of the drugs by accelerating their clearance. Prediction of drug–drug interactions associated with CYP is an important issue in drug discovery and development as well as in clinical applications. Bonnabry et al.[13] compiled information on CYP substrates, inhibitors, and inducers and developed a database to make both qualitative deductions and quantitative predictions of drug–drug interactions. In the present study, we aimed to characterize the structure–activity relationship of drug metabolism mediated by 6 different CYP isoforms based on the data set of Bonnabry et al. Our present informatics-driven analysis supports interpretations in earlier reviews of the characteristics of CYP substrates.[14,15]

METHODS

Data Set. A set of metabolic stability data for 161 clinically relevant drugs involving 6 CYP isoforms (1A2, 2C9, 2C19, 2D6, 2E1, and 3A4) was taken from the article of Bonnabry et al.[13] Bonnabry et al. collected information on CYP metabolism, inhibition, and induction by in-house measurement or literature search and semiquantified their importance. Only a CYP metabolism data set was subjected to the present analysis. For each drug, a profile of its metabolic stability was expressed as a categorical vector with 6 elements corresponding to the CYP isoforms, where each element had three categories (none, moderate, and strong). Molecular descriptors of the drugs were calculated using ADMET Predictor Ver. 2.0.1 (Simulations Plus, Inc. Lancaster), which include constitutional descriptors, topological and electrotopological descriptors, and descriptors relating to hydrophobicity, electronic properties, hydrogen bonding, and molecular ionization.

Multiobjective Recursive Partitioning Analysis. Recursive partitioning is an exploratory data mining technique, which successively splits a data set into increasingly homogeneous subsets. In the present study, an extension of recursive partitioning techniques to deal with multiobjective problems was developed. As in the case of conventional techniques, our method looks at all possible splits for all variables included in the analysis and ranks each splitting rule in order on the basis of a quality-of-split criterion. We defined the following quality-of-split criterion

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

$$\mathrm{IG}_i(S) = \mathrm{IE}_i(S) - \left(\frac{N_{S1}}{N_S}\mathrm{IE}_i(S_1) + \frac{N_{S2}}{N_S}\mathrm{IE}_i(S_2)\right) \tag{2}$$

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

where N_j^i is the number of elements of which the ith 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 the 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, the quality-of-split criterion $O(S)$ was defined as the sum of the information gains for each objective attribute, in order to obtain the best compromise of multiple objective variables. After all possible binary splits for all explanatory variables were subjected to calculation of their $O(S)$ values, the rule that gave the maximum $O(S)$ was regarded as the best one for splitting group S. By performing the procedure recursively, a binary classification tree was developed.

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. An entire data set was divided into 10 roughly equal subsets. The analysis took the first 9 subsets of the data as a training data set and obtained the best decision tree model which predicted the remaining 1/10 of the data most accurately. The same process continued until each subset of the data was used as a test sample. The number of terminal groups giving the minimum misclassification rate was regarded as optimal.

Visualization of Hierarchically Structured Data. Hierarchical data obtained through recursive partitioning analysis were displayed using a data visualization technique HeiankyoView.[8,9] The technique represents leaf nodes of hierarchical data as square icons and nonleaf nodes as nested rectangular frames. Figure 1A shows a processing flow of the display layout of the nodes in HeiankyoView. 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. It has been described elsewhere how to solve the rectangle packing problem of finding the optimal display layout of leaf and nonleaf nodes.[8,9] 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 is to be minimized; and (3) the aspect ratio of the rectangular region enclosing the placed rectangles is to be optimized.

In the present study, HeiankyoView was extended to visualize multiple attributes of the data simultaneously (Figure 1B). The leaf-node icon was divided into subregions according to the number of attributes, each of which has

