Identification of Endocrine Disruptor Biodegradation by Integration of Structure–activity Relationship with Pathway Analysis

TADASHI KADOWAKI,[§]
CRAIG E. WHEELOCK, TETSUYA ADACHI,[†]
TAKU KUDO,[‡] SHINOBU OKAMOTO,
NOBUYA TANAKA,
KOICHIRO TONOMURA,
GOZOH TSUJIMOTO,[†]
HIROSHI MAMITSUKA, SUSUMU GOTO,
AND MINORU KANEHISA*

Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan, kanehisa@kuicr.kyoto-u.ac.jp

Received November 18, 2006. Accepted September 18, 2007.. Revised manuscript received September 11, 2007

We present a SAR method that can predict estrogen-like endocrine disrupting chemical (EDC) activity as well as key biodegradation steps for detoxification. This method is based on a recent graph-mining algorithm developed by Kudo et al., which generates a set of descriptors from all potent chemical fragments (including rings). This method is novel in that it achieves chemical diversity in the training data set by sampling another data set of larger diversity. The model achieved an 83% accuracy prediction rate, and identified 1291 EDC candidates from the KEGG database. From this set of candidate compounds, bisphenol A was chosen for assay validation and biodegradation pathway analysis. Results showed that bisphenol A exhibited estrogen-like activity and was degraded in three distinct reactions. The prediction model provided information on the mechanism of the ligand-target binding, such as key functional groups involved. We focused on the enzyme commission number, which is useful for analyses of biodegradation pathways. Results identified oxygenases, ether hydrolases, and carbon-halide lyases as being important in the biodegradation pathway. This combined approach provided new information regarding the biodegradation of EDCs, and can potentially be extended to applications with transcriptomic, proteomic, and metabolomic data to provide a quick screen of biological activity and biodegradation pathway(s).

Introduction

Concern has been raised regarding the effects of exposure to environmental chemicals that interact with the endocrine system (*1*). To date, particular attention has been paid to compounds that are capable of affecting estrogen functions. Potential endocrine disrupting chemicals (EDCs) include organochlorine compounds such as 1,1,1-trichloro-2,2-bis(*p*-chlorophenyl)ethane (DDT) and its metabolites, as well as polychlorinated biphenyls (PCBs) and dioxins. Exposure to EDCs can perturbate endocrine functions, resulting in abnormal physiological states that can lead to adverse effects including reproductive disruption, hormonal imbalance, and some cancers (*2*).

A number of taskforces are currently attempting to identify endocrine disrupters, including the Endocrine Disruptor Screening Program at the U.S. Environmental Protection Agency, and similar programs in Japan (Strategic Programs on Environmental Endocrine Disruptors; SPEED'98) and Europe (the Cluster of Research into Endocrine Disruption in Europe; CREDO). For example, Walker and co-workers have identified approximately 58000 chemicals that have been selected for assay validation (*3*).

The computational-aided prioritization of the screened chemicals by structure–activity relationship (SAR) is a challenging component of these programs. Although quantitative structure–activity relationship (QSAR) models have been developed for steroid hormone receptors, (e.g., estrogen (*4*), progesterone, and androgen (*5*)), these models cannot be utilized for large-scale screening because they are too specific for congeneric chemical structures. A wide range in chemical diversity in the training data set is required to predict EDCs from a large-scale virtual chemical library. For this reason, SAR and QSAR models of EDCs have been developed with training data sets that consist of a variety of chemical structures. These models are based on improved (Q)SARs, such as CoMFA (*6, 7*), GRID/GOLPE (*8*), multidimensional QSAR (*9*), kNN QSAR (*10*), MCASE (*11*), and Decision Forest (*12*). A large-scale study to predict estrogen receptor binding affinity was performed using a virtual library of 58000 potential EDCs, with results estimating that 80–83% of the chemicals were nonbinders (*13, 14*).

The environmental fate of potential EDCs is also important, with more persistent compounds potentially presenting greater risk. It is therefore important to develop methods to estimate biodegradation process for these compounds. For instance, QSAR/SAR and the prediction of biotransformation pathways have been previously applied to this problem with some success (*15*). A similar integrated approach for ADME/Tox research has been proposed (*16*), in which the biological activities of a drug and its metabolites, interaction between compounds and genes, and gene expression data are all combined to provide an overall model of the toxicity/biological activity of a compound and known or predicted metabolites. This type of approach provides a more complete picture of the toxicology and pathological profile of a compound. It also provides important mechanistic information that is useful for furthering our understanding of the biological processes behind any observed adverse effects.

Recently, an efficient algorithm, hereafter called the Kudo algorithm/classifier, for classifying given graphs (i.e., chemical compounds) was proposed in the artificial intelligence area (*17*). As opposed to existing SAR models which consist of only linear or tree fragments, the Kudo algorithm can include ring structures in its descriptors by searching through all possible input graph substructures/fragments as potential molecular descriptors. This feature is expected to be an advantage over other existing SAR models.

The goal of this study was to obtain new knowledge from the combination of SAR and pathway data. First, we developed a new SAR algorithm based on the Kudo algorithm,

* Corresponding author phone: +81-774-38-3270; fax: +81-774-38-3269 ; e-mail: kanehisa@kuicr.kyoto-u.ac.jp.

