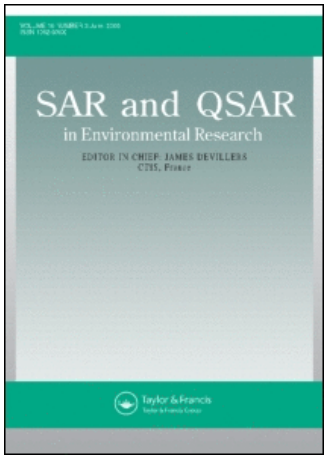

SAR and QSAR in Environmental Research
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t716100694

How can structural similarity analysis help in category formation?

To cite this Article: Jaworska, J. and Nikolova-Jeliazkova, N. , 'How can structural similarity analysis help in category formation? ', SAR and QSAR in Environmental Research, 18:3, 195 - 207
To link to this article: DOI: 10.1080/10629360701306050
URL: http://dx.doi.org/10.1080/10629360701306050

SAR and QSAR in Environmental Research,
Vol. 18, Nos. 3–4, May–June 2007, 195–207

How can structural similarity analysis help in category formation?§

J. JAWORSKA*† and N. NIKOLOVA-JELIAZKOVA‡

†Procter & Gamble, Eurocor, Central Product Safety, 100 Temselaan,
1853 Strombeek-Bever, Belgium
‡Institute for Parallel Processing, Bulgarian Academy of Sciences, Sofia, Bulgaria

(*Received 19 May 2006; in final form 3 October 2006*)

Chemical category is a regulatory concept facilitating filling safety data gaps. Practically, all chemical management programs like the OECD HPV Program, EU REACH, or the Canadian DSL Categorization are planning to use or are already using categorization approaches to reduce resources including animal testing. The aim of the study was to discuss the feasibility to apply computational structural similarity methods to augment formation of a category. The article discusses also how this understanding can be translated into computer readable format, an ultimate need for practical, broad scope applications. We conclude that for the skin sensitization endpoint, used as a working example, mechanistic understanding expressed as chemical reactivity can be exploited by computational structural similarity methods to augment category formation process. We propose a novel method, atom environments ranking (AER), to assess similarity to a reference training set representing a common mechanism of action, as a potential method for grouping chemicals into reactivity domains.

Keywords: Category formation; Similarity assessment; Mechanism of action; Atom environments; Skin sensitization

1. Introduction

A number of chemical categories have already been successfully assessed within the OECD HPV program and more are in preparation [1]. Practically, all chemical management programs including EU REACH and Canadian DSL Categorization are planning to apply categorization approaches to use resources more efficiently and reduce animal testing when filling in data gaps. This is potentially a big gain for the industry. The current definitions of category allow for a variety of approaches for its formation. Specifically, in the OECD Manual for Investigation of HPV Chemicals Chapter 3, section 2 we read: '*A chemical category is a group of chemicals whose*

*Corresponding author. Email: jaworska.j@pg.com
§Presented at the 12th International Workshop on Quantitative Structure-Activity Relationships in Environmental Toxicology (QSAR2006), 8–12 May 2006, Lyon, France.

SAR and QSAR in Environmental Research
ISSN 1062-936X print/ISSN 1029-046X online
DOI: 10.1080/10629360701306050

physicochemical and toxicological properties are likely to be similar or follow a regular pattern as a result of structural similarity, these structural similarities may create a predictable pattern in any or all of the following parameters: physicochemical properties, environmental fate and environmental effects, and human health effects.' Similar definitions can be found in Annex IX of REACH guidance documents: '*Substances for which physicochemical, toxicological and ecotoxicological properties are likely to be similar or follow a regular pattern as a result of structural similarity may be considered as a group, or ''category'' of substances*'.

According to the OECD manual for investigation of HPV chemicals (manual currently in revision by OECD Existing Chemicals Task Force) the rationales for a category include: a common functional group (e.g., aldehyde, epoxide, ester, metal ion), an incremental and constant change across the category (e.g., a chain-length category), the likelihood of common precursors and/or breakdown products via physical or biological processes which result in structurally similar chemicals (''metabolic pathway approach'') and, explored to a lesser extent, common mechanism of action.

In specific cases, constitution of a category can be complicated and boundaries of a category are not straightforward. To improve sharing experience, and to build confidence and consensus among various stakeholders, an OECD workshop on forming categories was held in Brussels in January, 2004 [2]. Several case studies were presented with a focus on key issues related to formation of a category. Fragment similarity was identified as the leading approach used. Workshop participants recommended the development of other approaches to improve scientific basis of a category such as dividing category into subcategories according to properties and endpoints, relating observed properties to an underlying theory (mechanism of action) to increase confidence in category and, if justified, allowing for extrapolation, and exploring the use of computational chemistry/QSARs for the category justification, specifically to help to formalize categories and its boundaries. It was also noted when establishing trends within a category, that there is a need to consider underlying variability in experimental data and that it is easier to identify outliers if the category contains many substances with data. Thus, developing approaches allowing large categories was recommended, if possible.

