Methodology for grouping the SIN List and development of the SINimilarity tool

BACKGROUND AND CONTEXT

The substitution of hazardous chemical substances by less hazardous alternatives is not a simple task. About 150,000 chemical substances have been pre-registered under the EU chemicals regulation REACH and most of these have unknown hazardous properties. One way to assess the hazards of untested chemical substances is to compare the chemical structures with the structures of tested compounds. This concept is based on the hypothesis that similar chemical compounds have similar biological activities, a hypothesis that is widely accepted in chemical toxicology.

The SIN List contains chemicals that ChemSec has identified as Substances of Very High Concern based on the criteria defined within REACH. The SIN List is very well scientifically investigated and is therefore suitable as a reference source to assess the hazards of other chemical substances.

ChemSec has now created the SINimilarity tool, which can detect whether a chemical substance that is not on the SIN List contains the same group-specific structural elements as SIN substances and/or if it has structural similarity to SIN substances. A structural element is a part or different parts of a molecular structure, connected or placed in a specific way that are important for a certain property. In the context of the SIN List, these structural elements are thought to be responsible for the hazardous properties.

AIMS

The SIN List has been grouped with the aim to make it more user friendly and create a foundation for the SINimilarity tool. Since the substances have been listed on the SIN List because of their hazardous properties, this is also what we have chosen to base the grouping on.

The aim of the SINimilarity tool is to make it easier for companies and other users of the SIN List to avoid substances that may have similar hazardous properties as substances on the SIN List.

GENERAL INFORMATION ABOUT GROUPING

A chemical group is selected based on the hypothesis that the properties of a series of compounds with common structural features will show coherent trends in their toxicological effects or environmental fate properties. A group may be justified on more than one basis, for example a common functional group, common precursors and/or the likelihood of common breakdown products due to physical and biological processes that result in structurally similar compounds. The presence of common behaviour or coherent trends is generally associated with a common underlying mechanism of action. The justification of groups has traditionally been conducted manually by careful examination of chemical structures and their effects. Computational methods have the potential to facilitate this procedure by complementing more traditional approaches by providing mathematical descriptions based on structural features and relating these to measured activity data. The best-known are quantitative structure-activity relationship models (QSAR). (ECHA 2008, OECD 2011, JRC 2007).

THE METHODOLOGY BEHIND GROUPING OF THE SIN LIST CHEMICALS

Structural information for all SIN List substances has been retrieved according to the method described in the next chapter about SINimilarity and stored in a database (ISIS Base), which allows the necessary viewing of the structures. The structures that could not automatically be identified were manually retrieved from the IUPAC names or by Internet searches and manually drawn using ISIS Draw (ISIS Draw). Mixtures containing easily identifiable structural isomers were represented by a single structure containing the most important parts of the molecules in each mixture.

To get the most accurate division for our needs, we chose to group the SIN List manually. For each substance, structural elements were identified by consulting scientific literature

and experts from the Department of Chemistry and Molecular Biology, University of Gothenburg and the Swedish Environmental Institute (IVL). Information on the molecular mechanisms behind hazardous properties and information on the reason for inclusion on the SIN List were important factors. 31 groups were created using this process.

An important aim of grouping is to make the SIN List easier to view and understand. In line with this aim, we have based the group names, when possible, on common names that are also familiar to non-chemists, e.g. "Phthalates" and "Parabens". We have also kept the numbers of groups down by using broad definitions of the structural elements, e.g. different types of aromatic amines could have been divided into several similar groups based, for instance, on the substitution pattern on the nitrogen or on the aromatic ring, but are all placed in the "Aromatic amines" group since this structural element is the most important for the hazard properties of this type of compounds.

Finally, the SIN List substances were assigned to one or more of the newly created SIN groups. About 60 substances could not be placed in any group since a distinct structural element could not be identified in the structures of these substances. Many compounds contain several group-specific structural elements and can therefore belong to multiple groups.

After all SIN substances had been assigned to one or more groups, the structures were manually analysed and definitions of the specific structural elements were created, i.e. the number and type of halogens, how many carbons in a chain, how many rings etc. These rules are used by the SINimilarity tool in assessing SIN-group similarity and can be found in the chapter "Group descriptions".

chemsec
SIN LIST

THE METHODOLOGY BEHIND THE SINIMILARITY TOOL

From the European Chemicals Agency (ECHA) website we have extracted text information about the 150,000 substances that have been pre-registered under REACH. CAS and IUPAC names (International Union of Pure and Applied Chemistry) were used to retrieve structural information primarily by the use of the "Chemical Identifier Resolver" (NCI/CADD Group, 2009) and verified by JChem for Excel (ChemAxon) and/or Chemspider (Chemspider). The structural information was stored as SMILES strings (Anderson, 1987). Of the 150,000 substances, chemical structures could be identified for about 80,000 substances and these are now used in the SINimilarity tool. Structures for the remaining substances could not be retrieved since they only had EC numbers or consisted of complex mixtures, or the structural information could not be retrieved using the available tools. In addition to these 80 000 substances 400 000 further substances have now been added from PubChem as well as a selection from the ZINC database increasing the total dataset to more than 480 000 unique substances. All 480 000 substances do not have a CAS/EC number but can still be identified by name or SMILES notation.

