

# Fragment-based prediction of skin sensitization using recursive partitioning

Jing Lu · Mingyue Zheng · Yong Wang ·  
Qiancheng Shen · Xiaomin Luo · Hualiang Jiang ·  
Kaixian Chen

Received: 13 January 2011 / Accepted: 2 September 2011 / Published online: 20 September 2011  
© Springer Science+Business Media B.V. 2011

**Abstract** Skin sensitization is an important toxic endpoint in the risk assessment of chemicals. In this paper, structure–activity relationships analysis was performed on the skin sensitization potential of 357 compounds with local lymph node assay data. Structural fragments were extracted by GASTON (GrAph/Sequence/Tree extractiON) from the training set. Eight fragments with accuracy significantly higher than 0.73 ( $p < 0.1$ ) were retained to make up an indicator descriptor *fragment*. The *fragment* descriptor and eight other physicochemical descriptors closely related to the endpoint were calculated to construct the recursive partitioning tree (RP tree) for classification. The balanced accuracy of the training set, test set I, and test set II in the leave-one-out model were 0.846, 0.800, and 0.809, respectively. The results highlight that *fragment*-based RP tree is a preferable method for identifying skin sensitizers. Moreover, the selected fragments provide

useful structural information for exploring sensitization mechanisms, and RP tree creates a graphic tree to identify the most important properties associated with skin sensitization. They can provide some guidance for designing of drugs with lower sensitization level.

**Keywords** Skin sensitization · LLNA · SAR · Fragment · Recursive partitioning tree · Substructure mining algorithm

## Introduction

Skin sensitization is an immunologic reaction induced by an external substance that is inhaled or penetrates the skin. Local lymph node assay (LLNA) provides a reliable and quantitative criterion for skin sensitization assessment. It locates in the induction phase of skin sensitization and measures the estimated concentration (namely EC<sub>3</sub> values) of test chemicals required to induce a threefold or greater increase in lymph node cell proliferation activity in treated groups relative to the control [1]. Under current regulatory frameworks, LLNA is recommended for using fewer animals and reducing animal pain. In 2002, the LLNA became OECD Guideline 429 [2]. However, with the advent of the EU cosmetics directive [3] and the recently adopted Registration, Evaluation, Authorization and Restriction of Chemicals (REACH) [4–6], there is a drive to develop non-animal alternative methods such as structure–activity relationships (SAR).

SAR approaches are useful for toxicity evaluation because they can not only be used for toxicity prediction, but also provide insights into toxicity mechanisms [7, 8]. However, there are two main factors that may cause difficulty in the models' interpretation. Firstly, models derived from some sophisticated data mining techniques

**Electronic supplementary material** The online version of this article (doi:10.1007/s10822-011-9472-7) contains supplementary material, which is available to authorized users.

J. Lu · M. Zheng · Y. Wang · Q. Shen · X. Luo (✉) ·  
H. Jiang (✉) · K. Chen  
Drug Discovery and Design Center, State Key Laboratory  
of Drug Research, Shanghai Institute of Materia Medica,  
Chinese Academy of Sciences, 555 Zuchongzhi Road,  
Shanghai 201203, China  
e-mail: xmluo@mail.shnc.ac.cn

H. Jiang  
e-mail: hljiang@mail.shnc.ac.cn

H. Jiang  
School of Pharmacy, East China University of Science  
and Technology, Shanghai 200237, China

such as support vector machine (SVM) and artificial neural network (ANN) usually behave like a black box. The relationship between the toxicity and the descriptor is not evident. Secondly, physical meanings of some numeric descriptors are not clear, and it is difficult to provide intuitive structural information for toxicologists and chemists [9, 10]. To avoid these problems, we developed a *fragment*-based skin sensitization prediction model. In this model, molecular fragments of statistical significance were used as indicators, and recursive partitioning tree (RP tree), a readily interpreted modeling method, was used to construct the final SAR models. This method is similar to recursive partitioning in conjunction with binary-valued descriptors as described by Rusinko et al. [11].

Fragments used in this study are molecular functional groups or substructures that can be directly linked to mechanism knowledge of skin sensitization, and thus are more accessible to toxicologists. In the present study, substructure mining algorithm was used to extract discriminating fragments that made up an indicator descriptor. Using this *fragment* descriptor and just a few descriptors closely related to the toxicity mechanism, we built a reliable SAR model using RP tree. RP tree can automatically select the most important descriptors, choose the appropriate cutoff points, and create a graphical tree that can be interpreted as straightforward rules to classify chemicals. In the end, the robustness of the RP model was systematically evaluated by various metrics.