In this article we attempt to address these recommendations by exploring the feasibility to use computational structural similarity to group chemicals into reactivity domains to augment formation of a category. Until now, assessment of reactivity domain required human expert decision and reference sets for different reactivity domains were constructed manually for skin sensitizers [3]. We propose a novel method to convert reference sets derived with expert knowledge into computer readable format and use structural similarity methods to assess similarity of a new chemical to a particular reference set or reactivity domain. Ability to capture expert knowledge into computer readable format is critical for practical broad scale applications. In addition it adds transparency to the process. As a working example, we discuss the skin sensitization endpoint.

1.1 *Current use of structural similarity concepts in toxicology*

Since the definitions of category frequently mention structural similarity, our intention was to explore a well-developed field in computational chemistry, structural similarity assessments for its application to forming categories. The objective of assessing

chemical similarity in toxicology is to automatically determine the chemicals with similar biological activities. It is based on the hypothesis that similar compounds have similar biological activities. The hypothesis has numerous supportive, but also many contradictory examples, since there are cases where a small change in chemical structure leads to a drastic change in the biochemical activity ('similarity paradox') [4, 5]. Recall from philosophy that similarity is not an absolute concept, but a relative one, and this has important consequences to the precise definition of chemical similarity. Thus, as two objects cannot be similar in absolute terms, but only with respect to a property of the object, two chemicals cannot be similar *per se*, but only with respect to some measurable key feature [6, 7]. The above mentioned 'similarity paradox' suggests that there exists no single universally appropriate similarity measure, but rather its choice depends on the particular endpoint.

The idea of customizing similarity measures to specific activities is becoming increasingly accepted. Rosenkranz and Cunningham demonstrated that traditional organic chemical classes are not able to separate chemicals into toxic and non-toxic groups [8]. Others make use of various supervised learning techniques in order to select the relevant chemical representations and classify toxicity classes [9, 10]. Genetic algorithm was employed [9] to derive weights for the edit distance that is used as a similarity measure between reduced graph representations of chemical structures. When the algorithm is trained on multiple activity classes, different similarity measures with respect to the activity are obtained. Basak presents a comparison of property-based, arbitrary, and 'tailored' similarity spaces [10]. 'Tailored' space is a space comprising of properly selected descriptors, in this case via ridge regression, and is reported to be superior to the arbitrary similarity methods in log K_{ow} estimation [11]. Another example of ''tailored'' similarity, this time structural, is the method for extracting toxicity indicating structural patterns based on fragment frequency [12].

An approach regaining attention recently and one that could be used for ''tailoring'' structural similarity to the endpoint is one based on mechanism of action. This development is interesting for potential applications in category formation because it would allow formulating a toxicological hypothesis, a possibility to group initially larger number of chemicals, use more available data and overall provide higher reliability for a category. Research regarding how precisely a mechanism of action needs to be defined for computerized application is ongoing and the chemical reactivity basis, type of interactions with proteins, -omics data are among many possibilities.

2. Methods

2.1 *The skin sensitization data*

The data set used for the analysis is the local lymph node assay (LLNA) data set consisting of 211 chemicals [13]. These chemicals were rated based on their EC3% potency into 5 classes: non-sensitizers and weak, moderate, strong and extreme sensitizers. Of these, Aptula *et al.* [3] analyzed a subset of 41 compounds to derive reaction mechanistic applicability domains for structure-activity relationships. Based on expert knowledge, they grouped chemicals according to chemical reactivity and defined 5 reactivity domains: Michael acceptors (MA), Schiff base formers (SCH), acylating

agents (Acyl), SN2 electrophiles (SN2) and no evident reactivity mechanism (DIV). The chemicals assigned to each group are listed in table 1.

Until now, efforts to analyze the LLNA data set were focused on developing binary classification models to separate non-sensitizers from sensitizing chemicals [14, 15]. Li *et al*. [14] developed a binary classification model for differentiating strong sensitizers *versus* non-sensitizers using a random forest modeling approach and GETAWAY descriptors, while Devillers [15] developed a neural network with K_{ow}, one topological index and 12 structural alerts to separate non-sensitizers from all other sensitizer classes. Both models achieved high predictivity of 80% [14] and 83–94% [15] in the internal validation. Since contacts allergens vary substantially with regard to the relative potency with which they are able to induce skin sensitization there is a growing consensus that consideration of potency will become a significant factor in the classification of skin sensitizers [16] and as a precedent a weak sensitizers was accepted in a detergent [17]. To meet these needs, Dimitrov *et al*. [18] proposed separating chemicals into non-, weak and strong sensitizers (the last group included moderate, strong and extreme sensitizers) into a model TIMES. The authors used a complex methodology, including skin metabolism simulator, conformational analysis, pattern recognition and 3D QSAR using physicochemical, steric, and electronic parameters. However, such a complex procedure, based on mechanistic rationale, still does not provide much better results compared [14, 15] yielding correct predictions for 70% non-sensitizers, 20–40% for weak sensitizers, and 80% for other sensitizers (moderate, strong and extreme). In our analyses we will attempt to explain these results by studying chemical structures of the LLNA set [13].