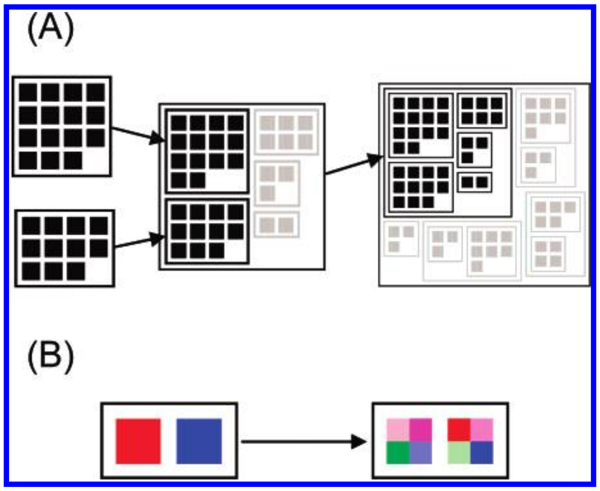


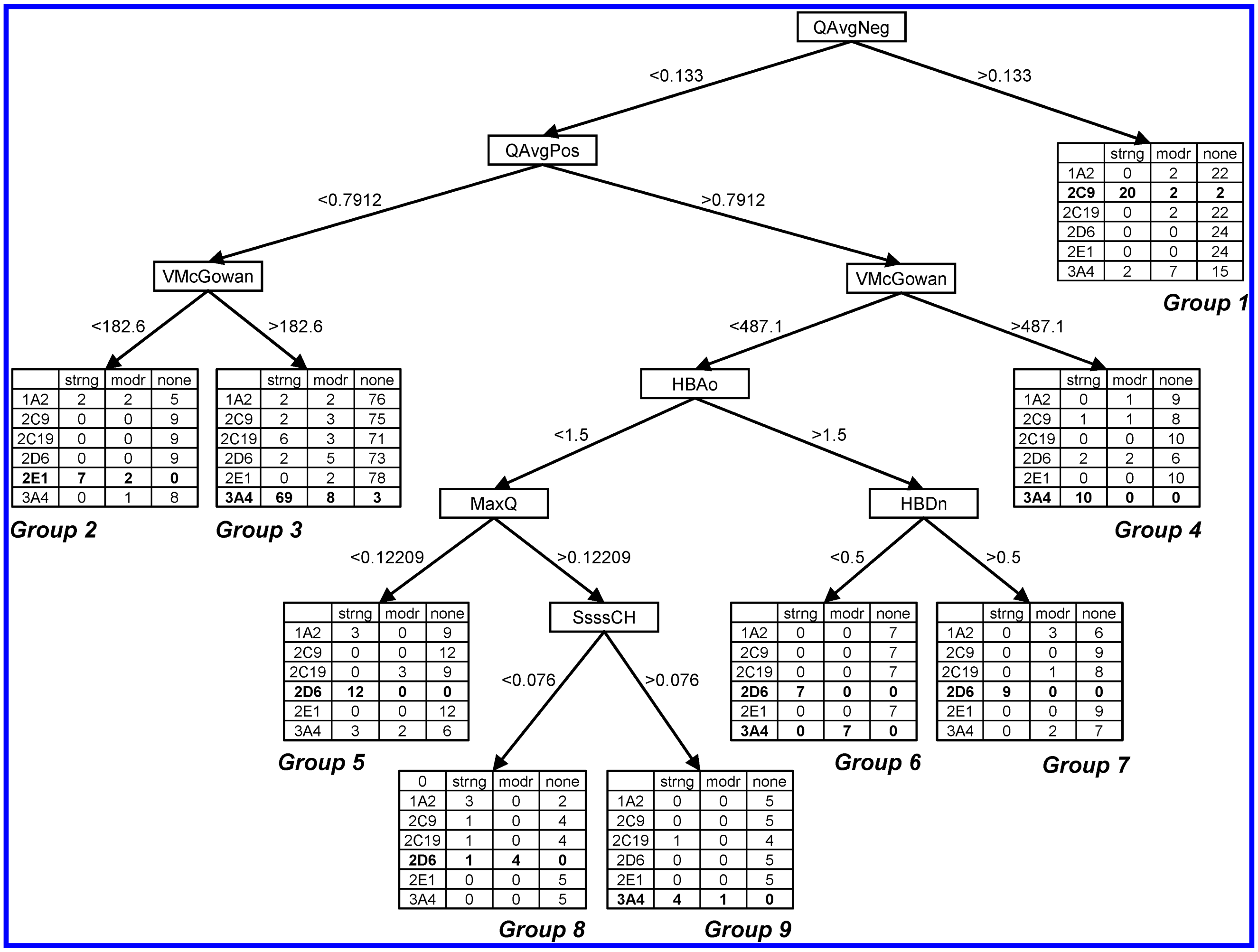
Figure 1. Representation of hierarchically structured data by HeiankyoView. Panel A indicates a processing flow of display layout of nodes in HeiankyoView, while panel B indicates an extension for visualization of multiple attributes of leaf-nodes.

individual colors corresponding to the attributes. The attribute measures were expressed by the brightness of the colors.