## Materials and methods

### Dataset and descriptors calculation

LLNA data for a set of compounds, which comprise Michael acceptors,  $S_N2$  and  $S_NAr$  electrophiles, Schiff base formers, acyl transfer agents, and so on, were collected from literature [11–22]. The compounds were categorized according to their  $EC_3$  values. The compounds with  $EC_3 \leq 100\%$  (w/v) were classified as sensitizers (255 compounds); otherwise, as non-sensitizers (102 compounds). To validate the prediction ability of our model, we chose 25 compounds which were identical to the Ref. [14] as test set I (20 sensitizers and 5 non-sensitizers). However, test set I possessed a bias with a large ratio of positives to negatives (4:1), so we chose 37 additional test

compounds (19 sensitizers and 18 non-sensitizers) [18–22] as test set II to assess the balanced ability of our model. The remaining compounds were used as training set (216 sensitizers and 79 non-sensitizers). Details of each compound are listed in Supporting Information Table S1, S2, and S3.

The initial structures of compounds were drawn and optimized in Sybyl 6.8 (default parameters) [23]. Further structure optimization was performed using the SAM1/d semi-empirical method implemented in AMPAC 8.16 [24]. Although skin sensitization is a complex toxicological process and its biological processes have not been fully understood, previous studies have indicated that the ability of sensitizers to cause immune response is related to skin permeability, chemical reactivity, and molecular size [1, 25]. Sixteen descriptors reflecting these factors were calculated as shown in Table 1. Twelve descriptors such as *HOMO* and *LUMO* were calculated via Codessa 2.7.2 [26]. AlogP was obtained from Discovery Studio 2.5 [27]. Electronegativity ( $\chi$ ) and electrophilicity ( $\omega$ ) index were calculated using corresponding formula [28]. In addition, fragments were extracted using substructure mining algorithm described in the next section.

### Substructure mining

The types of atoms and bonds which are identical to the definition of Tripos' atom typing scheme are used to characterize compounds. Then, GASTON (GrAph/Sequence/Tree extraction) [29] was employed to extract fragments that satisfy predefined requirements. The process of substructure mining is described as following:

First, the algorithm was used to search for substructures. Examples of fragments include paths, trees, and graphs. The fragments are retained if they satisfy the two following requirements:

- In order to eliminate fragments that are not sufficiently well represented in the population of sensitizers, the frequency threshold that fragments occur in sensitizers is defined as 5% (10 times).
- In the training set, the ratio of sensitizers to compounds is  $216/295 = 0.73$ , so only fragments with accuracy  $Q$  significantly higher than 0.73 (the  $p$  value of binomial distribution is  $<0.1$ ) are selected.

$Q$  is defined as [30]

$$Q = \frac{\text{number of sensitizers containing the fragment}}{\text{number of compounds containing the fragment (sensitizers and non - sensitizers)}} \quad (1)$$

**Table 1** The 16 descriptors closely related to skin sensitization mechanism were selected to construct the classification models

Descriptors	Definitions [26–28]
<i>Fragment</i>	Substructures of compounds
AlogP	Log of the octanol–water partition coefficient using Ghose and Crippen's method
Molecular volume	The 3D volume for each molecule using the current 3D coordinates
<i>HOMO</i>	Energy of the highest occupied molecular orbital
<i>LUMO</i>	Energy of the lowest unoccupied molecular orbital
<i>HOMO–LUMO</i> energy gap	<i>HOMO–LUMO</i>
WPSA-3	Surface weighted CPSA (PPSA-3*TMSA/1000) [Quantum-chemical PC]
FPSA-3	Fractional CPSA (PPSA-3/TMSA) [Quantum-chemical PC]
DPSA-3	Difference in CPSAs (PPSA3-PNSA3) [Quantum-chemical PC]
RPCG	Relative positive charge [Quantum-chemical PC]
RPCS	Relative positive charged surface area (SAMPOS * RPCG) [Quantum-chemical PC]
Max partial charge for a C atom	Most positive atomic partial charges for a C atom [Zefirov's PC]
Max bond order of a C atom	Max bond order of a C atom
Max atomic state energy for a C atom	Max atomic valence state energy for a C atom
Electronegativity ( $\chi$ )	A measure of an atom to attract electrons towards itself [ $\chi = -1/2(HOMO + LUMO)$ ]
Electrophilicity index ( $\omega$ )	A measure of the second-order energy that an electrophile gets saturated with electrons. [ $\omega = \mu^2/2\eta$ ]

Altogether 48 fragments passed the above filtering procedure (see Table S4 in Support Information). Further investigation of these structures revealed that some fragments were substructures of others. To avoid overlapping issue in structural matching, only the non-redundant and smallest fragments were kept for making up the indicator descriptor *fragment*, which included 8 individual fragments. The indicator takes a value of 1 if at least one of the 8 fragments is present in the molecule; otherwise it takes 0. Following this definition, the compounds were examined on whether they contained any of the 8 fragments, and the value of *fragment* for each compound was obtained accordingly.

The substructure mining algorithm is only based on statistics, and it does not rely on a priori knowledge of mechanism of action.

