J. Pestic. Sci., **31**(3), 273–281 (2006)

Review

(Special Topic)

Structure–activity relationships and pathway analysis of biological degradation processes

Tadashi KADOWAKI,* Craig E. WHEELOCK,†,* Masahiro HATTORI, Susumu GOTO and Minoru KANEHISA**

Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611–0011, Japan

(Received May 29, 2006)

(Q)SARs estimate biological activity; however these models are insufficient to fully understand and predict the ADME-Tox processes of small molecules in biological systems. By integrating (Q)SARs with biological databases, the predictive capability of these models can be significantly improved. However, the techniques and methods for integrated analysis have not yet been sufficiently developed for these combined systems. In this review, we discuss standard (Q)SAR methods and biological database construction as well as provide an example of how SAR and metabolic pathway analysis can be combined to examine the biological degradation processes of endocrine disrupting chemicals. © Pesticide Science Society of Japan

Keywords: structure–activity relationships, pathway analysis, biodegradation, 1,1,1-trichloro-2,2-bis(*p*-chlorophenyl)ethane, 1,1-dichloro-2,2-bis(*p*-chlorophenyl)ethylene.

Introduction

Quantitative (and qualitative) structure–activity relationships, (Q)SARs, have been extensively developed for use in biochemistry and toxicology. Numerous applications exist in other fields, such as the physical sciences and ecological health; however they are beyond the scope of this review. In molecular biology, knowledge about genomes, proteomes, chemical structure and relationships between these molecules have been collected and stored in public databases.[1)] These two methods are utilized in different ways to understand biological systems. For example, in the development of drugs, a biological database is used to identify a drug target using data acquired *via* high-throughput approaches (*i.e.* microarray data) and then (Q)SAR analysis is used to identify small molecules that bind the drug target. However binding to a target is insufficient for full evaluation of a compound's biological activity and eventual utility as a pesticide or pharmaceutical. It is also vital to understand the compound's adsorption, distribution, metabolism, excretion and toxicity (ADME-Tox) properties. This means that (Q)SAR analysis is insufficient to understand the pharmacokinetics of compounds in a biological system. One solution to this problem is to consider drug metabolites on an individual basis, integrating the enzymes that are responsible for the degradation/metabolism of a given compound as well as its metabolites. This analysis can reveal the dynamics of xenobiotics in biological systems. For example, Ekins *et al.* integrated QSAR coupled to a gene network dataset to examine drug metabolism and toxicity.[2)] However methods to incorporate integrated (Q)SAR analysis with that of a biological database are not sufficiently advanced for wide-scale application. In this review, we give a brief overview of (Q)SAR and a publicly available biological database. We then show an example of a combined chemical and genomics analysis by using SAR and pathway analysis of the biological degradation processes of endocrine disrupting chemicals (EDCs).

Structure–activity relationships

1. *Overview of structure–activity relationships*

(Q)SARs are mathematical relationships that link chemical structure to biological (ecological, toxicological or pharmacological) activity for a series of compounds. There are a multitude of methods available that can be used in (Q)SAR analy-

* These authors equally contributed to this work.
** To whom correspondence should be addressed.
E-mail: kanehisa@kuicr.kyoto-u.ac.jp
† Current address: Microbiology and Tumorbiology Center, Karolinska Institute, Novels väg 16, SE-171 77 Stockholm, Sweden

ses, including various regression, multivariate and pattern recognition techniques.[3–5] An increase in the need for bioinformatics methods to cope with omics-related data sets has transferred to the QSAR field, with a number of advances in chemoinformatic techniques.[5,6] (Q)SARs are used extensively in the design of novel biologically active compounds, either pharmaceutical or agrochemical,[7] and can subsequently be useful in industrial research settings. However, these methods are also increasingly being used by regulatory agencies to predict acute toxicity, mutagenicity, carcinogenicity, and other health effects.[8] In addition, numerous QSARs have been developed to predict the physical properties, fate, and effects of many chemicals.[9] As increasing numbers of standard QSAR methods are developed and validated to predict health effects, ecologic effects and environmental fate of chemicals, it is anticipated that more regulatory agencies will employ these methodologies as alternatives to chemical testing.[10,11] It can be particularly useful to conduct screening risk assessments to assist in prioritizing or ranking chemicals on the basis of potential hazard and exposure assessment parameters, consequently focusing research and/or cleanup efforts on specific chemicals of concern.

2. *Applications of structure–activity relationships to endocrine disrupting activity*

QSAR analysis methods have been particularly beneficial in EDC research and a number of different models and methodologies have been developed.[12] EDCs are compounds that directly modulate steroid hormone receptor pathways (estrogens, antiestrogens, androgens, antiandrogens) and aryl hydrocarbon receptor (AhR) agonists, including 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (TCDD) and related compounds[13] and are thought to pose a number of distinct health risks to both humans and wildlife.[13,14] Some pesticides (*i.e.* 1,1-dichloro-2,2-bis(*p*-chlorophenyl)ethylene) or DDE, the metabolite of 1,1,1-trichloro-2,2-bis(*p*-chlorophenyl)ethane or DDT (Fig. 1) also contain EDC activity.[15,16] A few of the numerous QSAR models available to describe EDC activity are discussed below to illustrate the breadth of available methodologies and applications.