RESULTS

A set of metabolic stability data for 161 clinically relevant drugs against 6 CYP isoforms (1A2, 2C9, 2C19, 2D6, 2E1, and 3A4) was analyzed by the multiobjective recursive partitioning analysis developed in this study. First of all, we estimated the optimal number of terminal groups for the decision tree by the cross-validation procedure. The data set was arbitrarily divided into 10 subsets, each of which included 16 or 17 data. When a "leave-some-out" prediction was performed, the misclassification rate was minimal (12.6 ± 4.3%) at a terminal group number of 9 (Table 1).

Figure 2 shows a classification tree model for CYP-mediated drug metabolism. The model consists of 8 splitting rules and 9 terminal groups, with a resubstitution misclassification rate of 9.63% (Table 1). Figure 2 also presents the distribution of CYP metabolism levels in each terminal group of the decision tree. To obtain intuitive understanding of the trends of the tree structure, a visual image of the hierarchically structured data was presented using a novel data visualization technique (Figure 3). The 6-colored square icons indicate the compounds studied, where the color and its brightness represent metabolic susceptibility toward each CYP isoform. The trends found were as follows: (1) CYP2C9 and CYP2E1 substrates mostly belong to Groups 1 and 2, respectively; (2) CYP2D6 substrates belong to Groups 5−8; and (3) CYP3A4 substrates are detected in almost all groups, while CYP3A4 substrates belonging to Groups 3 and 4 are highly susceptible to the enzyme.



	strng	modr	none
1A2	0	2	22
2C9	**20**	**2**	**2**
2C19	0	2	22
2D6	0	0	24
2E1	0	0	24
3A4	2	7	15

Group 1

	strng	modr	none
1A2	2	2	5
2C9	0	0	9
2C19	0	0	9
2D6	0	0	9
2E1	**7**	**2**	**0**
3A4	0	1	8

Group 2

	strng	modr	none
1A2	2	2	76
2C9	2	3	75
2C19	6	3	71
2D6	2	5	73
2E1	0	2	78
3A4	**69**	**8**	**3**

Group 3

	strng	modr	none
1A2	0	1	9
2C9	1	1	8
2C19	0	0	10
2D6	2	2	6
2E1	0	0	10
3A4	**10**	**0**	**0**

Group 4

	strng	modr	none
1A2	3	0	9
2C9	0	0	12
2C19	0	3	9
2D6	**12**	**0**	**0**
2E1	0	0	12
3A4	3	2	6

Group 5

	strng	modr	none
1A2	0	0	7
2C9	0	0	7
2C19	0	0	7
2D6	**7**	**0**	**0**
2E1	0	0	7
3A4	**0**	**7**	**0**

Group 6

	strng	modr	none
1A2	0	3	6
2C9	0	0	9
2C19	0	1	8
2D6	**9**	**0**	**0**
2E1	0	0	9
3A4	0	2	7

Group 7

0	strng	modr	none
1A2	3	0	2
2C9	1	0	4
2C19	1	0	4
2D6	**1**	**4**	**0**
2E1	0	0	5
3A4	0	0	5

Group 8

	strng	modr	none
1A2	0	0	5
2C9	0	0	5
2C19	1	0	4
2D6	0	0	5
2E1	0	0	5
3A4	**4**	**1**	**0**

Group 9

Figure 2. Decision tree model for classifying CYP substrates. The decision tree was constructed by using the multiobjective recursive partitioning method developed in the present study. Each value indicates the count of compounds belonging to each category. Abbreviations: strng, strong; modr, moderate.

Table 1. Accuracy of Classification Tree Models for CYP-Mediated Drug Metabolism

	misclassification rate	
terminal groups[a]	leave-some-out cross-validation[b]	resubstitution
1	0.2017 ± 0.0467	0.2019
6	0.1357 ± 0.0386	0.1149
7	0.1303 ± 0.0426	0.1108
8	0.1292 ± 0.0425	0.1014
9	0.1261 ± 0.0431	0.0963
10	0.1293 ± 0.0443	0.0963
11	0.1314 ± 0.0500	0.0911
12	0.1335 ± 0.0484	0.0880

[a] Models with terminal groups of 2−5, 13, and larger were omitted. [b] The values indicate the average ± SD of a 10-fold leave-some-out cross-validation.