#### Recursive partitioning tree

Recursive partitioning method implemented in Pipeline Pilot (version7.5) [31] was used to build the classification model. A binary-classification tree consists of internal nodes and leafs: each internal node corresponds to a test on one descriptor and each leaf is assigned a classification label. *Gini* index, a measure for splitting criteria, is defined as:

$$Gini = \sum_{j \neq k} p_{ij} p_{ik} \quad (2)$$

where  $p_{ij}$  and  $p_{ik}$  are the proportions of  $j$  class and  $k$  class in node  $i$ , respectively [32]. At each node, the algorithm looks

for descriptor and cutoff point minimizing the *Gini* index to divide the compounds into branches. The splitting process continues until no more significant nodes are obtained or when a minimum number of samples per node are reached. In order to get better results and reduce the tree complexity, 'Minimum number of samples per node' and 'Maximum Tree Depth' were set to 6 and 5, respectively. Moreover, leave-one-out cross-validation was used for estimating the error rates of the tree.

#### Models evaluation

The performance of the prediction model was measured by sensitivity, specificity, and balanced accuracy. Sensitivity and specificity can assess the prediction ability of the model to correctly identify sensitizers and non-sensitizers, and are calculated as:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (3)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (4)$$

where  $TP$  is true positives,  $TN$  is true negatives,  $FP$  is false positives, and  $FN$  is false negatives. Accuracy is a common index for the overall classification performance, but it may be misleading if the class distribution of used datasets is highly unbalanced [33]. In our training set, the ratio of positive to negative compounds is about 3:1, so we replace accuracy with balanced accuracy, which is equivalent to the index proposed by Youden [34], as a performance measurement.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (5)$$

$$\text{Balanced accuracy} = \frac{(\text{sensitivity} + \text{specificity})}{2} \quad (6)$$