Mekenya *et al.* analyzed a chemically diverse training set of 151 chemicals with measured human alpha estrogen receptor binding affinities (as well as mouse uterine, rat uterine, and MCF7 cells). The training set was analyzed using the COmmon REactivity PAttern (COREPA) approach, which attempts to derive a binding affinity model through a 3D pattern recognition analysis. This multidimensional approach is not dependent upon a specific conformer alignment or a specified pharmacophore. Using the results of this analysis, an exploratory system was developed for use in ranking relative mammalian estrogen receptor binding affinity potential for large chemical data sets.[17] The COREPA approach was also used by Serafimova *et al.* to examine androgen receptor binding affinity using the interatomic distances between nucleophilic sites and charges to classify active *versus* non-active chemicals. These stereoelectronic characteristics were then used to predict the biological activity of pesticide formulation ingredients in an attempt to identify chemicals with potential androgen receptor binding affinity.[18]

Lill and coworkers investigated the influence of induced fit of the androgen receptor binding pocket on free energies of ligand binding using a multidimensional QSAR receptor-modeling tool. On the basis of a novel alignment procedure using flexible docking, molecular dynamics simulations, and linear-interaction energy analysis, they examined the binding of 119 molecules from six different compound classes.[19] They employed multi-dimensional QSAR and compiled a pilot system that included the 3D models of three receptors known to mediate endocrine-disrupting effects (the aryl hydrocarbon receptor, the estrogen receptor and the androgen receptor) and validated them against 310 compounds.[20] Zhao *et al.* tested a data set of 146 EDCs belonging to a broad range of structural classes for their relative binding affinity to the androgen receptor. QSARs were determined using three methods: multiple linear regression, radical basis function neural network and support vector machine (SVM). Comparison of the results showed that the SVM method exhibited the best overall performance. Moreover, six linear QSAR models were constructed for some specific families based on their chemical structures.[21] The general structural requirements for chemical binding to androgen receptor was explored by Fang *et al.* who measured the binding activity of 202 natural, synthetic, and environmental chemicals.[16]

Jacobs *et al.* combined QSAR methods with crystal structures and homology modeling as well as molecular dynamics simulations to examine receptor-ligand endocrine disruption dynamics.[22] QSAR models were constructed using multivariate partial least squares techniques and specific descriptors. Multivariate PCA techniques were employed to predict those cell lines that would be best for performing risk assessment studies for EDCs. Results showed that in silico methods could be used effectively in endocrine disruptor risk assessment for prescreening potential endocrine disruptors, improving experimental *in vitro* screening assay design and facilitating more thorough data analyses. A novel computational technology derived from gene structure for screening, selecting, and designing pharmaceutical candidates was developed by Hendry *et al.* Endocrine pharmacophores, composites of the van der Waals surfaces and hydrogen bonding functional groups, were created by docking known active structures into specific sites in partially unwound DNA.[23]

A number of different groups have examined the utility of CoMFA methods for examining EDCs. Tong *et al.* compared the ability of three different QSAR methods to screen large chemical data sets for endocrine disrupting activity: CoMFA, classical QSAR, and Hologram QSAR (HQSAR). HQSAR attempts to correlate molecular structure with biological activity for a series of compounds using molecular holograms

constructed from counts of sub-structural molecular fragments. The statistical quality of the QSAR models constructed using CoMFA and HQSAR techniques were comparable and were generally better than those produced with the classical QSAR methods.[24] Yu *et al.* constructed CoMFA models based on biological data for a structurally diverse set of compounds spanning eight chemical families. Results indicated that flexible field-fit alignment offered improved models over atom-fit alignment as the structural diversity of the data set increases.[25] Hong *et al.* used CoMFA to examine androgen receptor-ligand binding affinities using a training data set of 146 structurally diverse chemicals with a 10^6-fold range in relative binding affinity.[26]

3. Future of structure–activity relationships

Research trends are moving towards the integration of larger data sets in the expectation of increasing our understanding of biological processes. The implementation of omics technologies has increased the need for informatics methods capable of analyzing these data sets and extracting meaningful information. The use of high-throughput screening of large compound collections and combinatorial libraries has increased the amount of data exponentially and subsequently the complexity of the data analysis. For example, the number of potential lead compounds available (either agrochemical or pharmaceutical) has greatly increased, but so has the cost of pursuing each of those leads in the development of a commercial product. These points highlight the need to collect efficacy, pharmacokinetic and metabolism data as early as possible in order to "kill" a compound early in the development process. These ADME-Tox properties can be estimated by a range of *in vivo* and *in vitro* methods, most of which have been adapted to a high-throughput format.[27,28] The task then becomes adapting the in silico methods to keep pace with the experimental data acquisition.

A number of studies are pushing the boundaries of QSAR through the development of methodologies for integrating high-throughput data with QSAR techniques in an effort towards achieving in silico system biology. For example, progress has been made in the development of in silico methods using various QSAR and molecular modeling techniques that employ a range of recently introduced descriptors tailored to estimating ADME-Tox. These in silico approaches are promising filters for virtual libraries to aid synthesis as well as the selection of compounds for acquisition and screening in the early stages of drug discovery.[27] Lapinsh *et al.* have created QSAR models that incorporate a recently developed proteo-chemometrics approach, which is based on the combined analysis of series of receptors and ligands. In this method, descriptions of ligands, proteins, and ligand-protein cross-terms are correlated with interaction activities. The compounds are characterized by structural descriptors, including three-dimensional grid-independent descriptors, topological descriptors, and geometrical descriptors.[29] Ekins and coworkers have combined QSAR and systems biology methods to develop a novel computational approach called MetaDrug. This method predicts metabolites based on the chemical structure of the parent molecule and predicts the biological activity of the original compound and its metabolites using various ADME-Tox models. Results are incorporated into a predictive model that is coupled with human cell signaling and metabolic pathways as well as networks, which in turn integrates networks and metabolites with relevant toxicogenomic or other high throughput data.[2]