[†] Department of Genomic Drug Discovery Science, Graduate School of Pharmaceutical Sciences, Kyoto University, Kyoto 606-8501, Japan

[‡] Google Japan, Inc., Cerulean Tower 6F, 26-1 Sakuragaoka-cho, Shibuya, Tokyo, Japan, 150-8512.

[§] Present address: Eisai Co., Ltd. 5-1-3, Tokodai, Tsukuba, Ibaraki 300-2635, Japan

10.1021/es062751s CCC: $37.00
Published on Web 10/25/2007

and built a model based on the input chemical structures from large chemical space. Thus, our method should achieve chemical diversity in the training data set. Then, we performed an integrated study of SAR prediction and the metabolic pathway data set. By using the pathway data, chemicals and genes (enzymes) were linked, thereby establishing a relationship between an enzyme and its substrate or product. A metabolic pathway is a network that is expressed with connections of these chemical–gene (chemical–enzyme) relationships. Network analysis focused on the biodegradation process and provided the dominant Enzyme Commission (EC) numbers involved in the detoxification of EDCs. This approach would be generally useful for identifying enzyme groups involved in a specific type of reaction. Naturally, once these relationships are assigned, other high-throughput data, such as transcriptomic and proteomic data, could also be integrated, which would be useful for understanding the molecular mechanisms behind endocrine disruption.

Material and Methods

Data Sets. In this study, two different chemical data sets were used. The first was a small training data set, which contained chemical structures and their corresponding biological activities. The second was a test data set, which contained chemical structures only. The training data set was obtained from the Endocrine Disruptor Knowledge Base (EDKB, http://edkb.fda.gov/databasedoor.html). The data set contains six types of bioassays, of which the E-SCREEN assay was most suited to the purpose of this study (*18*). The data set contains 120 different chemicals, 59 of which are biologically active and 61 of which are inactive. The estrogen-like activities of the E-SCREEN assays were measured in units of log translated relative proliferative potency (log RPP).

The large-scale chemical data test set was obtained from the COMPOUND database, a component of the KEGG database (http://www.genome.jp) (*19*). The COMPOUND database contains approximately 12000 chemicals (accessed in August 2005), consisting of metabolites, as well as drugs and xenobiotic compounds. The database also contains pathway information, consisting of relationships between chemicals and enzymes, which enables a combined analysis of SAR and pathway data.

The EDKB database does not contain sufficient chemical diversity because of knowledge-based filtering in the selection of assay chemicals. For instance, a large number of endogenous chemicals, such as carbohydrates, fatty acids, etc. are not assayed for estrogen-like activity even though large-scale chemical libraries contain these obviously inactive chemicals. However, these inactive chemicals should be included in our data set for model generation to maximize chemical diversity. We therefore added 30 inactive chemicals randomly sampled from the COMPOUND database to the EDKB database (the training data set) and integrated 100 classifiers that were trained from a different data set to retrieve chemical diversity. A total of 3000 inactive chemicals were used for our model. The formulation of our model is similar to the Random Forest algorithm, which achieves more robust prediction than a model lacking randomness.

Kudo Classifier. Our model is shown in Figure 1 and has three layers. The first and second layers correspond to the Kudo algorithm, which is a graph-mining version of the AdaBoost algorithm. The first layer of the Kudo classifier is a so-called decision stump, which is trained by finding a substructure/fragment that most discriminates the input graphs and predicts the class (positive or negative) of an input graph by checking whether or not it contains the substructure. (In this process, atoms and connections are considered, but the bond strength is not.) The second layer combines the outputs of the decision stumps, considering the weights computed by the Kudo algorithm. (See ref *17* for more details. Briefly, decision stumps with low error rates gain increased weight, whereas those with high error rates lose weight.) By integrating the decision stumps, the performance of the Kudo classifier can be improved even if the classification ability of the decision stump is weak. The third layer is the integration of classifiers to involve a larger number of training chemicals. Similarly to AdaBoost, the parameters of this model were only numbers of decision stumps and Kudo classifiers, which were optimized by a cross validation test.

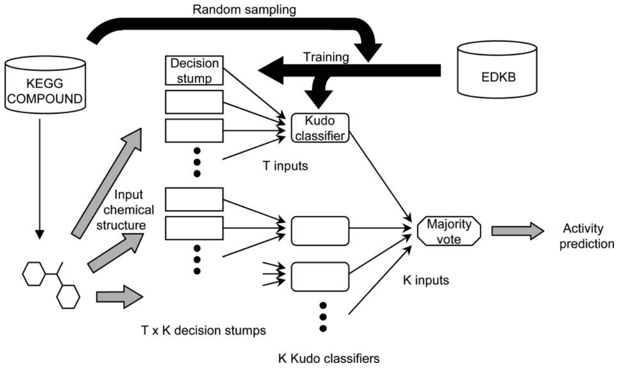

FIGURE 1. Overview of the prediction system. Each decision stump determines if a given chemical contains a specific substructure. A Kudo classifier integrates the weighted *T* inputs and provides an output consisting of a binary prediction, (e.g., active or inactive). These *K* outputs are used to generate the final output. Decision stumps and Kudo classifiers are trained with the EDKB and COMPOUND database.