The QAvgNeg descriptor is the population average across all ionized species of the net formal negative charge calculated at pH 7.4. Considering that most of the CYP2C9 substrates belong to Group 1, it seems that CYP2C9 preferentially metabolizes anionizable compounds. On the other hand, many CYP2D6 substrates belong to Groups 5−8, in which the QAvgPos, i.e., the population average across all ionized species of the net formal positive charge calculated at pH 7.4, is greater than 0.7912. In contrast to CYP2C9 substrates, many CYP2D6 substrates appear to be cationic compounds.

CYP2E1 substrates belong to Group 2, for which the splitting rule is a McGowan molecular volume (VMcGowan) of less than 182.6. This suggests that CYP2E1 substrates are smaller compounds. In contrast, CYP3A4 substrates appear to be larger compounds, taking into account the fact that many of them belong to Groups 3 and 4. In fact, when the box-whiskers plot was applied to all drugs (Figure 4), there was a positive correlation between CYP3A4 susceptibility and the McGowan molecular volume.

DISCUSSION

In the present study, we proposed a novel recursive partitioning method for solving multiobjective optimization problems. The method presents a decision tree model consisting of simple splitting rules, so that patterns and trends in data structure can be easily identified. In particular, it becomes more powerful when combined with a large-scale data visualization technique. We extended HeiankyoView,