In this post-genomic era of high-throughput analyses and combinatorial library screening, the future of QSAR studies and computational approaches in general lies in applications that combine multiple methods. Methods need to be able to predict the efficacy, activity, metabolism or toxicity of a pesticide or drug and integrate a range of data from multiple systems including: *in vitro* assays, *in vivo* animal models, high-throughput genomics, proteomics and metabolomics techniques, as well as computational methods. These approaches are each limited in the sense that in order to understand the complexity of biological systems and predict biological activity, it is necessary to focus on a wide range of methods/applications as opposed to a single isolated method. Only by combining the information collected by all of these different methods can an integrative picture of a molecule's biological activity and ADME-Tox properties be achieved.

Chemical issues to be solved with KEGG

The Kyoto Encyclopedia of Genes and Genomes (KEGG) database is a composite database for comprehensive bioinformatics and chemical genomics analyses.[1] A user can access nearly all information stored in KEGG *via* a web browser. The genomic data of KEGG come from completely sequenced genomes and comprehensively annotated genes. The chemical contents of KEGG are limited to the known fundamentally important chemical events in the living cell, however, the amount of chemical data are greater than any other resources in the bioinformatics area. The details on KEGG web interfaces and data types are reviewed in another article of this journal,[30] therefore this review focuses on applications to the analysis of metabolic pathways for incorporation into (Q)SAR model development.

1. Metabolic network information in KEGG PATHWAY

KEGG PATHWAY is a collection of manually drawn pathway maps representing well-known facts regarding molecular interactions and reaction networks. The contents of this database can be divided into two major classes: the metabolic pathways and the regulatory pathways. The current list of pathways maintained by KEGG is discussed by Aoki-Kinoshita in this issue.[30] Regulatory pathways consist of many types of biomolecules such as DNA, RNA, proteins, and chemical compounds and many types of interactions including direct binding, phosphorylation, ubiquitination, glycosyla-

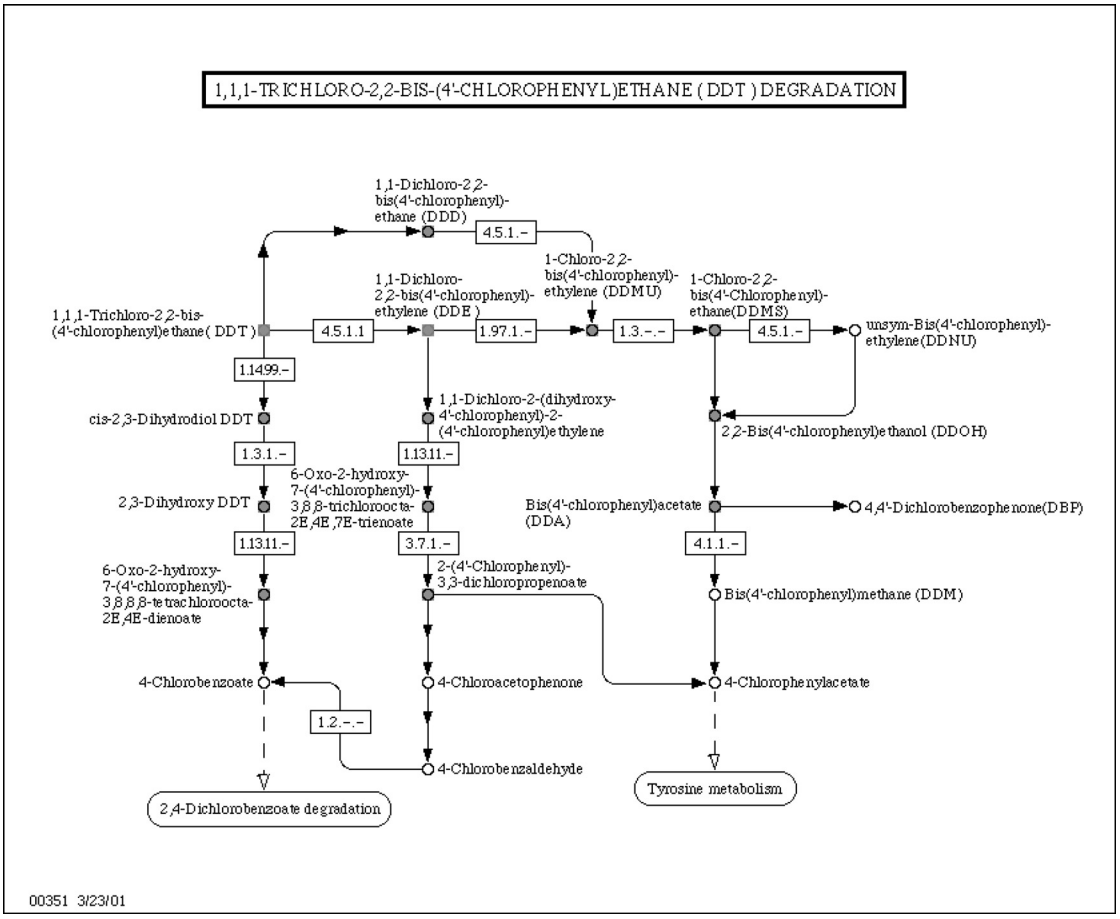


Fig. 1. DDT biodegradation pathway. Circles and boxes indicate compounds and enzymes, respectively. Numbers inside boxes are EC (Enzyme Commission) numbers. Arrows through boxes indicate reaction direction. Filled circles indicate that a compound is predicted to be active by a SAR model, and open circles indicate inactivity. An interactive clickable pathway map is available from http://www.genome.jp/kegg/pathway/map/map00351.html.

