cheminformatics data management system

Nina Jeliazkova[1], Joanna Jaworska[2], Bruno Hubesch[3]

[1]IdeaConsult, Sofia, Bulgaria, [2]Procter & Gamble, Brussels, Belgium, [3]Cefic, Brussels, Belgium

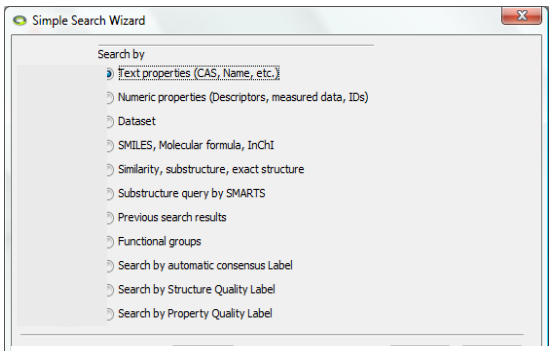
OVERVIEW

AMBIT software is a cheminformatics data management system. It consists of a database and functional modules allowing a variety of simple and complex queries. Data can be organized in hierarchical templates using predefined ontologies.

Unique features of AMBIT:

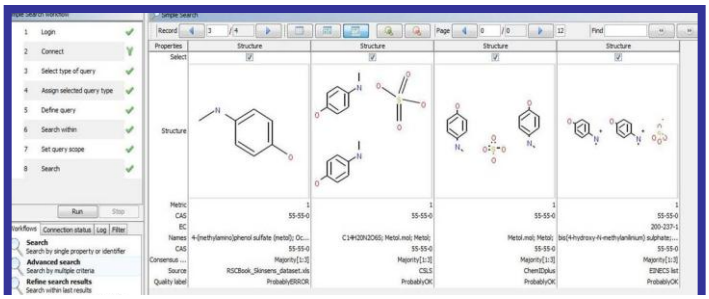
- able to store and query multifaceted information about chemicals (structure, data, text)
- workflow engine to facilitating often repeated user's actions (e.g Analogue identification, PBT assessment are completed and can be applied in REACH).
- data provenance and automated quality assurance
- internal pKa calculation

Querying capabilities



The above queries can be combined by OR and AND operators to form a more complex query.

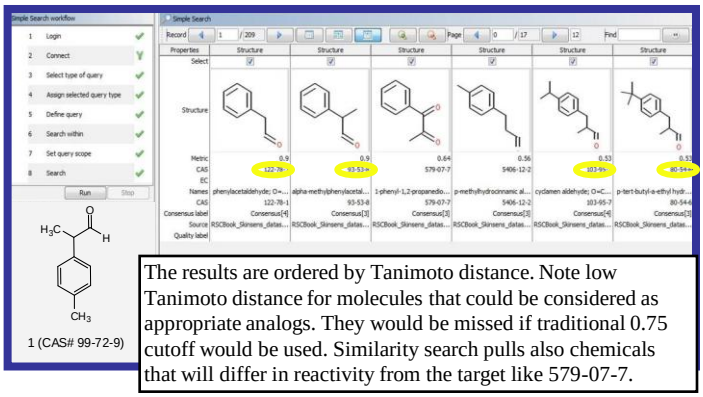
Automated quality assurance and data provenance



When using different data sources or importing new structures it is useful to verify the structure. Ambit keeps track of the source of the chemical and compares structures having the same identifier such as CAS. In the example structure with CAS 55-55-0 imported in a set RSCBook is checked against CSLS, ChemIDPlus and EINECS. Automatic comparison of a search results for chemical with CAS 55-55-0 generates a quality label. Typical errors are illustrated in this example: ignoring full structure of the salt, presenting structure as neutral while it is charged.

Analogue identification for CAS # 99-72-9 for skin sensitization

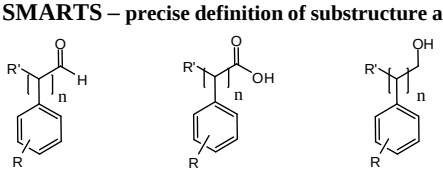
Fingerprints similarity



1 (CAS# 99-72-9)

The results are ordered by Tanimoto distance. Note low Tanimoto distance for molecules that could be considered as appropriate analogs. They would be missed if traditional 0.75 cutoff would be used. Similarity search pulls also chemicals that will differ in reactivity from the target like 579-07-7.

SMARTS – precise definition of substructure and atom types



2 3 4

The search strategy for the target **is formulated by an expert** considering reactivity, metabolism and steric effects, and is shown in compound **2** where R and R' can be H, Me or Eth groups. The n is 1-2 carbons. When n is 2 the R' group is connected to the carbon which is next to the aldehyde functional group. Additionally, if searches around **2** do not bring back sufficient data, it may be appropriate to consider carboxylic acid **3** and alcohol **4**, which would be the potential metabolites of **1** via the *C*-oxidation and reduction pathways. In the compound **3** and **4**, the R and R' groups and n should be the same as that of the compound **1**.

The whole strategy can be written as the following SMARTS: [$(c1cc([$([#1][CH3]),$([CH2][CH3])])ccc1[$(C(C)C(=O)),$(CC(C)C(=O))]), $(c1c([$([[#1]CH3]),$([CH2][CH3])])cccc1[$(C(C)C(=O)),$(CC(C)C(=O))]),$(c1([$([[#1]CH3]),$([CH2][CH3])])ccccc1[$(C(C)C(=O)),$(CC(C)C(=O))])] and was executed as one single query.

The search yields 122-78-1 as a single hit. For verification purpose the SMARTS searches were relaxed to allow for arbitrary substituents at para – position. The relaxed search yielded the following hits 103-95-7, 93-53-8, 80-54-6. The chemical assessed for skin sensitization potential is evaluated for its potential to act as an electrophile towards nucleophilic groups on skin protein. The analogue candidates 103-95-7, 93-53-8, 80-54-6 are aliphatic aldehydes with para-substituted aromatic ring. The substitutents such as methyl, ethyl groups at para-position would not significantly change the electronic nature and reactivity of these molecules. In contrast, the steric hinder groups such as isopropyl, tert-butyl and isobutyl groups may have steric effect and tend to affect the physicochemical properties, ability of absorption or skin penetration. By considering above thetarget chemical is more likely to be close to 122-78-1 and 93-53-8 (Moderate) rather than 103-95-7 and 80-54-6 (Weak), because the steric effects of the substituents of the latest two will likely to lower absorbtion/penetration ability and therefore skin sensitisation potential.

In general similarity based search yields a mix of suitable and not suitable chemicals for read across and needs to be consulted with chemist for evaluation. In homogenous data sets the search results by SMARTS and similarity will give similar results , however this is not to be expected searching diverse, large libraries. As such SMARTS based search is far more efficient, and read-across more transparent and verifiable.

http://ambit.sourceforge.net

Acknowledgments