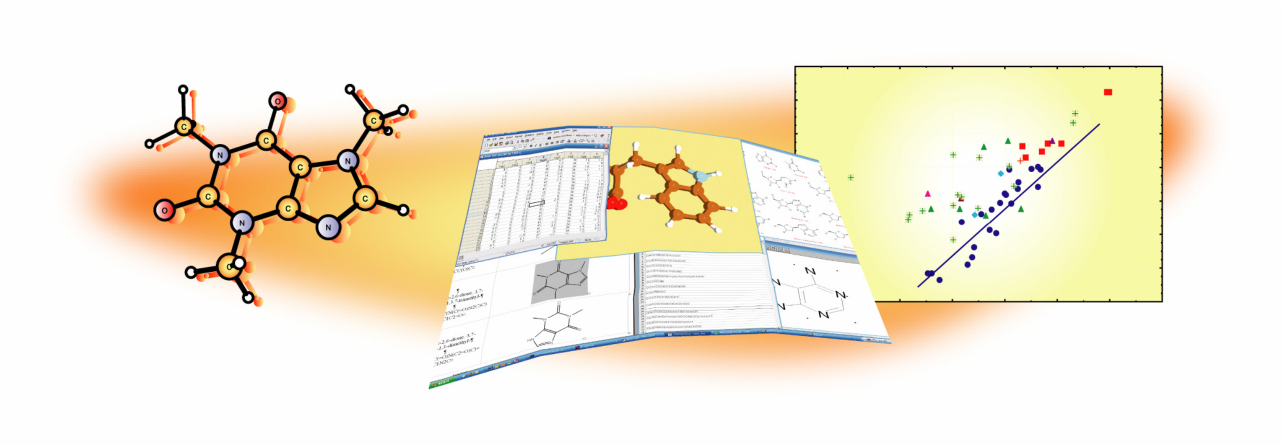
EUROPEAN COMMISSION
DIRECTORATE-GENERAL
Joint Research Centre

# Development and Beta Testing of the Toxmatch Similarity Tool

Ana Gallegos Saliner & Andrew P. Worth



IHCP

2007

EUR 22854 EN

The mission of the IHCP is to provide scientific support to the development and implementation of EU policies related to health and consumer protection.

European Commission
Directorate-General Joint Research Centre
Institute for Health and Consumer Protection

Contact information: Andrew Worth
Address: European Chemicals Bureau, TP581
E-mail: andrew.worth@ec.europa.eu
Tel.: +39 0332 789566
Fax: +39 0332 786717

http://ihcp.jrc.cec.eu.int/; http://ecb.jrc.it/; http://ecb.jrc.it/qsar/
http://www.jrc.cec.eu.int

Legal Notice
Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use which might be made of this publication.

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 22854 EN
ISSN 1018-5593

Luxembourg: Office for Official Publications of the European Communities

Printed in Italy

Quality control insert

| | *Name* | *Signature* | *Date* |
|---|---|---|---|
| **Report Prepared by:** | **Ana Gallegos Saliner** | | 25.06.07 |
| **Reviewed by: (Scientific level)** | **Andrew Worth** | APWorth | 20.06.07 |
| **Approved by: (Head of Unit)** | **Steven Eisenreich** | | 20.06.07 |
| **Final approval: (IHCP Director)** | **Elke Anklam** | | 20/6/07 |

# ABSTRACT

Toxmatch is a software tool that provides a means of grouping chemical substances according to various measures of chemical similarity. It is designed to support the formation of chemical categories and the application of read-across within the hazard and risk assessment of chemicals. Such a tool will be useful for scientific researchers, for end-users in industry, for regulatory authorities, and for the EU Chemicals Agency.

Toxmatch was developed by Ideaconsult Ltd (Sofia, BG) under the terms of a JRC-ECB contract. It is a flexible user-friendly, computer-based open source application which is accessible via internet. It encodes and applies a range of different structural and descriptor based chemical similarity indices. The novelty of this software lies in its ability to calculate similarity measures that are tailored for specific activities/toxicities. Thus, relevant chemical representations can be selected for a given activity and the chemicals of interest can hence be classified into toxicity classes.

The present document summarises the beta testing of Toxmatch, reporting general comments and suggestions for further improvement.

# LIST OF ABBREVIATIONS

| | |
|---|---|
| AE | Atom environments |
| BfR | (German) Federal Institute for Risk Assessment |
| ECB | European Chemicals Bureau |
| EINECS | European Inventory of Existing Commercial Chemical Substances |
| EPA | Environmental Protection Agency |
| EU | European Union |
| JRC | Joint Research Centre |
| *k*NN | *k*-Nearest Neighbours |
| MCS | Maximum Common Substructure |
| MoA | Mode of Action |
| OECD | Organisation for Economic Co-operation and Development |
| (Q)SAR | (Quantitative) Structure-Activity Relationship |
| REACH | Registration, Evaluation and Authorisation of Chemicals |
| RIP | REACH Implementation Project |