tion, methylation, and the formation of protein complexes. In a KEGG PATHWAY, each molecule is described as a rectangle or similar object-like cartoon, and each interaction is depicted as an arrow in order to render the graphical displays more intuitive. Regulatory pathways in a living cell are extremely complicated and the network itself is not-yet-understood, making it difficult to develop an algorithm to describe the system. Subsequently, the regulatory pathways section in KEGG PATHWAY may be modified according to new developments available in the scientific literature. On the other hand, metabolic pathways that consist of chemical compounds and enzymatic reactions have been sufficiently established through exhaustive experimentation. In KEGG PATHWAY, each chemical compound and reaction is also described as a small circle and an arrow in each metabolic pathway, respectively. (In addition, red filled circles indicate predicted active compounds that is described in the next section.) Enzyme commission (EC) numbers of enzymatic reactions assigned by the EC committee are also provided near the reaction arrows in boxes (see Fig. 1).

We divide metabolic pathways into two categories on the basis of their biological meaning: the first is common metabolic processes such as carbohydrate metabolism, energy production, lipid synthesis, and those reactions relating to nucleotides, amino acids, glycans, or other essential compounds. These types of processes can be considered as the core metabolic reactions of biology. The second type of metabolic process is species-specific metabolic reactions associated with secondary metabolites or xenobiotics. The metabolites involved in the core metabolic reactions are quite different from those involved in secondary metabolism, with the former ones being very common and the latter ones being restricted to environmental reactions. Therefore, we can assume that these secondary pathways may be used to adapt to environmental changes in which each organisms grows, and that the secondary metabolic pathways cause the functional diversity of organisms with a variety of genes. Of course, all environmental chemicals and their degradation pathways, such as bio-remediation activities in some bacteria, cannot be completely elucidated by current experimental techniques. It is very impor-

tant for us to investigate enzymes, enzymatic reactions and chemical compounds included in those pathways in order to understand the chemical aspects of biological systems.

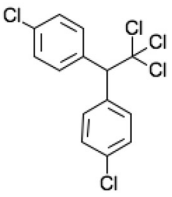
2. Chemical compound information in COMPOUND/ DRUG parts of KEGG LIGAND

KEGG LIGAND is also a composite database and contains our knowledge on chemical substances and reactions that are relevant to biological processes. LIGAND consists of COMPOUND, DRUG, GLYCAN, REACTION, RPAIR, and ENZYME parts for chemical compounds, effective drug compounds, carbohydrates chains, enzymatic reactions, atom-level reaction mechanisms, and enzyme nomenclature information, respectively.[31] Here, we used COMPOUND and DRUG categories of LIGAND to obtain the chemical compound structures. Each chemical compound structure in those databases is drawn as a 2-dimensional graph, and stored in a KCF (KEGG Chemical Function) format[32] and a MDL/MOL format. Here, the 2-dimensional graph means the graph theoretical object with nodes (vertices) and edges with only two-axis coordination of atoms, that is, any stereo chemical features of chemical compounds are not considered in the structure files. A KCF format is the original data format developed by KEGG to represent several types of chemical information like glycan structures, chemical compounds, drug structures, and enzymatic reactions. Particularly in the KCF format of chemical compounds, the information of physicochemical environment of atoms has been added into the representation of atoms and used for the further computation.[32,33] However, the other informational content represented in KCF format is essentially the same as that of the MDL/MOL format, thus a user can use either format if only structural information on chemical compounds is required (see Fig. 2)

The current number of chemical compounds in KEGG LIGAND Release 38.0+/05–02, May 02 is 14,092 entries for chemical compounds and 2894 entries for drugs. These num-

KEGG COMPOUND: C04623

Help

Entry	C04623 Compound
Name	1,1,1-Trichloro-2,2-bis(4-chlorophenyl)ethane; 1,1,1-Trichloro-2,2-bis-(4'-chlorophenyl)ethane; DDT
Formula	C14H9Cl5
Mass	351.9147
Structure	 C04623 Mol file KCF file DB search
Reaction	R04522 R05260 R05476 R05492
Pathway	PATH: map00351 1,1,1-Trichloro-2,2-bis(4-chlorophenyl)ethane (DDT) degradation
Enzyme	1.14.99.- 4.5.1.1
Other DBs	CAS: 50-29-3 PubChem: 7213 ChEBI: 16130 3DMET: B00739
LinkDB	All DBs
KCF data	Show

=> **Original format**

DBGET integrated database retrieval system, GenomeNet

Fig. 2. Example of a chemical entry in KEGG LIGAND. In this entry view, 'Entry' is the accession number, which is the unique identifier in KEGG. 'Name' is the recommended name, usually a generic name, and alternative names including systematic nomenclatures of the compound. 'Formula' and 'Mass' are the molecular formula and mass of the compound. 'Structure' is the chemical structure (structural formula) of the compound, which is stored in an MDL/MOL file and a KCF file. In the lower parts of this view, 'Reaction', 'Pathway', 'Enzyme', 'Other DBs', or 'LinkDB' are all link information to the REACTION database entries, to KEGG pathway maps, to the ENZYME database entries, to outside databases, or to LinkDB that is a part of the DBGET system, respectively. The final 'KCF data' section is to display the KCF representation of the chemical compound structures.

bers are not as extensive as the Cambridge Structural Database, which is a comprehensive database of solved crystal structures of small molecules.[34] However, KEGG LIGAND focuses only on the biological events with chemical processes, and their data is very useful to validate the chemical aspects of biological processes, especially, by correlating with other biological information obtained from genomic data. For instance, each chemical compound is linked to the PATHWAY, REACTION, and ENZYME databases with pathway map numbers (*e.g.* mapNNNNN), reaction numbers (*e.g.* Rnnnnn), and EC numbers, respectively. Thus, when we want to know the biological function of a chemical compound, we can use the information linked to intra-KEGG entries as well as the following BRITE information.