## Results and discussion

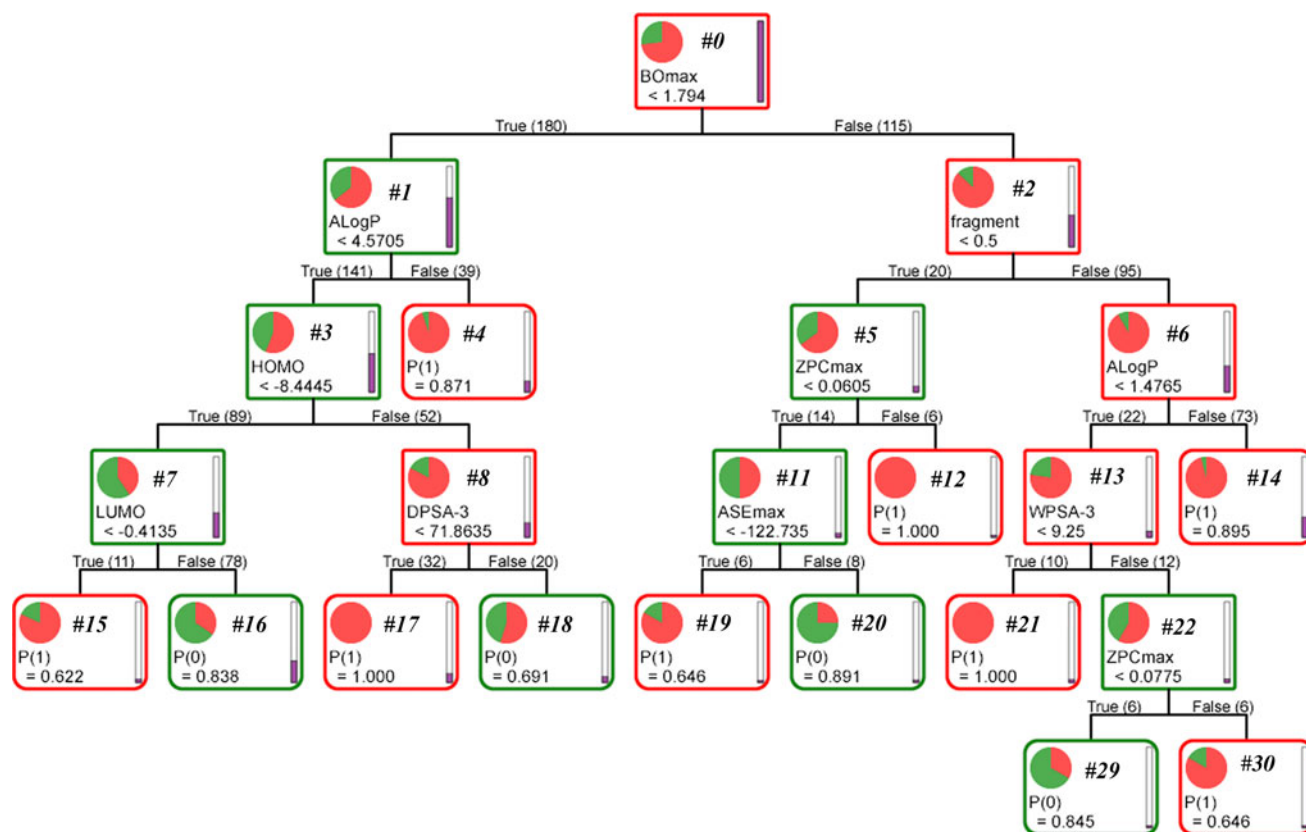
### Performance of models

Figure 1 shows the constructed *fragment*-based RP tree (FG-RP tree) with 9 selected descriptors (The RP tree without *fragment* is shown in Figure S1 in Support Information). In this tree, max bond order of a C atom (hereafter called ‘BOmax’), AlogP, *fragment*, HOMO, max partial charge for a C atom (‘ZPCmax’), LUMO, DPSA-3, WPSA-3, and max atomic state energy for a C atom (‘ASEmax’) played important roles to distinguish sensitizers from non-sensitizers.

Li et al. investigated expert-system (Derek for Windows), logistic regression [25], Fisher’s linear discriminant analysis (LDA) and SVM with different kernel functions

(linear, cubic, RBF and sigmoid) for skin sensitization. RBF-SVM with 23 descriptors and linear-SVM with 24 descriptors were deemed to provide preferable performance [14]. Thus we listed results of these two best models to compare our FG-RP model in Table 2. Moreover, in order to investigate the importance of fragments, we also constructed an RP model without the *fragment* (NFG-RP) with the same training set and test sets.

The results shown in Table 2 demonstrate that our FG-RP model outperforms the reference models. Firstly, the SVM models have a low specificity contrast to high sensitivity (0.766 to 0.931 and 0.745 to 0.916), which may have resulted from the substantially unbalanced number of positive and negative compounds [25]. Likewise, NFG-RP tree exhibits unbalanced performance (0.962–0.662). By contrast, *fragment* helps our FG-RP tree model to overcome the difficulty and obtain a better balance (0.886–0.806). Secondly, there are large drops of balanced accuracy in the test set I for both SVM models, which indicate that these models are overtrained. The FG-RP model obtains better results compared to the two models. NFG-RP tree cannot yet balance sensitivity and specificity (0.6 and 1) in spite of



**Fig. 1** The simplified figure of the FG-RP tree with 23 nodes and 9 descriptors. In nodes, pie chart colors show the class: positive (red), negative (green). *p* values are for winning class and reflect sample weights. 1 means positive and 0 means negative. In the model, the training set contains 216 positives and 79 negatives, so each positive

sample is assigned a weight of 79 and each negative sample is assigned a weight of 216. For example, in node 4 (#4), there are 37 positives and 2 negatives, so *p*(1) can be calculated by the algebraic expression:  $79 \cdot 37 / (216 \cdot 2 + 79 \cdot 37)$

**Table 2** Performance of RP tree models versus the two best models in the Ref. [14]

	Methods	TP	FN	TN	FP	Balanced accuracy	Sensitivity	Specificity	No of descriptors
Training set	SVM23 <sup>a</sup>	122	9	36	11	0.849	0.931	0.766	23
	SVM24 <sup>b</sup>	120	11	35	12	0.830	0.916	0.745	24
	FG-RP	174	42	70	9	0.846	0.806	0.886	9
	NFG-RP	143	73	76	3	0.812	0.662	0.962	10
Test set I	SVM23	12	8	3	2	0.600	0.600	0.600	
	SVM24	14	6	3	2	0.650	0.700	0.600	
	FG-RP	16	4	4	1	0.800	0.800	0.800	
	NFG-RP	12	8	5	0	0.800	0.600	1.000	
Test set II	SVM23	–	–	–	–	–	–	–	
	SVM24	–	–	–	–	–	–	–	
	FG-RP	17	2	13	5	0.809	0.895	0.722	
	NFG-RP	9	10	16	2	0.682	0.474	0.889	

<sup>a</sup> SVM23 means RBF-SVM with 23 descriptors<sup>b</sup> SVM24 means linear-SVM with 24 descriptors

good balanced accuracy. Finally, AlogP, *HOMO*, and *LUMO* selected from our models highlight the importance of hydrophobicity and reactivity, which are relevant to the ability of a chemical to react covalently with the protein [35]. However, the numbers of compounds of these two training sets have slight difference, so we made a more direct comparison: FG-RP model and the SVM models based on the reference's smaller training set. Our model indeed shows better prediction result: the balanced accuracy of training set is 0.869 of our model versus 0.849 of the SVM model and test set I is 0.725 versus 0.650. Moreover, the training set and test set I possess the same imbalance, so we prepared a set of balanced compounds as test set II to validate the models against a balanced dataset. From the results, the *fragment* descriptor significantly helped to increase sensitivity, and thus better balanced sensitivity and specificity. However, specificity dropped somewhat, a possible reason was that the smallest fragments selected matched too many compounds and increased false positives. Another advantage of fragments is that they are easier to understand by toxicologists and provide a platform for dialogue between model developers and toxicologists.