2.2 *Structural similarity methods used in the study*

The variety of numerical representations (graphs, descriptors, wave function, etc.) of molecular structure and ways to define a comparative measure between them resulted in several approaches to measure similarity between chemical compounds. In this study, we used Daylight-like fingerprints combined with Tanimoto coefficient (DFT), Atom Environments combined with Tanimoto (AET) and Atom Environments Ranking based on random walk theory (AER). The details of each approach are described below.

2.2.1 Daylight-like fingerprints combined with Tanimoto. Fingerprint generation was based on the fingerprint implementation by open source cheminformatics library, The Chemistry Development Kit (CDK), [19] and follows the ideas of Daylight fingerprint theory [20] that states: (1) for a given molecule all possible paths for a predefined length (default is 7) are generated, (2) the path is submitted to a hash function which uses it as a seed to a pseudo-random generator, (3) the hash function outputs a set of bits, and (4) the set of bits thus produced is added (with a logical OR) to the fingerprint. We use 1024 bit fingerprints in this study. The Tanimoto coefficient is calculated as $\text{Tanimoto} = \text{NA} \cap \text{NB}/(\text{NA} + \text{NB} - \text{NA} \cap \text{NB})$, where NA is the number of bits ''on'' in fingerprint A, NB is the number of bits ''on'' in fingerprint B and $\text{NA} \cap \text{NB}$ is the number of bits ''on'' in both fingerprints. Since Tanimoto distance is a pair-wise measure, and here the objective is to assess the similarity to the set of molecules, we generate a consensus fingerprint, which is again 1024 bit fingerprint where each bit is set

Table 1. Chemicals assigned to the flowing reactivity domains Michael Acceptors (a), SN2 electrophiles (b), Schiff base formers (c), Acylating agents (d) and chemicals without specific reactivity profile (e) based on [12] (total $n = 41$) and personal communication with A. Aptula ($n = 4$).

CAS	*SMILES*	*Name*
(a) Michael acceptors (MA) $n = 12$		
106-51-4	O=C1C=CC(=O)C=C1	1,4-Benzoquinone
123-31-9	Oc1ccc(O)cc1	Hydroquinone
106-50-3	Nc1ccc(N)cc1	Phenylene diamine
1166-52-5	CCCCCCCCCCCCCOC(=O)c1cc(O)c(O)c(O)c1	Lauryl gallate
97-54-1	CC=Cc1ccc(O)c(OC)c1	Isoeugenol
104-55-2	O=CC=Cc1ccccc1	Cinnamic aldehyde
122-57-6	O=C(C=Cc1ccccc1)C	Benzylidene acetone
101-86-0	CCCCCCC(C=O)=Cc1ccccc1	Hexyl vinnamic aldehyde
104-27-8	CCC(=O)C=Cc1ccc(OC)cc1	1-*p*-Methoxyphenyl-1-penten-3-one
923-26-2	OC(C)COC(=O)C(=C)C	2-Hydroxypropyl methaacrylate
92-48-8	Cc1ccc2OC(=O)C=Cc2(c1)	6-Methyl coumarin
97-00-7	O=N(=O)C1CCC(Cl)C(C1)N(=O)=O	1-Chloro-2,4-dinitrobenzene
(b) SN2 electrophiles (SN2) $n = 7$		
3508-00-7	CCCCCCCCCCCCCCCCCBr	1-Bromoheptadecane
112-82-3	CCCCCCCCCCCCCCCCBr	1-Bromohexadecane
111-25-1	CCCCCCBr	1-Bromopentane
55965-84-9	O=C1N(C)SC(Cl)=C1	Methylisothiazoline
57-57-8	O=C1OCC1	beta-Propiolactone
629-62-9	CCCCCCCCCCCCCCCBr	1-Bromopentadecane
109-65-9	BrCCCC	1-Bromobutane
(c) Schiff base formers (SCH) = 7		
111-30-8	O=CCCCC=O	Glutaraldehyde
50-00-0	O=C	Formaldehyde
122-78-1	O=CCc1ccccc1	Phenyacetaldehyde
80-54-6	O=CC(C)Cc1ccc(cc1)C(C)(C)C	Lilial
103-95-7	O=CC(C)Cc1ccc(cc1)CCC	Cyclamen aldehyde
107-75-5	O=CCC(C)CCCC(O)(C)C	Hydroxycitronellal
13706-86-0	O=C(C(=O)CC(C)C)C	5-Methyl-2,3-hexanedione
(d) Acylating agents (Acyl) $n = 5$		
15646-46-5	c1ccc(C2=NC(=COCC)C(=O)O2)cc1	Oxazolone
119-84-6	O=C2Oc1ccccc1CC2	3-Propelidenephtalide
112-67-4	ClC(=O)CCCCCCCCCCCCCCCC	3,4-Dihydrocoumarin
93-99-2	O=C(Oc1ccccc1)c2ccccc2	Propylidene phthalide
17369-59-4	O=C1OC(=CCC)c2ccccc12	Phenyl benzoate
(e) Chemicals without specific reactivity profile (DIV) $n = 12$		
109-55-7	NCCCN(C)C	3-Dimethylaminopropylamine
591-27-5	Nc1cccc(O)c1	3-Aminophenol
97-53-0	C=CCc1ccc(O)c(OC)c1 O=C(O)C2(C)(CCCC3(C)(C1C(C=C	Eugenol
514-10-3	(CC1)C(C)C)=CCC23)	Abietic acid
78-70-6	OC(C=C)(C)CCC=C(C)C O=C(Cc1ccccc1)NC3C(=O)N2C3(S	Linalool
61-33-6	C(C)(C)C2(C(O)=O))	Penicillin G
108-90-7	Clc1ccccc1	Chlorobenzene
1459-93-4	O=C(OC)c1cccc(c1)C(=O)OC	Dimethylisophtalate
56-81-5	OCC(O)CO	Glycerol
110-54-3	CCCCCC	Hexane
99-96-7	O=C(O)c1ccc(O)cc1	1,4-Dihydrobenzoic acid
67-63-0	OC(C)C	2-Propanol
50-21-5	CC(O)C(O)=O	Lactic acid
119-36-8	O=C(OC)c1ccccc1(O)	Methyl salicylate

