CheS-Mapper: New Developments

Martin Gütlein
mguetlein@fdm.uni-freiburg.de

Andreas Karwath & Stefan Kramer

OpenTox Euro 2013, October 2
CheS-Mapper
- Chemical Space Mapping and Visualization in 3D

- 3D viewer for small molecule datasets
- Published in Journal of Cheminformatics, March 2012, >6000 accesses in 18 months
- Project homepage: http://ches-mapper.org
- Open-source java software
- Uses: Jmol, CDK, WEKA, OpenBabel, R
- Compatible to OpenTox dataset services
Artificial test dataset with 10 compounds
CheS-Mapper Workflow

Dataset

Chemical Space Mapping

3D Structures → Extract Features → Clustering → 3D Embedding → 3D Alignment

3D-Visualization

OpenTox Euro 2013 - Martin Gütlein - CheS-Mapper
Chemical Space Mapping

3D Structures → Extract Features → Clustering → 3D Embedding → 3D Alignment → 3D-Visualization

Caffeine
THC
Lidocaine
Nicotine
Diazepam
Sucrose
Glucose
Citric acid
Sulfuric acid
Phosphoric acid
Chemical space mapping can be configured with wizard.
CheS-Mapper Workflow

Dataset

Chemical Space Mapping

3D builders:
- OpenBabel
- CDK

3D Structures → Extract Features → Clustering → 3D Embedding → 3D Alignment

3D-Visualization

OpenTox Euro 2013 - Martin Gütlein - CheS-Mapper
CheS-Mapper Workflow

Feature Extraction:
- Included in the dataset
- PC-Descriptors (computed with CDK or OpenBabel)
- Structural Fragments
  - Subgraph mining
  - Smarts matching

3D Structures → Extract Features → Clustering → 3D Embedding → 3D Alignment

3D-Visualization

OpenTox Euro 2013 - Martin Gütlein - CheS-Mapper
Physico-chemical descriptors:
- MW (Molecular weight)
- logP (Octanol water coefficient)
- abonds (Number of aromatic bonds)
- HBD (Number of hydron bond donors)
- ...
Show that compounds have openbabel features

Highlight features

Show that similar compounds are close to each other

Show that glucose is closer to acids than to sucrose

Embedding is based on PC descriptors, molecular weight is selected

Dataset:
- Num compounds: 10
- Cluster algorithm: No Dataset Clustering
- 3D Embedding: Sammon 3D Embedder (R)
- 3D Embedding Quality: excellent (CCC: 1, r^2: 0.99)

**Compound Glucose**
- Smiles: OCC1OC(O)...
- dbonds: 0
- HBD: 5
- logP: -3.22
- MR: 35.74
- HBA1: 6
- HBA2: 6
- TPSA: 110.38
- sbonds: 12
-_atoms: 12
- bonds: 12
- MW: 180.16
- abonds: 0

**Feature**: MW
- Values: 193.16 ±83.94
- Description: Molecular Weight filter (OB Descriptor)
- Usage: Used for clustering and/or embedding.
- Missing values: 0

**Graph**
- X-axis: #compounds
- Y-axis: MW
- Bars: Glucose
- Dataset: 100, 150, 200, 250, 300, 350

Embedding is based on PC descriptors, molecular weight is selected.
CheS-Mapper Workflow

MACCS list (166 SMARTS fragments):
- P
- [#8]~[#6]~[#8]
- *~[CH2]~[#8]
- ...

OpenTox Euro 2013 - Martin Gütlein - CheS-Mapper
Show embedding with structural features

Show that glucose is now close to sucrose

Show feature that only these glucose and sucrose match

Show one more easy feature OCO

Show that match of SMARTS is highlighted

Same dataset, but embedding is based on structural fragments (MACCS list)
<table>
<thead>
<tr>
<th>MW</th>
<th>LogP</th>
<th>abonds</th>
<th>...</th>
<th>HBD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caffeine</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>THC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glucose</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sulfuric acid</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **3D-Visualization**
- **3D Structures**
- **Extract Features**
- **Clustering**
- **3D Embedding**
- **3D Alignment**
CheS-Mapper Workflow

3D embedding algorithms:
- PCA (fast)
- Sammon embedding (non-linear, configurable distance)
- SMACOF
- t-SNE

3D-Visualization

OpenTox Euro 2013 - Martin Gütlein - CheS-Mapper
Quality of the 3D Embedding

<table>
<thead>
<tr>
<th>Compounds</th>
<th>MW</th>
<th>LogP</th>
<th>abonds</th>
<th>...</th>
<th>HBD</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caffeine</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>THC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glucose</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sulfuric acid</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Global embedding quality: Correlation between distance matrixes (CCC, R²)

Embedding stress for each compound:
1 – correlation of the corresponding matrix rows

How well do the 3D positions reflect the feature values?

