

# A Structural Similarity Evaluation by SimScore in a Teratogenicity Information Sharing System

Kumiko SAKAMOTO<sup>a</sup>, Aiko YAMAUCHI<sup>a\*</sup>, Mikio SASAKI<sup>b</sup> and Hiroshi CHUMAN<sup>a</sup>

<sup>a</sup>Graduate School of Pharmaceutical Science, The University of Tokushima  
1-78, Shomachi, Tokushima 770-8505, Japan

<sup>b</sup>Saila Systems, Inc.

Gotanda Yamakatsu Building 3-16-44, Higashi Gotanda, Shinagawa-Ku Tokyo 141-0022, Japan

\*e-mail: aiko@ph.tokushima-u.ac.jp

(Received: January 23, 2007; Accepted for publication: February 2, 2007; Published on Web: April 4, 2007)

Many efforts have been devoted to the development of computer-aided prediction of drug toxicity over the decades, but at present, systems and programs available for predicting teratogenicity from chemical structures do not always give satisfactory answers yet, mainly because of the complex and unknown mechanism of reproductive and developmental toxicity. We developed a novel algorithm and implemented in the program “SimScore” to evaluate quantitatively the structural similarity score of a target compound with the teratogenic drugs which are defined as serious human teratogens by the United States Food and Drug Administration. In SimScore, a molecular structure is divided into its skeletal and substituent parts in order to perform similarity comparison for these parts independently. This idea is based on that compounds with the same or similar skeleton show a similar biological activity, but their activity strengths depend on the variation of substituents. We demonstrated the usefulness of SimScore by applying it to an example.

SimScore will be used in our web-based information system about teratogenicity to predict the potential risk of query compounds.

**Keywords:** SimScore, Structural similarity matching, Teratogenic toxicity prediction, Teratogenicity information sharing system, Molecular structural information

## 1 Introduction

Recent advances in laboratory automation and experimental techniques such as combinatorial chemistry and high-throughput screening resulted in the growing efficiency of the identification of novel drug candidates. However it became clear that currently one of the biggest challenges is the early determination of unfavorable ADMET properties. Since many failures due to toxicity have been recognized in the development stage [1, 2], the prediction of toxicity *in silico* has become desired before the screening assay *in vitro* [3, 4]. Because the teratogenic adverse events cannot be tested in the human body, the computer-aided screening for the reproductive and developmental toxicity is especially important to ensure the safety of drug candidates.

Expert systems such as DEREK [5], HazardExpert [6] and TOPKAT [7] have been currently available to predict the teratogenicity or reproductive toxicity of new

chemical compounds. However, to obtain satisfactory results for the practical use, these systems have some limitations due to the following reasons. The knowledge necessary for prediction of teratogenic activity from chemical structures is not essentially adequate. The complicated and unknown molecular mechanism of teratogenicity makes it very difficult to properly identify the mode of action of molecules, and this fact makes the development of consistent prediction models a very challenging task [8–10]. Another obstacle in the way of proper statistical analyses is the insufficient biological and/or clinical data and the highly diverse nature of the compound sets.

In this paper, we briefly introduce a novel similarity algorithm and an expert system, SimScore for the predicting teratogenicity of a given compound by comparing its chemical information with that of each human teratogen categorized by the United States Food and Drug Administration (FDA). In SimScore, a molecular structure is divided into its skeletal and substituent parts and

the similarity matching for each part is executed independently. The idea is that compounds with the same or similar skeleton show a similar biological activity, but their activity strengths depend on the variation of substituents.

We have been constructing a web-based drug safety information community system [11] to share the teratogenic information among community members [12]. SimScore is one of the subsystems and intended mainly for drug discovery researchers.

