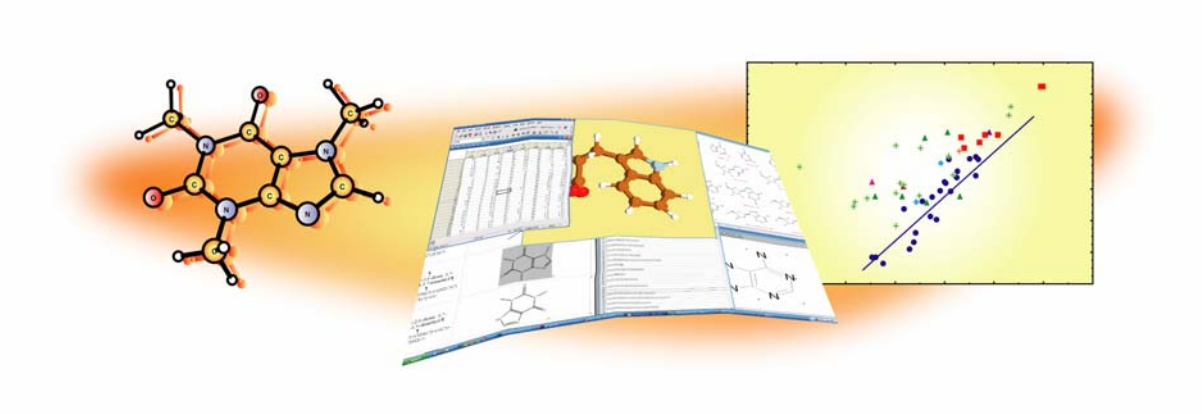


A Similarity Based Approach for Chemical Category Classification

Ana Gallegos Saliner, Grace Patlewicz, Andrew P. Worth

2005 **EUR 21867 EN**

EUROPEAN COMMISSION
DIRECTORATE-GENERAL
Joint Research Centre

EUROPEAN COMMISSION
DIRECTORATE GENERAL
JOINT RESEARCH CENTRE

Institute for Health and Consumer Protection
Toxicology and Chemical Substances Unit
European Chemicals Bureau
I-21020 Ispra (VA) Italy

A Similarity Based Approach for Chemical Category Classification

Ana Gallegos Saliner, Grace Patlewicz, Andrew P. Worth

2005 **EUR 21867 EN**

LEGAL NOTICE

A great deal of additional information on the
European Union is available on the Internet.
It can be accessed through the Europa server
(http://europa.eu.int)

EUR 21867 EN

Printed in Italy

Abstract

This report aims to describe the main outcomes of an IHCP Exploratory Research Project carried out during 2005 by the European Chemicals Bureau (Computational Toxicology Action). The original aim of this project was to develop a computational method to facilitate the classification of chemicals into similarity-based chemical categories, which would be both useful for building (Q)SAR models (research application) and for defining chemical category proposals (regulatory application).

Preparatory work to investigate the notion of chemical similarity and explore how it could be applied to both the development of chemical categories as well defining the domain of applicability for SAR models, e.g. structural alerts was conducted.

The state of the art of chemical similarity indices was reviewed, and a selection of those chemical similarity indices that showed greatest promise for describing toxicological and ecotoxicological effects were described in further detail.

A scoping study to explore the utility of similarity measures for describing the applicability domain of structural alerts was conducted. A set of skin sensitisation structural rules that are currently encoded into the Derek expert system were explored. Recommendations for further research work have been proposed.

A multi-stakeholder workshop, involving academic scientists, regulators and industry participants, was organised to discuss various issues surrounding chemical similarity, in particular how such indices could be applied in the formation of chemical categories that are appropriate for regulatory use.

Finally the development of a software tool capable of calculating and applying similarity indices is outlined.

Table of Contents

List of Abbreviations

(Q)SAR	(Quantitative) Structure-Activity Relationships
(Q)SPR	(Quantitative) Structure-Property Relationships
(Q)STR	(Quantitative) Structure-Toxicity Relationships
CADD	Computer Aided-Drug Design
CAMD	Computer-Aided Molecular Design
CEFIC	European Chemical Industry Council
DEREK	Deductive Estimation of Risk from Existing Knowledge
ECB	European Chemicals Bureau
FIRM	Formal Inference-based Recursive Modelling Analysis
HTS	High Throughput Screening
ICCA	International Council of Chemical Associations
JRC	Joint Research Centre
LDA	Linear Discriminant Analysis
LHASA	Logic and Heuristics Applied to Synthetic Analysis
LLNA	Local Lymph Node Assay
Log Kp	Logarithm of the permeability coefficient
Log P	Logarithm of the octanol/water partition coefficient
MW	Molecular Weight
PCA	Principal Component Analysis
QSI	Quantum Similarity Indices
QSM	Quantum Similarity Measures
REACH	Registration, Evaluation, and Authorisation of Chemicals
RIPs	REACH-implementation projects
SMILES	Simplified Molecular Input Line Entry Specification
TTC	Thresholds of Toxicological Concern

Introduction: General Background

Under the current legislation for New and Existing Chemicals in the European Union, the regulatory use of structure-activity relationships (SARs) and quantitative structure-activity relationships (QSARs), collectively referred to as (Q)SARs, for the assessment of chemicals is somewhat limited. On 29 October 2003, the European Commission adopted a legislative proposal that foresees the introduction of a new regulatory system called REACH (Registration, Evaluation, and Authorisation of Chemicals) [1]. This calls for equivalent information requirements to be applied to New and Existing Chemicals. The proposed REACH legislation is expected to result in some 30,000 chemicals requiring evaluation for toxicity, ecotoxicity and environmental fate, over a period of 11 years. For reasons of cost, practicality, and animal welfare, this assessment exercise cannot be achieved by applying traditional test methods. Instead, the REACH proposal foresees greater use of non-testing approaches so called *in silico* methods, such as QSARs, SARs, read-across and chemical categories. Analyses carried out by the ECB have shown that such non-testing approaches have the potential to provide an efficient means of obtaining the required information on chemicals whilst reducing testing costs and the amount of (animal) testing necessary [2,3].

Guidance on the use of (Q)SARs is provided in Annex IX of the proposed REACH legislation. It states that (Q)SARs may be used to indicate the presence or absence of a certain dangerous property if the following conditions are met [4]:

- results are derived from a (Q)SAR model whose scientific validity has been established
- results are adequate for the purpose of classification and labelling and risk assessment
- adequate and reliable documentation of the method is provided

Annex IX also states that chemicals may be classified on the basis of their (eco)toxicological hazard by applying chemical grouping approaches (e.g. read-across, chemical categories [5]).

To date, the acceptance of (Q)SARs has been limited due to a lack of understanding in how to evaluate the scientific validity of the models. Recently several initiatives have emerged to explore ways of evaluating validity. The first was a Workshop organised by CEFIC/ICCA in Setubal in 2002 [6] which established principles for the validity of (Q)SARs. These were then evaluated by the OECD Ad hoc group for (Q)SARs and are now referred to as the 'OECD principles for (Q)SAR validation'. According to these principles, "to facilitate the consideration of a (Q)SAR model for regulatory purposes, it should be associated with the following information:

- a defined endpoint
- an unambiguous algorithm
- a defined applicability domain (see [7])
- appropriate measures of goodness-of-fit, robustness and predictivity
- a mechanistic interpretation, if possible"

The principles provide a useful framework and practical guidance of how to demonstrate concordance for a (Q)SAR is under development [5]. Limited guidance for developing chemical categories (in essence a group of "similar" chemicals with respect to their properties and hence could be viewed as an extension of a SAR) does exist but the tools for their practical implementation are still lacking [8].

In practice, the acceptance and use of (Q)SARs under REACH will depend on the availability of technical guidance and tools. Indeed, Annex IX of the REACH proposal indicates that the Chemicals Agency, in collaboration with the Commission, Member States and interested parties will develop and provide guidance in assessing which (Q)SARs will meet the above-mentioned conditions and provide examples. The development of such guidance and tools is being carried out and coordinated by European Commission's Joint Research Centre (JRC). Within the JRC, the European Chemicals Bureau (ECB) [9] is responsible for:

a) providing scientific and technical support to the European Commission and EU Member States in relation to current legislation on chemicals, biocides and plant protection products;

b) coordinating the scientific and technical preparations needed for the implementation of the REACH legislation – the so-called REACH-implementation projects (RIPs); and

c) coordinating the JRC activity on computational toxicology, which is providing input into the development of technical guidance for REACH, such as guidance on the use of (Q)SARs and related estimation approaches, and guidance on integrated testing strategies.

Chemical category development is dependent on grouping chemicals on the basis of their structural similarity but there is emerging evidence that structural similarity does not always leads to similarity in activity. There is a need to provide guidance for how to encode "similarity in activity" in a meaningful way that will assist in category development.

The OECD has developed some guidance on how to group chemicals [5] and some examples of chemical categories have been provided by the US EPA [10]. However the OECD guidance is written at a very generic level and does not explain how chemical similarity should be interpreted in

a context-dependent and scientifically-meaningful way. The ECB has been keeping a watching brief on developments in this active research field in QSAR.

Specifically it has performed a feasibility exercise to identify quantitative measures of chemical similarity and apply them to evaluate the applicability domain for a number of skin sensitisation structural alerts. The outcome of this work has helped to illustrate the challenges in describing the applicability domain of structural alerts. In addition it has provided some tangible proposals for how to justify a read across and how chemical categories could be formulated.

Literature Review on Chemical Similarity

A literature review on chemical similarity indices and available approaches for encoding chemical similarity has been carried out. Special focus has been made on the application of similarity indices for encoding toxicological and ecotoxicological activity. A range of different approaches for encoding similarity have been illustrated with special attention on how such indices can be used in the development of chemical categories.

Introduction

The definition of a similarity measure between two chemicals has been an ancestral question in theoretical chemistry. Chemical similarity attempts to answer the question: “how similar is a given molecule to another?”. In general, it is assumed that the similarity principle holds, that is, similar compounds have similar activities. This assumption has been the driving force for the development of a pool of computer-based methods for toxicity prediction, such as Quantitative Structure-Activity Relationships (QSARs. The application of molecular similarity concepts in QSAR analysis is reviewed in the following sections.