#### Eight extracted fragments by substructure mining

The eight fragments involve Michael addition, nucleophilic substitution, acylation, and oxophilic  $S_N2$ . In Table 3, we illustrate a few sensitizers associated with eight fragments and sensitization-dependent mechanisms [12, 36–39].

In addition, we can predict some activity trends of sensitizers from the linking group of fragments. For instance, the sensitization level of simple alkyl halides is highly dependent on their hydrophobicity, with longer chain alkyl (e.g., fragment 6) halides being stronger sensitizers (this

relationship holds true until hydrophobicity becomes so large that it prevents penetration through the skin and thus the sensitizing potency of very long chain alkyl halides diminishes quickly) [40]. Moreover, when the carbon atom attaches an electron-withdrawing group (e.g., fragment 5),  $S_N2$  electrophiles (e.g., fragment 4) would be easier to be attacked by nucleophilic protein as compared to an attached alkyl chain (e.g., fragment 6). For example, 2-bromotetradecanoic acid has much stronger sensitizing effect than 1-bromotridecane due to an extra carboxyl group on the bromo-substituted carbon atom.

At the same time, it should be noted that the above mentioned 'fragments-reaction' relationships are interchangeable. For example, for fragment 1, if an  $\alpha,\beta$ -unsaturated carbonyl has steric hindrance effects (e.g., conformational constraint or presence of methyl groups at the site of nucleophilic attack), the compound would be prone to react by Schiff base reaction [41]. Farnesal and citral that match the pattern are classified to Schiff base domain [12].

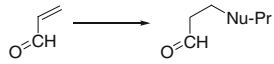
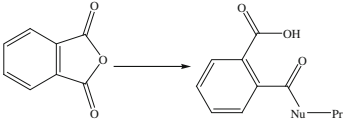
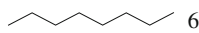
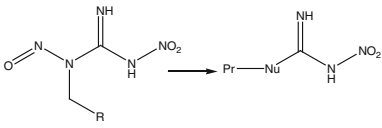
Although trigonal planar nitrogen  $N_{(p13)}$  is less common in compounds, all 11 compounds containing C– $N_{(p13)}$  are sensitizers in our study. According to the conditions set in the substructure mining ( $Q > 0.73$  ( $p < 0.1$ )), C– $N_{(p13)}$  remained.

#### Eight physicochemical descriptors selected to skin sensitization

In the process of substructure mining, we selected the eight smallest fragments to construct the *fragment* descriptor. But the smallest fragments may result in the model's overgeneralization. For example, Michael acceptors that have an electron-deficient double bond are susceptible to nucleophilic attack, but not all compounds with an alkene



**Table 3** Eight fragments selected from substructure mining, related reaction mechanisms and some compound-examples

Fragments-reaction	Potential Reaction	Example
$\text{C}=\text{C}$ 1 $\text{C}_2-\text{C}_2$ a 2	 [36]	5,5-dimethyl-3-methylenedi hydro-2(3H)-furanone [12], hexyl cinnamic aldehyde [12], p-benzoquinone [12]
$\text{C}_2-\text{O}-\text{C}_2$ 3	 [37]	phthalic anhydride [16], oxazolone and alkyl azlactones, which can be regarded as aza-analogues of cyclic anhydrides [12].
$\text{C}-\text{Br}$ 4 $\text{C}_2$ 5  6	$\text{X}-\text{C}-\text{C}-\text{Nu-Pr}$ X=halogen or other leaving group [7]	2-bromotetradecanoic acid [12], 1-bromohexadecane [12], 1-bromotridecane [12]
$\text{C}=\text{N}$ 7	 c.[38]	1-methyl-3-nitro-1-nitrosog uanidine [12], 1-Ethyl-3-nitro-1-nitrosogu anidine [11], 1-Propyl-3-nitro-1-nitrosog uanidine [11]
$\text{C}-\text{N}(\text{pl}_3)$ b 8		

<sup>a</sup>  $\text{C}_2$  means the  $\text{sp}^2$  carbon

<sup>b</sup>  $\text{N}_{(\text{pl}_3)}$  means trigonal planar nitrogen [39]

<sup>c</sup> Although the reactivity of these compounds toward DNA and proteins is not doubtful, the underlying mechanistic chemistry is quite complex [38]. In this table, only one possibility was listed for protein binding

( $\text{C}=\text{C}$ ) are prone to induce sensitization. Moreover, the chemical function groups only reflect a trend of being sensitizers, and it is unlikely that a single fragment will strictly define the activity of a molecule [42]. Thus it is necessary to consider physicochemical properties of chemicals.

