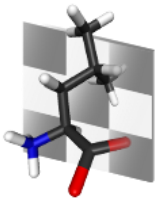
Workshop Handout: Visualize and Explore Chemical Feature Space of a Dataset

Martin Gütlein (Uni-Freiburg, mguetlein@fdm.uni-freiburg.de)
Andreas Karwath (Uni-Mainz, karwath@uni-mainz.de)

OpenTox EURO2013
30. Sept. - 2. Oct. 2013

This workshop gives an introduction to

CheS-Mapper - Chemical Space Mapping and Visualization in 3D



Abstract

Visualization is essential when analyzing chemical datasets in order to understand the relationship between the structure of chemical compounds, their physico-chemical properties, and biological or toxic effects.
In this workshop, you will learn how to use CheS-Mapper [1], a 3D molecular viewer. The tool arranges compounds in 3D space such that compounds with similar feature values will be close to each other. A small test dataset will be utilized to introduce the basic workflow of CheS-Mapper, and to apply clustering, feature computation and highlighting techniques. Subsequently, two real world QSAR datasets will be explored with CheS-Mapper.
We would like to encourage you to bring your own datasets and visualize them with CheS-Mapper. Alternatively, you could directly visualize datasets from OpenTox compliant dataset web services.

Prerequisites

- Java (Version 6 or 7) Runtime environment (http:// java.com, http://openjdk.java.net)
- Recommended: a mouse (alternatively, use your touchpad)
- Optional: Open Babel (http://openbabel.org) for advanced 3D builder and feature computation
- Optional: R (http://r-project.org.de) for advanced cluster and embedding algorithms

Introduction

The presenters will give a brief introduction to outline the workflow and functionalities of CheS-Mapper.

First steps with CheS-Mapper

Start the CheS-Mapper program

Visit http://ches-mapper.org and click on 'Run CheS-Mapper' (requires Java Webstart). Alternatively, go to the 'Download' section, save the file 'ches-mapper-complete.jar' (e.g. to your desktop) and double click on the downloaded file.

Have a first look at a small test dataset

Use the **Wizard** to load the dataset and compute basic PC-descriptors:

- 'Load dataset': Insert http://opentox.informatik.uni-freiburg.de/ches-mapper/data/demo.smi and press 'Load dataset'
- Skip 'Create 3D structures'
- 'Extract features': select either 'CDK descriptors' → 'constitutional' or 'OpenBabel descriptors' and click on 'Add feature'.
- 'Cluster dataset': select 'No'
- Skip further wizard steps and click 'Start mapping'

Investigate the dataset with the **3D-Viewer**:

- *Rotate*: hold down the left mouse button and move the mouse.
- *Zoom*: use the mouse wheel to zoom in and out (if you have no mouse wheel, hold down shift and the left mouse button and move the mouse up/down).
- *Select a compound*: click on the compound, click somewhere next to the compound to zoom out again.
- *Increase compound sizes*: use the slide bar on the bottom left.
- *Select features*: use the dropdown menu on the bottom left and select MW (molecular weight, on the bottom half of the list). Note how the compound list on the top left changes.
- Enable '*Highlight features with spheres*' (in the 'Highlighting' Menu) to show atom coloring and feature values at the same time.

Enable 3D Building and Clustering

Restart the **Wizard** with right-click → 'new dataset / mapping':

- 'Create 3D structures': select OpenBabel builder (if available), else select CDK builder (the latter will produce warnings, as it fails on some compounds).
- 'Cluster dataset': select 'Yes' and click 'Start mapping'

Examine the clusters with the **3D Viewer**:

- *Zoom into a cluster*: by clicking on it. Zoom out by clicking on 'All clusters' on the top left.
- *To directly focus compounds within a cluster* hold down the shift key while moving the mouse over the compound.
- Please take note that the compounds are no longer flat.
- *What drives this cluster?* Click on a cluster and look at the feature list (right side of the screen): it shows the median feature values (±standard deviation). It is sorted by specificity, i.e. the features at the top of the list are the import ones for this cluster. Click on the top feature and take a look at the graph (bottom right).