E-SCREEN Assay. Estradiol 17-beta (E2), bisphenol A (BPA), 4-hydroxybenzoic acid (4-HBAc), 4-hydroxybenzaldehyde (4-HBAl), 4-hydroxyacetophenone (4-HAP) and DMSO were purchased from Sigma-Aldrich (St Louis, MO). MCF-7 cells, a breast cancer cell line, were obtained from the American type Culture Collection (Manassas, VA). Cells were cultured in Dulbecco's modified Eagle's medium (Invitrogen Corp., Carlsbad, CA) with 1.2 g/L sodium bicarbonate and 10% fetal bovine serum, 100 penicillin unit/mL; and 100 streptomycin μg/mL at 37 °C in a 5% CO_2 atm, with fluid renewal every 2 days.

Cells (1×10^4 cells/well) were plated on 96-well microtiter plates and incubated overnight at 37 °C. On the following day, MCF-7 cells were treated with E2, BPA, 4-HBAc, 4-HBAl, and 4-HAP (from 10 nM to 10 μM) in DMSO for 48 h. Control cells were treated only with DMSO. The final concentration of DMSO never exceeded 0.1%. For assay culture, bovine serum was treated with charcoal dextran (Sigma-Aldrich). The proliferation of MCF-7 cells treated with these chemicals was determined by WST-1 assay, which involved obtaining absorbance at a test wavelength of 450 nm in a colormetric assay using (4-[3-(4-iodophenyl)-2-(4-nitrophenyl)-2H-5-tetrazolio]-1,3-benzene disulfonate) (WST-1, Dojindo, Kumamoto, Japan) (*20*), with eight replicates for each chemical. Relative cell proliferation was calculated after proliferation in DMSO treatment was normalized to 1 (*21*). Statistical evaluation was performed using a Student's *t* test using the R environment (http://www.R-project.org). All data are shown as mean ± standard error.

Results

Model Construction and Validation. As discussed previously, the training data set does not contain sufficient chemical diversity. Therefore, the third layer in our model is designed to compensate for this problem. Each Kudo classifier is trained with a data set of 59 active chemicals and 60 inactive chemicals, with 50% of the inactive chemicals sampled from the EDKB and the rest sampled from the COMPOUND

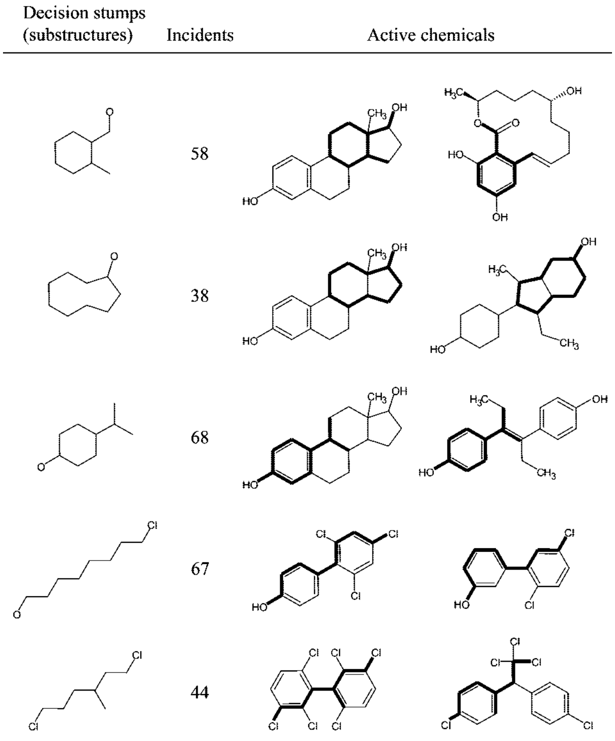

FIGURE 2. Decision stumps for active chemicals commonly used in Kudo classifiers. Matched substructures are drawn with thick lines. The frequencies of decision stumps in 100 Kudo classifiers are also displayed.

database. It is important to point out that the activities of the compounds in the COMPOUND database are unknown, but of the majority are assumed to be inactive, enabling the sampled chemicals to be used for the inactive data set. In the third layer we generated 100 independent Kudo classifiers by using a randomly generated different data set for each classifier. We assigned an activity of an input chemical by a majority vote of the 100 classifiers.

The performance of our model was measured using a 5-fold cross validation test. Two model parameters were optimized by sweeping the number of decision stumps for each Kudo classifier (*T*) from 5 to 50 at intervals of 5, and the number of Kudo classifiers for a majority vote (*K*) with 1, 10, 100. The result (Supporting InformationFigure S2a) shows that the accuracy was improved when the number of decision stumps and Kudo classifiers was increased. The highest performance obtained was 83% at a majority vote of 100 Kudo classifiers with 40 decision stumps. By using these conditions, the system predicted 1291 chemicals (Supporting InformationTable S2) that possess potential estrogen-like activity out of the 12109 chemicals in the COMPOUND database. Figure 2 shows the typical chemical substructures included in our prediction system. These were distinct fragment structures selected from commonly used decision stumps for active chemicals around 100 Kudo classifiers by intelligible structure. According to weights in classifiers, they are classified into two groups: active and inactive subgraphs.