AlogP is a measure of molecular hydrophobicity, and it is important for absorption, permeability, and distribution of chemicals in vivo [43]. It also models the bioavailability at the location where the protein-binding reaction leads to sensitization [35]. Thus AlogP is a common physicochemical descriptor in QSAR studies of Skin Sensitization [38, 44, 45]. In our RP tree, AlogP occurred twice, which illustrated the descriptor's significance.

*HOMO* and *LUMO* are the criteria to evaluate the ability of a molecule to undergo nucleophilic and electrophilic reactions, respectively [2]. If the toxic process is initiated by a nucleophilic attack on the chemical to form covalent binding, *HOMO* could also determine whether this is a favorable process [2].

Bond order is the number of chemical bonds between a pair of atoms, which indicates the stability of a chemical bond. When the value of 'BOmax' was <1.794, 180 chemicals were classified correctly. The application of this descriptor manifested indirectly that the importance of bond fission was favorable for skin sensitization.

Partial atomic charge plays an important role to analyze the polarization effects of molecules and the electrostatic interactions in chemical reactions [46]. In our model, 'ZPCmax' based on Zefirov's approach [47, 48] was used. The appearance of 'ZPCmax' hinted that apart from covalent-mode, skin sensitization may be related to some weak interactions. Some related references [49–51] also made similar statements.

Charged partial surface areas (CPSAs) is a hybrid set of descriptors that combine the surface areas of the geometric descriptors with partial charges of the electronic descriptors [52]. DPSA-3 and WPSA-3 encode the features responsible for polar interactions between sensitizer and

**Table 4** Correlation analysis of 8 physicochemical descriptors

	BOmax	AlogP	HOMO	LUMO	ZPCmax	DPSA-3	WPSA-3	ASEmax
BOmax	1							
AlogP	−0.019	1						
HOMO	−0.003	−0.067	1					
LUMO	−0.316	−0.106	0.218	1				
ZPCmax	0.550	−0.145	−0.086	−0.284	1			
DPSA-3	0.023	0.179	0.008	−0.218	0.222	1		
WPSA-3	−0.096	0.548	0.012	0.008	−0.044	0.732	1	
ASEmax	0.076	−0.102	−0.119	−0.287	0.129	−0.150	−0.125	1

protein, which indicate that non-covalent interactions are also important for skin sensitization.

The descriptor ‘ASEmax’ describes the atomic valence state energy of a carbon atom in the molecule. The atomic valence state energy characterizes the perturbation magnitude of an atom in the molecule as compared to be an isolated atom [26].

Moreover, correlation matrix of these eight physicochemical descriptors in the FG-RP model is listed in Table 4. As shown, the correlation coefficient value of each pair of descriptors was less than 0.75, which meant that the selected descriptors were independent.

From the above discussion, we can conclude that the factors influencing the potency of skin sensitization mainly include reactivity, hydrophobicity and electrostatic interactions. The descriptors in the developed model are closely related to these properties and provide insights into the sensitization mechanisms. Moreover, *fragment* describes a set of substructures affecting allergic reaction and provides useful structural information to chemists.

## Conclusions

The goal of this study was to develop an effective model to distinguish sensitizers from non-sensitizers. FG-RP model highlights some advantages through several aspects: Firstly, the dataset which is not limited to a single mechanism of action has more practical application. Secondly, the *fragment* descriptor intuitively helps toxicologists and chemists to investigate related mechanisms. Thirdly, the FG-RP model provides the relative important descriptors to skin sensitization and the criteria for developing SAR models. Finally, the model is more reliable compared to the previously published models. Thus we expect the developed model to be a useful tool to identify structural features of sensitizers and provide some instructions for designing of drugs with lower sensitization level.

**Acknowledgments** This work was supported by Hi-TECH Research and Development Program of China (Grant 2006AA020402), National

S&T Major Project (Grants 2009ZX09301-001, 2009ZX09501-001), and the State Key Program of Basic Research of China (Grant 2009CB918502).

## References

- Golla S, Madhally S, Robinson RL Jr, Gasem KA (2009) Quantitative structure–property relationship modeling of skin sensitization: a quantitative prediction. *Toxicol In Vitro* 23:454–465
- Warne MA, Nicholson JK, Lindon JC, Guiney PD, Gartland KP (2009) A QSAR investigation of dermal and respiratory chemical sensitizers based on computational chemistry properties. *SAR QSAR Environ Res* 20:429–451
- EU, Directive 2003/15/EC of the European Parliament and the Council of 27 February 2003 Amending Council Directive 76/768/EEC on the Approximations of Laws of the Member States Relating to Cosmetic Products. *Official Journal of the European Union, L: Legislation* 66:26–35. <http://www.eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2003:066:0026:0035:en:PDF>
- Commission of the European Communities, White Paper on the Strategy for a Future Chemicals Policy. COM 88 (2001) Brussels, Belgium. <http://www.europa.eu.int/comm/environment/chemicals/whitepaper.htm>
- Commission of the European Communities, Proposal for a Regulation of the European Parliament and of the Council Concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), Establishing a European Chemicals Agency and Amending Directive 1999/45/EC and Regulation (EC) on Persistent Organic Pollutants (2003). <http://www.europa.eu.int/comm/enterprise/chemicals/chempol/whitepaper/reach.htm>
- Commission of the European Communities, Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, Amending Directive 1999/45/EC and Repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC. *Official Journal of the European Union, L: Legislation* 396:49:1–849 (pp 14, 374, 375). [http://www.europa.eu.int/comm/enterprise/reach/index\\_en.htm](http://www.europa.eu.int/comm/enterprise/reach/index_en.htm)
- Patlewicz G, Aptula AO, Roberts DW, Uriarte E (2008) A minireview of available skin sensitization (Q)SARs/Expert systems. *QSAR Comb Sci* 27:60–76
- Netzeva T, Pavan M, Worth AP (2008) Review of (quantitative) structure–activity relationships for acute aquatic toxicity. *QSAR Comb Sci* 27:77–90