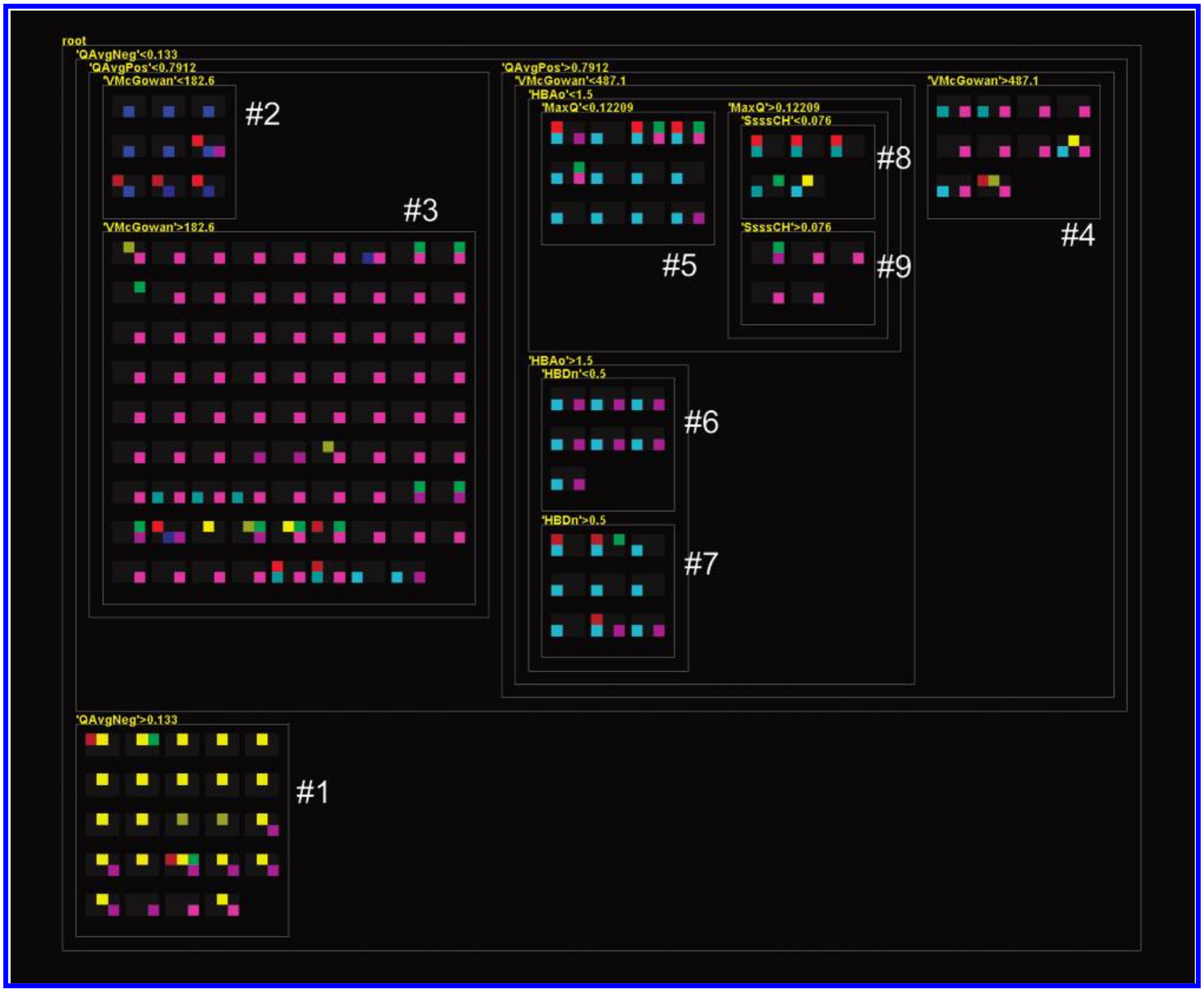


Figure 3. Extended HeiankyoView image of hierarchically structured data involving CYP-mediated drug metabolism. Each 6-colored rectangular icon indicates compounds where the brightness of each color represents the metabolic susceptibility toward each CYP inform: CYP1A2 (red), CYP2C9 (yellow), CYP2C19 (green), CYP2D6 (cyan), CYP2E1 (blue), and CYP3A4 (magenta). Rectangular borders represent hierarchically organized group structures based on the splitting rules of the decision tree. Each number indicates the group ID in the classification tree model (Figure 1).

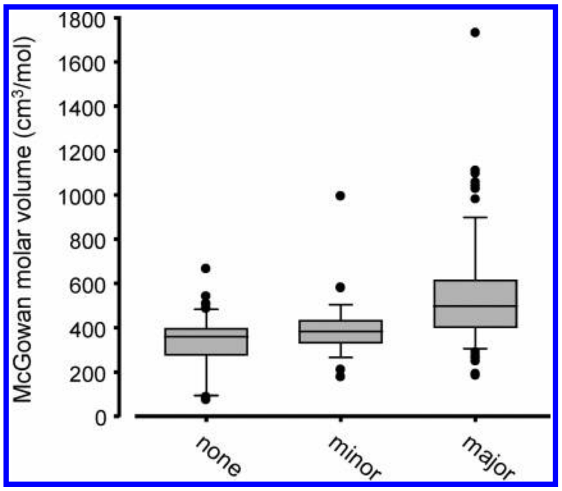


Figure 4. Box-whiskers plot of CYP3A4 susceptibility and McGowan molar volume. Boundaries of the box indicate the 25th and 75th percentiles, and a line within the box marks the median. Whiskers above and below the box indicate the 90th and 10th percentiles. Outlying points are also plotted.

which is a data visualization technique for hierarchically structured data,[8,9] in order to present multiple attributes simultaneously.

The technique was applied to structure−activity relationship analysis of CYP-mediated drug metabolism. As a result, we successfully constructed a decision tree model for classifying compounds of varying metabolic susceptibility to six different CYP isoforms. The decision tree model exhibited a reasonable accuracy of prediction in cross-validation tests, with a misclassification rate of approximately 12%. In an external validation test where one-tenth of the original data set was removed for testing, the misclassification rate for prediction of the external data was as low as 15.7% (data not shown). The present analysis provided us with structure−activity relationships 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. Interestingly, these findings are essentially the same as those reported by Smith et al.[14] and Lewis.[15] This implies that the present informatics-driven approach successfully produces a summary of information on CYP substrates without any preconceptions.

It is beyond doubt that data visualization by an extended HeiankyoView played an important role in acquiring such information on the characteristics of CYP substrates. The visual image (Figure 3) gave us a guide as to what to look for. The usefulness of HeiankyoView in large-scale data visualization has been demonstrated in various fields, such as detection of computer network intrusions[8] and QSAR analysis.[9] The technique represents all large-scale data in one display space without any focus-and-context techniques. The technique also ensures comfortable image viewing, by representing data objects as equishaped icons and involving squarish images of subspaces with a low aspect ratio configuration. Therefore, HeiankyoView helps us not only to overview large-scale hierarchical data but also to discover interesting local characteristics. In the present study, HeiankyoView was extended to represent multiple attributes using different colors. It greatly helps us to acquire information efficiently and effectively from a limited display space. In user experiments with a limited number of examinees, ∼20 attributes of ∼300 terminal nodes were cognizable at 1024 × 768 pixel resolution.