Historical concept of similarity

In human consciousness, the intuitive concept of similarity is strongly attached to knowledge. On a daily basis, humans unconsciously make associations from visual perception of objects or situations and in doing so establish common characteristics and differences through applying latent criteria. Instinctively, the human mind continuously compares new knowledge with existing knowledge, using criteria from experience. A new concept is reached when some similarities and/or dissimilarities are processed between the new information received and the previous one [11].

The similarity concept is rooted in science but has also been the subject of study in both psychology [12] and philosophy [13]. The first contributions to similarity date back to ancient Greek philosophy when comparative measures between geometrical shapes were already proposed and established.

Similarity is undoubtedly an important geometrical and spatial concept in mathematics. Mathematicians term "similar" as objects that have the same shape but not necessarily the same size; thus, proportional objects with the same ratio [14]. Pythagoras applied the similarity of triangles to formulate his theorem, based on the similarity of triangles.

The underlying assumption in chemistry is that similar molecules possess similar properties, and this has been the foundation of empirical relationships between structure and activity. In 1869 Mendeleev [15] formulated the periodic table of elements through observation and comparison of the similar chemical behaviour and reactivity of elements. By systemically considering the atomic properties, Mendeleev was able to classify all the elements into a table leaving gaps for substances still unknown. By noting patterns between the combinations of well-classified elements, he was able to predict both undiscovered elements as well as their physico-chemical properties.

In contrast, the systematization of cognitive processes leading to the evaluation of similarity has proven to be much more difficult. In the chemistry domain, different proposals attempted to measure the similarity between two molecules, in order to obtain a sound definition of unbiased and unambiguous quantitative measures of molecular similarity.

Current applications of chemical similarity in toxicity prediction

Nowadays, computer-based similarity techniques are mainly directed to the development of rational molecular design strategies in the drug discovery process.

For a long time, medicinal chemists have systematically modified lead compounds. The process of synthesising new drugs implies discovery of a potential active. Once a candidate structure is identified, analogue compounds with the optimal desired properties are investigated. These should have an improved biological activity and pharmacokinetic characteristics, but diminished adverse effects such as toxicity. The biological phase will include comprehensive animal and human testing, specificity, bioavailability, lack of toxicity. It may take months to synthesise a new compound for biological testing using traditional techniques. The high expense in resources in the drug discovery process have prompted a drive to supplement conventional drug discovery technologies with molecular and drug design strategies [16].

The potential of being able to design new useful compounds with well-defined properties virtually in silico and thus reduce the high costs of experimental synthesis has recently promoted investment in theoretical research. Methods, such as biostructural research, computer-assisted data handling, data storage, retrieval and processing from chemical databases [17], and, especially, structure-based design, structure-function correlation studies, and other statistical techniques, are of special relevance in the discovery and development of compounds with specific pharmaceutical properties.

Hence, the effective design of chemical structures with the desirable therapeutic properties is directed towards Computer-Aided Molecular Design (CAMD), also more specifically called Computer Aided-Drug Design (CADD) [18-23]. These techniques comprise new methodologies, such as molecular modelling, computer simulation, and the discipline of Quantitative Structure-Activity Relationships (QSAR).

The strategy of structure-based molecular design has been proven to be very successful in the pharmaceutical industry [24]. Where structural information about the biological target is lacking, the strategy of lead finding still involves the synthesis and testing of widely diverse compounds. The systematic variation of substituents in a molecule has been the subject of various studies in the past. As it is not straightforward to select a representative subset of substituents that adequately covers the multidimensional parameter space, relevant properties are selected from large sets of property descriptors by using statistical techniques.

In combinatorial chemistry, enormous libraries of millions of compounds are analysed by High Throughput Screening (HTS) methods. HTS screens large numbers of compounds selected from a library against a biological target, i.e. a protein playing a fundamental role for a particular disease. Nowadays, it is possible to assemble chemical building blocks in all combinations, generating large virtual libraries of structurally related compounds by means of automated procedures [25]. High throughput screening methods screen these databases with a defined query, usually a pharmacophore, "testing" hundreds to millions of compounds, and looking for relevant information. Afterwards, data mining techniques identify novel patterns in the data, potentially useful to analyse the data sets. Combinatorial approaches seek to maximise the structural diversity of the final library, i.e. the degree of heterogeneity, that is, the structural range or dissimilarity, to ensure the coverage of the largest possible expanse of chemical space in the search for bioactive molecules [26]. These computational tools improve molecular diversity and the chance of lead discoveries. The ready availability of chemical structure databases plays an important role in enhancing the drug discovery approach [27]. These databases find increasing use in environmental, inorganic, and organic chemistry. The combinatorial chemistry technologies have increased the number of compounds synthesised and tested for every new chemical entity and have also provided a far more cost-effective approach to the discovery of bioactive compounds, in comparison with traditional approaches.

Both molecular modelling techniques and quantitative statistical methods may be useful in elucidating structural information of active compounds. Since a biological effect seldom depends on just one or two chemical properties, the multidimensional problem takes into account a large number of factors, rationalised to cover a broad parameter space. In order to be able to deal with

complex data sets, consisting of more than one biological activity and many descriptors, advanced statistical and computational tools have been developed in the field of chemometrics, the discipline that uses statistical and mathematical methods for selecting and optimizing procedures for the analysis and interpretation of data. These techniques allow the rapid retrieval and prediction of molecular and biological properties by means of multivariate methods and artificial intelligence techniques [29].

Structure-function correlation studies aim to broaden understanding of relationships between molecular intrinsic chemical features and physicochemical or biological properties. Such studies are Quantitative Structure-Activity Relationships (QSAR), Quantitative Structure-Property Relationships (QSPR), or Quantitative Structure-Toxicity Relationships (QSTR).

Chemical Similarity

The definition of a chemical similarity measure depends on the molecular feature under analysis, such as functional groups or common substructures. In general, it is widely assumed that the characteristics and behaviour of substances are partially conditioned by their structure, and the description of quantitative measures for molecular similarity has been carefully examined in the bibliography [30].

The definition of similarity with respect to molecules consists of mapping the chemical space, i.e. a representation of the molecule in terms of relevant descriptors in one-dimensional space with real numbers. The definition also depends on the representation of the molecules under consideration in descriptor space. In the general case of chemical similarity, molecules may be represented using a range of different depictions.

Representation of Chemical Structures

The characterization of chemical structure has long been of great interest even though the term was not properly described until 1861 by the Russian chemist Butlerov [31]. Butlerov defined chemical structure as the type and manner of the mutual binding of atoms in a compound, without specifying the nature of bonding. The links existing between atoms in molecules were depicted as dotted or continuous lines [32], solid rods [33], or even as tubes of force [34]. Structural formulas drawn with straight lines connecting the bonded atoms were first published in 1858 by Couper [32], and in 1864 by Crum Brown [35-37]. Since those times, several tiers of characterising molecular structures have been described, from simple enumeration of atoms to complex metabolic simulations.

The characterization of a structure may be represented as an ordered set of components with information concerning the relationship between those components. This information may be in the form of a list i.e. the labelling of atoms and bonds (molecular codes), or in the form of the count of

components of various types describing the mathematical properties of a structure (structural invariant). Different structural molecular description levels, ordered by degree of information provided are listed below:

1) List of type of atoms that constitute the molecule.

2) Empirical formula, that is, the simplest stoichiometric formula indicating the proportion of different atoms.

3) Molecular formula, indicating the number of atoms of each type. This corresponds to the formula needed to calculate the exact molecular mass.

4) In contrast to the one-dimensional constitutional information provided by the preceding formulas, the two-dimensional structural formula represents the arrangement of atoms using the topology of the molecule and the connectivity of the constituting atoms. The graph, a variant of the structural formula, omits the type of atom and nature of bonding. It is worth stating that alternative representations at a similar level have been designed such as the Simplified Molecular Input Line Entry Specification (SMILES) [38] and InCHI codes [39].

5) Three-dimensional structure describe the structure of the molecule as a three-dimensional entity with the atoms situated in specific positions in the space (x,y,z, coordinates), thus providing geometrical and spatial information.

6) The resolutions of the Schrödinger equation, which include a description of the charge distribution. These constitute the most accurate descriptions (depending on the level of theory used to solve them) but are typically computationally intensive.

In general, the representation of a chemical can be considered in terms of constitution, configuration, and conformation. Constitution provides information about the sequence of bonding of atoms and is expressed by topological descriptors, presence and absence of fragments, and descriptors that account for the two-dimensional features of a molecule. Configuration is defined by a three-dimensional or spatial arrangement of atoms, characterized by angles, and is expressed by shape descriptors and approaches accounting for the three-dimensional arrangement of atoms. Finally, conformations represent thermodynamically stable spatial arrangements of the atoms of a molecule.

A number of methods for the quantitative description of molecular structures have been proposed and applied to date. Different descriptors can be employed for the formulation of structure-function relationships depending on the theoretical basis adopted for the description of the structure of molecules.

A common issue in QSAR is how to describe molecules and their properties. The nature of the descriptors used and the extent to which they encode the structural features related to the biological activity is a crucial part of a QSAR study [40]. It has been estimated that more than 3,000 molecular descriptors are now available [41-42]. Most of them can be theoretically calculated by using commercial software packages such as DRAGON [43], ADAPT [44-45], OASIS [46], and CODESSA [47], among others.

From the extensive available bibliography, some of the most widely used in order of increasing complexity are the topostructural, topochemical, geometrical, relativistic, and biodescriptors. The main descriptors used to characterise chemical compounds can be arbitrarily classified in different groups:

1) Empirical parameters derived from organic chemistry. These are used in classical QSAR models, for example Hansch analysis. Initially, these models were based on several varieties of physicochemical descriptors, classified into electronic, hydrophobic, and steric. Subsequently other descriptors were also included, i.e. experimental properties like solubility, melting point, boiling point, spectroscopic descriptors, etc.

2) Theoretically determined properties. This group includes topological descriptors as well as parameters derived from computational chemistry. The main advantage of these descriptors is that they can be calculated.

3) Three-dimensional descriptors. These parameters, used in 3D-QSAR techniques, take into account the three-dimensional structure of molecules and they may require a molecular superposition procedure. This group includes molecular similarity indices and topological quantum similarity indices.