## 2 Database and similarity matching algorithm

### 2.1 Molecule database of teratogens (TeraMol DB)

The FDA used five categories to classify thousands of compounds with existing data (either human or animal) about teratogenic effects. These categories are as follows; A (controlled studies show no risk), B (no evidence of risk in humans), C (risk cannot be ruled out), D (positive evidence of risk), or X (contraindicated in pregnancy) [13]. Positive evidences of fetal abnormalities for the drugs belonging to the categories D and X have been confirmed by the epidemiological studies in pregnant women.

In this study, we have constructed the molecular database TeraMol DB, which contains the chemical structures of the drugs classified into the FDA categories D and X. In TeraMol DB, each chemical structure together with other information was stored as the MDL format [14]. The skeleton structure of each molecule in TeraMol DB was defined by the maximum common substructure among the structurally similar teratogens. The substituent parts were defined as the rest of the whole structure. The skeleton structure was not necessarily defined as a unique chemical structure, but some changes of atom and bond types were allowed. This definition was somewhat arbitrary, but various alternative skeletal substructures were prepared and stored. Structural information for the skeletal and substituents structures was stored in an extended MDL format in TeraMol DB.

### 2.2 Structure matching of skeletons

In SimScore, the atomic information in each molecule is expressed as an atom code array, which consists of the following eight atom codes; element, element group, hybridization type, ring, adjacent atom, hydrogen bonding, atomic charge and stereo codes. These atom codes are easily obtained from the molecular connection table in the MDL file. Based on the connection table and the above atomic codes, the structural similarity scores for the skeletal and substituent parts are calculated between a given molecule (Mol\_G) and each molecule (Mol\_T) in TeraMol DB. To compare the substructure of Mol\_G,

first, the atom codes (array  $A$ ) and connection arrays (array  $B$ ) of it in Mol\_G are extracted. It is ensured that the size of the substructure is comparable to those of the skeletal substructure of Mol\_T in the database. The arrays  $A$  and  $B$  are compared to those of the already defined skeletal structure of Mol\_T. The extracted substructure in Mol\_G and skeletal substructure in Mol\_T are referred as SsG and SsT, respectively. The array  $A(i, k)$  indicates the  $k$ -th atom code of atom  $i$  ( $i=1, \dots$ , a number of the skeletal atoms in SsT). The array  $B(i, j)$  indicates a number of bonds from atom  $i$  to an atom with the  $j$ -th atomic number. If both of the arrays of SsG are identical with the corresponding ones of SsT, or they are subsets of SsT, then the skeletal similarity score is calculated, as defined in the next section. Otherwise the score is set to be zero. There are generally a huge number of ways to extract the arrays  $A$  and  $B$  from Mol\_G. The above procedure is repeated until all of the possible substructures are extracted. The best match of the skeletal atoms between the two substructures is the one with the highest similarity score.

### 2.3 Similarity score

The similarity scores of skeletal and substituents parts (noted as  $SkSS$  and  $BSS$ , respectively) between Mol\_G and Mol\_T are calculated independently and the total similarity score  $SSS$  is defined from  $SkSS$  and  $BSS$ .

$SkSS$  is defined by eq. 1.

$$SkSS = \left[ \sum \sum Ss(k, i)^2 / (8ns) \right]^{1/2} \quad (1)$$

where  $Ss(k, i)$ , explained below, is the similarity score between the  $i$ -th skeletal atom in SsT and its best matching atom  $i$  in SsG,  $k$  is the  $k$ -th atom code,  $ns$  is the number of skeletal atoms in SsT ( $i=1, 2, \dots, ns$ ), and the summation is taken over all of the atom codes and matching atoms. For the element and element group codes, the  $Ss(k, i)$  value takes unity when the  $k$ -th atomic codes of atom  $i$  in SsG and SsT are the same and otherwise takes zero. For the hybridization, ring, adjacent, hydrogen bonding atom codes, their scores between 0 and 1 are assigned depending on their similarity.