Use structural features instead of PC-descriptors

Remove all features in **wizard** step 'Extract features' (click on one selected feature, press 'Ctrl+a', click on 'Remove feature'). Instead add feature 'Structural Fragments' → 'Smarts list: MACCS'. Click on 'Load feature values'. Click on settings for fragments and reduce min-frequency to 1. 'Press start mapping' and explore with the **3D-Viewer** how the different features selection changes embedding and clustering results. Select structural fragment features to highlight SMARTS matches in each compound.

Cluster and 3D-align ligands of a real world dataset

The dataset contains about 450 structurally similar inhibitors (ligands) of the enzyme COX-2 [2]. **Wizard** settings:

- 'Load dataset': http://opentox.informatik.uni-freiburg.de/ches-mapper/data/cox2.sdf
- 'Create 3D structures': Disable 3D builder (3D structures are already available in the SDF)
- 'Extract features': use again structural features (e.g. MACCS). Do NOT select the feature IC50_uM that is included into the dataset!
- 'Cluster dataset': Force the clustering algorithm to divide the data into at least 5 clusters (e.g. 5-10), to cluster the dataset into small groups that share common structures.
- Enable 'Align compounds' → 'MCS aligner' to 3D align compounds.
- Click 'Start mapping', if you are using CDK (instead of OpenBabel), the embedding will take a couple of minutes.

Using the **3D-Viewer:**

- Decrease the compound size (bottom left slider).
- You might want to disable 'Highlight features with spheres' to improve the rendering performance.
- Select the feature MCS (Maximum common subgraph). Investigate the MCS of the clusters. Enable 'Superimpose' (on the left) to align the cluster compounds according to the MCS.

- Select the IC50_uM endpoint feature. In the 'Highlighting' menu, click the item 'Highlighting colors for numerical endpoints' to 'Enable log highlighting' and to reverse the color gradient with '↔'. Now, active compounds (with low feature values) will be highlighted in red.
- **Challenge:** Try to find substructures that match mostly active compounds.

Inspect the correlation of pc-descriptors and the endpoint fish toxicity

The Fathead Minnow Acute Toxicity Database File contains about 600 industrial organic chemicals [3]. A slightly modified[i] version of the dataset is available at the AMBIT dataset service from *Ideaconsult Ltd.*: http://apps.ideaconsult.net:8080/ambit2/dataset/1873416. CheS-Mapper can directly load datasets from OpenTox compatible dataset services, just fill in the URI and press 'Load dataset'.

Challenge: Visually verify that regression models using pc-descriptors work quite well on this dataset (i.e. that pc-descriptor feature values correlate to the endpoint LC50).

More

- Apply CheS-Mapper to a dataset of your choice.
- Try the "Search ChEMBL" functionality (first Wizard step).
- If you have KNIME installed on your system, try the CheS-Mapper extension (http://tech.knime.org/book/ches-mapper).

References

[1] CheS-Mapper - Chemical Space Mapping and Visualization in 3D, M. Gütlein, A. Karwath and S. Kramer, Journal of Cheminformatics 2012, 4:7

[2] Spline-Fitting with a Genetic Algorithm: A Method for Developing Classification Structure−Activity Relationships, Jeffrey J. Sutherland, Lee A. O'Brien, and, and Donald F. Weaver, Journal of Chemical Information and Computer Sciences 2003 43 (6), 1906-1915

[3] Predicting modes of action from chemical structure: Acute toxicity in the fathed minnow (Pimephales promelas). Russom, C.L., S.P. Bradbury, S.J. Broderius, D.E. Hammermeister, and R.A. Drummond (1997) Environmental Toxicology and Chemistry, 16(5): 948-967.

[i] 37 Compounds without endpoint value and one large outlier compound have been removed.