Predicted Chemicals and Degradation Pathways. Predicted chemicals were mapped on to the KEGG PATHWAY database. Metabolic pathways and enzymes associated with the predicted chemicals can be collected. Twenty-eight metabolic pathways were mapped with predicted chemicals (Table 1). If chemicals with predicted estrogen-like activity are placed together in a chain of reactions, the prediction can be considered more accurate, and reaction chain boundaries of such active chemical clusters provide information on critical reactions of activation/synthesis and deactivation/degradation. Once mapping is performed for the KEGG pathways, active chemical clusters and their boundaries are simply identified. An example pathway is shown in Figure 3a, which displays the BPA degradation pathway. The model predicted that BPA is deactivated in three steps, in other words, the model reports the number of metabolic reaction steps required to deactivate BPA and its metabolites (i.e., remove EDC activity). The model also identifies which enzymes are required for the predicted deactivation metabolism, which in the case of BPA involved only oxidoreductases (EC 1.-.-.-). It is not possible to extract this information directly from the SAR and pathway database, and it could only be achieved through a combination of both approaches.

Statistical tests on the metabolic pathways with predicted estrogen-like activities reveal characteristics of the chemicals in pathways. The significant groups of EC numbers involved in degradation pathways were calculated. Enzymes whose substrates or products are predicted to be estrogen-like active compounds were considered. By taking into account the relationship between activity and reaction direction, the enzymes were categorized into four groups: activation, nonchange (to keep active), deactivation, and bidirection. This study focused on biodegradation processes of EDCs, because this information will be the most useful in understanding detoxificational deactivation pathways. Accordingly, we analyzed enzymes that catalyze deactivation reactions and their four previous steps.

TABLE 1. Predicted Pathways and Chemicals Mapped onto the KEGG Metabolic Pathways[a]

	pathway		chemical	
metabolic pathway category	active	total	active	total
1.1 carbohydrate metabolism	0	(17)	0	(602)
1.2 energy metabolism	0	(8)	0	(136)
1.3 lipid metabolism	5	(12)	131	(519)
1.4 nucleotide metabolism	0	(2)	0	(150)
1.5 amino acid metabolism	2	(16)	4	(675)
1.6 metabolism of other amino acids	0	(9)	0	(184)
1.7 glycan biosynthesis and metabolism	0	(18)	0	(151)
1.8 biosynthesis of polyketides and nonribosomal peptides	3	(9)	37	(301)
1.9 metabolism of cofactors and vitamins	2	(11)	11	(326)
1.10 biosynthesis of secondary metabolites	8	(16)	77	(566)
1.11 xenobiotics biodegradation and metabolism	8	(21)	50	(628)
total	28	(139)	310	(4238)

[a] Displayed metabolic pathways are those currently available in the KEGG database. The total number of pathways and chemicals in a category is expressed in parentheses.

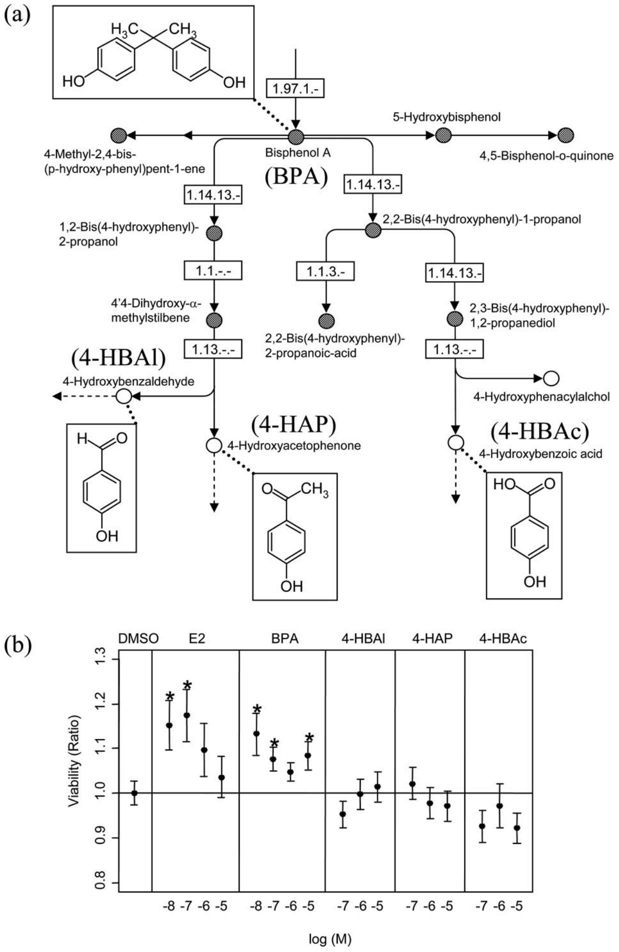

FIGURE 3. (a) BPA biodegradation pathway. Circles and boxes indicate compounds and enzymes, respectively. Numbers inside boxes are EC (Enzyme Commission) numbers. Arrows through boxes indicate reaction direction. Red filled circles are predicted as positive, and open circles are negative. The compounds experimentally validated in this study are displayed at the appropriate place in the metabolic pathway. (b) E-SCREEN (MCF-7 cell proliferation) assay. Cells are counted following 48 h treatment with estradiol 17-beta (E2) (0.01, 0.1, 1, 10 μM), bisphenol A (BPA) (0.01, 0.1, 1, 10 μM), 4-hydroxybenzaldehyde (4-HBAl) (0.1, 1, 10 μM), 4-hydroxyacetophenone (4-HAP) (0.1, 1, 10 μM), 4-hydroxybenzoic acid (4-HBAc) (0.1, 1, 10 μM). An asterisk indicates a significant change relative to DMSO control (p < 0.05).