3. Functional information of each compound in KEGG BRITE

KEGG BRITE is a collection of hierarchical classifications that represents our knowledge of various aspects of biological systems. Specifically, it incorporates many different types of relationships rather than only molecular interactions and reactions stored in KEGG PATHWAY. For instance, the functional classification of chemical compounds or drugs is described in a hierarchical tree graph and its entire view can be displayed as a text-based representation (for example, see http://www.genome.jp/dbget-bin/get_htext?Metabolite). The tree in the BRITE database can be expanded or shrink its branches at any branch level or any tree node that a user selects. In most cases, each final node of this tree, namely a leaf, represents just one entry of molecules. In the functional classification tree of chemical compounds, each chemical compound is classified according to those biochemical classifications. With this tree, we can find relationships between structural components or building blocks of chemical compounds and their biological function or biological meanings within each metabolic reaction network. When we choose the functional hierarchical classification of drugs, all drug compounds in the DRUG database are classified on every leaf of the tree in accordance with their therapeutic effects. From this type of tree, we can identify a set of drug molecules that have a specific effect on an organism. In addition, each drug entry contains structural information as well as data on side-effects and drug composition that can be examined with several information techniques.

It is currently difficult to use the hierarchical tree of function in a computational manner because it has only been prepared in a text-based format on the KEGG web site, as no sophisticated format such as XML has been developed. However, most of the established knowledge in the life science should be accumulated into some of the categorized trees of KEGG BRITE. This is why it is still useful for us to consider how a genetic/chemical molecule works in biological systems, and from those observations we may have insight into the further/higher meanings of small molecules.

(Q)SAR and pathway analysis*

Biological network databases, especially metabolic pathway network databases, such as KEGG,[1] BioCyc[35] and BRENDA[36] contain genes (enzymes) and chemicals as distinct entities as well as the relationships between them (*e.g.*, reactant, biological catalyst and product). These relationships compose a large connected gene-chemical network, which describes the flow of metabolites in a cell. The combination of (Q)SAR and gene-chemical network analyses gives a different profile of chemical biological activity, and thus can reveal the roles of genes and chemicals in a cell. High-throughput data (*i.e.* microarray data) have been used in combination with a pathway database to develop novel hypotheses.[37] Therefore an analogous approach, (Q)SAR with a pathway database, should be useful for understanding the molecular biology processes involved in biologically active chemicals. In the ADME-Tox field, a similar approach has been proposed (see the second section).

Our analysis has two distinct steps; the first step is to identify active chemicals in a database by using SAR method, and the second step is to analyze pathways, in which active chemicals are mapped. Because of the large diversity of chemicals in a pathway database, the SAR used in our analysis was required to reduce the level of false-positives throughout the large chemical space. Once the active/inactive biological information is mapped in the pathway database, various methods of pathway analysis based on statistics and graph-mining technique can be applied. For instance, deactivate reaction/enzyme is useful information for designing a compound.

This approach is a new application of (Q)SAR analysis that is expected to increase in interest and use in the near future. This combination of approaches will be especially useful for large datasets based upon omics-related studies and methodologies. In order to explain the utility and application of this approach, we will show an example of the combination analysis with SAR and KEGG pathway database focused on EDCs. The analysis provides the enzyme group(s) related to EDC detoxification as well as the characteristics of the predicted active chemicals.

1. Datasets

The training dataset was obtained from the Endocrine Disruptor Knowledge Base (EDKB) provided by the National Center for Toxicological Research (http://edkb.fda.gov/databasedoor.html). The dataset contains six types of bioassays, estrogen and androgen receptor binding assays,[38–40] reporter gene assay,[40] uterotrophic assay,[41] 2 cell proliferation assays.[38,42] The E-SCREEN data, based upon a cell proliferation assay, were most suited to the purpose of this study, as cell-based assays incorporate *in vivo* transport properties as

* Detailed description of the method and its results are in preparation.

well as a greater range of metabolic processes then *in vitro* assays. The dataset contains 120 different chemicals, 59 of which are biologically active and 61 of which are inactive. The biological activities of the E-SCREEN assays were measured in units of log translated relative proliferative potency (log RPP), with values ranging from −4.2 to 3.0. The large-scale chemical data test set was obtained from the COMPOUND database, a component of the KEGG database. The database contains approximately 12,000 chemicals, consisting of metabolites, as well as drugs and xenobiotic compounds. The median molecular mass is 378 g/mol and the molecular mass of >97% of the chemicals is under 1000 g/mol.