The influence of structural characteristics on activity may be localised to the whole molecule or a part of it. This is another commonly employed classification pattern of descriptors.

a) Substituent constants or parameters based on fragment constants or physicochemical parameters. A significant number of these descriptors belong to the category of empirical parameters derived from physical organic chemistry. These parameters focus on how chemical reaction rates depend on differences in molecular structure. The characterization of these differences in structure on account of differing substitutions of functional groups on a fixed core pattern has led to the development of substituent constants. These constants relate the effect of substituents on a reaction centre from one type of process to another. Some examples are electronic substituent constants, hydrophobic substituent constants, and steric substituent constants

b) Whole molecule representations or descriptors derived from entire molecular structures are either extensions of the substituent constant approach or completely novel descriptors. Several are based on the spatial conformation of compounds and therefore require a molecular superposition process. Other examples include electronic whole molecule descriptors, polar descriptors, energetic descriptors, geometric descriptors, topological descriptors, information-content indices, as well as quantum similarity indices. The latter are derived from quantum mechanical calculations which take into account three-dimensional conformational information.

Similarity indices

There are many different types of similarity indices, which can be derived from the similarity matrix $\{Z_{AB}\}$, where *A* and *B* are the two molecules being compared:

$$\mathbf{Z} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \vdots \\ j \\ \vdots \\ n \\ \\ \mathbf{z}_1 \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & 2 & 3 & \cdots & i & \cdots & n \end{array} \\ \begin{pmatrix} z_{11} & z_{12} & z_{13} & \cdots & z_{1i} & \cdots & z_{1n} \\ & z_{22} & z_{23} & \cdots & z_{2i} & \cdots & z_{2n} \\ & & z_{33} & \cdots & z_{3i} & \cdots & z_{3n} \\ & & & \ddots & & & \vdots \\ & & & & z_{ii} & & z_{jn} \\ & & & & & \ddots & \vdots \\ & & & & & & z_{nn} \end{pmatrix} \\ \begin{array}{ccccccc} \downarrow & \downarrow & \downarrow & & \downarrow & & \downarrow \end{array} \\ \begin{array}{ccccccc} \mathbf{z}_1 & \mathbf{z}_2 & \mathbf{z}_3 & & \mathbf{z}_i & & \mathbf{z}_n \end{array} \end{array}$$

where *n* is the number of molecules, and *Z* the similarity matrix.

Some of the more commonly used indices are:

Distance-like dissimilarity indices are measures of dissimilarity between objects or measures of the distance in a multidimensional geometric space. Their metrics has the following properties:

$D_{AB} \geq 0$
$D_{AA} = D_{BB} = 0$
$D_{AB} = D_{BA}$
$D_{AB} \leq D_{AC} + D_{CB}$
$A \neq B \leftrightarrow D_{AB} > 0$

The general definition of a distance dissimilarity index, which is comprised between the value of zero for identical molecules and infinity, can be expressed as:

$$D_{AB}(k,x) = \left[k\left(Z_{AA} + Z_{BB}\right)/2 - xZ_{AB}\right]^{1/2} \qquad D_{AB} = [0,\infty)$$

- Euclidean Distance Index ($k = x = 2$) can be defined according to the classical definition of distance [48]:

$$D_{AB} = \sqrt{Z_{AA} + Z_{BB} - 2Z_{AB}}$$

D_{AB} is comprised within the interval [0,∞) but, conversely to the previous case, values close to zero imply a greater similarity between the compared objects. Hence if the two compared objects are identical, D_{AB}=0.

Geometrically, this index may be interpreted as the norm of the difference between the density functions of the compared objects. The Euclidean distance index can be defined as a distance or dissimilarity index, also called a **D-class** index.

Another distance-coefficients are the Hamming distance, and the Soergel distance.

Correlation-like similarity indices

The general definition for such an index can be expressed as:

$$V_{AB}(k,x) = (k-x) Z_{AB} D_{AB}^{-2}(k,x) \qquad V_{AB} = [0,1]$$

Some examples of such indices are the following:

- Hodgkin – Richards Index [49] ($k = 2; x = 0$)

$$H_{AB} = 2Z_{AB}\left[Z_{AA} + Z_{BB}\right]^{-1}$$

- Tanimoto Index [50] ($k = 2; x = 1$)

$$T_{AB} = 2Z_{AB}\left[Z_{AA} + Z_{BB} - Z_{AB}\right]^{-1}$$

- Cosine-like similarity index or Carbó Index.

$$C_{AB} = Z_{AB}\left[Z_{AA} Z_{BB}\right]^{-1/2}$$

C_{AB} varies in the interval (0,1]. The nearer to the unit, the more similar are the compared objects, while a value approaching to zero indicates that the two objects are dissimilar. The exact unity value is only obtained when both compared objects are the same, that is, in the case of self similarity measures, where $C_{AB} = 1$, that is, an object is identical to itself.

Geometrically, the Carbó index can be interpreted as the cosine of the angle subtended by the involved electronic density functions, considered in turn as vectors. The Carbó index is a correlation-like or cosinus index, also called **C-class** index.

Some studies comparing the Quantum Similarity Measures (QSM) generated by different operators and several Quantum Similarity Indices (QSI) have been reported in the literature [51-53].

The use of similarity measures in defining the applicability domain of skin sensitisation SARs

Validation of Sensitisation Rules within the DEREKfW Expert System

DEREK is a knowledge-based expert system that identifies the structural features of a chemical that may result in the manifestation of toxicity. It was developed and it is still being enhanced by LHASA, Ltd and the members at the School of Chemistry, University of Leeds, UK. The system contains over 320 rules for endpoints such as skin sensitisation, mutagenicity, carcinogenicity, skin and eye irritation [54].

Of these endpoints skin sensitisation and mutagenicity are perhaps the most well developed within DEREK. Skin sensitisation is an endpoint of great interest within REACH and other forthcoming legilislation such as the 7th Amendment for Cosmetics. The latter in particular since a ban on all animal testing for cosmetics will come into effect in 2009 for a number of endpoints including skin sensitisation. Currently testing is conducted using an in vivo test named the Local Lymph Node Assay (LLNA). There are no in vitro strategies that have been developed that are sufficienly robust to assess skin sensitisation hazard. The use of structure activity techniques in this area shows greatest promise and for this reason, the work carried out here was focused on skin sensitisation as a priority endpoint.

Additionally there have been a number of efforts to collate and harmonise available data on skin sensitisation that could be useful for the development of new in vitro techniques as well as facilitate the development of new in silico models. Gerberick et al [55] compiled a list of some 41 compounds that could be useful as one source of data. A second more extensive dataset of 211 chemicals has also been compiled by Gerberick et al [55]. This dataset provided a good starting point for evaluating some of the existing alerts within DEREK.

DEREK is a knowledge based expert system comprising a number of structural rules that aim to encode structure-toxicity information with an emphasis on mechanisms. The toxicity predictions made by DEREK are the result of two processes. The program checks whether any alerts in the knowledge base match toxicophores in the query structure. The reasoning engine then assesses the likelihood of a structure being toxic. There are 9 levels of confidence: certain, probable, plausible, equivocal, doubted, improbable, impossible, open, contradicted. The reasoning model considers the following information:

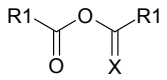
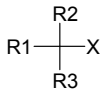
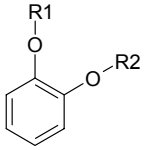
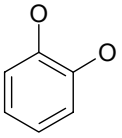
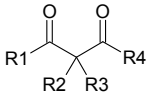
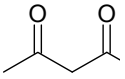
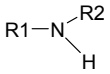
- The toxicological endpoint
- The alerts that match toxicophores in the query structure
- The physicochemical property values calculated for the query structure
- The presence of an exact match between the query structure and a supporting example within the knowledge base

For skin sensitisation and photoallergenicity, DEREK uses a calculation of skin permeability, which is estimated by Log Kp derived from the Log P (octanol/water partition coefficient) value and molecular weight. DEREK uses an estimated calculation of the Log P developed by Moriguchi [56]. A Clog P (BioByte Corp, USA) plug in can be used to override the Moriguchi calculation of Log P. Human log Kp values are calculated from the molecular weight and log P values of a chemical by using the Potts and Guy equation [57]. This equation is derived from a data set of ninety three chemicals with a molecular weight range of 18 to >750, and a log P range of -3 to +6.

The objective of this study was to investigate the feasibility of utilising different similarity measures as a means of evaluating the scope of several of the structural alerts within the DEREK system.

Method. The dataset of 211 chemicals [55] was processed through DEREK Version 7 to identify any skin sensitisation structural alerts. This dataset will be referred to as the Mastertable. These alerts were prioritised to evaluate those alerts possessing the greatest number of chemicals. A short list of five alerts was selected and the LHASA was contacted to provide the training set for each of the five alerts. The dataset for each of these alerts was compared again mastertable to verify the extent of overlap of the chemicals used to develop the alerts. Each alert was evaluated separately. For each alert, SMILES codes were generated for the mastertable chemicals (the testset dataset) and the training set chemicals used to develop the structural alert. The testset and training set chemicals were imported into TSAR (Version 3.3) and labelled accordingly. A range of descriptors were calculated using the TSAR (Accelrys) software. Table 1 shows the number of compounds in each of the training and the test sets as well as the main functional group underpinning each alert.

Table 1. Number of compounds in the training and the test set, and functional groups underpinning each alert.

Structural Requirement	Fragment Search	Alert	Training Set N.Compounds	Test Set N.Compounds
Acid anhydride or analogue  X = O, S, NR2 R1 = C, H R2 = any		405	7	8
Haloalkane  R1-R3 = any except F, Cl, Br, I X = Cl, Br, I	—F —Cl —Br —I	413	70	17
Catechol or precursor  R1 = H, acyl, alkyl R2 = H, acyl		418	54	8
1,3-Diketone  R1, R4 = C R2, R3 = any		420	5	12
Aromatic primary or secondary amine  R1 = C (aromatic) R2 = H, C, Not C=O	—N	427	98	8
TOTAL			234	53

A set descriptors for all the studied chemicals were calculated by using TSAR version 3.3 molecular spreadsheet (2000, Accelrys, Oxford, England). These included molecular attributes, such as the octanol-water partition coefficient (log P), and molecular weight (MW), topological indices based on graph representations, and atom and group counts.

Formal Inference-based Recursive Modelling Analysis (FIRM) [58] was performed by using the descriptors calculated [59]. FIRM analysis was carried out for each alert, taking into account both the compounds present in the training and the test sets. The predictor variables selected by FIRM that split the two data sets were: number of N atoms, number of halogen atoms, group count for acid anhydride, number of halogen atoms, 6-membered aromatic rings, number of Br atoms, number of halogen atoms, 5-membered aliphatic rings, 6-membered aromatic rings, and number of H-bond acceptors. The results of the FIRM analysis model revealed accuracy for classification higher than 80 %.