$BSS$  is defined by eq. 2.

$$BSS = \left[ \sum \sum Sb(k, i)^2 / (8nbi) \right]^{1/2} \quad (2)$$

where  $Sb(k, i)$  is the similarity score between the substituent atoms attached to the  $i$ -th skeletal atom in SsT and those to the  $i$ -th atom in SsG,  $k$  is the  $k$ -th atom code,  $nbi$  is the number of substituent atoms attached to the  $i$ -th atom in SsT, and the summation is taken over all of the atom codes and substituent atoms in SsT. The numbers of the  $k$ -th atom code in the substituent atoms attached to the  $i$ -th atom in SsT and SsG are stored in the two arrays  $VecT(k, i)$  and  $VecG(k, i)$ , respectively. Then, the value of  $Sb(k, i)$  is calculated as the Tanimoto similarity coefficient [15] between  $VecT(k, i)$  and  $VecG(k, i)$ .

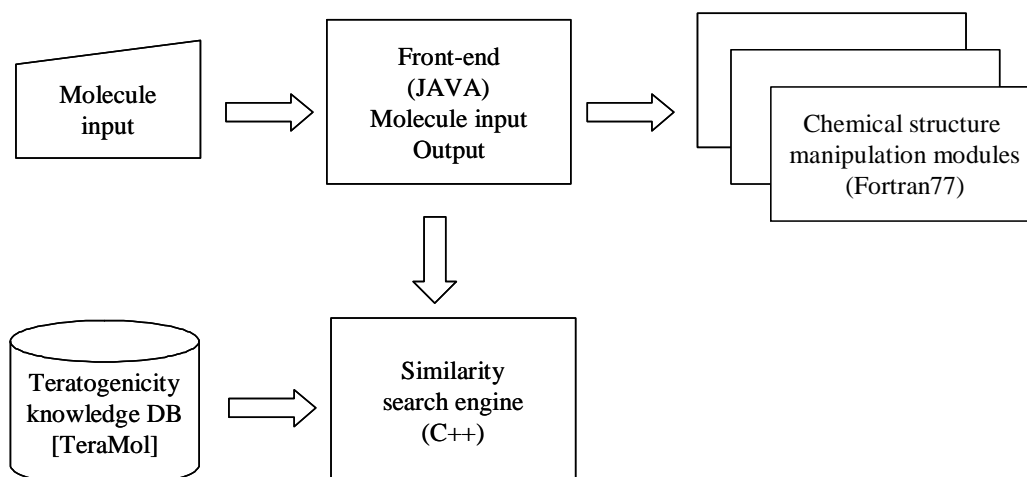


Figure 1. Architecture of SimScore

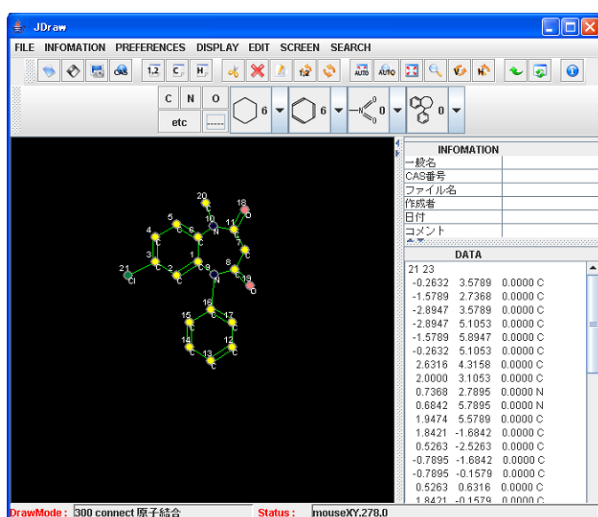


Figure 2. Screen shot of input window in SimScore

Total similarity score  $SSS$  is defined by eq. 3 according to the Tanimoto similarity score.

$$SSS = \frac{SkSS + BSS}{2 + SkSS^2 + BSS^2 - (SkSS + BSS)} \quad (3)$$

$SSS$  varies in the range between 0 and 1 and its score of unity represents the perfect similarity between Mol\_G and Mol\_T.

