



Prediction of Carcinogenicity for Some Compounds from the CosIng Database using the VEGA Platform (CAESAR) Program



Alja Plošnik, Marjan Vračko, Jure Zupan

National Institute of Chemistry, Hajdrihova 19, 1000 Ljubljana, Slovenia

alja.plosnik@ki.si

Introduction

Carcinogenicity of chemical compounds is one of the frequently discussed topics in relation to their potential negative effects on human health. Chemicals are defined as carcinogenic if they induce tumours (benign or malignant), increase tumour incidence, or shorten the time to first tumour occurrence. Ordinary conventional experimental test takes a lot of time, money and animal casualties. Therefore, for economic, social and ethical pressures new approaches like Quantitative Structure-Activity Relationships (QSAR) carcinogenicity models are proposed [1,2,3].

CAESAR was an EC funded project (Project no. 022674 - SSPI), which was specifically dedicated to develop QSAR models for the REACH legislation. The models have been assessed according to the five OECD principles for validation of (Q)SAR models used for regulatory purposes (a defined endpoint, an unambiguous algorithm, a defined domain of applicability, appropriate measures of goodness-of-fit, robustness and predictivity, and a mechanistic interpretation, if possible). Five endpoints with high relevance for REACH have been addressed within CAESAR: bio-concentration factor, skin sensitisation, carcinogenicity, mutagenicity and developmental toxicity. For a predicted compound the CAESAR models provide, beside the prediction, a comprehensive information on applicability domain and a set of six most similar compounds from the training set (the similarity set) [1,2,3].

The main purpose of this work is to analyse the results of carcinogenicity predictions, their applicability domain, and similarity sets for 558 randomly chosen compounds from the CosIng database.

Methods

The investigated dataset contains 558 chemicals which were randomly chosen from the CosIng database of 20,000 substances. The Inventory CosIng (<http://ec.europa.eu/consumers/cosmetics/cosing>) includes the data for cosmetics ingredients since the adoption of the Cosmetics Directive in 1976. CAS registry numbers and chemical names were taken from CosIng data base, while the single line structure notation SMILES codes were taken from publicly available databases PubChem system.

The CAESAR model for carcinogenic potency prediction was built on a training data set of 807 compounds applying the Counter Propagation Artificial Neural Networks (CP ANN) as a modelling method. The task of CP ANN is to adjust the weights in the neurons according to similarity between subjects i.e., similar objects are located near each other in the network [2,4].

In our investigation the compounds were represented with vectors, which were constructed from the similarity sets. Technically, it means that each compound was represented with a 807 dimensional vector where each dimension indicates one compound of training set. The components of vector are zero or one, if a compound belongs to the similarity set.

The Kohonen layer of CP ANN model is shown in Figure 2. Compounds are divided into three groups, cyclic, aromatic and aliphatic compounds.