9. The CAESAR model for skin sensitization. <http://www.caesar-project.eu/workshop/info.htm>, December 5, 2010
10. Yuan H, Huang J, Cao C (2009) Prediction of skin sensitization with a particle swarm optimized support vector machine. *Int J Mol Sci* 10:3237–3254
11. Rusinko A 3rd, Farmen MW, Lambert CG, Brown PL, Young SS (1999) Analysis of a large structure/biological activity data set using recursive partitioning. *J Chem Inf Comput Sci* 39(6): 1017–1026
12. Roberts DW, Patlewicz G, Kern PS, Gerberick F, Kimber I, Dearman RJ, Ryan CA, Basketter DA, Aptula AO (2007) Mechanistic applicability domain classification of a local lymph node assay dataset for skin sensitization. *Chem Res Toxicol* 20:1019–1030
13. Gerberick GF, Vassallo JD, Foertsch LM, Price BB, Chaney JG, Lepoittevin JP (2007) Quantification of chemical peptide reactivity for screening contact allergens: a classification tree model approach. *Toxicol Sci* 97:417–427
14. Li S, Fedorowicz A, Andrew ME (2007) A new descriptor selection scheme for SVM in unbalanced class problem: a case study using skin sensitisation dataset. *SAR QSAR Environ Res* 18:423–441
15. Miller MD, Yourtee DM, Glaros AG, Chappelow CC, Eick JD, Holder AJ (2005) Quantum mechanical structure–activity relationship analyses for skin sensitization. *J Chem Inf Model* 45:924–929
16. Natsch A, Emter R, Ellis G (2009) Filling the concept with data: integrating data from different in vitro and in silico assays on skin sensitizers to explore the battery approach for animal-free skin sensitization testing. *Toxicol Sci* 107:106–121
17. Patlewicz G, Dimitrov SD, Low LK, Kern PS, Dimitrova GD, Comber MI, Aptula AO, Phillips RD, Niemela J, Madsen C, Wedebye EB, Roberts DW, Bailey PT, Mekenyan OG (2007) TIMES-SS—a promising tool for the assessment of skin sensitization hazard. A characterization with respect to the OECD validation principles for (Q)SARs and an external evaluation for predictivity. *Regul Toxicol Pharmacol* 48:225–239
18. <http://www.toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?HSDB>, May 23, 2011
19. <http://www.inchem.org/pages/sids.html>, May 24, 2011
20. Kreiling R, Hollnagel HM, Hareng L, Eigler D, Lee MS, Griem P, Dreessen B, Kleber M, Albrecht A, Garcia C, Wendel A (2008) Comparison of the skin sensitizing potential of unsaturated compounds as assessed by the murine local lymph node assay (LLNA) and the guinea pig maximization test (GPMT). *Food Chem Toxicol* 46(6):1896–1904
21. Bergstrom MA, Andersson SI, Broo K, Luthman K, Karlberg AT (2008) Oximes: metabolic activation and structure–allergenic activity relationships. *J Med Chem* 51(8):2541–2550
22. Fedorowicz A, Singh H, Demchuk E (2004) QSAR study of skin sensitization using local lymph node assay data. *Int J Mol Sci* 5:56–66
23. Sybyl, Tripos Inc.: St. Louis, MO 63144-2913
24. AMPAC, Semichem, Inc.: Shawnee, KS 66216
25. Fedorowicz A, Singh H, Soderholm S, Demchuk E (2005) Structure–activity models for contact sensitization. *Chem Res Toxicol* 18:954–969
26. Comprehensive Descriptors for Structural and Statistical Analysis (CODESSA), Semichem, Inc.: Shawnee, KS 66216
27. Accelrys Discovery Studio, Accelrys Software Inc.: San Diego, CA
28. Chattaraj PK, Sarkar U, Roy DR (2006) Electrophilicity index. *Chem Rev* 106:2065–2091
29. Nijssen S, Kok J (2004) A quickstart in frequent structure mining can make a difference. In: 10th ACM SIGKDD international conference on knowledge discovery and data mining (KDD), pp 647–652
30. Kazius J, Nijssen S, Kok J, Back T, Ijzerman AP (2006) Sub-structure mining using elaborate chemical representation. *J Chem Inf Model* 46:597–605
31. Pipeline Pilot, Accelrys Software Inc.: San Diego, CA
32. Li S, Fedorowicz A, Singh H, Soderholm SC (2005) Application of the random forest method in studies of local lymph node assay based skin sensitization data. *J Chem Inf Model* 45:952–964