Linear discriminant analysis using descriptors accounting for the size (molecular mass, molecular surface area, and molecular volume), the lipophilicity (total dipole moment, and log P), and two indicator variables mapping the structural alerts (number of halogen atoms, and number of N atoms), provided a significantly lower accuracy. This was done to check if there was any significant alternative classification model for this data set.

Principal component analysis (PCA) was then used as a statistical technique for exploratory structure similarity data analysis. PCA was performed for the descriptors chosen for each of the alerts. The results considering all the alerts simultaneously are presented below. As seen, it was possible to distinguish differentiated clusters, representing the different alerts. The grouping of chemicals in the descriptor space indicates that the compounds belonging to different alerts display differentiated structural characteristics. Thus, this trend suggests treating the alerts separately, in order to be able to discriminate the important features for each alert.

Figure 1. Score plot of the two first principal components differentiating each alert.

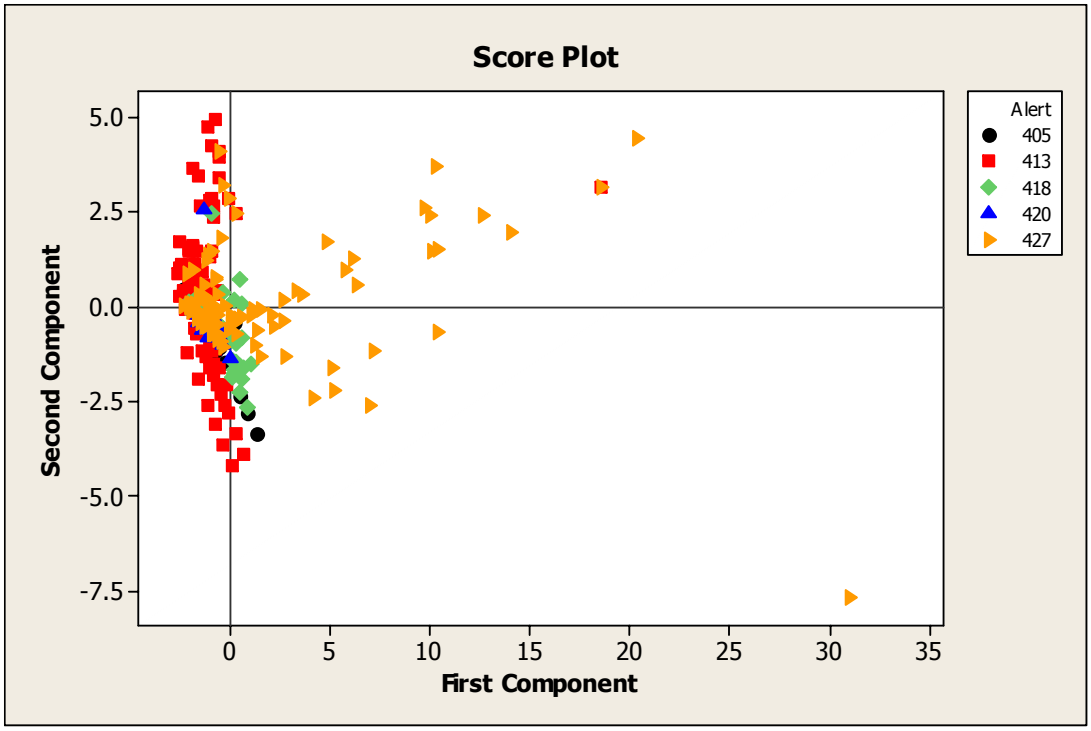
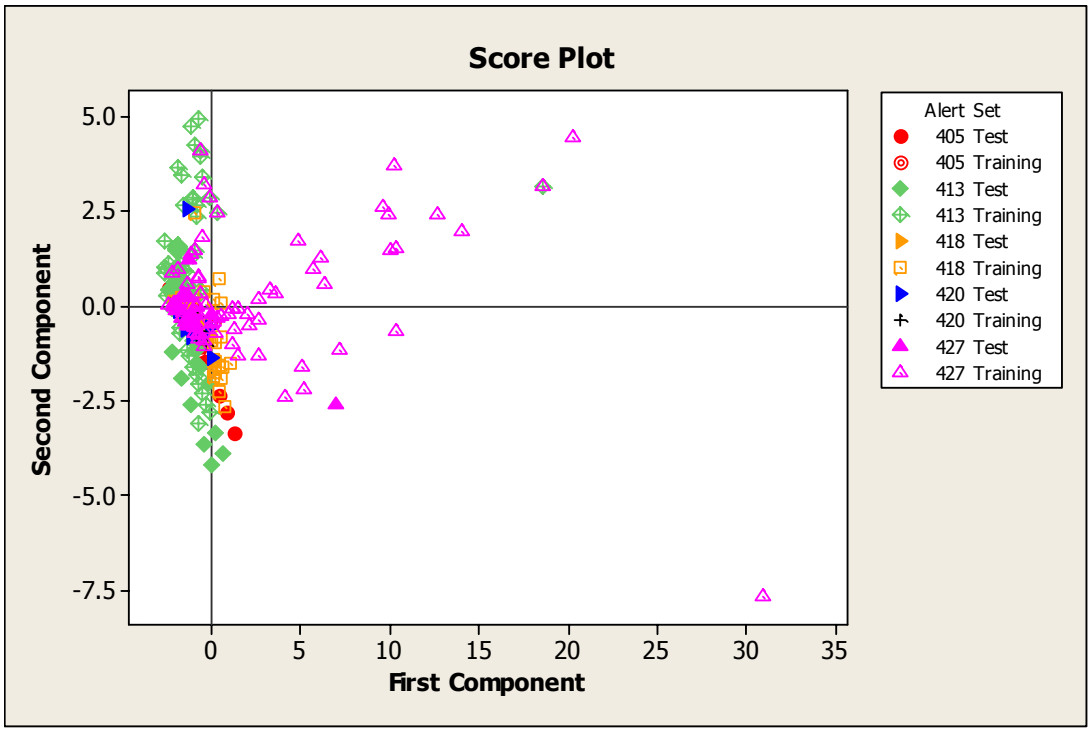


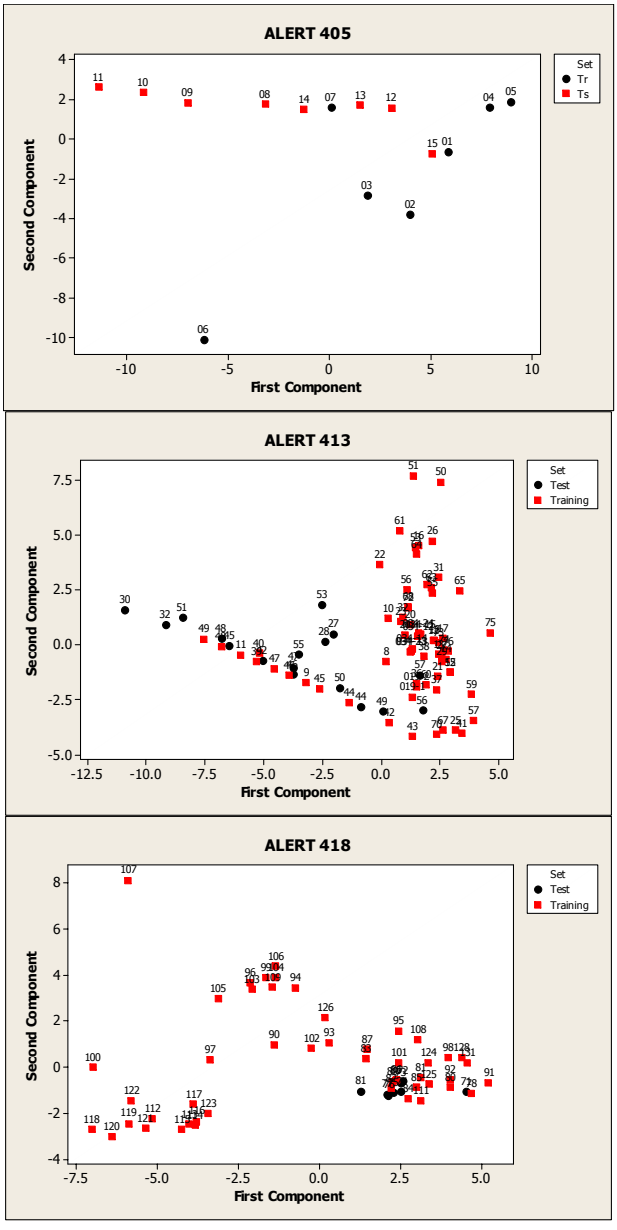
Figure 2. Score plot of the two first principal components differentiating the training and the test set for all the alerts.

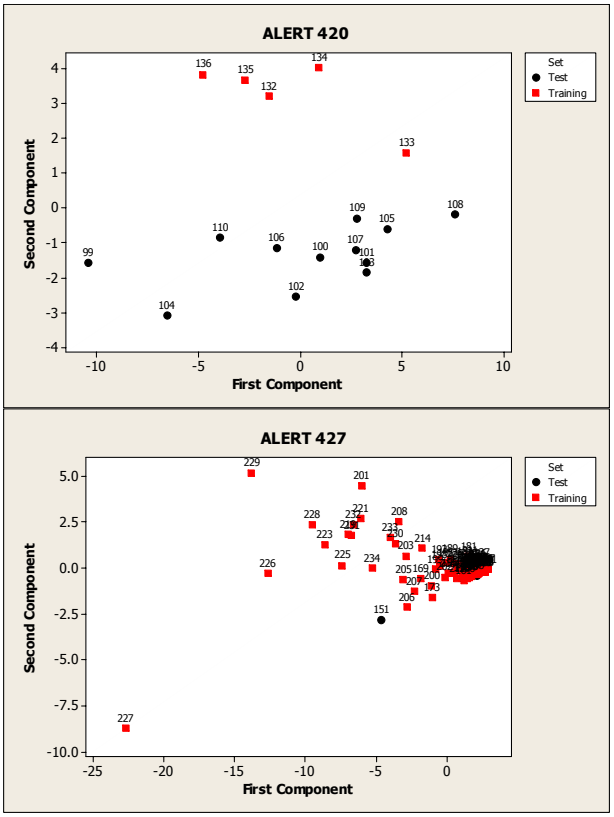


The PCA analysis was also performed for each alert separately. The first two components were generally found to describe a satisfactory amount of the information in the dataset (higher than