Martin Gütlein - CheS-Mapper
THC is the compound with the highest embedding stress.
CheS-Mapper Workflow

Cluster Algorithms:
- k-Means (num clusters configurable)
- Hierarchical Clustering (dynamic num-clusters detection)
- Expectation maximization
- ...

OpenTox Euro 2013 - Martin Gütlein - CheS-Mapper
COX-2 dataset

- 467 COX-2 Inhibitors
- Inhibition of the enzyme Cyclooxygenase-2 (COX-2) is investigated in cancer studies
- Compounds are structurally very similar (docking dataset)
- Endpoint: IC$_{50}$ µMol (half maximal inhibitory concentration)

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

OpenTox Euro 2013 - Martin Gütlein - CheS-Mapper
Compounds are colored according to cluster assignment.
Sorting of features

- Feature values are shown for the selected compound or cluster
- Features are sorted according to the p-value of
  - \(X^2\) test for nominal features
  - ANOVA test for numerical features

- The most 'important' features are listed first: the feature values of this compound/cluster differ the most from the complete dataset

<table>
<thead>
<tr>
<th>Feature</th>
<th>Values</th>
<th>Description</th>
<th>Usage</th>
<th>Missing values</th>
</tr>
</thead>
<tbody>
<tr>
<td>logP</td>
<td>4.35 ± 1.05</td>
<td>octanol/water partition coefficient (OB Descriptor)</td>
<td>Used for clustering and/or embedding.</td>
<td>0</td>
</tr>
<tr>
<td>MR</td>
<td>67.71 ± 15.04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MW</td>
<td>246.39 ± 76.69</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TPSA</td>
<td>20.31 ± 16.52</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HBA1</td>
<td>1 ± 0.72</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HBA2</td>
<td>1 ± 0.79</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DSSTox_CID</td>
<td>2080 ± 447.85</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HBD</td>
<td>0 ± 0.57</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LC50_mmol</td>
<td>0 ± 0.09</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nF</td>
<td>0 ± 0.23</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Easy to find important features for clusters
CheS-Mapper Workflow

3D Alignment:
- Detect common subgraph of each cluster:
  - Compute MCS (maximum common subgraph)
  - Use largest fragment
  - Manually specify subgraph
- Align compounds according to common subgraph:
  - Obfit (OpenBabel)
  - Kabsch Alignment (CDK)
Cluster compounds aligned according to MCS
Clustering and embedding separates active and inactive compounds.
The CheS-Mapper extension for KNIME

- KNIME:
  - graphical workbench for data access, investigation and predictive analysis
  - various extensions to process chemical data
  - CheS-Mapper integration as visualization node
EPAFHM dataset

- EPAFHM: US Environmental Protection Agency (EPA) Fathead Minnow Acute Toxicity Database File
- 617 industrial organic chemicals
- Endpoint: LC50 mMol ((lethal) concentration that kills 50%)

Predicting modes of action from chemical structure: Acute toxicity in the fathed minnow (Pimephales promelas).
Environmental Toxicology and Chemistry, 16(5): 948-967.
Fish-Toxicity is well correlated to PC-descriptors (used for embedding)
Highlight two features at once to detect high prediction errors.
More features

- Export clusters/compounds/features
- Export high-res images
- Access ChEMBL database
- Save and share embedding settings
- Data tables to browse through raw compound/feature/cluster data
- Command line interface
- Configurable highlight and view settings
  - Adjust highlight color gradient
  - Enable log highlighting
  - Switch between sphere and atom-color highlighting
  - ...

OpenTox Euro 2013 - Martin Gütlein - CheS-Mapper
Quotes about CheS-Mapper

... a very nice piece of software.
   *Basil Hartzoulakis, PhD, Xention Ltd, Cambridge*

You help to make me look like a genius to my bosses ... this is the only open source tool I know of that is usable by a regular bench chemist.

   *Kerry W. Fowler, Ph.D., Senior Scientist
    Kineta Inc, Seattle*

CheS-Mapper has come in handy.

   *Kaushik Hatti, Vittal Mallya Scientific Research Foundation, Banagalore*

I am very impressed by the ches-mapper software ... the tool is very effective to spot trends of a chemical group

   *Hiroshi Nara, Organic chemist, Japan*

thanks for the gorgeous ches-mapper.

   *Santi Villalba, University College Dublin*

Many compliments because we have found that it is highly useful and well working.

   *Prof. Paola Gramatica, University of Insubria*