**TABLE OF CONTENTS**

1. Background to the project .......... 1

2. Description of the software .......... 4

2.1. Similarity indices .......... 4

3. Work packages of the project .......... 8

3.1. Work package 1: Review .......... 8

3.2. Work package 2: Compilation of toxicity datasets .......... 8

3.3. Work package 3: Examination of existing literature .......... 8

3.4. Work package 4: Development of a user friendly software tool .......... 8

3.5. Work package 5: Capacity building and beta testing .......... 8

3.6. Work package 6: Dissemination of results .......... 9

4. Important dates and deliverables .......... 10

4.1. First deliverable: review ‘Similarity in toxicology’ .......... 10

4.2. Second deliverable: first face-to-face meeting and minutes .......... 10

4.3. Third deliverable: beta-prototype .......... 10

4.4. Fourth deliverable: Release of different versions .......... 11

4.4.1. Version 1.00 .......... 11

4.4.2. Version 1.01 .......... 11

4.4.3. Version 1.02 .......... 11

4.4.4. Version 1.03 .......... 11

4.5. Fifth deliverable: training course on the software .......... 12

4.6. Sixth deliverable: Beta-test of the software .......... 12

5. Overview of the evaluation of Toxmatch .......... 13

6. Identification of technical imprecision and errors .......... 14

6.1. Version 1.01 .......... 14

6.2. Version 1.02 .......... 14

6.3. Version 1.03 .......... 15

7. Comments and suggestions on different issues .......... 17

7.1. Classification by using sets of rules .......... 17

7.2. Graphical interface .......... 17

7.3. Help file and capacity building .......... 18

7.4. Import/Export........19
7.5. Installation........19
7.6. Large datasets........19
7.7. Similarity indices and measures........20
7.8. Stability........20
7.9. User configuration file........20
7.10. User-friendliness........20
8. Overall performance of Toxmatch / Conclusions........21
9. References........23

## 1. Background to the project

Toxmatch was developed as a result of a proposal approved within the JRC Innovation Project Competition in 2005. The aim of the project proposal was to develop the prototype of a software tool for supporting the risk assessment of chemical substances. Such a tool will be useful for scientific researchers, for end-users in industry, for regulatory authorities, and in the future EU Chemicals Agency.

This tool will be especially useful in views of the forthcoming REACH legislation, which will result in some 30,000 chemicals requiring evaluation for toxicity, ecotoxicity and environmental fate, over a period of 11 years [1]. For reasons of cost, practicality, and animal welfare, this assessment exercise cannot be achieved by applying traditional test methods. To address this issue, the REACH proposal foresees greater use of *in silico* approaches, such as (Quantitative) Structure-Activity Relationships [Q)SARs], read across and chemical categories. Analyses carried out by ECB have shown that these non-testing approaches can provide an efficient means of obtaining the required information on chemicals while reducing testing costs and the amount of (animal) testing necessary [2]-[3].

The chemical grouping approach is based on the premise that, for a given chemical property, the experimental data for one chemical can be used to predict the same property for a similar chemical. This is called ‘read-across’ and is useful when reliable indicators of chemicals similarity are used. Furthermore, the grouping of three or more similar chemicals into a ‘chemical category’ allows for multiple ways of predicting the unknown properties of chemicals (i.e. filling the data gaps). However, the scientific basis of grouping methods is only partially developed, and tools are urgently needed to apply such methods for the purposes of REACH. This was concluded in a recent multi-stakeholder evaluation of the workability of the REACH legislation [4]. The overall objective of the proposal for the software development was to contribute to the science of chemical similarity and to provide an internet-accessible and user-friendly tool for classification of chemical substances into chemical categories. Due to the partially-developed state of the science, and the absence of such a computer-based tool, this project adds significant scientific and technological value.

Some guidance on how to group chemicals has been developed by the OECD [5], and some examples of chemical categories are provided by the US EPA [6]. The

OECD guidance is written at a very general level, referring to the main steps that Industry needs to consider when developing a category proposal for regulatory consideration. However, the guidance does not give detailed, practical advice on how to formulate a category (e.g. specifying what information is necessary, which methods should be applied, and what tools would facilitate the process). Furthermore, the OECD guidance does not explain how to interpret chemical similarity in a context-dependent and scientifically-meaningful way. In fact, an active field of research in QSAR is dedicated to this very question. More recently, the ECB has been following developments in this field and has provided a more practical guidance on how to formulate and use chemical categories [7]-[8]. ECB is also carrying out some novel research to identify and evaluate quantitative measures of chemical similarity applicable to specific regulatory endpoints.

The proposed approach to deal with these issues was to develop the prototype of a user-friendly, internet-accessible, computer-based tool that could enable stakeholders to group chemicals within their own databases, or within a regulatory inventory of chemicals, such as the EINECS list of existing EU substances. To code the software tool a bid for tender was launched. Nina Jeliazkova (Ideaconsult Ltd., Sofia, Bulgaria) was the external contractor selected to develop the similarity tool on behalf of ECB as a part of the exploratory research project.

The project involved research to develop concepts and methods for assessing similarity, as well as the coding of new software to implement the methods for transparent use by the non-specialised user. By making the software tool readily available, the JRC will promote a consistent approach to grouping by stakeholders involved in chemical risk assessment process. This is crucial for the harmonised implementation of EU legislation across the Member States. The successful provision of such a tool will also enhance the visibility of the JRC, and demonstrate its commitment to providing technical support to the future Chemicals Agency.