Table 2. Atom types used to generate AEs as provided by CDK [19].

H	N	P3	F
Hplus	N.sp2	P4	F–
Hminus	Nplus	S2	Cl
C.sp3	Nplus.sp3	S2–	Cl–
C.sp2	O.sp2	S4	Br
C.sp	O.sp3	S	Br–
C.default	Oplus		I
Cplus.sp2	Ominus		I–
Cminus.sp2			Misc
Caromatic.sp2			
Cminus			

if at least one percent of the compounds in the dataset has a bit set. Then, the Tanimoto distance is calculated between a queried compound and the consensus fingerprint.

2.2.2 Atom environments with *k*-nearest neighbor (*k*-nn) and Tanimoto. Atom Environments (AE) [21, 22] can be regarded as fragments, surrounding each atom in a molecule, up to a predefined level. The calculation procedure is as follows. First, atom types to be included in the generation of AEs are selected.

We use 34 atom types, listed in table 2, which are very similar to Sybyl atom types that have been recommended in Bender *et al.* [22]. The choice is based on the available atom type parameterization in CDK library. Next, a vector of length $(34 * L + 1)$ is constructed for each atom, where L is the maximum level for generating atom environments and $L = 3$ by default. Third, for each atom, neighbors at level 1, 2, 3 are identified and corresponding counts stored in the vector. An example of a string representation of the result for a single atom (C.sp2) is:

C.sp2, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1

Note that if there are several C.sp2 atoms with the same neighbors up to 3rd level in the molecule, they will have the same string representation. We will refer to this representation as a "fragment". AEs could be compared by Tanimoto distance (see above), where NA is the number of fragments in molecule A, NB is the number of fragments in molecule B and $NA \cap NB$ is the number of common fragments between the two molecules. Here, we take average Tanimoto distance for the nearest neighbors, instead of defining consensus fingerprint. For each molecule, the similarity measure is the averaged Tanimoto distance between the molecule and its 5 nearest molecules.

2.2.3 Atom environment ranking (AER). First Atom Environments are extracted for each molecule in the reference set. The novel approach is to represent the reference set not as sum of independent molecules, but as a set of weighted molecules and fragments to represent their importance and connectivity within the set. The weights are derived using random walk theory. Our idea of AER method is inspired by the success of Link Analysis Ranking methods, which are underlying current Web search engines, most notably Google. There are a vast number of references following the 1998 seminal

papers by Kleinberg [23] and Brin and Page [24]. In Link Analysis Ranking, linked structures are used to determine the relative importance of a Web page and rank Web search results. The Google ranking algorithm (PAGERANK) [25] is usually interpreted as performing a random walk that simulates the behavior of a "random surfer". The surfer starts from some node chosen according to some distribution D (usually assumed to be the uniform distribution). At each step, the surfer proceeds as follows: with probability $1-\varepsilon$, one of the hyperlinks on the current page is picked uniformly at random, and the surfer moves to a new page, and with probability ε, the surfer jumps to a random page chosen according to distribution D. The "jump probability" ε is a parameter of the algorithm. In our work we used recommended as default $\varepsilon = 0.85$. The weight of node i is the fraction of time that the surfer spends at node i. That is, it is proportional to the number of visits to node i during the random walk (i.e., visit rate). Random walks are Markov processes which are extensively studied [26]. For ergodic Markov processes, there are unique long-term visit rates for every state, known as stationary or steady-state probabilities that are independent of the starting state of the process. This means that the node weights depend only on how the nodes are linked and algorithms exist to find the unique weighting, or ranking, for each set of nodes.