The details for the matching algorithm and similarity scores will be reported elsewhere.

## 2.4 Development environments

SimScore was developed on Windows XP. Its front-end for molecule input, similarity calculations, and other sub-programs were coded in Java 1.5, Visual C++ 6.0 and Visual Studio NET2005, and Fortran77, respectively. The

schematic overview of the SimScore system is depicted in Figure 1. Multithreading technique is used in the search engine for achieving an efficient structural matching search.

## 2.5 Graphic interface

By the use of an interface screen as shown in Figure 2, a query molecular structure is directly drawn there and also a MDL mol file is loadable. The results of SimScore are visualized in the window lists of matched structures in TeraMol DB and their similarity scores as shown in Figure 3.

## 3 Validation of algorithm

Clobazam is a benzodiazepine agent that is used orally as an anticonvulsant, and it was approved by Japanese Ministry of Health, Labor and Welfare in 2000. While benzodiazepine compounds have generally two nitrogen atoms at the 1, 4-positions (Figure 4) in the heterocyclic ring, clobazam is the first benzodiazepine in which the nitrogen atoms are in the 1, 5-positions in the heterocyclic ring. The fetal risk of this drug has not been evaluated by the FDA risk classification system. For critical validation of SimScore, the teratogenic possibility of clobazam was searched using SimScore. As a result, twelve teratogenic drugs in TeraMol DB exhibited nice matching with clobazam and their structures are listed in Table 1 with their similarity scores and FDA pregnancy category codes. The total similarity score of diazepam was computed to be 0.982, which was the highest score among the matched drugs. When the structural differences among clobazam and matched ones were compared, it was confirmed that the similarity scores reflected well their chemical differences, as shown in Figure 4.