The end-users of the final product will be: a) industrial companies in the EU, who will be required to register chemicals under the REACH legislation, and who will need a tool for grouping similar chemicals; b) EU regulators in the Competent Authorities responsible for the implementation of REACH, who will need to review the registration dossiers submitted by Industry; c) the future Chemicals Agency, which will need a tool to explore groupings of chemicals produced by multiple industries, with a view to putting different companies in touch with each other, so that they could form consortia that pool their data; and d) scientific researchers.

The JRC retains the intellectual property rights of the tool, but as an open-source software application, it is made freely available via the ECB website [9].

## 2. Description of the software

Toxmatch (Ideaconsult Ltd.) is a flexible and user-friendly open source application specifically commissioned by ECB, which encodes and applies a range of different structural and descriptor based chemical similarity indices. The novelty of this software lies in its ability to calculate similarity measures that are tailored for specific activities/toxicities. Thus, relevant chemical representations can be selected for a given activity and the chemicals of interest can hence be classified into toxicity classes.

Toxmatch is aimed at performing standard calculations of pairwise similarity measures using a wide range of different similarity indices. Furthermore, it also allows the prediction of an endpoint on the basis of similarity measures, the classification in groups/categories, and the calculation of the similarity to groups. On one hand, the similarity to the entire dataset allows the prediction of the activity based on nearest neighbours. On the other hand, the calculation of the similarity to pre-defined dataset groups allows the classification between the groups, and each compound is assigned a probability to belong to each group. Finally, it is also possible to perform read across on a case by case basis, by calculating the weighted similarity of nearest neighbours or most similar compounds. The implementation of prediction and classification is based on Weka data mining software [10].

### *2.1. Similarity indices*

Toxmatch allows the calculation of several types of similarity indices (described below) [11]-[12]. Similarity indices can be arranged in a symmetric matrix, where each element $\mathbf{Z}_{AB}$ corresponds to the pairwise comparison of molecule *A* with molecule *B*, and the diagonal is composed by self-similarities. Toxmatch provides the upper triangle of the similarity matrix.

### a) Descriptor-based similarity indices

Descriptor-based similarity indices are calculated on the basis of the descriptors. The strength of this method relies on the use of descriptors relevant for the mechanism of toxicity of interest. For the descriptor-based similarity, the calculated distance is the averaged distance between the point and its *k* nearest neighbours.

**a1) Distance-like dissimilarity (D-class) indices** range from 0 (total similarity) to infinity (complete dissimilarity), if they are not normalised. The generalised formula can be expressed as follows:

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

where $Z_{AB}$ is the distance between two objects $A$ and $B$, and $k$ and $x$ are scalars.

When $k$=$x$=2, the previous equation reduces to the **Euclidean distance index**:

$$D_{AB}=\sqrt{Z_{AA}+Z_{BB}-2Z_{AB}}\ \text{ or }\ D_{AB}=\sqrt{\sum_{j=1}^{j=n}\left(x_{jA}-x_{jB}\right)^2}\ , \tag{2}$$

Where $x_{jA}$, and $x_{jB}$ are the coordinates of position of the molecule and $A$ and the molecule $B$, respectively.

**a2) Correlation-like similarity (C-class) indices** range from 0 (total dissimilarity) to 1 (complete similarity). Tanimoto and Hodgkin – Richards coefficients can be extracted from the following generalised formula:

$$V_{AB}(k,x)=(k-x)Z_{AB}D_{AB}^{-2}(k,x);\ V_{AB}=[0,1], \tag{3}$$

where $D_{AB}$ is the previously given definition of a D-class index.

When $k$=2 and $x$=1, the previous equation corresponds to the **Tanimoto coefficient**:

$$T_{AB}=Z_{AB}\left[Z_{AA}+Z_{BB}-Z_{AB}\right]^{-1}\ \text{ or }\ T_{AB}=\frac{\sum_{j=1}^{j=n}x_{jA}x_{jB}}{\sum_{j=1}^{j=n}\left(x_{jA}\right)^2+\sum_{j=1}^{j=n}\left(x_{jB}\right)^2-\sum_{j=1}^{j=n}x_{jA}x_{jB}} \tag{4}$$

For $k$=2 and $x$=0 the **Hodgkin – Richards coefficient** holds:

$$H_{AB}=2Z_{AB}\left[Z_{AA}+Z_{BB}\right]^{-1}\ \text{ or }\ H_{AB}=\frac{2\sum_{j=1}^{j=n}x_{jA}x_{jB}}{\sum_{j=1}^{j=n}\left(x_{jA}\right)^2+\sum_{j=1}^{j=n}\left(x_{jB}\right)^2} \tag{5}$$

Alternatively, the **Cosine-like similarity index**, so-called **Carbó index**, can be calculated as:

$$C_{AB} = Z_{AB}\left[Z_{AA}Z_{BB}\right]^{-1/2} \text{ or } C_{AB} = \left[\frac{\sum_{j=1}^{j=n} x_{jA}x_{jB}}{\sum_{j=1}^{j=n}\left(x_{jA}\right)^2 \cdot \sum_{j=1}^{j=n}\left(x_{jB}\right)^2}\right]^{1/2} \tag{6}$$

**b) Structure-based similarity indices**

In this case, the toxicity information can be extracted by specifying structural patterns based on fragment frequency. Three different structure-based similarity indices have been implemented in Toxmatch:

**b1) Fingerprints**: bit strings which account for the presence or absence of molecular fragments within a molecule.

**b2) Atom environments (AE)**: arrays that contain counts of molecular features instead of binary bits, which are present at a certain topological distance. AE can be regarded as fragments, surrounding each atom in a molecule, up to a predefined level [13]-[15].