80%). This facilitated visualisation of the distribution of the chemicals in both training sets and test sets which enabled a rapid inspection on the degree of similarity between compounds using a distance measure as the discriminator for similarity. The PCA plot presents a picture of the diversity of the chemicals using a number of non-specific descriptors. Overlaying the same descriptors for the test set chemicals allows a rapid assessment to be made to what extent these chemicals are similar to the training set chemicals. The similarity represented is with respect to the parameters chosen and does not necessarily indicate that these chemicals are likely to behave similarly with respect to sensitisation. The principal components of the training and test sets for each alert are displayed below:

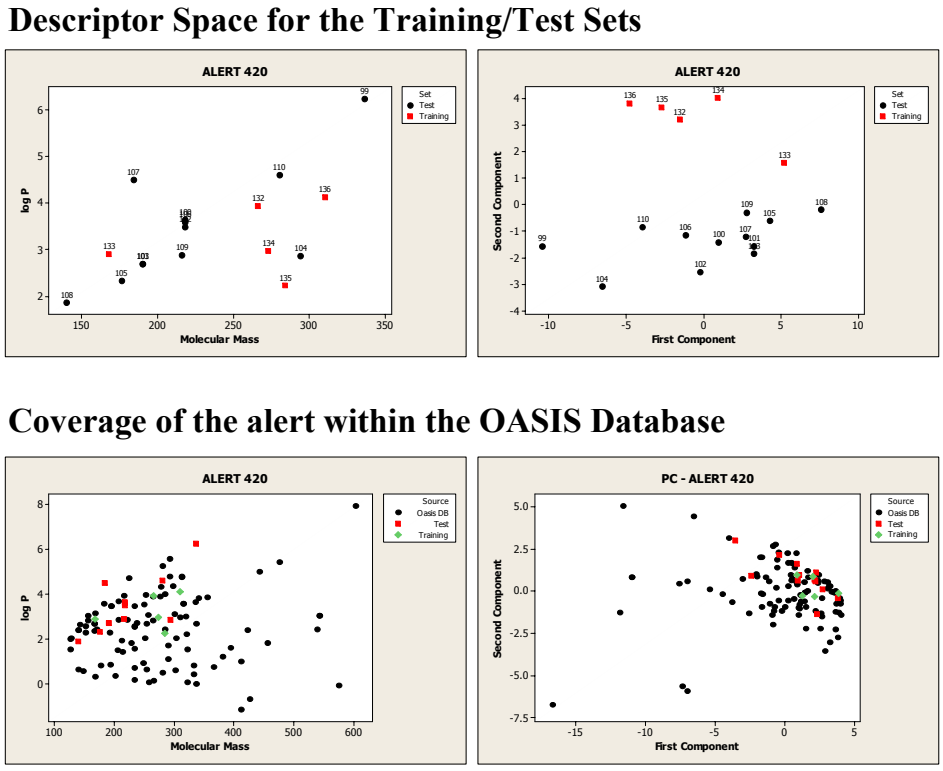
Figure 3-7. Principal components for each alert.





As can be seen from the various PCA plots, using an empirical set of different descriptors failed to provide any insight about the extent of similarity between the test set chemicals and those in each training set. This led us to suspect that the chemicals in the test set were too different to be useful in the assessment of domain or indeed in the evaluation of an alert's validity. On account of this reasoning, an attempt was made to source additional information. The OASIS software (OASIS by LMC, Bourgas, Bulgaria) which contains 160,000 chemicals with predictions for a range of endpoints was used to identify chemicals for one of the alert. Alert 420 was evaluated in more detail. The plots below reflect the limited breadth of the training set of compounds and how many different chemicals could fit in this rule. The descriptor space was examined using two non-relevant descriptors (log P and molecular mass), and using the first and the second principal components (PCs) of the calculated descriptors. From the plot it can be observed that the PCs split the training and the test sets into two separate groups. The coverage of compounds belonging to alert 420 with the OASIS database shows that a non negligible number of chemicals are located in near a region in the descriptor space. This could be useful to detect other chemicals with similar patterns, and to have a greater number of compounds in each alert.

Figure 8. Plots of the more relevant descriptors, the two principal components for alert 420; coverage of the descriptor space of alert 420 with the compounds underpinning the same alert in the OASIS database.



There are a variety of characteristics that determine whether a chemical is likely to function as a skin sensitizer including its ability to penetrate the skin, to react with proteins and be recognized as antigenic by immune cells. The correlation with protein reactivity with skin sensitization is well established. That is to say if a chemical is capable of reacting with protein either directly or after metabolism then it has the potential to act as a sensitizer.

Consideration of the chemical properties of a wide variety of other known sensitisers and comparison with non sensitisers led to the conclusion that binding to a protein takes place by the protein acting as a nucleophile and the sensitiser acting as an electrophile. In considering whether or not a given compound is likely to sensitise or in trying to predict whether a given compound is the active component in a sensitizing mixture, the approach has been to look for electrophilic characteristics in the molecular structure. This implies that the hapten has chemical reactivity that allows it to form bonds with side chains of amino acids and that these reactions with protein are likely to be selective for particular amino acids units depending on the chemical functionality for the sensitising chemical.

The TOPS-MODE (topological substructural molecular descriptors) approach has been used to derive QSAR models for understanding the molecular structural contribution to skin sensitization [60]. A data set of 93 compounds was used in the development of the discriminant models. The

models developed possess high predictivity and have been validated through the use of cross-validation and external validation sets. Various classes of chemicals and their mechanisms for skin sensitization were presented on the basis of bond contributions. The new mechanisms proposed or modified thereafter were validated by experimental findings supporting them.

This descriptors generated from this model were calculated for each of the training and test sets and the PCA was conducted once more. The hope here was that the relevant descriptors underpinning skin sensitisation were used instead and that this would provide a more meaningful comparison of chemical similarity with respect to sensitisation. The PCA plots shown below reveal a very different perspective and reinforces how important both the context and molecular representation can be.

Leadscope

The Leadscope datamining tool (www.leadscope.com) was also used to study the training set and test sets for each alert. Leadscope possesses a unique chemical hierarchy containing over 27,000 chemical fingerprints. These fingerprints represent functional groups, chemical groupings, and pharmacophores that provide a presentation of a database/dataset/inventory in terms of its actual chemistry. The hierarchy can be exploited to group chemicals according to a specific level of concern through the use of structural rules.

Leadscope was used to cluster the different compounds into similar classes according to structural fingerprints, 42 different clusters were obtained, most of them corresponding to structural alerts, or fragments of them. Why – what did this tell us?

The Leadscope tool was also used to assess the domain of the test set with respect to the training set, revealing that in general test set compounds are very different from the training set.

Conclusions

Our preliminary findings confirm how context-dependent chemical similarity truly is . This is particularly important for defining the applicability domain of SARs in a meaningful way. Future work should seek to identify additional test data (chemicals) to supplement the training set of chemicals as well as to explore other means of encoding similarity for sensitisation through the use of appropriate descriptors and fingerprints, and to establish whether the ADs of selected SARs (structural alerts) can be defined in a quantitative manner by using cut-off values.

Publication. The results obtained have been summarised in the poster presented in the CTW Berlin world congress (see Appendix 1).

Workshop on Chemical Similarity and TTC approaches

In order to draw further guidance on the use of similarity measures in chemical categories and validation of SAR rules, ECB organised a workshop on *Chemical Similarity and Thresholds of Toxicological Concern (TTC) approaches,* on 7 – 8 November 2005 in Ispra. The main aims of the workshop were to discuss the terminology and review existing approaches for the categorisation of chemicals both for the development of thresholds (TTC) and chemical categories. The meeting also communicated some of the work undertaken in the area of TTC. The agenda of the meeting can be seen in Appendix 2.

Follow-up of the meeting

Minutes of the meeting have been drafted (see Appendix 3). In addition a summary and a detailed report of the meeting, with the input from the participants will be written. The target audience will be composed of QSAR researchers as well as risk assessors in Regulatory agencies and in Industry.

The discussion sessions in the meeting were organised to obtain as an input on what guidance for the development of chemical categories should look like. Outcomes from this meeting will be communicated to the QSAR working group as proposals under consideration.

Further Work

The preliminary work carried out in this exploratory research project highlights the need of specific expertise to calculate a broad spectrum of molecular similarity indices. The ECB has elected to fund the development of a standalone easy-to-use software tool for this purpose. This tool would encode a variety of similarity indices to facilitate systematic and transparent justification for read across as well as chemical categories, with specific reference to current OECD guidance on the formation of categories [5].

References

1. European Commission (2003) 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}. Website: http://europa.eu.int/comm/enterprise/chemicals/chempol/whitepaper/reach.htm
2. Van der Jagt, K., Munn, S., Tørsløv, J., de Bruijn, J. (2004) Alternative approaches can reduce the use of test animals under REACH. Addendum to the report "Assessment of additional testing needs under REACH. Effects of (Q)SARs, risk based testing and voluntary industry initiatives". JRC Report EUR 21405 EN, 25pp. Ispra, Italy: European Commission, Joint Research Centre.
3. Pedersen, F. de Bruijn, J. Munn, S., van Leeuwen, K. (2003) Assessment of additional testing needs under REACH. Effects of QSARs, risk based testing and voluntary industry initiatives. JRC Report EUR 20863 EN, 33pp. European Commission, Joint Research Centre: Ispra, Italy.
4. Website: http://europa.eu.int/comm/enterprise/reach/overview.htm
5. Section 3.2 of the OECD Manual for Investigation of HPV Chemicals: Chapter of guidance document on the formation and use of chemical categories. http://www.oecd.org/document/7/0,2340,en_2649_34379_1947463_1_1_1_1,00.html
6. (Q)SARs for Human Health and the Environment - Workshop on Regulatory Acceptance. (2002) CEFIC, ACC, ECETOC.
7. Netzeva T.I., Worth A.P. Aldenberg T., Benigni R., Cronin M.T.D., Gramatica P., Jaworska J.S., Kahn S., Klopman G., Marchant C.A., Myatt G., Nikolova-Jeliazkova N., Patlewicz G.Y., Perkins R., Roberts D.W., Schultz T.W. Stanton D.T., van de Sandt J.J.M., Tong W., Veith G., Yang C. (2005) Current Status of Methods for Defining the Applicability Domain of (Quantitative) Structure-Activity Relationships. The Report and Recommendations of ECVAM Workshop 52. *ATLA-Alternatives to Laboratory Animals* **33**, 155-173.
8. Report of the SPORT (Strategic Partnership on REACH Testing) pilot project: "The SPORT Report: Making REACH work in practice". Website: http://www.sport-project.info
9. ECB website: http://ecb.jrc.it
10. US Environmental Protection Agency. Site on chemical categories http://www.epa.gov/oppt/newchems/chemcat.htm
11. Johnson, M.A., Maggiora, G. (Eds.) (1990) *Concepts and Applications of Molecular Similarity*. John Wiley & Sons Inc.: New York.
12. Tversky, A. (1977) Features of Similarity. *Psychological Reviews* **84**, 327 – 354.