The innovative idea is to represent the relationship between molecules and fragments by a Markov process. The process starts from a molecule, picks a fragment with probability $1-\varepsilon$ and then selects another molecule containing this fragment or jumps to a completely different molecule from the set and repeats the walk again. The weight of a molecule and of a fragment is proportional to the number of visits. Fragments with higher weights could be considered more important, and correspondingly, the molecule with a higher weight will be a better representative of the set. According to the Markov chain theory, the iterative procedure converges and we use the obtained weights to order molecules and fragments by their importance. Note that in this case we do not calculate pair-wise similarities, but similarity of each molecule to the entire set of molecules. Therefore, there is no need to invent "consensus" fingerprints or make arbitrary decisions about the number of nearest neighbors.

To obtain the weight of a query molecule, its atom environments (i.e., fragments) should be generated and then the weight is calculated according to the first formula above. The molecule is considered similar to the training set if its weight is equal or higher to the smallest weight of the molecule from the training set. Since the weights come from a steady state probability distribution, generated during an optimization process, the weight can be interpreted as probability that a chemical is similar to a set.

2.3 *Analyses*

To explore whether structural similarity is equivalent to biological activity, i.e. similarity in activity, we compared the structural similarity of extreme, strong, moderate and weak sensitizers and non-sensitizers from the LLNA set [13]. Next, to explore if there is structural similarity in the chemicals acting via a common mechanism, and dissimilarity between chemicals acting via different mechanisms, we compared structural similarity between chemicals in different reactivity domains as defined by Aptula *et al*. [3]. We also compared the performance of different structural similarity methods to recommend a specific method for practical application.

Table 3. Performance of (a) DFT, (b) AET, (c) AER methods, expressed as% of compounds similar to the potency classes: nonsensitizers (NS), weak (W), moderate (M), strong (S), extreme (E) from set in [13].

%	NS	W	M	S	E
(a) DFT					
NS	100	93	100	100	81
W	98	100	100	100	88
M	100	97	100	100	96
S	50	95	95	100	90
E	84	85	100	100	100
(b) AET					
NS	100	36	45	10	0
W	33	100	57	29	0
M	29	52	100	19	0
S	15	20	25	100	0
E	8	8	23	0	100
(c) AER					
NS	100	26	12	0	0
W	23	100	22	2	0
M	15	19	100	3	0
S	5	5	10	100	0
E	0	0	0	0	100

3. Results

The results are presented in tables 3 and 4. The tables report percent of compounds from each potency class (rows) found similar to the given potency class (column). The ideal similarity measure should yield 100% only on diagonals (the set is similar to itself) and zeroes everywhere else. Nonzero off-diagonal entries could be interpreted as error rates.

Table 3(a) presents results from similarity assessment by Daylight-like fingerprints and Tanimoto distance to a consensus fingerprint derived for each potency class. The non-sensitizers (NS) column provides a measure of how many compounds are found similar to non-sensitizers. Obviously, non-sensitizers are similar to themselves (100% at NS–NS cell), but so are weak (98% at W–NS cell) and moderate (100% at M–NS cell) sensitizers. Half of the strong sensitizers (50% at S–NS cell) are found similar to non-sensitizers, and surprisingly, 84% of extreme sensitizers are found similar to non-sensitizers. Similarly, bad performance is observed when comparing to the rest of potency classes, highlighting the lack of discriminating power of this method. It is important to note that comparing non-sensitizers to extreme sensitizers (row NS – column E) is not the same as comparing extreme sensitizers to non-sensitizers (row E – column NS), since the procedure involves comparison to a different reference set. Namely, the consensus fingerprints for NS and E based on [13] are different thus the comparisons are not symmetric.

The atom environments combined with averaged Tanimoto distance (AET) perform better table 3(b). They are able to correctly discriminate extreme sensitizers as no compounds from other groups are predicted to be similar to extreme sensitizers. However, there is a considerable structural similarity between compounds from all groups when compared particularly to weak and moderate

Table 4. Performance of (a) DFT, (b) AET, (c) AER methods, expressed as% of compounds, to assess similarity between different reactivity domains Michael Acceptors (MA), SN2 electrophiles (SN2), Schiff base formers (SCH), Acylating agents (Acyl) and chemicals without specific reactivity profile (DIV) based on grouping provided in [3].

%	*SN2*	*SCH*	*Acyl*	*MA*	*DIV*
(a)					
SN2	100	100	0	14	86
SCH	14	100	0	57	86
Acyl	20	100	100	100	100
MA	17	100	43	100	100
DIV	7	100	28	57	100
(b)					
SN2	100	0	0	0	0
SCH	0	100	0	0	0
Acyl	20	0	100	0	0
MA	8	0	0	100	0
DIV	0	0	0	0	100
(c)					
SN2	100	0	0	0	0
SCH	0	100	0	0	0
acyl	0	0	100	0	0
MA	0	0	0	100	0
DIV	7	0	7	7	100

sensitizers (columns W and M). There is an observable trend of decreasing similarity with groups with more dissimilar potencies (e.g., column NS – 33% of weak sensitizers are predicted to be similar to NS, but only 15% strong sensitizers are predicted to be similar to NS). The same trends, more defined but still with non-zero error rates (12–26%), are obtained by Atom environment ranking (table 3c).