**b3) Maximum Common Substructure (MCS)**: accounts for the largest substructure in common between two molecules. The MCS is calculated as the number of common elements provided by the matching conditions [16]. A measure of similarity obtained by the MCS between two compounds *s* and *t* is given by:

$$SI_{st} = \frac{(A+B)_{MCS}}{(A+B)_s} . \frac{(A+B)_{MCS}}{(A+B)_t} \tag{7}$$

where (*A*+*B*) is the sum of atoms and bonds in the MCS of compounds $s^{th}$ and $t^{th}$, respectively.

**c) Set of rules for specific activities**

Verhaar scheme performs classification into Mode of Action (MoA) groups by applying structural rules to predict aquatic toxicity [17].

The BfR physicochemical and structural rules for the prediction of skin irritation and corrosion have been extensively reported in the literature [18]-[19]. The physicochemical rulebase is used to predict the absence of skin irritation, and the structural alerts are toxicophore fragments that predict skin irritation.

### d) Clustering method

Supervised learning techniques can be used in order to select the relevant chemical representations and classify them into toxicity classes. In this case, the composite (averaged) similarity measure between a compound and a selected subset of compounds is performed by *k*-Nearest Neighbours (*k*NN). This clustering algorithm calculates the average similarity between a query compound and the nearest *k* compounds, where *k* is arbitrarily chosen. The implementation is based on Weka Data Mining software kNN clustering [10].

## 3. Work packages of the project

### *3.1. Work package 1: Review*

The first step of the project was to write a scoping literature-based review on the field of similarity measures, and on the use of particular models/descriptors to predict a limited number of selected regulatory endpoints [20]. As a starting point, four endpoints were considered: aquatic toxicity, bioconcentration, skin sensitisation and skin irritation, where the latter two endpoints are important for animal-welfare reasons.

### *3.2. Work package 2: Compilation of toxicity datasets*

The software tool provides training data sets for the four predefined toxicity endpoints. In particular, datasets were compiled containing experimental data obtained for the fathead minnow test (fish toxicity), the bioconcentration factor, the mouse local lymph node assay (sensitisation), and the EU classification phrases for skin irritation and corrosion. This work package included an evaluation of chemical diversity and data quality within the datasets, and the structuring of the data in an appropriate platform.

### *3.3. Work package 3: Examination of existing literature*

Existing literature was examined to identify the structural/chemical factors that drive the endpoints.

### *3.4. Work package 4: Development of a user friendly software tool*

Based on the evaluation in work package 3, the most promising rules were coded into a computer platform for automated use. This work requires considerable computer programming expertise, and was outsourced to a contractor (Nina Jeliazkova, Ideaconsult Ltd., Sofia, Bulgaria). Training documentation on how to use the software was also developed.

### *3.5. Work package 5: Capacity building and beta testing*

A training course on Toxmatch was organised by the contractor to a limited number of experts, who tested the functionalities of the software, performed the beta test report on the performance of the tool, and suggested improvements to the user manual.

### *3.6. Work package 6: Dissemination of results*

In the final stage, innovative scientific work applying different functionalities of the tool has been carried out and presented at several scientific conferences with international scope [21]-[23]. The prototype will subsequently be disseminated as a downloadable program from the ECB website [9], following any necessary modifications.

## 4. Important dates and deliverables

The official start date of the contract was 22/12/2005 and the formal end date was 21/09/2006 (i.e. nine months later). Later the contract was postponed three extra months 21/12/2006.

### *4.1. First deliverable: review 'Similarity in toxicology'*

Completed by 3rd April 2006.

As a first deliverable of the project, the contractor conducted a review of chemical similarity and their application in chemical categories for specific endpoints [20].

### *4.2. Second deliverable: first face-to-face meeting and minutes*

Held on 19th and 20th April 2006.

Discussion and dialogue to feed in ideas and recommendations from the various REACH Stakeholders such as those within the EU QSAR Working group, OECD QSAR group, and RIP 3.3.

The aim of the visit of the contractor was to consider recommendations and discuss the next steps of how the tool would be developed. The topics treated in this meeting were:

- Presentation on the similarity review (slides summarising the review and highlighting areas of specific interest or difficulties were presented)
- Proposed recommendations and ideas for next steps (interim report)
- Discussion and agreement on next steps including foreseen milestones/ deliverables
- Contract timelines and postponement of the contract
- Publications arising out of the contract work
- Discussion of parallel synergistic activities on categories and impact on this work

As a result of the first face-to-face meeting, some detailed minutes were compiled.

### *4.3. Third deliverable: beta-prototype*

11th October 2006

The similarity tool was in good shape and with many features running (some screenshots circulated) but not yet released.

**Plan to implement unambiguous BfR rules**. It was agreed that physicochemical exclusion rules could be either entered manually one by one or read from a file specified by the end-user (i.e. experimental values available), or calculated on the fly by using batch processing or single entry mode [24] (i.e. no experimental data available).