2. Model construction

A SAR model was constructed based on a novel graph-mining algorithm proposed by Kudo *et al*.[43] Our model has three layers from an input as chemical structure to a binary output as active or inactive. The first and second layers correspond to the Kudo 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. The second layer combines the outputs of the decision stumps, considering the weights computed by the Kudo algorithm. 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 in our model is designed to integrate many Kudo classifiers. Each Kudo classifier is trained by a subset of the training dataset. In the third layer we generated 100 independent Kudo classifiers by using a randomly generated different dataset for each classifier. We assigned the label (activity) of an input graph (chemical compound) by a majority vote of the 100 classifiers.

Fragments identified by the Kudo algorithm can include ring structures in its descriptors by searching through all possible input graph substructures/fragments as potential molecular descriptors. Therefore our molecular descriptors are more informative than other molecular descriptors in models that consist of only linear or tree fragments. For example, a fragment with 9-membered ring was found, suggesting a combined 6-membered and 5-membered ring fragment. This fragment can be observed in steroid skeletons. Another example is a phenol-like structure, which is included in well-known active chemicals, such as estradiol-17beta, bisphenol A, diethylstilbestrol, *etc.* Our model finds these fragments in an input chemical, and then sums up scores assigned to the fragments. If the score is over a defined threshold, the chemical is predicted to be biologically active.

3. Mapping chemicals on metabolic pathways

The system predicted that 1291 chemicals out of the 12,109 chemicals in the KEGG COMPOUND database to be potentially biologically active. The predicted biological activities of chemicals were mapped on to the KEGG PATHWAY database. The KEGG Markup Language (KGML) describes the relationship between chemicals and enzymes on metabolic pathways in a computer readable format. By using the KGML, metabolic pathways and enzymes associated with the predicted chemicals can be collected. Mapping on to the KEGG PATHWAY database reveled that 28 of 139 pathways contained at least one active chemical in its map, and 310 of 4238 chemicals were predicted as active.

Relations between a degradation process and activities of chemicals involved in the process were obtained. For example, a map that explained the degradation process of DDT is shown in Fig. 1. In the training dataset DDT and DDE were included in the active chemical group, and the prediction results were accurate. Some of the compounds in the degradation pathway were predicted as active, moreover, the figure showed a detoxification point in the DDT detoxification pathway. Other examples were in the flavonoid and alkaloid biosynthesis pathways. In the flavonoid biosynthesis pathway, genistein, a component of soybeans, is an active chemical in the training dataset. Some metabolites of genistein were predicted to be active chemicals (Fig. 3). There were no active chemicals mapped on to the alkaloid biosynthesis pathway in the training dataset, but we found tree candidates (data not shown). Alkaloids are nitrogenous organic chemicals, and some of them have pharmacological effects on human and other organisms.

In addition to the results of individual maps, network analysis extracted characteristics of the degradation processes. In this analysis, enzymes involved in a degradation process were defined as the enzymes involved in the "inactivatation" reaction as well as upstream reactions. These enzymes should subsequently be involved in the EDC detoxification process. The EC numbers were used for finding specific enzymatic groups involved in this process, and results identified the most frequent family of enzymes was dioxygenases (EC1.14.-.-) (data not shown). Oxygenases include members of the well-known cytochrome P450 family, which is broadly used in bioremediative microorganisms.[44,45]

4. Mapping chemicals on functional hierarchies

Predicted active chemicals were also mapped on to the functional hierarchies of drugs and compounds in the KEGG BRITE database. Functional hierarchies have tree structures, in which each node represents a certain chemical function. The deeper a level goes, the more specific the function it defines. However, statistics in the top level give an overview of the function of predicted chemicals. Table 1 shows the statistics of predicted lipids. In total, 318 of 1259 chemicals were active candidates, with sterol lipids, prenol lipids and polyketides being dominant. Many chemicals in these categories are biologically active substances, and predicted EDCs also share these categories. This method is therefore useful for quickly screening large quantities of chemical compounds and identi-

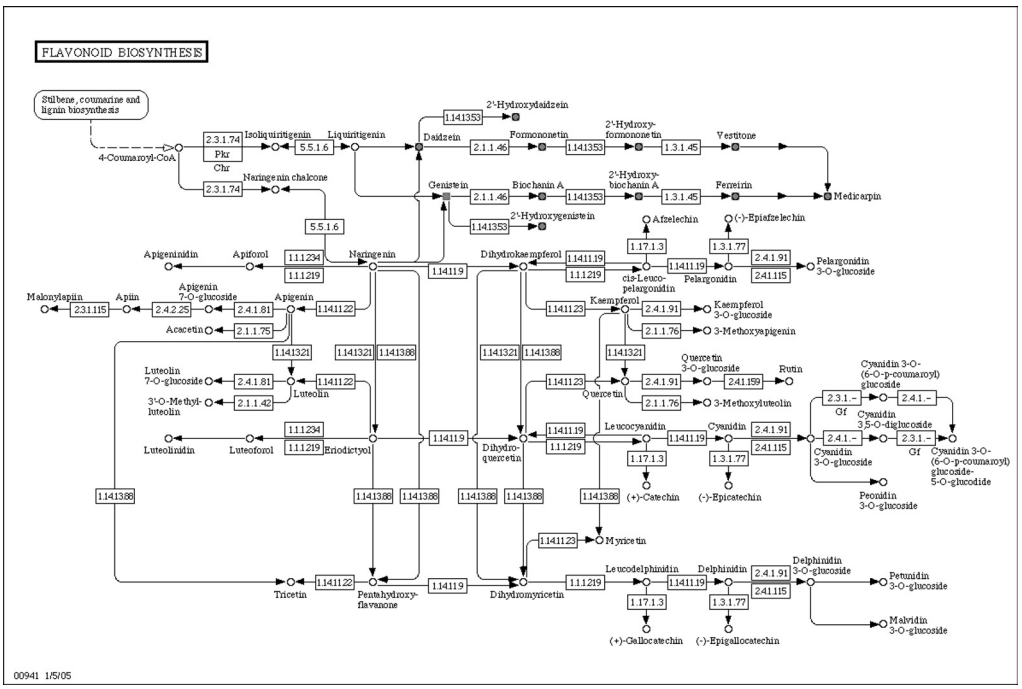


Fig. 3. Flavonoid biosynthesis pathway. Filled circles indicate that a compound is predicted to be active by a SAR model, and open circles indicate inactivity. Interactive clickable pathway maps are available from http://www.genome.jp/kegg/pathway/map/map00941.html.