A search in the SINimilarity tool generates a query structure, keyed as SMARTS patterns (Daylight Inc.), which is matched against the SMARTS patterns of the structural elements from the SIN groups. A match is reported if the query structure contains the structural element for a given group. Some groups, like the electrophiles, contain several different structural elements and each of these is matched against the query compound. Matching is performed with publically available tools from the OpenBabel toolkit (O'Boyle et.al 2011).

Determination of similarity between a query structure and available compounds from the SIN List is performed using binary FP2 fingerprints, keyed as 1 or 0. The fingerprints consist of an array of presence or absence of a set of 1021 predefined chemical features, created from atom types, bond types and rings, up to a path length of seven atoms (O'Boyle et.al 2011). The fingerprint from the query compound is compared to all compounds on the SIN List, and every compound that has a match above a certain percentage is reported. The fingerprints used in SINimilarity do not take explicit account of molecular size, only the presence of "paths" of up to seven atoms. The size will then by implication be taken into account, because in a large molecule there are more long paths.

The match is calculated as: the number of features common to the query compound and the reference compound divided by the total number of features in the query and reference compound. In other words a value of 1 corresponds to a perfect match, and 0 means no similarity at all. Since the fingerprints are built from paths of atoms and bonds, compounds that do not have formal bonds or consist of multiple fragments, e.g. inorganic compounds, are not suitable for this type of similarity comparison and will therefore give a very low similarity match.

GROUPS NOT INCLUDED IN OR GIVING UNRELIABLE RESULTS IN SINIMILARITY

The SIN groups "Petroleum" and "Mineral fibres" contain substances of very complex chemical composition. Substances in these groups are for this reason not used in the SINimilarity tool. As stated in the previous chapter, inorganic compounds and many salts are not suited for the similarity methods used in SINimilarity. This is especially true for many compounds in the metal groups. The similarity will be too low to be shown. If the substance contains a group specific metal, it will be identified, which is often the most useful information from SINimilarity on metals.

chemsec
SINIMILARITY

GROUP DESCRIPTIONS

• Alkylphenols
Contains phenols with lipophilic alkyl groups of at least four carbons attached to the aromatic ring. The aromatic ring can contain other functional groups as well. The phenol oxygen is unsubstituted or ethoxylated.

• Aminocarbonyl compounds
Contains amides, carbamates and similar aminocarbonyl compounds.

• Antimony compounds
Contains salts and complexes of oxidized antimony.

• Aromatic amines
Contains compounds with benzene rings substituted with amino, alkylamino, amide, phenylhydrazine or phenylazo groups.

• Arsenic compounds
Contains salts and complexes of oxidized arsenic.

• Azo compounds
Contains both aromatic and non-aromatic azo compounds.

• Beryllium compounds
Contains oxidized and metallic beryllium.

• Bisphenols
Contains compounds in which two phenols are bridged with one carbon or heteroatom. The bridge-atom can be oxidized or substituted with hydrogen, alkyls, phenyl and esters. The phenol oxygens are unsubstituted.

• Boron compounds
Contains salts and complexes of oxidized boron.

• Cadmium compounds
Contains salts and complexes of oxidized cadmium.

• Chromium compounds
Contains salts and complexes of chromium(VI).

• Cobalt compounds
Contains salts and complexes of oxidized cobalt.

• Electrophiles
Contains compounds with the following reactive electrophilic groups; anhydrides, carbamoyl chlorides, carbonyl chlorides, epoxides, aziridines, alkyl sulphates, sulphamoyl chlorides, primary alkylbromides, allylic and benzylic halogens, cationic triaryl-methanes, a, b-Unsaturated carbonyl compounds, mustard-type compounds, dialkyl sulphates, sulphamoyl chlorides, 1,3-propane sultones, diazo alkyls, chloromethyl ethers and 1,2-dihalo alkyls.

• Glycol ethers
Contains ethers and esters of 1,2-ethanediols and 2-hydroxyacetic acid. The carbon backbone can be further substituted with alkyl groups.

• Hydrazines
Contains hydrazine and alkylated hydrazines.