A binomial test can determine significant EC numbers used within a group of degradation processes. An observed number in the binomial test is defined as the total observation of certain EC numbers in the degradation processes. The total number of reactions on the PATHWAY database and the average observation ratio are parameters of the test. Enzymes involved in biodegradation processes are summarized in Table 2. EC 1.-.-.- were the most common enzymes involved in metabolic biodegradation, especially, oxygenases, EC 1.13.-.- and 1.14.-.-. The statistical significance (p value) for these enzymes was less than 1%. Ether hydrolases and carbon-halide lyases were also determined to be significant (p < 0.01).

Predicted Chemicals and Functional Hierarchies. Predicted chemicals were also mapped on to the functional hierarchies of drugs and compounds in the KEGG BRITE database. The significance of a function defined in a hierarchical tree can also be derived through a binomial test. Predicted active chemicals were sorted into categories according to their functional classification (Supporting InformationTable S1). For example, the categories of “sterol lipids”, “terpenoids (including phytosterols)”, and “endocrine drugs” contained significant large numbers of predicted active chemicals.

Experimental Verification. The prediction of estrogen-like activity for chemicals in Figure 3a was confirmed experimentally (Figure 3b). The tested chemicals were BPA, 4-hydroxybenzaldehyde, 4-hydroxyacetophenone, and 4-hydroxybenoic acid, where BPA was predicted to be active and the others inactive. In addition, estradiol 17-beta and dimethyl sulfoxide (DMSO) were used for positive and negative controls, respectively. A proliferation ratio larger than 1 means that a chemical has activity and 1 means no activity. A ratio less than 1 means that a chemical possesses inhibition activity, which was not observed in these samples. Therefore, estrogen-like activity is discriminated by the ratio being greater than 1 or not. Based on this criterion, the positive estradiol 17-beta control and BPA were active, whereas the other three chemicals were inactive.

To further validate the compounds that were identified to have EDC activity; we performed a literature search on resveratrol, which is synthesized from cinnamic acid in three reactions (see map 00940 in Supporting Informationfigure S1). Consistent with our prediction, the activities of the two chemicals were already reported (*26, 27*), in which resveratrol is active and cinnamic acid is inactive in the ER binding assay.

Discussion

In this study, we have developed a new graph mining method and applied it to make a SAR model of EDCs. Cross validation analysis showed that our method achieved 83% prediction accuracy. We believe that this figure is improved over the results of a similar study (*11*), in which the prediction accuracy was 74% for the RPP data set, since most compounds of their data set (110/122) were used in our analysis. Another advantage of our method is its ability to search all possible substructures/fragments including ring structures (bicyclic as well as polycyclic), as opposed to previous studies that were unable to include ring structures (*11, 22*) In addition, the descriptors we identified can be more general than those of existing methods. For instance, we identified five- and six-membered rings, phenol-like structures, and other substructures, by which we can predict the organic functional group that binds to its receptor (Figure 2). These descriptors can be useful for understanding the activity of the pharmacophore and the binding mode between a chemical compound and its receptor. For example, two oxygen atoms in estradiol 17-beta were identified by two ring-structure descriptors as important for estrogen-like activity. Crystal-structure analyses of the estrogen receptor alpha complex with estradiol 17-beta showed that these two atoms are involved in hydrogen bond formation with residues in the receptor (PDB code 3erd) (*23*). Another important example is the phenol-like structure as a chemical descriptor. Diethylstilbestrol, which is the most potent EDC identified to date (*18*), has two phenol-like structures and was identified by our method as possessing estrogen-like activity. Namely, the chemical has also two oxygen atoms at a defined distance from each other, and it is most likely bound to the estrogen receptor alpha. This prediction was confirmed by crystal-structure analysis (PDB code 1ere) (*24*). The other two substructures in Figure 2 show that there is a distance constraint on the chlorines and oxygen in the polychlorinated compounds.

In the development of the new method, the chemical diversity of the training data set was limited. In other words, the data set was biased. This problem is quite common in

TABLE 2. Prediction of Enzymes Involved in Biodegradation Processes[a]

EC number	$n = 1$	$n = 2$	$n = 3$	$n = 4$	all	p-value	function
1.1	0	8	14	16	(570)	0.31	
1.2	0	4	4	5	(192)	0.28	
1.3	0	6	6	6	(219)	0.09	
1.11	1	1	1	1	(25)	0.30	
1.13	6	6	6	6	(121)	7.2E-3	monooxygenase
1.14	9	14	20	23	(439)	5.5E-6	dioxygenase
1.97	0	0	1	2	(20)	0.24	
2.1	0	1	1	1	(232)	0.96	
2.3	0	2	2	2	(292)	0.91	
2.6	1	1	1	1	(109)	0.78	
2.8	0	0	0	2	(73)	0.80	
3.1	0	1	3	3	(309)	0.86	
3.3	3	3	4	4	(17)	7.7E-5	Ether Hydrolases
3.5	1	1	1	1	(234)	0.96	
4.1	4	5	6	6	(202)	0.065	
4.5	1	1	2	3	(10)	8.1E-3	Carbon-Halide Lyases
5.3	0	0	2	2	(64)	0.22	
5.5	1	1	1	1	(24)	0.29	
6.2	2	2	2	2	(65)	0.23	
others	0	0	0	0	(2557)	1	
total	29	57	77	87	(5530)		