Clearly, in the data set [13] there is a high structural similarity between non-sensitizers and weak and moderate sensitizers. However, there is a trend indicating that increasing differences in potency correlate with decreasing structural similarity. The clarity of the trend highly depends on the similarity method used. Data in table 3 support the general opinion that similarity in activity is not equivalent to structural similarity. This finding can, in part, explain why other studies failed to develop predictive models using the LLNA data set [13] reliably differentiating between non-sensitizers and weak and moderate sensitizers. Another possible reason is noisy data in this activity range that could be potentially reduced by a larger dataset.

Next, we analyzed structural similarity between reactivity domains – Michael Acceptors (MA), SN2 electrophiles (SN2), Schiff base formers (SCH), Acylating agents (Acyl) and chemicals without specific reactivity profile (DIV) defined by Aptula *et al*. [3].

Separation by structural similarity between reactivity domains worked much better (table 4) than separation between potency classes (table 3). Similar to the previous analysis, a remarkable difference in performance was observed between different methods with AER yielding lowest errors (0–7% errors), AET in the middle (0–20% error) and DFT yielding an unacceptable error of 0–100%.

4. Discussion

In the above analyses we demonstrated that grouping chemicals by mechanism of action within a given endpoint using computational structural similarity methods is feasible. Attention needs to be paid to the choice of a particular similarity method as different method yield differing results. We also demonstrated that we can directly capture expert knowledge, represented as a reference set into computer readable language, to screen quickly a large number of chemicals. Atom Environments Ranking performed best as assessed by the lowest error rates. Next was AET, and the worst performance was recorded for DFT. While our evidence is limited to one data set, we offer an explanation based on the principles of each method.

The idea of modeling the training set as a Ranked Atom Environments provides a natural representation of the training set molecules as one entity. Molecules' and fragments' weights are obtained by an iterative procedure, accounting for interactions between the fragments in a given training set. This novel representation of the entire training set has an important consequence for screening new chemicals because the similarity comparisons with the training set are straightforward, in contrast to the classical pair-wise similarity assessments and the need to choose arbitrary averaging statistics.

Atom environments provide a more efficient representation of 2D structural features than fingerprints because they do not suffer from hashing procedures [19]. The results obtained by averaging Tanimoto distance between atom environments of the 5 nearest neighbors are of higher quality than DFT. In addition to the absence of hashing procedure, local similarity by nearest neighbors is better exploited. Still, the choice of the number of nearest neighbors is arbitrary and the information of how fragments are distributed within the entire training set is not accounted for.

We ascribe the fact that similarity by Daylight-like fingerprints did not work well for the following reasons. The fingerprint generation procedure includes hashing, meaning that the same bin of the fingerprint can refer to more than one structural feature. This may lower the discriminating ability of the hashed fingerprints. Another drawback is the involvement of the consensus fingerprint as a representative for the dataset. For example, if most of the bits in the consensus fingerprint are set, then the Tanimoto distance between consensus fingerprint and query molecule fingerprint will provide similar results for all molecules. The overarching goal of forming categories is to fill in, reliably and efficiently, data gaps. Because of the considered scope of potential applications involving thousands of chemicals, there is a need to develop computerized procedures for categories formation. While the majority of efforts to screen large chemical inventories focused to date on QSARs, we would like to bring attention to broader exploitation of similarity methods because they may be better suited, especially for complex human health endpoints. Given that many currently existing and developed databases like ChemIDplus [27], ZINC [28], and PubChem [29] provide simple structural similarity searches (although, not connected to toxicity information), their direct use may be very tempting by risk assessors. However, as demonstrated with our skin sensitization case study, relying only on structural similarity is not going to be sufficient. There is a critical need to introduce endpoint specific context and knowledge to guide similarity assessments.

The advantages of using mechanistic understanding as a guiding principle in similarity assessment application to form categories appear manifold and meet many

expectations expressed in the current definitions of the category. Namely, within a category with common mechanism of action, a commonality such as similarity in structures can be observed and characterized. This will allow more effective grouping of chemicals than grouping based on chemical class or same activity as we observed in the skin sensitization case study. Because of the common mechanism of toxicity, a hypothesis supporting a given category can be formulated. With mechanistic hypothesis comes reliability in predictions as was confirmed in modeling a variety of endpoints including fish toxicity [30, 31].