33. Zheng M, Luo X, Shen Q, Wang Y, Du Y, Zhu W, Jiang H (2009) Site of metabolism prediction for six biotransformations mediated by cytochromes P450. *Bioinformatics* 25:1251–1258
34. Youden WJ (1950) Index for rating diagnostic tests. *Cancer* 3(1):32–35
35. Roberts DW, Aptula AO, Patlewicz G, Pease C (2008) Chemical reactivity indices and mechanism-based read-across for non-animal based assessment of skin sensitisation potential. *J Appl Toxicol* 28:443–454
36. Langton K, Patlewicz GY, Long A, Marchant CA, Basketter DA (2006) Structure–activity relationships for skin sensitization: recent improvements to Derek for Windows. *Contact Derm* 55:342–347
37. Enoch SJ, Roberts DW, Cronin MT (2009) Electrophilic reaction chemistry of low molecular weight respiratory sensitizers. *Chem Res Toxicol* 22:1447–1453
38. Roberts DW, Aptula AO, Patlewicz G (2007) Electrophilic chemistry related to skin sensitization. Reaction mechanistic applicability domain classification for a published data set of 106 chemicals tested in the mouse local lymph node assay. *Chem Res Toxicol* 20:44–60
39. [http://www.sdsc.edu/CCMS/Packages/cambridge/pluto/atom\\_types.html](http://www.sdsc.edu/CCMS/Packages/cambridge/pluto/atom_types.html)
40. Gerberick GF, Ryan CA, Kern PS, Schlatter H, Dearman RJ, Kimber I, Patlewicz GY, Basketter DA (2005) Compilation of historical local lymph node data for evaluation of skin sensitization alternative methods. *Dermatitis* 16:157–202
41. Patlewicz GY, Wright ZM, Basketter DA, Pease CK, Lepoittevin JP, Arnau EG (2002) Structure–activity relationships for selected fragrance allergens. *Contact Derm* 47:219–226
42. Burton J, Danloy E, Vercauteren DP (2009) Fragment-based prediction of cytochromes P450 2D6 and 1A2 inhibition by recursive partitioning. *SAR QSAR Environ Res* 20:185–205
43. Gulyaeva N, Zaslavsky A, Chait A, Zaslavsky B (2001) Measurement of the relative hydrophobicity of organic compounds without organic solvent Effects of salt composition and pH on organic acids and nonionic compounds. *J Pharm Sci* 90(9): 1366–1374
44. Roberts DW, Aptula AO, Patlewicz G (2006) Mechanistic applicability domains for non-animal based prediction of toxicological endpoints. QSAR analysis of the schiff base applicability domain for skin sensitization. *Chem Res Toxicol* 19:1228–1233
45. Roberts DW, Basketter DA (2000) Quantitative structure–activity relationships: sulfonate esters in the local lymph node assay. *Contact Derm* 42:154–161
46. Lee JG, Jeong HY, Lee H (2003) An efficient method to compute partial atomic charges of large molecules using reassociation of fragments. *Bull Korean Chem Soc* 24:369–376
47. Zefirov NS, Kirpichenok MA, Ismailov FF, Trofimov MI (1987) Calculation schemes for atomic electronegativities in molecular graphs within the framework of Sanderson principle. *Dokl Akad Nauk SSSR* 296:883–887
48. Kirpichenok MA, Zefirov NS (1987) Electronegativity and molecular geometry. I. General principles of the method and analysis of the effect of short-range electrostatic interactions



- on bond lengths in organic molecules. *Zh Org Khim* 23: 673–691
49. Basketter D, Dooms-Goossens A, Karlberg AT, Lepoittevin JP (1995) The chemistry of contact allergy: why is a molecule allergenic? *Contact Derm* 32:65–73
50. Divkovic M, Pease CK, Gerberick GF, Basketter DA (2005) Hapten-protein binding: from theory to practical application in the in vitro prediction of skin sensitization. *Contact Derm* 53:189–200
51. Pichler WJ (2001) Predictive drug allergy testing: an alternative viewpoint. *Toxicology* 158:31–41
52. Stanton DT, Jurs PC (1990) Development and use of charged partial surface area structural descriptors in computer-assisted quantitative structure–property relationship studies. *Anal Chem* 62:2323–2329