[a] Enzymes are taken from deactivating reactions (expressed as $n = 1$) and traced back reactions ($n = 2$ to 4). The total number of enzymes for each category is shown in parentheses. Statistical analyses were performed using a binominal test ($n = 3$).

many classification studies. Our method used (potentially) inactive compounds in a general chemical compound database like COMPOUND combined with the inactive compounds in a given database such as EDKB. This approach presented general substructure decision stumps in terms of the chemical space, and improved the predictive performance by comparing to the case where only the original EDKB data set was used (see Supporting InformationFigure S2b). In addition, a model trained with only the EDKB data set predicted that approximately one-third of the chemicals in the COMPOUND data set were active (data not shown), whereas our model predicted approximately one-tenth (1291/12109). The number of active chemicals in the test data set is unknown, but the former figure of ~33% is extremely large. The later figure is consistent with other studies using a different large-scale data set, in which active compounds were at most 17–20% of the total (*13, 14*).

In addition, combined analysis with SAR and the pathway database was performed. The analyses revealed the significance of predicted chemicals in specific functions, such as endocrine drugs steroids, and phytosterols (Table 2). These functions were strongly related to the endocrine system. Moreover, pathway analysis provides various information, for instance, synthesis and biodegradation of active chemicals are easily recognized from the visualization of active chemicals in the pathways (Figure 3a and Supporting InformationFigure S1). We focused especially on the biodegradation process of EDCs, which is important for the detoxification of EDCs in the environment. The BPA pathway analysis showed that three biodegradation reaction steps are required for detoxification. This result was validated experimentally, supporting the validity of our prediction method. Activities of two intermediate metabolites from BPA and detoxified chemicals were also confirmed by Kitamura et al. (*25*). Their study complements our prediction of BPA detoxification in its biodegradation pathway.

Finally, we performed comprehensive analysis in the PATHWAY database, by identifying dominant EC numbers in biodegradation pathways. EC 1.-.-.- are observed most frequently, especially oxygenases (EC 1.13.-.- and EC 1.14.-.-). Oxygenases include members of the well-known cytochrome P450 family, which is broadly used in bioremediative microorganizms (*28, 29*) In addition, aromatic-ring-hydroxylating dioxygenases have been found in microorganizms (*30, 31*). Ether hydrolases and carbon-halide lyases were also significant in our analysis. These enzymes catalyze ring-opening and dehalogenation reactions. These results demonstrate that our method provides reasonable predictions of active chemicals, and that the analysis can elucidate chemical-gene relationships and/or chemical-enzyme relationships. In a separate study, we surveyed all known enzyme-catalyzed reactions and extracted characteristic reaction patterns in biodegradation processes, which were then used to predict new biodegradtion pathways (*32*).

The predicted reaction steps of BPA biodegradation and resveratrol synthesis were validated experimentally or via literature search, respectively; however, it is not known how many predicted EDCs in our analysis are actually active. It is, therefore, necessary to confirm the prediction accuracy in a large chemical space. Screening projects on the initiative of EPA and other agencies have been producing biological assay data for over a decade. These data will be useful for confirmation and improvement of the prediction model. In addition, eventually results from these types of models will be useful for focusing screening efforts of these consortiums. As funds are limited to perform screening for EDC activity, it is logical to use computation methods to focus resources on those compounds that have the greatest potential to evidence EDC activity.

Our method can be applied to other types of potentially toxic chemicals, such as carcinogens. Expanding the applications of this model will be useful for validating the new method. However, some current limitations in the model still need to be addressed, specifically the model does not consider aromaticity, and does not quantify the activity. Model output is strictly binary, resulting in a designation of "active" or "inactive". In other words, it is a SAR model, not a QSAR model. Developing a QSAR approach for our model will be useful in identifying the contribution of substructures to estrogen-like activity. Our model also does not recognize the subtype-selectivity of estrogen receptors (*33*), which requires some extension of our algorithm for nonbinary (multiclass) prediction. Regarding the pathway database, current chemical reaction information does not sufficiently cover all degradation processes, so that use of reaction prediction will be important for applications. Further integrated analysis with other "omics" data, such as transcriptomic, proteomic, or metabolomic studies, can be performed by assigning this information to genes or chemicals in the pathway, and promises to increase our understanding of biological mechanisms.

Acknowledgments

This study was supported by the 21st Century COE program "Genome Science", a grant-in-aid for scientific research on the priority area from MEXT Japan, and the Japan Science and Technology Agency. C.E.W. was funded by a Japanese Society for the Promotion of Science (JSPS) postdoctoral fellowship. Computational resources were provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University.