Results

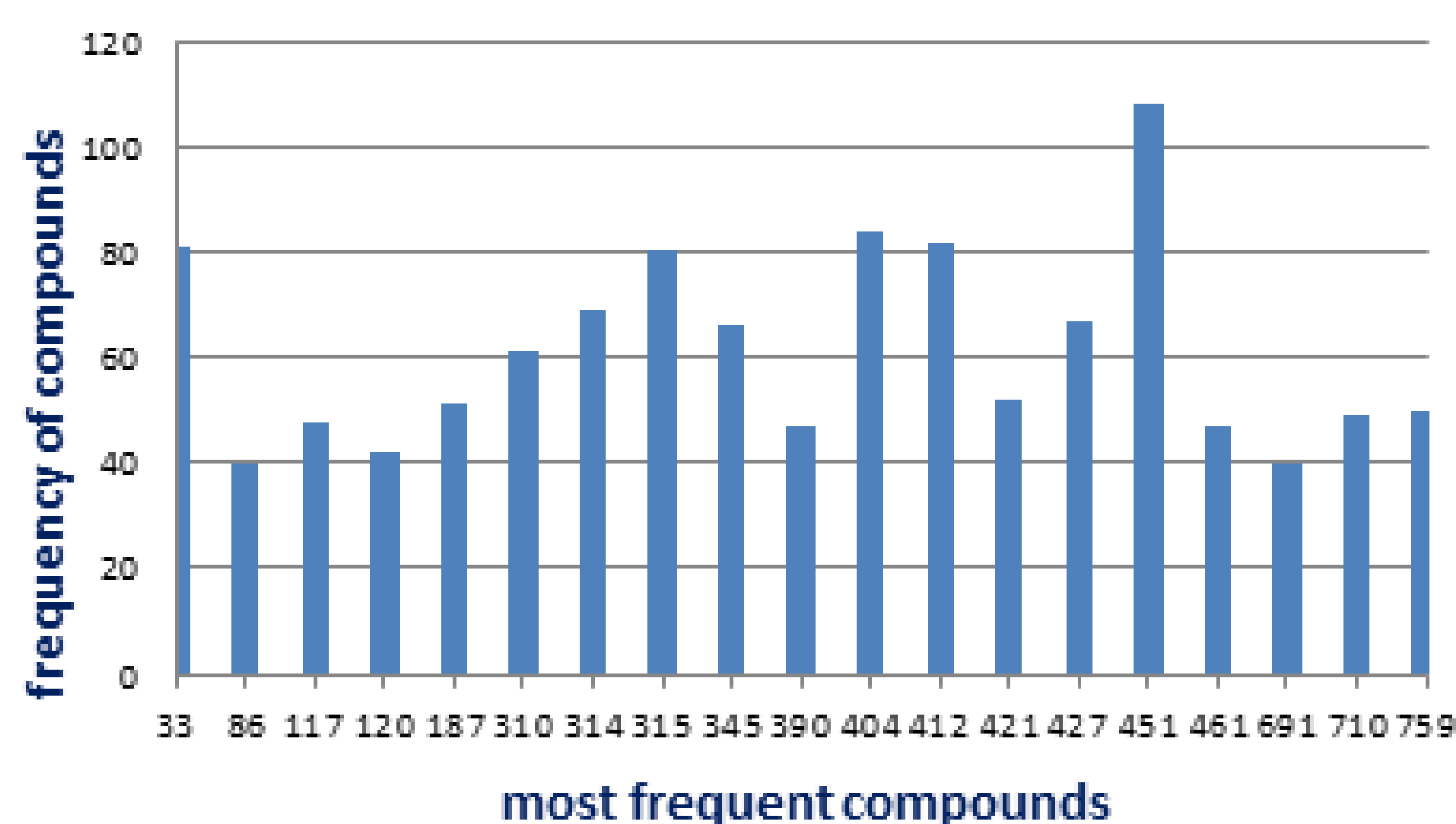


Figure 1. Frequency of nineteen most represented compounds from the training set.