## *4.4. Fourth deliverable: Release of different versions*

### 4.4.1. Version 1.00

Released on 1st December 2006.

The first downloadable version with a Windows installer was released from http://ambit.acad.bg/Toxmatch/ (user: jrc, password: Toxmatch).

### 4.4.2. Version 1.01

Released on 18th January 2007.

The BfR rules were additionally implemented as a Toxtree plugin[1] and could be tested also within Toxtree interface [25]. The updated Toxtree 1.12 was also downloadable from the same website [9]. Toxtree and Toxmatch are independent, and the BfR rules can be applied from one program or another.

### 4.4.3. Version 1.02

Released on 5th February 2007.

Further improvements were incorporated.

### 4.4.4. Version 1.03

Released on 23rd March 2007.

Improvements resulting from the first comments from the beta testing procedure were incorporated.

[1] Toxtree is a flexible and user-friendly open-source application that places chemicals into categories and predicts various kinds of toxic effect by applying decision tree approaches, developed by Ideaconsult, Ltd.

### *4.5. Fifth deliverable: training course on the software*

Delivered on 7th February 2007 by Nina Jeliazkova (Ideaconsult Ltd), in the ECB (Ispra, Italy).

### *4.6. Sixth deliverable: Beta-test of the software*

It started after the release of version 1.02 and has been used as a feedback to improve the subsequent versions of Toxmatch. It has been compiled in the present document.

## 5. Overview of the evaluation of Toxmatch

The following items have been addressed for the evaluation of Toxmatch software, based on its different released versions:

- Evaluation of the software technical content (implemented descriptor and structure-based similarity measures)
- Evaluation of the user-friendliness of the software
- Evaluation of the tool utility (basic statistics analysis, similarity indices, graphical user interface)
- Identification of possible errors
- Discussion of practical issues, such as the software general performance, time processing, capability to process big data matrices
- Additional comments and suggestions for further improvement

During the process of evaluating Toxmatch, there has been a close contact with the contractor Nina Jeliazkova (Ideaconsult, Ltd). Errors and various technical problems during the evaluation process have been continuously corrected. Although various aspects of the performance of Toxmatch have been evaluated during this process, further errors or technical problems may inevitably be discovered when the software will be applied for specific problems.

As a consequence of the close contact with Ideaconsult Ltd. during the evaluation process, this report will summarise the findings that have been corrected in further version(s) of Toxmatch. It should be noted that these are not error prohibitive for a proper operation of the software, but solely suggestions to improve the user friendliness and performance. The release of Toxmatch to a broader audience is assumed to further contribute to the release of further improved version(s) of the tool.

## 6. Identification of technical imprecision and errors

### *6.1. Version 1.01*

When an arbitrary (not predefined) training set file is opened, some functionalities coded in the configuration option are missing:

- In order to import a given activity as the endpoint for the training set, the program expects a *Result* field to exist in the training set file (i.e. a column in CSV/XLS file or a field in SDF file). This is already fixed in version 1.02.

- The Euclidean distance index and maximum common substucture methods are not implemented. This is already fixed in version 1.02.

- Euclidean, Cosine, Hodgkin-Richards, and Tanimoto indices allow discriminating between similarity to the group and to the dataset, while fingerprints and atom environments only allow calculating the similarity to the dataset. This is due to the internal implementation and is changed in version 1.02.

### *6.2. Version 1.02*

It mainly addresses the following problems:

- When loading an arbitrary training set, a window asking to transfer and configure the meaning of fields from the file appears (three tabs are available: *Identifiers*, *Descriptors* and *Result*). For a field to be considered as the endpoint, it should be moved into the *Results* tab. However, when using a test set, it does not recognise the endpoint field to be predicted, even if it was previously transferred into *Results* tab.

- When opening a test file, it is possible to establish a correspondence between the test set and the training set descriptors. To do this, the user should click on the second column, and this should bring a list with the available descriptor names in the training set.

- Euclidean distance index and maximum common substucture methods are implemented.

- Nearest neighbours classification and prediction to structural similarity methods are added. Thus, fingerprints, atom environments and maximum common substucture methods are not only able to calculate similarity values as in the previous version, but are also able to behave in the same way as descriptor based similarity, i.e. predict endpoint values or perform classification.

- Some misleading headings are changed, i.e. *Similarity to entire data set* by *Calculate smilarity and predict activity* and *Similarity to groups* by *Calculate similarity and classify into groups*.

- When selecting groups for the test set it is not possible to change the colour of the groups. This is amended in version 1.03.

- It is not possible to run the BfR rules implemented in Toxtree because of the naming of the descriptors. This is amended in version 1.03.

- The statistics provided when trying to predict the activity are the same in all the similarity calculations, irrespectively of the similarity index used. Thus, the statistics report only for the first index calculated, and for the next calculation the *Status* window is not updated. This is amended in version 1.03.

- The diagonal of the similarity matrix is zero for Euclidean, Cosine, Hodgkin-Richards, and Tanimoto indices. However, for correlation-like indices (i.e. Cosine, Hodgkin-Richards, and Tanimoto) it should be the unity (1 for complete similarity and 0 for complete dissimilarity). This is done for compatibility with the underlying data mining engine, which needs dissimilarity rather than similarity functions (i.e. 1-*Index* instead of *Index*). This is amended in version 1.03.