Supporting Information Available

List of predicted chemicals, functional categories, pathway maps, and cross-validation results, Table S1, functional categories of predicted EDCs, Table S2, a list of predicted EDCs in the KEGG COMPOUND database, Figure S1, pathway maps with predicted EDCs (the pathway maps are provided separately), Figure S2, and prediction accuracy by cross-validation test. This information is available free of charge via the Internet at http://pubs.acs.org.

Literature Cited

(1) Colburn, T.; Clement, C. *Chemically induced alterations in sexual and functional development: the wildlife/human connection*; Princeton Scientific Publishing: Princeton, 1992.

(2) Newbold, R. R.; McLachlan, J. A. Transplacental hormonal carcinogenesis: diethylstilbestrol as an example. In *Cellular and Molecular Mechanisms of Hormonal Carcinogenesis: Environmental Influences*; Huff J;, Boyd J.; Barrett J. C., Eds.; Wiley-Liss: NY, 1996; pp 131–147.

(3) Walker, J. D.; Waller, C. W.; Kane, S. The Endocrine Disruption Priority Setting Database EDPSD): A Tool to Rapidly Sort and Prioritize Chemicals for Endocrine Disruption Screening and Testing. In *Handbook on Quantitative Structure Activity Relationships (QSARs) for Predicting Chemical Endocrine Disruption Potentials*; Walker, J. D., Ed.; SETAC press: Pensacola, FL, 2001.

(4) Waller, C. L.; Minor, D. L.; McKinney, J. D. Using three-dimensional quantitative structure-activity relationships to examine estrogen receptor binding affinities of polychlorinated hydroxybiphenyls. *Environ. Health Perspect.* **1995**, *103*, 702–707.

(5) Loughney, D. A.; Schwender, C. F. A comparison of progestin and androgen receptor binding using the CoMFA technique. *J. Comput. -Aided Mol. Des* **1992**, *6*, 569–581.

(6) Waller, C. L.; Oprea, T. I.; Chae, K.; Park, H. K.; Korach, K. S.; Laws, S. C.; Wiese, T. E.; Kelce, W. R.; Gray Jr, L. E. Ligand-based identification of environmental estrogens. *Chem. Res. Toxicol.* **1996**, *9*, 1240–1248.

(7) Shi, L. M.; Fang, H.; Tong, W.; Wu, J.; Perkins, R.; Blair, R. M.; Branham, W. S.; Dial, S. L.; Moland, C. L.; Sheehan, D. M. QSAR models using a large diverse set of estrogens. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 186–195.

(8) Sippl, W. Binding affinity prediction of novel estrogen receptor ligands using receptor-based 3-D QSAR methods. *Bioorg. Med. Chem.* **2002**, *10*, 3741–3755.

(9) Lill, M. A.; Dobler, M.; Vedani, A. In silico prediction of receptor-mediated environmental toxic phenomena-application to endocrine disruption. *SAR QSAR Environ Res* **2005**, *16*, 149–169.

(10) Asikainen, A. H.; Ruuskanen, J.; Tuppurainen, K. A. Performance of (consensus) kNN QSAR for predicting estrogenic activity in a large diverse set of organic compounds. *SAR QSAR Environ Res* **2004**, *15*, 19–32.

(11) Cunningham, A. R.; Cunningham, S. L.; Rosenkranz, H. S. Structure-activity approach to the identification of environmental estrogens: the MCASE approach. *SAR QSAR Environ Res* **2004**, *15*, 55–67.

(12) Tong, W.; Xie, Q.; Hong, H.; Shi, L.; Fang, H.; Perkins, R. Assessment of prediction confidence and domain extrapolation of two structure-activity relationship models for predicting estrogen receptor binding activity. *Environ. Health Perspect.* **2004**, *112*, 1249–1254.

(13) Shi, L.; Tong, W.; Fang, H.; Xie, Q.; Hong, H.; Perkins, R.; Wu, J.; Tu, M.; Blair, R. M.; Branham, W. S.; Waller, C.; Walker, J.; Sheehan, D. M. An integrated "4-phase" approach for setting endocrine disruption screening priorities--phase I and II predictions of estrogen receptor binding affinity. *SAR QSAR Environ. Res.* **2002**, *13*, 69–88.

(14) Hong, H.; Tong, W.; Fang, H.; Shi, L.; Xie, Q.; Wu, J.; Perkins, R.; Walker, J. D.; Branham, W.; Sheehan, D. M. Prediction of estrogen receptor binding for 58000 chemicals using an integrated system of a tree-based model with structural alerts. *Environ. Health Perspect.* **2002**, *110*, 29–36.

(15) Jaworska, J. S.; Boethling, R. S.; Howard, P. H. Recent developments in broadly applicable structure-biodegradability relationships. *Environ. Toxicol. Chem.* **2003**, *22*, 1710–1723.

(16) Ekins, S.; Andreyev, S.; Ryabov, A.; Kirillov, E.; Rakhmatulin, E. A.; Sorokina, S.; Bugrim, A.; Nikolskaya, T. A combined approach to drug metabolism and toxicity assessment. *Drug Metab. Dispos.* **2006**, *34*, 495–503.

(17) Kudo, Taku; Maeda, Eisaku; Matsumoto, Yuji An Application of Boosting to Graph Classification. In *Advances in Neural Information Processing Systems*; Saul, L. K., Weiss, Y., Bottou, L., Eds.; MIT Press: Cambridge, MA, 2005; Vol. 17, pp 729–736.