Results and Conclusions

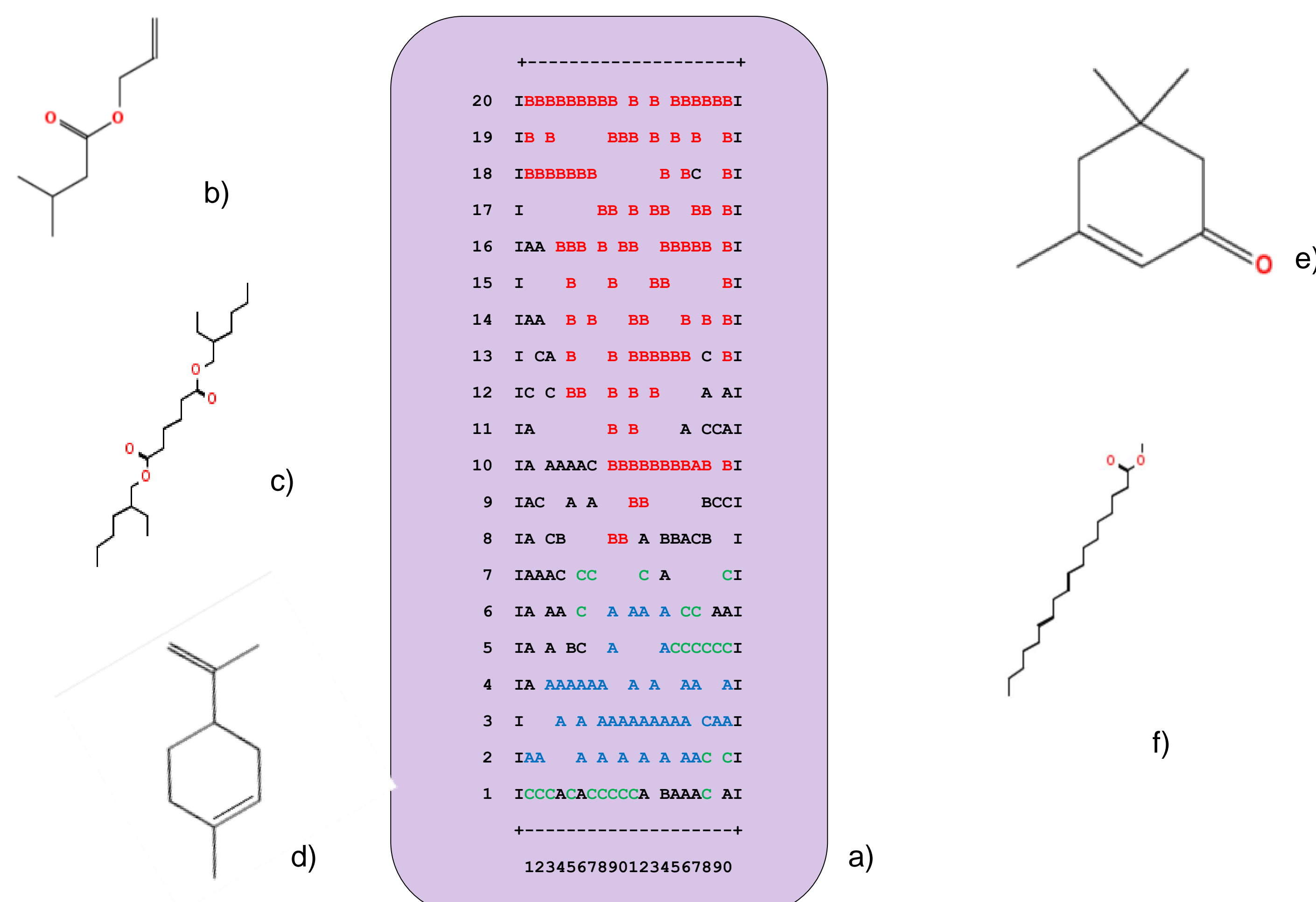


Figure 2. a) Distribution of compounds from the training and test set on the Kohonen layer of CP ANN (A-aliphatic compounds, B-aromatic compounds, C-cyclic compounds). The dimension of CP ANN is 20 X 20 and 200 learning epochs. b) prop-2-enyl 3-methylbutanoate, c) bis(2-ethylhexyl) hexanedioate, d) 1-methyl-4-prop-1-en-2-yl cyclohexene, e) 3,5,5-trimethylcyclohex-2-en-1-one, f) methyl octadeca-9,12-dienoate.

IUPAC name	CAS	Experimental value	Predicted value	Incidence
prop-2-enyl 3-methylbutanoate	2835-39-4	Positive	Non-positive	81
bis(2-ethylhexyl) hexanedioate	103-23-1	Non-positive	Non-positive	80
1-methyl-4-prop-1-en-2-ylcyclohexene	5989-27-5	Positive	Non-positive	82
3,5,5-trimethylcyclohex-2-en-1-one	78-59-1	Positive	positive	84
methyl octadeca-9,12-dienoate	112-63-0	Positive	Positive	108

Table 1. Characterisation (incidence, prediction and experimental) of five most represented compounds from the training set.

Figure 1 and Table 1 show the compounds of the CAESAR training set, which most frequently appear in the similarity sets.

Figure 2 shows the Kohonen layer of CP ANN model neural network. One can recognise the area of aromatic compound above (B-blue range), the area of aliphatic compounds (A-red range) and the area of cyclic compounds (C-green range).

References

- [1] Benfenati E. The CAESAR project for in silico models for the REACH legislation. Chem Cent J 2010;4(Suppl 1):11.
- [2] Fjodorova N, Vračko M, Novič M, Roncaglioni A, Benfenati E. New public QSAR model for carcinogenicity. Chem Cent. J. 2010;4(Suppl 1):S3.
- [3] Cronin MTD, Jaworska JS, Walker JD, Comber MHI, Watts CD, Worth AP. Use of QSARs in International Decision-Making Frameworks to Predict Health effects of Chemical Substances. Environmental Health Perspectives 2003;111(10):1391-1401.
- [4] <http://www.caesar-project.eu> (cited 25.1.2013)