Simultaneous multiobjective optimization of chemical properties is required in the lead optimization process. Although drug candidates with good pharmacokinetic properties coupled to outstanding pharmacodynamic effects are ideal, these two objectives are often mutually incompatible. In this case, medicinal chemists must explore the best compromise among the properties required or identify new leads of better quality. Early ADME/Tox profiling of compounds provides a lot of information, but it is still difficult to make the right decision during the development of the candidates. The modeling approach proposed in the present study considers multiple target properties of compounds, and hierarchically classifies the different profiles of compounds based on simple rules involving structural features. Therefore, the classification model can answer several questions relating to exploratory chemical synthesis: e.g., “what group of compounds possesses a better quality?” or “what are the structural features that could improve the overall quality of the compounds?”. Moreover, intuitive understanding of these questions can be obtained by the aid of an extended HeiankyoView. Considering that uncertainty in weighting objective attributes makes it difficult to get distinct solutions for multiobjective optimization, human judgment and decision process would be needed. Here, it should be remembered that large-scale data visualization techniques encourage the innate ability of human beings to recognize patterns. While the usefulness of the data mining and visualization method was exemplified by the SAR analysis of CYP substrates, further application studies need to be carried out to confirm the effectiveness and user-friendliness of this technique.

CONCLUSION

In the present study, we propose a novel data mining method for multiobjective optimization of chemical properties. The method consists of the hierarchical classification and visualization of multidimensional data and provides a comprehensive understanding of the trends underlying the data sets from a graphical image. Strictly speaking, multiobjective SAR analysis might deviate from natural-scientific concepts, since the mechanisms underlying the SAR of each target property are different in nature. However, simultaneous multiobjective optimization of chemical properties in exploratory studies of drugs is an imperative. From a practical point of view, there is a clear need for a method that will help medicinal chemists identify the best direction for an exploratory chemical synthesis. If this idea is extended, the data visualization technique may help a team of pharmaceutical researchers make a consensus decision based on multidimensional data by effectively encouraging their pattern recognition ability. We believe that the present method will be of great use in research involving drug discovery and development.

ACKNOWLEDGMENT

This research was supported in part by Grant-in-Aids for Scientific Research and for a Leading Project for Biosimulation from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

REFERENCES AND NOTES

(1) Kerns, E. H.; Di, L. Pharmaceutical profiling in drug discovery. *Drug Discovery Today* **2003**, *8*, 316−323.

(2) Li, A. P. Screening for human ADME/Tox drug properties in drug discovery. *Drug Discovery Today* **2001**, *6*, 357−366.

(3) Yu, H.; Adedoyin, A. ADME-Tox in drug discovery: integration of experimental and computational technologies. *Drug Discovery Today* **2003**, *8*, 852−861.

(4) Rusinko, A., III; Farmen, M. W.; Lambert, C. G.; Brown, P. L.; Young, S. S. Analysis of a large structure/biological activity data set using recursive partitioning. *J. Chem. Inf. Comput Sci.* **1999**, *39*, 1017−1026.

(5) Ekins, S.; Balakin, K. V.; Savchuk, N.; Ivanenkov, Y. Insights for human ether-a-go-go-related gene potassium channel inhibition using recursive partitioning and Kohonen and Sammon mapping techniques. *J. Med. Chem.* **2006**, *49*, 5059−5071.

(6) Xia, X.; Maliski, E.; Cheetham, J.; Poppe, L. Solubility prediction by recursive partitioning. *Pharm. Res.* **2003**, *20*, 1634−1640.

(7) Zmuidinavicius, D.; Didziapetris, R.; Japertas, P.; Avdeef, A.; Petrauskas, A. Classification structure-activity relations (C-SAR) in prediction of human intestinal absorption. *J. Pharm. Sci.* **2003**, *92*, 621−633.

(8) Itoh, T.; Takakura, H.; Sawada, A.; Koyamada, K. Hierarchical Visualization of Network Intrusion Detection Data in the IP Address Space. *IEEE Comput. Graph. Appl.* **2006**, *26*, 40−47.

(9) 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.

(10) Rendic, S. Summary of information on human CYP enzymes: human P450 metabolism data. *Drug Metab. Rev.* **2002**, *34*, 83−448.

(11) Lin, J. H.; Lu, A. Y. Inhibition and induction of cytochrome P450 and the clinical implications. *Clin. Pharmacokinet.* **1998**, *35*, 361−390.

(12) Lin, J. H. CYP induction-mediated drug interactions: in vitro assessment and clinical implications. *Pharm. Res.* **2006**, *23*, 1089−1116.

(13) 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.

(14) Smith, D. A.; Ackland, M. J.; Jones, B. C. Properties of cytochrome P450 isoenzymes and their substrates part2: properties of cytochrome P450 substrates. *Drug Discovery Today* **1997**, *2*, 479−486.

(15) Lewis, D. F. 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.

CI700262Y