The other difficulty categories often encounter is a small number of chemicals. Our proposed approach can address it in the following way. First, we can group a relatively large number of chemicals with known activity data acting via a common mechanism. Second, we would model activity within the reactivity domain and only then extract chemicals for the category. In this way, a category will rely on much more data and potentially unclear trends within the category can be more easily elucidated by analyses of the whole group with a common mechanism. While this process may be seen as extrapolation because we are to extrapolate from outside a category into the category, in reality the process is much closer to interpolation, a procedure regarded as far more reliable, because a trend in the category is a piece of a larger trend for a group with a common mechanism and hypothesis.

In addition, grouping based on common mechanism will resolve, at least partially, the common problem of making predictions using data sets that are unbalanced towards active chemicals and that lead to a high number of false positives. The majority of chemical databases with safety information contain much more data on active, toxic chemicals than on non-active, non-toxic chemicals and is reflected in the fact that potency based classification models are not very reliable to identify weakly active and non-active chemicals. Grouping chemicals by common mechanism will group chemicals of whole spectrum of activities together, from highly toxic to non-toxic resulting in better balanced sets, allowing for better statistical analyses and more equal opportunities to identify active, weakly active and non-active chemicals. Finally, having the trend for the category derived from the trend of the group will allow better understanding of the trend and, therefore, should facilitate discussions of defining boundaries for a category for a given regulatory application, a point often deemed as very difficult in the practical application of the category concept.

So far, we discussed principles, but it is equally important to have a methodology and tools, in our case software that will allow efficient, practical applications. By creating benchmark AERs for reactivity domains, we demonstrated that we can directly capture expert knowledge about mechanistic understanding into computer readable language to quickly screen a large number of chemicals. This is a novel contribution that allows minimizing information loss and addresses several limitations of existing approaches to convert expert knowledge into computer readable information with the goal to screen chemicals.

The most popular approach to convert expert knowledge about toxicity mechanism to computer screening tool is to derive structural alerts. Structural alerts (SA) can be derived using knowledge-based and statistically-based approaches. Knowledge-based systems such as DEREK, OncoLogic, and HazardExpert [32] use rules about generalized relationships between structure and biological activity that are heuristically derived from human expert opinion and interpretation of toxicological data. On the other hand, MultiCASE [33] derives structural alerts, called biophores,

using a statistical mining technique from a training set of non-congeneric compounds in a formal manner. Structural alerts are geared towards finding an active substance as they ignore the remaining part of the molecule and, thus, they do not consider activating or detoxifying mechanisms in a balanced way. Moreover, structural alerts will not discriminate between non-toxic compounds and compounds for which there is no information on toxicity available.

A more refined approach to code mechanistic understanding is to combine in some logic manner structural alerts with molecular descriptors in characterizing mechanisms of toxicity (e.g., [18]). This approach allows capturing 3D information by using 3D descriptors, in addition to 2D information, that helps to consider whole molecule and better identifies non-active compounds. However, this approach is very data "hungry" because when dealing with heterogeneous datasets one needs to identify different descriptors for different subsets that are *nota bene*, in part, heuristically determined. In addition, a simple logical combination of SA with molecular descriptor may compromise the quality of prediction in the case of non-linear dependencies between them. Finally, by using descriptors, a chance of not capturing structure in all relevant ways remains because conversion to descriptors is always associated with some information loss.

We see reference AERs as the next refinement of expert knowledge conversion approaches into computer readable format. In the benchmark AER, all structural information about training set is retained and there is no need no extract descriptors. It is very transparent and straightforward as the AER algorithm identifies the fragments responsible for the similarity. Interactions, including non-linear interactions between SA and the rest of the molecule are captured, but do not have to be specified *a priori*. In fact, we offer a method that will allow for mechanistic screening without the need to explicitly formalize, many parameterize details that are not available, a situation frequently encountered in toxicology. The reference AERs are very easy to update if more data become available. Finally, reference AERs are less data "hungry" than other mentioned methods because more information about the training set is retained.

The fact that benchmark AER are less data hungry and that we demonstrated significant differences in performance benchmarks based on AER, AET, and especially DFTs leads to an interesting observation regarding the roadblocks to apply computer-based methods to screen chemicals. Too often, lack of data was seen as a major obstacle. The issue that might have been overlooked is the proper computer representation of expert knowledge and structural information minimizing information loss.

More research is needed on how to express mechanism of action to make it useful for computer screening since we only analyzed one set that was available to us. While we got promising results for skin sensitization based on chemical reactivity, it is possible that for other endpoints, different formulations will be more suitable.

Acknowledgements

This research was funded in part by CEFIC Long Research Initiative for funding AMBIT – software for chemoinformatic data management and QSAR applications and

ECB contract CCR.IHCP.C431607.X0 ''Development of a software tool to encode and apply chemical similarity indices''.