### *6.3. Version 1.03*

It mainly includes the following improvements:

- The *Open* menu offers two different possibilities: *Open > Predefined sets* for benchmark datasets already stored in the database of the program and a more general *Open* option for arbitrary datasets selected by the user.

- When loading a test set, the program tries to recognise automatically the headings corresponding to *Identifiers*, *Descriptors*, and *Endpoint*.

- The editing functionalities of the application to select groups have been significantly improved. It is possible to select the colour of the different groups for the training and for the test set. The options *Create groups for all available parameter values* and *Update values* when creating a new group have been added.

- The program deals in a more intelligent way with missing values than the previous version, assigning the value of zero to missing descriptor values.

- Yet another improvement when running the BfR rules is the provision of a single screen with all the missing values for each compound.

- Some functionalities of the *Similarity* menu have been restricted according to the nature of the endpoint. If the endpoint has a numerical continuous value, only the option *Calculate similarity and predict activity* is available. Similarly, if the endpoint has discrete (classification) values, only the option *Calculate similarity and classify into groups* is available.

- Some slight editing modifications have been added into the *Chart* area (more details are given in the *Graphical inferface* subsection).

- The appearance of the *Similarity histogram* area has been modified. Every time that a similarity calculation is run, the histogram is updated. In contrast to the previous version where the histogram displays all the calculations performed in the session by adding different similarity tabs corresponding to different similarity measures, in version 1.03, only the last similarity measure computed is visualised in the histogram.

- The histogram for similarity calculations when doing predictions is built upon distance values. When a classification is performed, the histogram displays the probability of chemicals to belong to each class, instead of the value of distance.

- The *Similarity matrix* area has been modified. It is updated every time that a similarity calculation is performed, and it includes the possibility to select the most similar compounds to a given chemical, by a user-defined threshold. This option is very useful to perform read across on a case by case basis. It is possible to export the similarity indices of the superior triangle of the pairwise similarity matrix as a CSV file, and also to save a PNG image file of it.

- The statistics are provided in the submenu *Similarity* > *Explore similarities*. Although this functionality is very useful, its location is not very intuitive.

## 7. Comments and suggestions on different issues

In the following section, a series of observations and suggestions for further improvement of Toxmatch (marked with ➢) are given in alphabetical order:

### *7.1. Classification by using sets of rules*

Toxmatch is able to handle datasets where one or more descriptor values are missing. However, when predefined classification systems are applied for specific activities (i.e. the application of the BfR rules), the program prompts the end-user for every single missing descriptor value. In the case of dealing with a large number of compounds, this is not very efficient. Indeed, it is not possible to cancel the window asking for missing descriptor values until the values for all the compounds have been examined.

➢ It would be desiderable that the program ignored the missing values and included a warning at the end of the calculation.

➢ The program does not provide any converter between units for a given descriptor, and what is more important, is not able to recognise the comma as a decimal separator. It would be useful to amend these problems or, at least, to specify this limitations in the documentation.

### *7.2. Graphical interface*

The main Toxmatch application window comprises training and test set data areas, and a visualisation area which is very useful. The implemented facility to investigate the coverage of the chemical space is also very useful.

Toxmatch software allows saving various parts of the output. The chart plots and the similarity histograms can be exported as PNG image files. Furthermore, it is possible to perform a zoom, and change range options.

The graphical interface allows changing the display of the title of the chart and plot properties, i.e. label, format, ticks, and range of x and y axes. The same holds for the similarity histogram.

Suggested improvements/modifications:

➢ When classification into groups is performed, the obtained graphic *Type_of_index.Group* versus *Type_of_index* is a horizontal line. This is because *Type_of_index.Group* is normally not a numeric value, but a label, and the chart can not display it and instead shows only zeros on that axis. Another kind plot should be designed to display classification results.

- It would be useful to include an option that allows manipulating the chart plots to a specific customized look (i.e. modifying the size of the symbols in the plots, setting different symbols for different groups, and changing the scale of the chart).

- In graphical representations, the identifier for each chemical (point) is temporally visible by pointing the point with the mouse and it can also be visualised by choosing a user-defined label. However, it is suggested to include an option that allows exporting subsets of chemicals, with the corresponding labels.

- The similarity histogram is cleared and updated automatically. However, it would be desirable to have the possibility to choose which/how many similarity measures are visualised in the histogram for comparison purposes. Secondly, it seems that the *Range* editing function for the axes is not working properly in all the cases. Third, the histogram for the training set and the histogram for the test set have been unified in a single histogram. This option should be not prescribed, but selected by the end-user.

### *7.3. Help file and capacity building*

Version 1.03 does not contain a proper help file. However, an installation and a use manual are provided. The user manual appropriately describes the various aspects of the software capabilities and guides to the operational use of the software. However, a comprehensive theoretical description on the similarity measures implemented, and some simple numerical examples including the interpretation of the results are still missing in the documentation.

Suggested improvements/modifications:

- Toxmatch is a powerful application, but is not intuitive in terms of the theory behind the calculation of similarity indices, and in terms of use. To exploit fully its functionalities training and more extensive supporting documentation are required.