13. Quine, W.V. (1977) Natural kinds. In, *Ontological relativity and other essays.* Columbia University Press: New York, NY.
14. Fuson, K.C. (1978) Analysis of research needs in projective, affine and similarity geometries, including an evaluation of Piaget's results in this area. In *Recent Research concerning the Development of Spatial and Geometric Concepts*. R. Lesh (Ed.) ERIC/SMEAC: Columbus, Ohio, pp 243-260.
15. Mendeleev, D.I. (1868–71) *Principles of Chemistry*, Vol. 2, tr. 1905,
16. Van de Waterbeemd, H. (1993) Recent progress in QSAR-technology. In *Drug Design and Discovery* **9**, 277-285.
17. 3D Virtual Chemistry Library. Imperial College of Science, Technology & Medicine. The Cambridge Structural Database System – from crystallographic data to protein-ligand applications. Stephen J. Maginn, avalaible via CrystalWeb, ICSD-WWW, ConQuest, ReactionWeb or ISIS. Available at the website: http: www.ccdc.cam.ac.uk/products/csd
18. Martin, Y.C. (1978) *Quantitative Drug Design: A Critical Introduction*. Marcel Dekker (Ed). New York.
19. Sanz, F., Martín, M., Pérez, J., Turmo, J., Mitjana, A., Moreno, V., Dearden, J.C., (Eds.) (1983) *Quantitative Approaches to Drug Design.* Elsevier: Amsterdam.
20. Franke, R. (Ed.) (1984) *Theoretical Drug Design Methods*. Elsevier: Amsterdam.
21. Codding, P.W. (Ed.) (1998) *Structure-based drug design: experimental and computational approaches*, Vol. 352. NATO ASI Series: Dordrecht.
22. Levy, M.D. (2000) The drug discovery and development process in the new millennium. *Canadian Journal of Gastroenterology* **14 (7)**, 641-643.
23. Richards, W.G. (1989) *Computer-Aided Molecular Design*. IBC Technical Services: London.
24. Gubernator, K. (Ed.) (1995) *Structure-Derived Ligand Design. Methods and Principles in Medicinal Chemistry*, Vol 4. Mannhold, R.; Krogsgaard-Larsen, P.; Timmerman, H. (Eds.) VCH: Weinheim.
25. Chaiken, I.M., Janda, K.D. (Eds.) (1996) *Molecular Diversity and Combinatorial Chemistry. Libraries and Drug Discovery.* American Chemical Society: Washington.
26. DeWitt, S.H., Czarnik, A.W. (Eds.) (1997) *A Practical Guide to Combinatorial Chemistry*. American Chemical Society: Washington.
27. Gillet, V.J., Wild, D.J., Willett, P., Bradshaw, J. (1998) Similarity and Dissimilarity Methods for Processing Chemical Structure Databases. *The Computer Journal* **41,** 547-558.
28. Spilker, B. (1989) *Multinational Drug Companies. Issues in Drug Discovery and Development*. Raven Press: New York.

29. Topliss, J.G. (Ed.) (1983) *Quantitative Structure-Activity Relationships of Drugs.* Academic Press: New York.
30. Rouvray, D.H. (1995) Similarity in chemistry: past, present and future. In *Topics in Current Chemistry*, Vol 173. Sen, K. (Ed.) Springer-Verlag: Berlin, pp 1-30.
31. Butlerov, A.M. (1861) *Z. Chem*. **4**, 549. (Tr. Kluge, F.F., Larder, D.F. (1971) *Journal of Chemical Education* **48**, 289).
32. Couper, A.S. (1858) *Ann. Chim. Phys*. **53**, 469.
33. Hofmann, A.W. (1865) *Proc. R. Inst. G.B. London* **4**, 414.
34. Stark, J. (1915) *Prinzipien der Atomdynamik, Part III. Die Elek- trizitat im Chemischen Atom*. Hirzel: Leipzig, pp 81.
35. Crum Brown, A. (1861) *The Theory of Chemical Combination.* M.D. Thesis, University of Edinburgh.
36. Crum Brown, A. (1864) *Trans. R. Sot. Edinburgh* **23**, 707.
37. Crum-Brown, A. Fraser, T.R. (1868-9) On the connection between chemical constitution and physiological action. Part 1. On the physiological action of the ammonium bases, derived from Strychia, Brucia, Thebaia, Codeia, Morphia and Nicotia. *Trans. Roy. Soc. Edinburgh* **25**, 151-203.
38. SMILES: Simplified Molecular Input Line Entry Specification. Daylight Chemical Information Systems, Inc. Website: http://www.daylight.com
39. InChI: IUPAC-NIST Chemical Identifier. Website: http://www.iupac.org/inchi
40. Downs, G.M. (2004) Molecular Descriptors. In *Computational Medicinal Chemistry for Drug Discovery*. Bultinck, P., De Winter, H., Langenaeker, W., Tollenaere, J. P., (Eds.). Marcel Dekker: New York, pp 515-538.
41. Devillers, J., Balaban, A.T. (1999) *Topological Indices and Related Descriptors in QSAR and QSPR*. Gordon Breach Scientific Publishers: Amsterdam, pp 811.
42. Karelson, M. (2000) Molecular Descriptors in QSAR/QSPR. Wiley-InterScience: New York.
43. Todeschini, R., Consonni, V., Pavan, M. (2005) DRAGON-Software for the Calculation of Molecular Descriptors. Release 5.3 for Windows. Website: http://www.talete.mi.it
44. Jurs, P.C. (2002) ADAPT-Automated Data Analysis and Pattern Recognition Toolkit. University Park, PA: Pennsylvania State University. Website: http://research.chem.psu.edu/pcjgroup/ADAPT.html
45. Stuper, A.J., Jurs, P.C. (1976) ADAPT: A computer system for auto-mated data analysis using pattern recognition techniques. *Journal of Chemical Information and Computer Sciences* **16,** 99–105.

46. Mekenyan, O., Bonchev, D. (1986) OASIS method for predicting bio-logical activity of chemical compounds. *Acta Pharm. Jugosl*. **36**, 225–237.
47. Katritzky, A.R., Lobanov, V.S., Karelson, M. (1994) CODESSA, Reference Manual. Gainesville, FL University of Florida. Website: http://www.semichem.com/codessarefs.html
48. Carbó, R., Domingo, L. (1987) LCAO-MO similarity measures and taxonomy. *International Journal of Quantum Chemistry* **32**, 517-545.
49. Hodgkin, E.E., Richards, W.G. (1987) Molecular similarity based on electrostatic potential and electric field. *International Journal of Quantum Chemistry* **32**, 105-110.
50. Tou, J.T., González, R.C. (1974) *Pattern recognition principles.* Addison-Wesley: Reading, M. A.
51. Carbó, R., Besalú, E., Amat, L., Fradera, X. (1996) On quantum molecular similarity measures (QMSM) and indices (QMSI). *Journal of Mathematical Chemistry* **19**, 47-56.
52. Robert, D. Carbó-Dorca, R. (1998) A formal comparison between molecular quantum similarity measures and indices. *Journal of Chemical Information and Computer Sciences* **38**, 469-475.
53. Lobato, M., Amat, L., Besalú, E., Carbó-Dorca, R. (1998) Estudi QSAR d'una familia de quinolones utilitzant índexs de semblança i índexs topològics de semblança. *Scientia Gerundensis* **23**, 17-27.
54. Estrada, E., Patlewicz, G., Gutierrez, Y. (2004). From knowledge generation to knowledge archive. A general strategy using TOPS-MODE with DEREK to formulate new alerts for skin sensitisation. *Journal of Chemical Information and Computer Sciences* **44**, 688-698.
55. Gerberick, G.F., Ryan, C.A., Kern, P.S., Schlatter, H., Dearman, R.J., Kimber, I., Patlewicz, G.Y., Basketter, D.A. (2005) Compilation of historical local lymph node data for the evaluation of skin sensitization alternatives. *Submitted to Dermatitis.*
56. Moriguchi, I., Hirono, S., Liu, Q., Nakagome, I., Matsushita, Y. (1992) Simple method of calculating Octanol/Water Partition Coefficient. *Chemical and Pharmaceutical Bulletin* **40**, 127-130.
57. Potts, R.O., Guy, R.H. (1992) Predicting Skin Permeability. *Pharmaceutical Research* **9**, 663-669.
58. Hawkins, D.M., Young, S.S., Rusinko III, A. (1997) Analysis of a large structure-activity data set using recursive partitioning. *Quantitative Structure-Activity Relationships* **16**, 296−302.
59. TSAR v3.3 (2000) Oxford Molecular Ltd. Oxford, U.K.
60. Estrada, E., Patlewicz, G., Chamberlain, M., Basketter, D., Larbey, S. (2003) Computer-aided knowledge generation for understanding skin sensitization mechanisms: the TOPS-MODE approach. *Chemical Research in Toxicology*, **16**, 1226-1235.

Appendix 1.

Poster presented in the 5th World Congress on Alternatives & Animal Use in the Life Sciences, held in Berlin, on 21-25 August.

2005

FIFTH WORLD CONGRESS

ALTERNATIVE CONGRESS TRUST

The use of Similarity Measures in defining the Applicability Domain of Skin Sensitisation SARs

A. Gallegos*, G. Patlewicz, A.P. Worth

European Chemicals Bureau (ECB), Institute for Health and Consumer Protection

European Commission - Joint Research Centre, 21020 Ispra, Italy

ABSTRACT

In the (Q)SAR field, the applicability domain (AD) is widely understood to express the scope and limitations of a model, i.e. the range of chemical structures for which the model is considered to be applicable. For QSAR models, the parameter space is typically represented by ranges of physicochemical descriptors. For SAR models in the form of structural alerts, the parameter space is typically represented by the structural feature that defines the presence of a hazard.