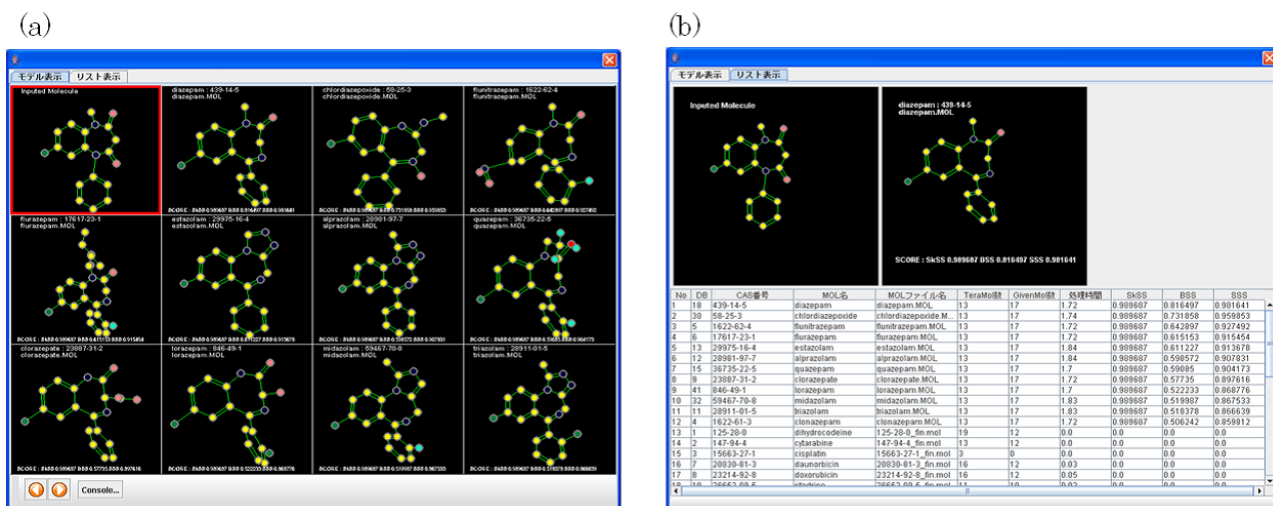
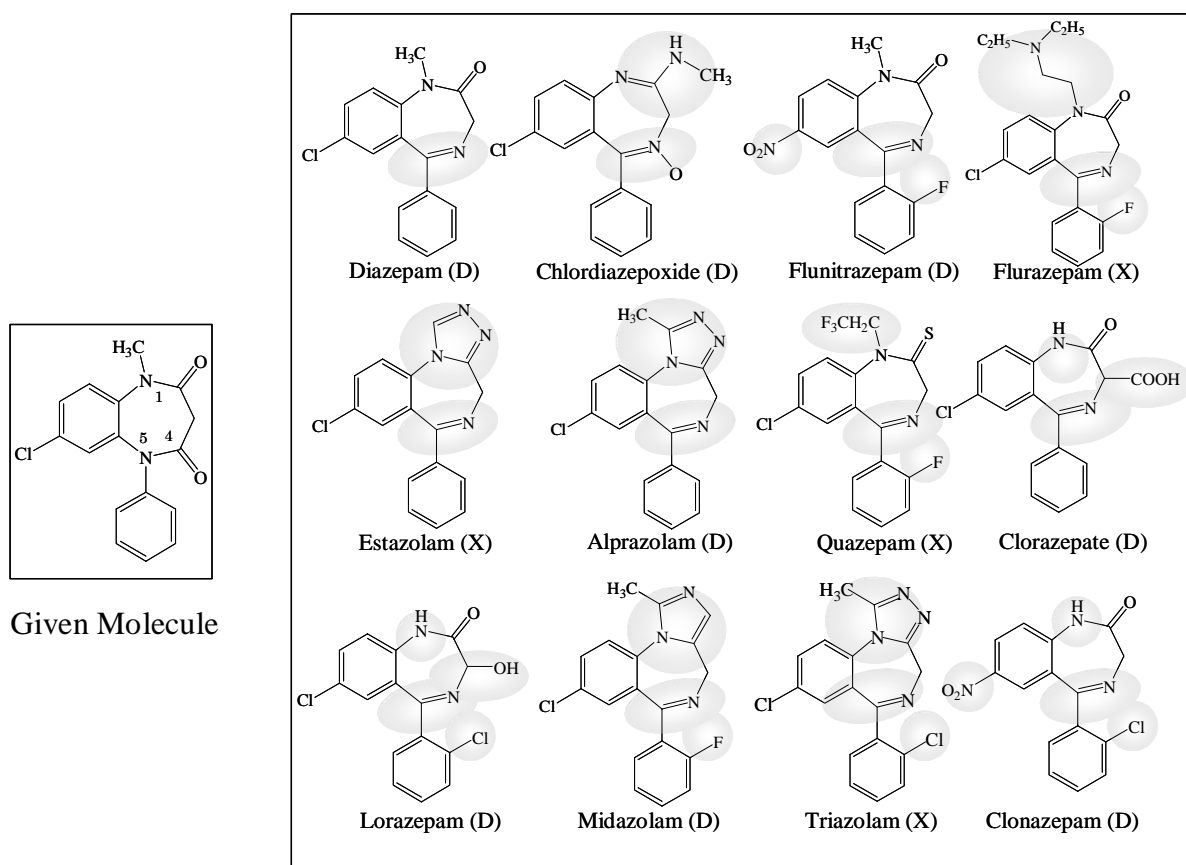


Figure 3. Screen shots of result windows in SimScore. (a) Structures of matched drugs. (b) Similarity scores.



### Teratogenic molecules in TeraMol DB

Figure 4. Structures of clobazam and matched drugs in the database. Characters in parentheses are the FDA pregnancy category codes.