- To include a more extensive and detailed help file/user manual with a theoretical scientific section that provides a simple explanation and the formulas used to calculate the similarity indices. The use of these indices possibly should be illustrated by selected numerical examples. Reference(s) to appropriate literature should be given thorough the help file.

### *7.4. Import/Export*

Data can be imported from and exported to a number of different formats: *.sdf, *.csv (comma delimited), *.smi (SMILES), *.txt (tab delimited), *.mol (MDL/MOL), *.ichi (ICHI), *.inchi (InCHi), *.cml (chemical markup language), *.hin, *.pdb, and *.xls (Microsoft Office Excel).

For the input, there are a number of fields which should be complimented (i.e. *Identifier*, *Descriptors*, and *Result*).

Suggested improvements/modifications:

➢ It would be advisable to have the possibility to export statistic results.

### *7.5. Installation*

The first beta version of Toxmatch 1.00 was received by the ECB by 1st December 2006, as a link to the appropriate site where from the software could be downloaded. Subsequently, beta versions 1.01, 1.02, and 1.03 have been installed. Eventually the lat version Toxmatch version 1.03 was received on 23rd March. In all cases the software was installed without problems.

### *7.6. Large datasets*

Toxmatch can import without problems large datasets; thus, datasets as large as 1341x16 has been tested and Toxmatch works without problems. However, it should be noted that the calculation of some similarity indices becomes cumbersome with large datasets.

Toxmatch software performs most similarity calculations in a reasonable time.

➢ The calculations *to predict skin irritation* on the 1341x16 dataset were performed within approx. 12 seconds (Euclidean distance, Cosine, Hodgkin Richards, Tanimoto), 14 seconds (fingerprints) and 50 seconds (atom environments), respectively, on an ordinary desk computer (Intel ® Pentium ® 4CPU 3.00 GHz 0.99 GB of RAM).

Suggested improvements/modifications:

➢ Maximum common substructure and atom environment calculations for classification are extremely slow, and almost impossible to calculate in reasonable time for a large number of structures.

### *7.7. Similarity indices and measures*

- Several descriptor-based similarity methods and classification techniques are included in Toxmatch. However, the definition of the similarity indices implemented is not available in the user manual. Similarly, simple examples demonstrating the applicability of similarity indices have not been included.
- When using non predefined sets, elements of the similarity matrix (i.e. Euclidean distance index) are above the unity some are. The values should be scaled to 1.

### *7.8. Stability*

- Toxmatch provides a stable platform for work and even if sometimes it might take a long time to complete a calculation, it is rarely found to crash. Only in few cases for cumbersome calculations the application must be stopped and reopened.

### *7.9. User configuration file*

- The user configuration option is an XML field which is not very user-friendly. It would be advisable to have the possibility to configure the various options by means of a window-based interface, exportable to the configuration file.

### *7.10. User-friendliness*

- When getting acquainted with the Toxmatch software, which is not straightforward, the software appears as quite user-friendly and easy to work with. However, more information is required for a comprehensive user manual, and also capacity building activities are strongly encouraged.

## 8. Overall performance of Toxmatch / Conclusions

The evaluation version of Toxmatch proved to be a valuable, powerful and robust software application for assessing the similarity of chemicals. Overall Toxmatch performs rather well. It operates in a user-friendly way and the results are often reached rapidly even in the case of larger data sets. However, Toxmatch is not very intuitive in terms of the theory behind the calculation of similarity indices, and is quite complex in terms of use.

Some of the comments and instabilities observed in the first versions have been already addressed; other comments may be more appropriate to be dealt with in future developments. A series of minor points for further improvement, applicability and user-friendliness has been already outlined.

## 9. 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}.

http://europa.eu.int/comm/enterprise/chemicals/chempol/whitepaper/reach.htm

[2] Pedersen, F. de Bruijn, J. Munn, S. and 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. Ispra, Italy: European Commission, Joint Research Centre.

[3] Van der Jagt, K., Munn, S., Tørsløv, J. and de Brujin, 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.

[4] Report of the SPORT (Strategic Partnership on REACH Testing) pilot project: "The SPORT Report: Making REACH work in practice" http://www.sport-project.info

[5] Section 3.2 of the OECD Manual for Investigation of HPV Chemicals.

http://www.oecd.org/document/7/0,2340,en_2649_34379_1947463_1_1_1_1,00.html

[6] US Environmental Protection Agency. Site on chemical categories.

http://www.epa.gov/oppt/newchems/chemcat.htm

[7] Gallegos, A., Patlewicz, G., and Worth, A. (2005) A Similarity Based Approach for Chemical Category Classification. JRC Report EUR 21867 EN, 42pp. Ispra, Italy: European Commission, Joint Research Centre.

[8] Guidance On The Grouping Of Chemicals (Including By Read-Across And Chemical Categories). RIP 3.3-2 Task 3 report, draft 2 February 2007.

[9] European Chemicals Bureau. Site on tools: http://ecb.jrc.it/qsar/qsar-tools/

[10] Witten, I.H., and Frank, E. (2005) Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco. http://www.cs.waikato.ac.nz/ml/weka/

[11] Robert, D., and Carbó-Dorca R. (1998) A Formal Comparison between Molecular Quantum Similarity Measures and Indices. *Journal of Chemical Information and Computer Sciences.* **38**, 620-623.