The aim of this work is to explore the utility of chemical similarity measures as a means of defining the applicability domain for a set of skin sensitization structural rules. Preliminary analysis confirms that chemical similarity is context dependent. Parameters that encode sensitisation are more meaningful than general descriptors.

INTRODUCTION

In the proposed REACH legislation[1], some 30,000 chemicals will require an evaluation for their toxicological and ecotoxicological profiles. Experimental testing for this number of chemicals is not feasible from both a time and cost perspective. However, in silico approaches such as (Q)SARs, read across, and chemical categories are thought to show promise from both an economic and animal welfare perspective.

The REACH proposal states that (Q)SARs may be used to indicate the presence or absence of a certain dangerous property if the following conditions are met:

- results are derived from a (Q)SAR model whose scientific validity has been established
- results are adequate for the purpose of classification and labelling and risk assessment
- adequate and reliable documentation of the method is provided

Uptake of (Q)SARs currently has been limited due to a lack of understanding in how to evaluate the scientific validity. Several initiatives in recent years have sought to explore ways of evaluating validity. The first was a workshop organised by CEFIC/ICCA in Setubal in 2002 which established principles for the validation of (Q)SARs. These were then evaluated and revised by OECD (by the Ad hoc group on (Q)SARs) and are now referred to as the 'OECD principles for the validation of (Q)SARs for regulatory purposes'.

To facilitate the consideration of a (Q)SAR model for regulatory purposes, it should be associated with the following information:

- a defined endpoint
- an unambiguous algorithm
- a defined domain of applicability
- appropriate measures of goodness-of-fit, robustness and predictivity
- a mechanistic interpretation, if possible

These principles provide a useful framework, and practical guidance on how to apply them to (Q)SARs is under development by Ad hoc (Q)SAR Group.

The AD is perhaps one of the most difficult concepts to apply. For SARs such as structural alerts, the domain may be represented by the structural feature that defines the presence of a hazard. However this definition presents difficulties as to when it is appropriate to use a structural alert or not. Consider the following example, an alert is expressed by the presence of a specific fragment together with one or more conditions associated with the immediate environment. With this in mind, how can the end-user be certain that it is appropriate to apply that alert to a new query structure; what are the boundaries of the given alert that dictate at which point the alert no longer is indicative of the effect. One way of evaluating this boundary is to explore whether chemical similarity indices provide a meaningful quantification of the boundary. The approach would be to examine the training set of chemicals used to define the alert and to explore whether any of the features describing the toxicity response enable cut-offs to be defined, which would provide a transparent means of determining when it is more or less reliable to apply the structural alert. The approach taken here was to consider different approaches for encoding chemical similarity and to explore their application to a set of structural alerts for skin sensitisation that are encoded into the DEREK expert system.

[1] http://europa.eu.int/comm/enterprise/reach/overview.htm

PRELIMINARY INVESTIGATIONS

Five alerts (acid anhydride or analogue, catechol or precursor, 1,3-diketone, aromatic primary or secondary amine, and haloalkane) were chosen from the DEREK for Windows skin sensitisation rulebase. The training sets of compounds used to derive the rules were supplied by LHASA Ltd. A dataset of compiled LLNA data was used to identify potential test set compounds that could be used to explore the scope of these five alerts.

SMILES (Simplified Molecular Input Line Entry System) codes were generated for both training and test sets of chemicals for each alert in turn. A range of descriptors (including Log P, MW, and a variety of molecular properties and indices) were calculated using the TSAR (Accelrys Ltd) software. Principal Components Analysis was performed on these descriptors. The first two components in each case was found to describe over 85% of the information in the dataset.

PRELIMINARY INVESTIGATIONS

The PCA plot (illustrated below for the 1,3-diketone alert 420) presents a picture of the diversity of the chemicals using a number of non-specific descriptors. Overlaying the same descriptors for the test set chemicals allows a rapid assessment to be made to what extent these chemicals are "similar" to the training set chemicals. The similarity represented is with respect to the parameters chosen and does not necessarily indicate that these chemicals are likely to behave similarly with respect to sensitisation. The PCA plot (Fig 1) reflects the limited breadth of the training set of compounds and how different the chemicals in the test set are. On the basis of this plot, it appears that the test set of chemicals may not be suitable for assessing the alert. Sensitisation results available for this test set of compounds suggest that the alert could be further refined to capture a greater diversity in chemical structure and response.

Figure 1

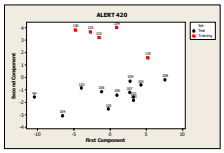
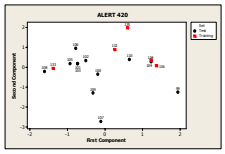


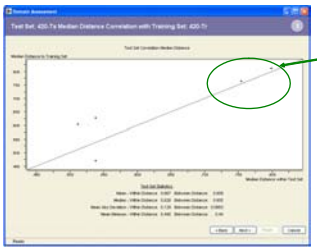
Figure 2



Ideally the similarity index should use parameters that are relevant to the sensitisation response. The second stage of this investigation was to use a general QSAR model published in the literature for sensitisation[3] and to calculate the descriptors used in this model. The descriptors included are those accounting for molar refractivity, hydrophobicity, various charges, van der Waals radii, polar surface area, and polarisability. A PCA was performed for alert 420 and a plot of the first two components (which account for 86% of the information) was drawn (Fig 2). This plot reflects a greater degree of similarity between the two datasets.

The two plots reflect context dependent differences in the way in which chemical similarity is defined. A third stage of the investigation was to explore the use of structural fingerprints to encode similarity. The Leadscope[4] tool was used to assess the domain of the test set with respect to the training set. Fig 3 reflects the median distance correlation with the training set. The plot reveals 2 points that are very different from the training set.

Figure 3



[2] in preparation

[3] Estrada et al. (**2003**) *Chem. Res. Toxicol. 16*, 1226-1235

[4] www.leadscope.com

CONCLUSIONS AND FURTHER WORK

Preliminary analysis confirms that chemical similarity is highly context-dependent. This is particularly important for defining the applicability domain of SARs in a meaningful way. Future work will seek to: a) identify additional test data (chemicals) to supplement the training set of chemicals; b) explore other means of encoding similarity for sensitisation through the use of appropriate descriptors and fingerprints; and c) establish whether the ADs of selected SARs (structural alerts) can be defined in a quantitative manner by using cut-off values.

ACKNOWLEDGEMENTS We gratefully acknowledge Carol Marchant and Kate Langton LHASA Ltd for providing training set information for each of the alerts.

CONTACT DETAILS : ana.gallegos@jrc.it

ihcp

ECB

EUROPEAN COMMISSION DIRECTORATE-GENERAL Joint Research Centre

2005.

Appendix 2.

Agenda of the consultation meeting on chemical similarity and TTC approaches, held in Ispra, on 7 – 8 November 2005.

Day 1 – Chemical Similarity (7 November)

09.00	**Start of the meeting Day 1**
09.00-09.20	Introduction of the participants. Aims and organisation of the meeting (Andrew Worth)
09.30-10.15	Chemical Similarity - an overview (Nina Jeliazkova)
10.15-10.45	Insights on Chemical Quantum Molecular Similarity Indices (Ana Gallegos)
10.45-11.15	**Coffee Break**
11.15-11.45	Chemical similarity in database searching (Val Gillet)
11.45-12.15	The concept of chemical categories (Brigitte Simon-Hettich)
12.15-12.45	Experiences in chemical series definition and chemical similarity (Aldo Benigni)
12.45-14.00	**Lunch**
14.00-14.30	From classification schemes for chemical structures to virtual biological profiling of chemical libraries (Jordi Mestres)
14.30-15.00	Introduction to the brainstorming and formulation of open questions (Grace Patlewicz)
15.00-15.30	**Coffee Break**
15.30-17.00	Discussion/ brainstorming on applicability of the techniques
17.00-17.30	Conclusions and recommendations.
17.45	**End of Day 1**

Day 2 – TTC (8 November)

9.00	**Start of Day 2**
9.00-9.30	Review of Day 1
9.15-10.00	TTC - an OFAS perspective (Andrew McDougal (conference call)
10.00-10.45	The Threshold of Toxicological Concern concept (Ian Munro)
10.45-11.15	**Coffee Break**
11.15-11.45	TTC - Literature review and applicability (Maria Wallén)
11.45-12.15	TTC - a SEAC perspective (Bob Safford)
12.15-12.45	TTC - Cramer classification scheme : a toolbox (Nina Jeliazkova)
12.45-14.15	**Lunch**
14.15-15.00	Overview of grouping (Chihae Yang)
15.00-15.30	Introduction to the brainstorming (Grace Patlewicz)
15.30-16.00	**Coffee Break**
16.00-16.30	Discussion/ brainstorming on applicability of the techniques
16.30-17.00	Conclusions and Recommendations – Report writing and next steps
17.00	**End of Day 2 and of the meeting – Transport to the airport**

Appendix 3.

Minutes of the consultation meeting on chemical similarity and TTC approaches, held in Ispra, on 7 – 8 November 2005.

Meeting Minutes

The meeting was chaired by **Grace Patlewicz** (ECB), who opened the workshop by welcoming the participants through a roundtable of introductions.

Presentations were made by several of the participants in order to provide an overview of ongoing activities from the perspective of different organisations (academia, industry, and regulatory organisations). This gave a perspective of some of the approaches available in the field of Chemical Similarity and TTC and how they were being applied. The presentations helped to structure the afternoon plenary discussions aimed at capturing potential strategies for Chemical Category development as well as research needs or opportunities.

Day 1 – Chemical Similarity

Andrew Worth (ECB) presented the aims, organisation and structure of the meeting. He briefly outlined the structure of the European Commission, the role of the JRC and that of ECB within the scientific and technical preparations for REACH. He presented the scope of the meeting namely, a review of approaches for chemical similarity and thresholds of toxicological concern. These approaches are of specific interest to the QSAR group since chemical similarity techniques could be potentially used to help classify chemicals into similarity-based chemical categories for read-across; and thresholds of toxicological concern for human health endpoints could help to evolve integrated testing strategies. He explained that the two topics (TTC and chemical similarity) had been combined into a single meeting, because they are basically both grouping approaches. Chemical similarity approaches provide a means of grouping chemicals for hazard identification (classification) purposes, whereas TTC approaches could be adapted to group chemicals according to their potency, i.e. provide a means of quantitative read-across.