References

[1] List of categories assessed in OECD HPV Program. Available online at: http://cs3-hq.oecd.org/scripts/hpv (accessed 31 August 2006).
[2] OECD Workshop of chemical categories. Brussels, 2004. Available online at: http://www.oecd.org/searchResult/0,2665,en_2649_34365_1_1_1_1_1,00.html (accessed 31 August 2006).
[3] A.O. Aptula, G. Patlewicz, D.W. Roberts. *Chem. Res. Toxicol.*, **18**, 1420 (2005).
[4] H. Kubinyi. *J. Brazil. Chem. Soc.*, **13**, 717 (2002).
[5] Y.C. Martin, J.L. Kofron, L.M. Traphagen. *J. Med. Chem.*, **45**, 4350 (2002).
[6] N. Nikolova, J. Jaworska. *QSAR Combi. Sci.*, **22**, 1006 (2004).
[7] A. Bender, R.C. Glen. *Org. Biomol. Chem.*, **2**, 3204 (2004).
[8] H.S. Rosenkranz, A.R. Cunningham. *Regul. Toxicol. Pharmacol.*, **33**, 313 (2001).
[9] K. Birchall, V.J. Gillet, G. Harper, S.D. Pickett. *J. Chem. Inf. Model.*, **46**, 577 (2006).
[10] S.C. Basak, B.D. Gute, D. Mills. *SAR QSAR Environ. Res.*, **13**, 727 (2002).
[11] B.D. Gute, S.C. Basak, D. Mills, D.M. Hawkins. *Internet Elec. J. Molec. Design*, **1**, 374 (2002).
[12] M. von Korff, T. Sander. *J. Chem. Inf. Model.*, **46**, 536 (2006).
[13] G.F. Gerberick, C.A. Ryan, P.S. Kern, H. Schlatter, R.J. Dearman, I. Kimber, G.Y. Patlewicz, D.A. Basketter. *Dermatitis*, **16**, 1 (2005).
[14] S. Li, A. Fedorowicz, H. Singh, S.C. Soderholm. *J. Chem. Inf. Model.*, **45**, 952 (2005).
[15] J. Devillers. *Toxicol. Methods*, **10**, 181 (2000).
[16] Anonymous. In *European Centre for Ecotoxicology and Toxicology of Chemicals (ECETOC)*. Technical Report No. 87, Contact Sensitisation: Classification According to Potency. Brussels, Belgium (2003).
[17] R. Roggeband, G. Helmlinger, I. Smith, K.-P. Wilhelm, C. Ryan, G.F. Gerberick. *Contact Dermatitis*, **46**, 211 (2002).
[18] S.D. Dimitrov, L.K. Low, G.Y. Patlewicz, P.S. Kern, G.D. Dimitrova, M.H Comber, R.D. Phillips, J. Niemela, P.T. Bailey, O.G. Mekenyan. *Intern. J. Toxicol.*, **24**, 189 (2005).
[19] Chemical Development Kit. Available online at http://cdk.sourceforge.net/ (accessed 31 August 2006).
[20] Available online at: http://www.daylight.com/dayhtml/doc/theory/theory.finger.html (accessed 31 August 2006).
[21] L. Xing, R.C. Glen. *J. Chem. Inf. Comput. Sci.*, **42**, 796 (2002).
[22] A. Bender, H.Y. Mussa, R.C. Glen, S. Reiling. *J. Chem. Inf. Comput. Sci.*, **44**, 170 (2004).
[23] J. Kleinberg. In *Proceedings of the Ninth Annual ACM-SIAM*, Symposium on Discrete Algorithms (1998), pp. 668–677.
[24] S. Brin, L. Page. In *Proceedings of the 7th International World Wide Web Conference*, Brisbane, Australia (1998).
[25] Available online at: http://en.wikipedia.org/wiki/PageRank (accessed 31 August 2006).
[26] Available online at: http://en.wikipedia.org/wiki/Markov_chain (accessed 31 August 2006).
[27] Available online at: http://chem.sis.nlm.nih.gov/chemidplus/ (accessed 31 August 2006).
[28] Available online at: http://blaster.docking.org/zinc/ (accessed 31 August 2006).
[29] Available online at: http://pubchem.ncbi.nlm.nih.gov/ (accessed 31 August 2006).
[30] M.T.D Cronin, T.W. Schultz. In *Designing Drugs and Crop Protectants: Processes, Problems and Solutions*, M. Ford, D. Livingstone, J. Dearden and H. van de Waterbeemd (Eds), pp. 294–296, Blackwell, Oxford (2003).
[31] M.T.D. Cronin. In *Quantitative Structure-Activity Relationship (QSAR) Models of Mutagens and Carcinogens*, R. Benigni (Ed.), pp. 235–258, CRC Press, Boca Raton, FL (2003).
[32] E.M. Hulzebos, P.C.J.I. Schielen, L. Wijkhuizen-Maslankiewicz. *(Q)SARs for Human Toxicological Endpoints: A Literature Search*. A Report by the RIVM (Research for Man and Environment), The Netherlands, RIVM Report 601516.001 (1999).
[33] G. Klopman, S.K. Chakravarti, H. Zhu, J.M. Ivanov, R.D. Saiakhov. *J. Chem. Inf. Comput. Sci.*, **44**, 704 (2004).