[12] Gallegos Saliner, A. (2006) Mini-review on chemical similarity and prediction of toxicity. *Current Computer-Aided Drug Design* **2(2)**, 105-122.

[13] Xing, L., and Glen, R.C. (2002) Novel Methods for the Prediction of logP, pKa, and logD. *Journal of Chemical Information and Computer Sciences* **42**, 796-805.

[14] Bender, A., Mussa, H.Y., Glen, R.C., and Reiling, S. (2004) Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection, and a Naïve Bayesian Classifier. *Journal of Chemical Information and Computer Sciences* **44(1)**, 170-178.

[15] Jaworska, J., and Nikolova-Jeliazkova, N. (2007) How can structural similarity analysis help in category formation. *SAR and QSAR Environmental Research* **18**, 195-207.

[16] Todeschini, R., and Consonni, V. (2000) Handbook of Molecular Descriptors. In the *Series of Methods and Principles in Medicinal Chemistry* (Eds. Mannhold, R., Kubinyi, H., and Timmerman, H.), Volume 11, WILEY – VCH.

[17] Verhaar H.J.M., van Leeuwen C.J., and Hermens J.L.M. (1992) Classifying environmental pollutants. 1. Structure-activity relationships for prediction of aquatic toxicity. *Chemosphere* **25**, 471-491.

[18] Gerner I., Schlegel K., Walker J.D., and Hulzebos, E. (2004) Use of physicochemical property limits to develop rules for identifying chemical substances with no skin irritation or corrosion potential. *QSAR and Combinatorial Science* **23**, 726-733.

[19] Walker, J.D., Gerner, I., Hulzebos, E., and Schlegel, K. (2005) The Skin Irritation Corrosion Rules Estimation Tool (SICRET). *QSAR and Combinatorial Science* **24**, 378-384.

[20] Jeliazkova, N. (2006) Review. Chemical Similarity In Toxicology. Service Contract: CCR.IHCP.C431607.X0 / 22.12.2005.

[21] Jeliazkova N, Gallegos Saliner A., Patlewicz G., and Jaworska J. (2007) Toxmatch - a tool to assess chemical similarity. Oral communication presented at *The Society of Environmental Toxicology and Chemistry (SETAC) Europe 17th Annual Meeting*, held in Porto (Portugal), on 20 – 24 May 2007.

[22] Gallegos Saliner A., Jeliazkova N, Patlewicz G, and Worth A.P. (2007) The use of a fragment-based approach to predict skin irritation using readily accessible software tools.

Poster presented at *The Society of Environmental Toxicology and Chemistry (SETAC) Europe 17th Annual Meeting*, held in Porto (Portugal), on 20 – 24 May 2007.

[23] Gallegos Saliner A., Jeliazkova N, Patlewicz G, and Worth AP. (2007) Application of Chemical Similarity Analysis in the Evaluation of a Rulebase for Predicting Skin Irritation. Poster presented at the *Occupational and Environmental Exposures of Skin to Chemicals* conference, held in Golden, Colorado (USA), on 17 – 20 June 2007.

[24] Dearden, J.C. (2003) Quantitative structure–property relationships for prediction of boiling point, vapor pressure, and melting point. *Environmental Toxicology and Chemistry* **22(8)**, 1696–1709.

[25] Ideaconsult, Ltd. (2007) Toxtree version 1.20, downloadable from http://ecb.jrc.it/qsar/qsar-tools/

European Commission

EUR 22854 EN – DG Joint Research Centre, Institute for Health and Consumer Protection

Title: Development and Beta Testing of the Toxmatch Similarity Tool

Authors: Ana Gallegos Saliner and Andrew P. Worth

Luxembourg: Office for Official Publications of the European Communities

2007 – 25 pp. – 21 x 29.7 cm

EUR - Scientific and Technical Research series; ISSN 1018-5593

## Abstract

Toxmatch is a software tool that provides a means of grouping chemical substances according to various measures of chemical similarity. It is designed to support the formation of chemical categories and the application of read-across within the hazard and risk assessment of chemicals. Such a tool will be useful for scientific researchers, for end-users in industry, for regulatory authorities, and for the EU Chemicals Agency.

Toxmatch was developed by Ideaconsult Ltd (Sofia, BG) under the terms of a JRC-ECB contract. It is a flexible user-friendly, computer-based open source application which is accessible via internet. It encodes and applies a range of different structural and descriptor based chemical similarity indices. The novelty of this software lies in its ability to calculate similarity measures that are tailored for specific activities/toxicities. Thus, relevant chemical representations can be selected for a given activity and the chemicals of interest can hence be classified into toxicity classes.

The present document summarises the beta testing of Toxmatch, reporting general comments and suggestions for further improvement.

EUROPEAN COMMISSION
DIRECTORATE-GENERAL
Joint Research Centre

The mission of the JRC is to provide customer-driven scientific and technical support for the conception, development, implementation and monitoring of EU policies. As a service of the European Commission, the JRC functions as a reference centre of science and technology for the Union. Close to the policy-making process, it serves the common interest of the Member States, while being independent of special interests, whether private or national.

Publications Office
*Publications.eu.int*