Table 1. Results of SimScore for clobazam

molecule	CAS No.	<i>SkSS</i> <sup>a</sup>	<i>BSS</i> <sup>b</sup>	<i>SSS</i> <sup>c</sup>	FDA Pregnancy Category <sup>d</sup>
Diazepam	439-14-5	0.990	0.816	0.982	D
Chlordiazepoxide	58-25-3	0.990	0.732	0.960	D
Flunitrazepam	1622-62-4	0.990	0.643	0.927	D
Flurazepam	17617-23-1	0.990	0.615	0.915	X
Estazolam	29975-16-4	0.990	0.611	0.914	X
Alprazolam	28981-97-7	0.990	0.599	0.908	D
Quazepam	36735-22-5	0.990	0.591	0.904	X
Clorazepate	23887-31-2	0.990	0.577	0.898	D
Lorazepam	846-49-1	0.990	0.522	0.869	D
Midazolam	59467-70-8	0.990	0.520	0.868	D
Triazolam	28911-01-5	0.990	0.518	0.867	X
Clonazepam	1622-61-3	0.990	0.506	0.860	D

<sup>a</sup> *SkSS* is the skeletal similarity score. <sup>b</sup> *BSS* is the substituent similarity score.  
<sup>c</sup> *SSS* is the total similarity score. <sup>d</sup> The FDA pregnancy category codes.

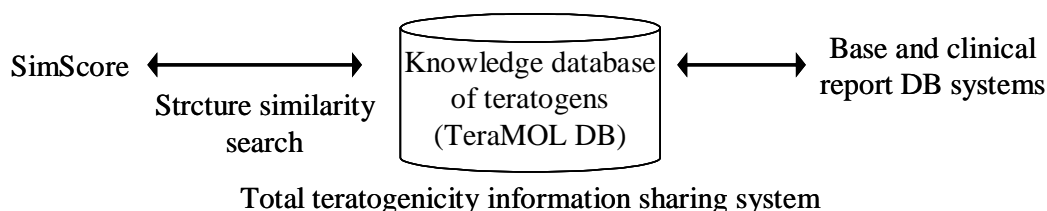


Figure 5. SimScore and teratogenicity information sharing systems

## 4 Conclusive Remarks

The prediction of drug teratogenicity induced in the human body has been one of the most serious difficulties in the drug discovery process. The algorithms in the available teratogenicity prediction systems can be classified into two classes. The softwares belonging to one class are using knowledge-based approaches, such as DEREK [5] and HazardExpert, which predict various toxic activities of a test compound [6]. The software of the other class utilizes statistical methods, like TOPKAT [7] which utilizes some quantitative structure-activity relationships (QSAR) models. There is also their hybrid type of system such as MCASE [16]. However, the reproductive and developmental toxicology is a very complicated phenomenon because many different and usually unknown mechanisms are involved. Therefore, it is not easy to identify “structural alerts” as substructures responsible for the toxicity. In fact, in DEREK only nine structural alerts are prepared for the reproductive toxicity endpoints [8]. The developmental toxicity prediction model in TOPKAT is based on data of rat studies [7, 17]. However, the chemicals, which have positive teratogenicity in laboratory animals, do not always induce the same result in humans [18]. Enhancement of prediction accuracy

for reproductive and developmental toxicity, especially in humans, has been one of the most important issues.

In the present study, SimScore was developed as a new type of knowledge-based system. The structural similarity search in SimScore utilizes chemical information in the whole structure of a given compound to overcome problems arising from the inconsistent fact data and high chemical diversity of possible teratogens. SimScore allows us to quantitatively predict the teratogenic possibility of a given compound by the structural similarity comparison between it and each human teratogen stored in the knowledge database. Furthermore, the other difference of SimScore, compared to other systems is that it is linked to another knowledge-database system which contains the documentary information such as drug-package inserts, scientific/clinical literatures and physicochemical properties of human teratogens, and so on. SimScore works as a part of our drug safety information community system on the web, as shown in Figure 5. Thus, SimScore guides us in the evaluation of the possible risk of human teratogenicity of a chemical from the comprehensive knowledge of chemistry and clinical fact data. SimScore will be applicable to the evaluation of other specific toxicities and activities of a candidate compound, if the corresponding database is prepared instead of the database of

teratogens. Thus, SimScore could be a potentially useful tool in pharmaceutical R&D and drug therapy. The details of the algorithm used in SimScore will be reported elsewhere.