fying those compounds that should be focused upon for further experimental analysis (for example, EDC activity[6]).

Conclusion

As the use of omics-technologies continues to increase, there will be a need for in silico screening methods to process the large quantities of data generated. It will be particularly important to identify key areas where laboratories should focus their experimental resources following bioinformatics screening. The coupling of (Q)SAR with gene/chemical network analysis will serve to both identify new potential lead compounds as well as place them within a biological and metabolic context within the target organism. This combination could be especially useful for determining environmental effects of developed compounds by predicting at which step of metabolism the compound lost its biological activity. Whether in the development of agrochemicals or pharmaceuticals, the methods of compound development and ADME-Tox properties as well as environmental fate analyses are extremely important. In this review, we have demonstrated how to integrate SAR and network analysis and what results are derived from the analysis using EDC compounds as a model system. As shown previously, this analysis lies in both genome and chemical space and is a more accurate way to understand biological systems than analyses of only one space.

A single approach to understanding ADME-Tox processes is not sufficient. In the bioinformatics field, there are a number of databases currently available or under construction and the development of analysis tools is an ever ongoing process. Many of these products will be useful to integrate and understand (Q)SAR results and ADME-Tox. By combining (Q)SAR analyses with pathway networking, we will greatly increase the information that we can glean from in silico screening methods and bioinformatic analyses.

Table 1. Predicted chemicals mapped on the lipid classification.

Category	Active	Total[a)]
FA Fatty acyls	1	(330)
GL Glycerolipids	0	(43)
GP Glycerophospholipids	0	(69)
SP Sphingolipids	0	(93)
ST Sterol Lipids	169	(228)
PR Prenol Lipids	126	(364)
SL Saccharolipids	0	(17)
PK Polyketides	22	(115)
Total	318	(1259)

[a)] The total number of pathways and chemicals in a category is expressed in parentheses.

Acknowledgments

This study was supported by the 21st Century COE program 'Genome Science' and a grant-in-aid for scientific research on the priority area from the Ministry of Education, Culture, Sports, Sci-

ence and Technology of Japan. The computational resources were provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University. CEW was funded by a Japanese Society for the Promotion of Science (JSPS) post-doctoral fellowship.