(18) Soto, A. M.; Sonnenschein, C.; Chung, K. L.; Fernandez, M. F.; Olea, N.; Serrano, F. O. The E-SCREEN assay as a tool to identify estrogens: an update on estrogenic environmental pollutants. *Environ. Health Perspect.* **1995**, *103 Suppl 7*, 113–122.

(19) Kanehisa, M.; Goto, S.; Hattori, M.; Aoki-Kinoshita, K. F.; Itoh, M.; Kawashima, S.; Katayama, T.; Araki, M.; Hirakawa, M. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* **2006**, *34*, D354–357.

(20) Ishiyama, M.; Tominaga, H.; Shiga, M.; Sasamoto, K.; Ohkura, Y.; Ueno, K. A combined assay of cell viability and in vitro cytotoxicity with a highly water-soluble tetrazolium salt, neutral red and crystal violet. *Biol. Pharm. Bull.* **1996**, *19*, 1518–1520.

(21) Adachi, T.; Okuno, Y.; Takenaka, S.; Matsuda, K.; Ohta, N.; Takashima, K.; Yamazaki, K.; Nishimura, D.; Miyatake, K.; Mori, C.; Tsujimoto, G. Comprehensive analysis of the effect of phytoestrogen, daidzein, on a testicular cell line, using mRNA and protein expression profile. *Food Chem. Toxicol.* **2005**, *43*, 529–535.

(22) Klopman, G. Artificial intelligence approach to structure-activity studies. Computer automated structure evaluation of biological activity of organic molecules. *J. Am. Chem. Soc.* **1984**, *106*, 7315–7321.

(23) Brzozowski, A. M.; Pike, A. C.; Dauter, Z.; Hubbard, R. E.; Bonn, T.; Engstrom, O.; Ohman, L.; Greene, G. L.; Gustafsson, J. A.; Carlquist, M. Molecular basis of agonism and antagonism in the oestrogen receptor. *Nature* **1997**, *389*, 753–758.

(24) Shiau, A. K.; Barstad, D.; Loria, P. M.; Cheng, L.; Kushner, P. J.; Agard, D. A.; Greene, G. L. The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. *Cell* **1998**, *95*, 927–937.

(25) Kitamura, S.; Suzuki, T.; Sanoh, S.; Kohta, R.; Jinno, N.; Sugihara, K.; Yoshihara, S.; Fujimoto, N.; Watanabe, H.; Ohta, S. Comparative study of the endocrine-disrupting activity of bisphenol A and 19 related compounds. *Toxicol. Sci.* **2005**, *84*, 249–259.

(26) Gehm, B. D.; McAndrews, J. M.; Chien, P. Y.; Jameson, J. L. Resveratrol, a polyphenolic compound found in grapes and wine, is an agonist for the estrogen receptor. *Proc. Natl. Acad. Sci. U. S. A.* **1997**, *94*, 14138–14143.

(27) Blair, R. M.; Fang, H.; Branham, W. S.; Hass, B. S.; Dial, S. L.; Moland, C. L.; Tong, W.; Shi, L.; Perkins, R.; Sheehan, D. M. The estrogen receptor relative binding affinities of 188 natural and xenochemicals: structural diversity of ligands. *Toxicol. Sci.* **2000**, *54*, 138–153.

(28) Wong, L. L. Cytochrome P450 monooxygenases. *Curr. Opin. Chem. Biol.* **1998**, *2*, 263–268.

(29) Gillam, E. M. J. Exploring the potential of xenobiotic-metabolising enzymes as biocatalysts: evolving designer catalysts from polyfunctional cytochrome P450 enzymes. *Clin. Exp. Pharmacol. Physiol.* **2005**, *32*, 147–152.

(30) Nojiri, H.; Maeda, K.; Sekiguchi, H.; Urata, M.; Shintani, M.; Yoshida, T.; Habe, H.; Omori, T. Organization and transcriptional characterization of catechol degradation genes involved in carbazole degradation by Pseudomonas resinovorans strain CA10. *Biosci., Biotechnol., Biochem.* **2002**, *66*, 897–901.

(31) Inoue, K.; Widada, J.; Nakai, S.; Endoh, T.; Urata, M.; Ashikawa, Y.; Shintani, M.; Saiki, Y.; Yoshida, T.; Habe, H.; Omori, T.; Nojiri, H. Divergent structures of carbazole degradative car operons

isolated from gram-negative bacteria. *Biosci., Biotechnol., Biochem.* **2004**, *68*, 1467–1480.

(32) Oh, M.; Yamada, T.; Hattori, M.; Goto, S.; Kanehisa, M. Systematic analysis of enzyme-catalyzed reaction patterns and prediction of microbial biodegradation pathways. *J. Chem. Inf. Model* **2007**, *47*, 1702–1712.

(33) Katzenellenbogen, J.; Muthyala, R.; Katzenellenbogen, B. S. The nature of the ligand-binding pocket of estrogen receptor alpha and beta: The search for subtype-selective ligands and implications for the prediction of estrogenic activity. *Pure Appl. Chem.* **2003**, *75*, 2397–2403.

ES062751S