• Lead compounds
Contains salts and complexes of oxidized lead, organolead compounds and metallic lead.

• Mercury compounds
Contains metallic mercury.

• Mineral fibres
Contains naturally occurring fibrous zeolites, asbestos-type minerals and synthetic mineral wools.

• Nickel compounds
Contains salts and complexes of oxidized nickel.

• Nitro compounds
Contains nitroaromatic and secondary nitroalkyl compounds.

• Nitrosamines
Contains alkylated and acetylated nitrosamines.

• Organotin compounds
Contains di- and triorganotin compounds.

• Parabens
Contain esters of parahydroxybenzoic acid. The alcohol part of the ester consists of a carbon chain with at least two carbons.

• Perfluorinated compounds
Contains compounds with fluorinated alkyl chains. At least six carbons in each chain are fully fluorinated. The fluorinated chain is substituted at one end with sulphonamides, sulphonic acid, phosphates, carboxylic acid or iodide. Some compounds contain unfluorinated carbons but are in this context treated as perfluorinated compounds (PFCs).

• Petroleum
Contains aliphatic hydrocarbons and aromatic compounds often in complex mixtures.

• Phthalates
Contains esters of phthalic acid.

• Polyaromatics
Contains compounds with two or more fused aromatic rings. The rings can be substituted with halogen, nitro and amino groups.

• Polyhalogenated alkanes
Contains brominated and/or chlorinated alkanes. The number of halogens is three or more, distributed over two or more carbons. At least two carbons are halogenated. There are at least one halogen/four carbons.

• Polyhalogenated alkenes
Contains brominated and/or chlorinated alkenes. The double bond is substituted with two or more halogens.

• Polyhalogenated aromatics
Contains brominated and/or chlorinated aromatic compounds with one benzene ring with at least three halogens or two or more benzene rings with at least two halogens on each ring.

• Thioaminocarbonyl compounds
Contains thioamides, thiocarbamates, and similar thioamino-carbonyl compounds.

Anderson, E; Veith, GD; Weininger, D (1987). "SMILES: A line notation and computerized interpreter for chemical structures". Duluth, MN: U.S. EPA, Environmental Research Laboratory-Duluth. Report No. EPA/600/M-87/021.

ChemAxon. Part of "J Chem for Office" sold by ChemAxon Kft., Záhony u. 7, Building HX, 1031 Budapest, Hungary. http://www.chemaxon.com.

ChemSpider is a free chemical structure database providing fast access to over 32 million structures, properties, and associated information.

http://www.chemspider.com.

Daylight Chemical Information Systems, Inc. http://www.daylight.com/dayhtml_tutorials/languages/smarts/

ECHA (2008). "Guidance on information requirements and chemical safety assessment Chapter R.6: QSARs and grouping of chemicals". European Chemicals Agency.

ISIS Base is a chemical database program developed by MDL Information Systems. MDL is now part of SYMYX TECHNOLOGIES INC, 3100 CENTRAL EXPRESS WAY, SANTA CLARA, CA 95051, USA.

ISIS Draw was a chemical structure drawing program for Windows, published by MDL Information Systems. It was available free of charge for academic and -personal use. The last version was 2.5 in 2002.

JRC (2007). Worth, A; Bassan, A; Fabjan, E; Saliner, AG; Netzeva, T; Patlewicz, G; Pavan, M; Tsakovska, I. "The Use of Computational Methods in the Grouping and Assessment of Chemicals - Preliminary Investigations". European Commission, Joint Research Centre, Institute for Health and Consumer Protection.

NCI/CADD Group (2009). "Chemical Identifier Resolver". http://cactus.nci.nih.gov/chemical/structure

O'Boyle, NM; Banck, M; James, CA; Morley, C; Vandermeersch, T; Hutchison, GR (2011). "Open Babel: An open chemical toolbox". Journal of Cheminformatics, 3:33

OECD (2011). OECD Series on Testing and Assessment Number 138. "Report of the Workshop on Using Mechanistic Information in Forming Chemical Categories". ENV/JM/MONO 8. Organisation for Economic Co-operation and Development. Paris, France.

The PubChem compounds database, https://www.ncbi.nlm.nih.gov/pccompound

A non-commercial database of almost 35 million substances, http://zinc.docking.org/

This project is supported by the Life+ project of the European Commission DG Environment (Child Protect project), Mistra and the Swedish Environmental Protection Agency.

MISTRA
The Swedish Foundation for Strategic Environmental Research

NATURVÅRDSVERKET
SWEDISH ENVIRONMENTAL PROTECTION AGENCY

chemsec
INTERNATIONAL CHEMICAL SECRETARIAT

www.chemsec.org