References

1) M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki and M. Hirakawa: *Nucleic Acids Res.* **34**, D354–357 (2006).
2) S. Ekins, S. Andreyev, A. Ryabov, E. Kirillov, E. A. Rakhmatulin, S. Sorokina, A. Bugrim and T. Nikolskaya: *Drug Metab. Dispos.* **34**, 495–503 (2006).
3) C. Hansch, D. Hoekman, A. Leo, D. Weininger and C. D. Selassie: *Chem. Rev.* **102**, 783–812 (2002).
4) J. J. Sutherland, L. A. O'Brien and D. F. Weaver: *J. Med. Chem.* **47**, 5541–5554 (2004).
5) A. Z. Dudek, T. Arodz and J. Galvez: *Comb. Chem. High Throughput Screen* **9**, 213–228 (2006).
6) J. Gasteiger and T. Engel: "Chemoinformatics," John Wiley & Sons, Wienheim, Germany, p. 680, 2003.
7) B. Bordas, T. Komives and A. Lopata: *Pest. Manag. Sci.* **59**, 393–400 (2003).
8) J. D. Walker, J. Jaworska, M. H. Comber, T. W. Schultz and J. C. Dearden: *Environ. Toxicol. Chem.* **22**, 1653–1665 (2003).
9) D. Mackay and E. Webster: *SAR QSAR Environ. Res.* **14**, 7–16 (2003).
10) M. T. Cronin, J. S. Jaworska, J. D. Walker, M. H. Comber, C. D. Watts and A. P. Worth: *Environ. Health Perspect.* **111**, 1391–1401 (2003).
11) M. T. Cronin, J. D. Walker, J. S. Jaworska, M. H. Comber, C. D. Watts and A. P. Worth: *Environ. Health Perspect.* **111**, 1376–1390 (2003).
12) P. K. Schmieder, G. Ankley, O. Mekenyan, J. D. Walker and S. Bradbury: *Environ. Toxicol. Chem.* **22**, 1844–1854 (2003).
13) J. Ashby, E. Houthoff, S. J. Kennedy, J. Stevens, R. Bars, F. W. Jekat, P. Campbell, J. Van Miller, F. M. Carpanini and G. L. Randall: *Environ. Health Perspect.* **105**, 164–169 (1997).
14) W. V. Welshons, K. A. Thayer, B. M. Judy, J. A. Taylor, E. M. Curran and F. S. vom Saal: *Environ. Health Perspect.* **111**, 994–1006 (2003).
15) K. F. Arcaro, D. D. Vakharia, Y. Yang and J. F. Gierthy: *Environ. Health Perspect.* **106** Suppl. 4, 1041–1046 (1998).
16) H. Fang, W. Tong, W. S. Branham, C. L. Moland, S. L. Dial, H. Hong, Q. Xie, R. Perkins, W. Owens and D. M. Sheehan: *Chem. Res. Toxicol.* **16**, 1338–1358 (2003).
17) O. Mekenya, V. Kamenska, R. Serafimova, L. Poellinger, A. Brouwer and J. Walker: *SAR QSAR Environ. Res.* **13**, 579–595 (2002).
18) R. Serafimova, J. Walker and O. Mekenyan: *SAR QSAR Environ. Res.* **13**, 127–134 (2002).
19) M. A. Lill, F. Winiger, A. Vedani and B. Ernst: *J. Med. Chem.* **48**, 5666–5674 (2005).
20) M. A. Lill, M. Dobler and A. Vedani: *SAR QSAR Environ. Res.* **16**, 149–169 (2005).
21) C. Y. Zhao, R. S. Zhang, H. X. Zhang, C. X. Xue, H. X. Liu, M. C. Liu, Z. D. Hu and B. T. Fan: *SAR QSAR Environ. Res.* **16**, 349–367 (2005).
22) M. N. Jacobs: *Toxicology* **205**, 43–53 (2004).
23) L. B. Hendry, L. W. Roach and V. B. Mahesh: *Steroids* **64**, 570–575 (1999).
24) W. Tong, D. R. Lowis, R. Perkins, Y. Chen, W. J. Welsh, D. W. Goddette, T. W. Heritage and D. M. Sheehan: *J. Chem. Inf. Comput. Sci.* **38**, 669–677 (1998).
25) S. J. Yu, S. M. Keenan, W. Tong and W. J. Welsh: *Chem. Res. Toxicol.* **15**, 1229–1234 (2002).
26) H. Hong, H. Fang, Q. Xie, R. Perkins, D. M. Sheehan and W. Tong: *SAR QSAR Environ. Res.* **14**, 373–388 (2003).
27) H. van de Waterbeemd: *Curr. Opin. Drug Discov. Devel.* **5**, 33–43 (2002).
28) H. van de Waterbeemd and E. Gifford: *Nat. Rev. Drug. Discov.* **2**, 192–204 (2003).
29) M. Lapinsh, P. Prusis, I. Mutule, F. Mutulis and J. E. Wikberg: *J. Med. Chem.* **46**, 2572–2579 (2003).
30) K. F. Aoki-Kinoshita: *J. Pestic. Sci.* **31**, 296–299 (2006).
31) S. Goto, Y. Okuno, M. Hattori, T. Nishioka and M. Kanehisa: *Nucleic Acids Res.* **30**, 402–404 (2002).
32) M. Hattori, Y. Okuno, S. Goto and M. Kanehisa: *J. Am. Chem. Soc.* **125**, 11853–11865 (2003).
33) M. Kotera, Y. Okuno, M. Hattori, S. Goto and M. Kanehisa: *J. Am. Chem. Soc.* **126**, 16487–16498 (2004).
34) F. H. Allen: *Acta Crystallogr.* **B58**, 380–388 (2002).
35) P. D. Karp, C. A. Ouzounis, C. Moore-Kochlacs, L. Goldovsky, P. Kaipa, D. Ahren, S. Tsoka, N. Darzentas, V. Kunin and N. Lopez-Bigas: *Nucleic Acids Res.* **33**, 6083–6089 (2005).
36) I. Schomburg, A. Chang, C. Ebeling, M. Gremse, C. Heldt, G. Huhn and D. Schomburg: *Nucleic Acids Res.* **32**, D431–433 (2004).
37) T. Ideker, V. Thorsson, J. A. Ranish, R. Christmas, J. Buhler, J. K. Eng, R. Bumgarner, D. R. Goodlett, R. Aebersold and L. Hood: *Science* **292**, 929–934 (2001).
38) N. Olea, R. Pulgar, P. Perez, F. Olea-Serrano, A. Rivas, A. Novillo-Fertrell, V. Pedraza, A. M. Soto and C. Sonnenschein: *Environ. Health Perspect.* **104**, 298–305 (1996).
39) C. L. Waller, B. W. Juma, L. E. Gray, Jr. and W. R. Kelce: *Toxicol. Appl. Pharmacol.* **137**, 219–227 (1996).
40) T. Nishihara, J. Nishikawa, T. Kanayama, F. Dakeyama, K. Saito, M. Imagawa, S. Takatori, Y. Kitagawa, S. Hori and H. Utsumi: *J. Health Sci.* **46**, 282–298 (2000).
41) A. G. Hilgar and J. Palmore, Jr.: "Endocrine Bioassay Data. Part VI: The Uterotrophic evaluation of steroids and other compounds - assay 2," ed. by A. G. Hilgar and L. C. Trench, National Cancer Institute, p. 181, 1968.
42) A. M. Soto, C. Sonnenschein, K. L. Chung, M. F. Fernandez, N. Olea and F. O. Serrano: *Environ. Health Perspect.* **103** Suppl. 7, 113–122 (1995).
43) T. Kudo, E. Maeda and Y. Matsumoto: *Adv. Neural Inform. Process. Systems* **17**, 729–736 (2005).
44) L. L. Wong: *Curr. Opin. Chem. Biol.* **2**, 263–268 (1998).
45) E. M. Gillam: *Clin. Exp. Pharmacol. Physiol.* **32**, 147–152 (2005).