This research was supported by the Research Institute of Science and Technology for Society, Japan Science and Technology Agency, and Grants-in-Aid for Scientific Research (No. 17590126) from the Ministry of Education, Culture, Sports, Science and Technology, Japan. We wish to thank Mr. Makoto Tani, Drs. A. Ammar Ghaibeh and Hiroki Gotoh in Saila Systems for coding the programs. We also thank Dr. Zsolt Lepp of the University of Tokushima for his linguistic suggestions.

## References

- [1] T. Kennedy, Managing the drug discovery / development interface, *Drug Discovery Today*, **2**, 436-444 (1997).
- [2] H. VD. Waterbeemd, E. Grifford, ADMET in silico Modelling: Towards Prediction Paradise?, *Mature Reviews, Drug Discovery*, **2**, 192-204 (2003).
- [3] N. Greene, Computer systems for the prediction of toxicity: an update, *Adv. Drug Deliv. Rev.*, **54**, 417-431 (2002).
- [4] J. C. Dearden, In silico prediction of drug toxicity, *J. Comput. Aided Mol. Des.*, **17**, 119-127 (2003).
- [5] DEREK, [http://www.lhasalimited.org/index.php?cat=2&sub\\_cat=64](http://www.lhasalimited.org/index.php?cat=2&sub_cat=64)
- [6] HazardExpert, <http://www.compudrug.com/>
- [7] TOPKAT, <http://www.accelrys.com/products/topkat/index.html>
- [8] L. Maślankiewicz, E. M. Hulzebos, T. G. Vermeire, J. J. A. Müller, A. H. Piersma, Can chemical structure predict reproductive toxicity?, *RIVM report* 601200005/2005
- [9] E. M. Hulzebos, R. Posthumus, (Q)SARs: gatekeepers against risk on chemicals?, *SAR QSAR Environ. Res.*, **14**, 285-316 (2003).
- [10] B. Simon-Hettich, A. Rothfuss, T. Steger-Hartmann, Use of computer-assisted prediction of toxic effects of chemical substances, *Toxicology*, **224**, 156-162 (2006).
- [11] A. Yamauchi, H. Chuman, K. Sakamoto, M. Sasaki, A. Ammar Ghaibeh, E. D. Nakata. A method to compute the similarity of chemical structure and to evaluate the safety of compound, and the drug safety information system. Patent application No. 2005348393; Japan.
- [12] A. Yamauchi, D. E. Nakata, H. Chuman, Construction of a growing information-community for teratogenic agents (in Japanese), *J. Comput. Chem. Jpn.*, **2**, 71-78 (2003).
- [13] G. G. Briggs, R. K. Freeman, S. J. Yaffe, editors, *Drugs in pregnancy and lactation*, 7th ed., Williams & Wilkins, Philadelphia (2005), pp.xxi-xxvi.
- [14] MDL mol file format, <http://www.graphics.cornell.edu/online/formats/mol/>
- [15] P. Willett, *Similarity-searching and clustering algorithms for processing databases of two-dimensional and three-dimensional chemical structures*, In *Molecular Similarity in Drug Design*, ed. by Dean, P.M., Blackie Academic and Professional, London (1995), pp.110-137.
- [16] MCASE, <http://www.multicase.com/>
- [17] V. K. Gombar, K. Enslein, B. W. Blake, Assessment of developmental toxicity potential of chemicals by quantitative structure-toxicity relationship models, *Chemosphere*, **31**, 2499-2510 (1995).
- [18] J. L. Schardein, B. A. Schwetz, M. F. Kenel, Species sensitivities and prediction of teratogenic potential, *Environ. Health Perspect*, **61**, 55-67 (1985).