Nina Jeliazkova (IDEA Consult Ltd.) presented a literature-based review on chemical similarity. She began by presenting similarity as an intuitive concept widely used in philosophy as well as many other disciplines. A meaningful, unambiguous and useful measure of similarity is needed to capture the resemblance in relation to the aspect to be described. She highlighted a myriad of different approaches for measuring the similarity between chemicals, from simple fingerprint counts, to 3D similarity including quantum chemistry field-based approaches. She stressed some of the main advantages and disadvantages of these different methods, depending on the numerical representation chosen for the molecular structures and the different types of similarity indices that are available. She concluded by highlighting several caveats for chemical similarity, in particular, how there is always a loss of information associated with any similarity measure; how some measures may not correctly represent the intuitive similarity between two chemicals; or even that structure may not be the sole factor for biological activity and that structurally similar molecules may still have differing mechanisms of action.

Ana Gallegos (ECB) presented some theoretical insights on the formulation of molecular similarity indices based on quantum mechanics calculations. She started by presenting the foundation of quantum similarity theory based on the characterisation of molecular structures by electronic density functions. She illustrated several approaches used to calculate first-order electronic density functions which minimise computational costs but preserve accuracy. The atomic shell approximation (ASA), and the promolecular ASA (PASA) are examples of these. She also presented different algorithms for molecular superposition, based on the maximal similarity alignment rule or the topo-geometrical superposition rule. She introduced topological quantum similarity measures based on the classical topological representation of molecular structures by molecular graphs. She stressed the novelty of this approach in that by substituting classical topological two-dimensional matrices with quantum derived matrices, important three-dimensional information can be accounted for.

Val Gillet (University of Sheffield) presented chemical similarity techniques used in database searching and applied in the pharmaceutical industry. These measures are based on the calculation of the pairwise similarity between a known active molecule and each database compound, and the subsequent ranking of the compounds according to their similarity to the known active. She presented similarity measures based on the representation of compounds by two-dimensional fingerprints (vectors with the binary values of 0 and 1, accounting for the absence or presence of certain fragments), and using the Tanimoto index as a quantitative measure of similarity. She also presented a novel method based on the assignment of four properties to each functional group, encoded by triplets of strings, and the use of reduced graphs. She finally illustrated the theoretical basis with several virtual screening, and data fusion experiments, based on the combination of different rankings on the same sets of molecules.

Brigitte Simon-Hettich (Merck Institute of Toxicology) provided an overview of the chemical category concept from a toxicological point of view, including some examples from the notification of new chemicals in the EU. She introduced the chemical category concept based on its use within the US EPA and the OECD. The main advantages of categories are their potential savings in cost, time, resources, and animal experimentation. She illustrated the principles of the US EPA approach and the OECD approach with some examples. The OECD approach groups compounds which show a predictable pattern in physicochemical properties, environmental fate, environmental effects or human health effects in order to identify and fill in data gaps for relevant endpoints. She raised some questions and concerns related to categories based on common functional groups, metabolic pathways, and incremental changes in groups. For example, the practicality and utility of forming categories based on metabolic pathways was questioned. She also highlighted the need for chemical categories based on common mechanisms of action.

Aldo Benigni (Istituto Superiore di Sanità) provided some practical insights based on the definition of chemical series and the use of chemical similarity in carcinogenic and mutagenic compounds. He started by raising the issue of why there is a need to define a valid chemical similarity measure and gradual scales of it. He highlighted the need for a subdivision between predictions of the biological activity of untested compounds from known QSAR into predictions within the spanned substituent space (SSS) and predictions outside the SSS. He illustrated this using the following classes of chemicals; benzaldehydes, camptothecins, and benzene derivatives.

Jordi Mestres (Municipal Institute of Medical Research) presented new challenges and achievements in the field of chemogenomics. He started by explaining the transition from mapping chemical and biological entities to obtain QSAR to using high throughput mapping techniques (virtual screening and profiling) to produce vast chemogenomic spaces. He showed several classification schemes for both chemical and biological entities and how this is important to facilitate extraction of knowledge from stored data. For biological entities, he presented unified classification schemes based on unique digit codes, illustrating their use for enzymes and nuclear receptors. For chemical entities, he presented a hierarchical classification scheme for chemical structures, based on the molecular equivalence number (MEQNUM) algorithm. This method uses graph chemical identifiers for different levels of description of molecules (scaffold, sidechains, links, ring systems, and rinks) to derive a unique chemical structure code. This classification scheme is very useful for storing data in databases and can enable filling of annotation gaps in the chemogenomic space.

Grace Patlewicz (ECB) introduced the plenary discussion. Using some open questions, she led the discussion on what might be the different steps in a process map for developing chemical categories. The discussion centred on endpoints of high priority within REACH, including skin sensitisation, mutagencity, carcinogenicity, endocrine disruption and reprotoxicity. The first three endpoints are perhaps better understood in terms of their “mechanisms” or at least there is more toxicity data associated with them that enables associations between chemical structure and effect to be made. For example there is a reasonable amount of public information available for mutagenicity and carcinogenicity from the Carcinogenicity Potency DataBase (CPDB), or US National Toxicology Program (NTP), whereas reprotoxicity data is substantially more limited. The suggestion was that knowledge about these endpoints (from toxicologists) could be formulated into simple structural rules, either by using statistical techniques on the available public datasets and cross checking the output with human experts or by interrogating the experts themselves and encoding their knowledge into a computer program. If data was more limited, surrogate assays could be promising tools in formulating mechanistic hypotheses e.g. the information derived from a peptide binding assay may provide sufficient information to enable some mechanistic information to be derived that can help in the formulation of groupings for skin sensitization. Additionally, metabolism information (using data derived from pharmacologists to determine which chemicals are activated, glucuronidated, sulphonated etc) could be used to understand more about the inherent behaviour of chemicals in order to formulate groupings.

Day 2 – Thresholds of Toxicological Concern

Grace Patlewicz (ECB) summarised the discussions carried out on the first day on chemical category formation.

Andrew McDougal (FDA) was unable to participate in person, but he provided a recorded presentation on how TTCs are applied within the US FDA's Office of Food Additive Safety (OFAS). He started by defining the concept of TTC and how it is used as a prioritisation tool within the FDA. He introduced the TOR (threshold of regulation) concept and explained that the Gold (CPDB) database had been used to define the TOR. He outlined current strategies for refining the TOR, such as using structural classes to identify chemicals of higher concern as well as the use of genetic assays that could lower the risk of carcinogenicity. A combination of the Ames test, mouse lymphoma assay and chromosome aberration assay helped to lower the incidence of carcinogens. An OFAS perspective on chemical similarity was provided – focussing on the (Q)SAR tools used, as well as current efforts to organise historical data into structure searchable databases. Following the recorded presentation, Andrew dialled in from the US to take any questions.

Ian Munro (CANTOX Health Sciences International) presented the concepts and assumptions underpinning TTC. He provided an extensive history of TTC and its evolution from the sixties to the present time. He presented an analysis of the threshold values for the carcinogenic compounds in the Gold database, and for non-carcinogenic endpoints. He also presented the Cramer classification tree as a means of classifying substances into one of 3 structural classes which could be used to define different human exposure thresholds. Finally he illustrated how these thresholds have been applied in the safety evaluation of flavouring ingredients by JECFA, an international expert scientific committee administered jointly by the Food and Agriculture Organization of the United Nations (FAO) and the World Health Organization (WHO).

Maria Wallén (Swedish Chemicals Inspectorate) presented a concise literature review and summary of different TTC approaches that had been carried out by KeMI. In particular she highlighted the advantages, limitations, and uncertainties of this approaches.

Bob Safford (SEAC, Unilever) presented current in house work being undertaken in the area of TTC. He presented the TTC as a useful approach in cases of low consumer exposure; such as contaminant incidents, indirect food additives or flavour components in food. Given that the premise of TTC is that 20% of chemicals are carcinogenic, he discussed whether the use of additional information (in silico, in vitro) could lower the incidence of carcinogens. Using the Gold (CPDB) dataset as a starting point, he used the Cramer classification scheme implemented in ToxTree to classify the chemicals into one of three classes. DEREK was used to identify any structural alerts for mutagenicity and carcinogenicity and Ames or mouse lymphoma data (MLA) was taken from the literature. Each piece of information helped to lower the incidence of carcinogens but the MLA was the most effective. DEREK was comparable to the Ames test in reducing the incidence of carcinogens whereas the Cramer classification scheme was found to be over conservative.

Nina Jeliazkova (IDEA Consult Ltd.) gave an overview of the Cramer scheme and demonstrated how this had been encoded into a new piece of software called Toxtree Version 1. She outlined some of the challenges she had encountered in building the software and approaches to resolve these. She also gave a demonstration of the software, showing how easy it was to process one or many structures and how to view the results generated. The software development was funded by ECB and the application will shortly be made available as a free download from the ECB website.

Chihae Yang (Leadscope® Inc.) presented an overview of grouping, adapted to the outcomes and discussions of the workshop. She started presenting a classification of grouping methods, from knowledge-based methods, to supervised and unsupervised methods. She exemplified the different grouping methods implemented in Leadscope software, i.e. expert rules that group chemicals into pre-defined hierarchical classes (more than 27000 fragments), Tanimoto, and Jaccard distance similarity coefficients calculated on fingerprints, unsupervised agglomerative nesting methods, supervised recursive partitioning, recursive partitioning with simulated annealing, new measures being currently developed such as bitset, and the modified Tanimoto coefficient, and analogue (surrogate) based grouping techniques.

Grace Patlewicz (ECB) introduced the second brainstorming session and led the plenary discussion on the basis of a number of issues and questions that arose from the morning's presentations. Discussion points included what modifications if any should be undertaken for the Cramer classification tool, whether TTC could be applied for other endpoints such as skin sensitisation, and what aspects of TTC could be applied in the context of REACH. It was generally agreed that the TTC concept could be difficult to apply in the context of industrial chemicals, since the necessary exposure information is rarely available, and there can be a complex chain of uses down the supply chain. She summarised some of the consensus conclusions and recommendations and outlined the next steps in drafting a report. The participants were thanked for their attendance and contribution and the workshop was closed with a final coffee break.

Ana Gallegos

Grace Patlewicz

16 November 2005