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

http://opentox.informatik.uni-freiburg.de/ches-mapper

Martin Gütlein[1]*, Andreas Karwath[1], Stefan Kramer[2]

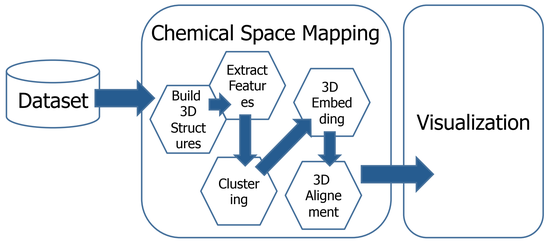
*guetlein@informatik.uni-freiburg.de

[1]Institute for Computer Science • Albert-Ludwigs-Universität Freiburg • Germany, [2] Institute for Computer Science I12 • Technische Universität München • Germany

Abstract

Scientific researchers in the field of chemoinformatics, are often overwhelmed by the size and the sheer complexity of chemical datasets. Therefore, the need for visualization tools, is one of the uttermost requests.
Our recently developed 3D molecular viewer CheS-Mapper (Chemical Space Mapper) includes many techniques, like state-of-the-art structural clustering, and multi-dimensional embedding techniques. Large datasets are divided into clusters of similar compounds and consequently arranged in 3D space, such that their spatial proximity reflects their chemical similarity. This intuitively provides essential information to the user, while making the dataset more easily accessible and allowing easy and understandable access to a large number of chemical structures within seconds.
The different clustering approaches employed in our tool utilize common substructures as well as quantitative chemical descriptors of the compounds. These features can be highlighted within CheS-Mapper, which aids the chemist to better understand the underlying scientific knowledge. As a final function, the tools can also be used to select and export specific part of a given dataset for further analysis.

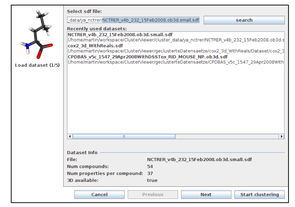
CheS-Mapper Workflow



- The workflow is divided into Mapping and Visualization
- Mapping:
  - Is a preprocessing step where the data is clustered and arranged in 3D space
  - Easy to use: the novice user can employ default settings
  - All steps can be configured manually
  - Developers can plug in own algorithms
- Visualization:
  - The dataset is presented in a 3D viewer
  - The clustering and embedding provides relational information about similarity and makes the data easily accessible

Wizard Dialog to Control Mapping

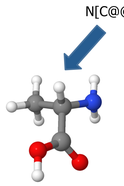
- A wizard dialog guides through the Mapping process
- Suitable for novice and expert users
- Single Steps:
  - Load dataset
  - Build 3D structure
  - Extract features
  - Clustering
  - 3D Embedding
  - 3D Alignment
- Automatic detection and plug in of new methods and algorithms



Chemical Space Mapping

Build 3D Structure and Extract Features

- Select input dataset
  - Various dataset formats are supported (sdf/mol/smiles/...)
  - Dataset can be directly loaded from the web
- 3D structure is built
  - 3D structure can be build with Chemical-Development-Kit (CDK) or OpenBabel
  - External libraries like Corina can be plugged in easily
- Extract features
  - Features are required for clustering and embedding
  - Automatic extraction of dozens of descriptors with CDK



Feature	Value
LogP	-2.824
Weight	89.048
Wiener path number	29.0
Wiener polarity number	4.0
...	...

3D Embedding (of Clusters & Compounds)

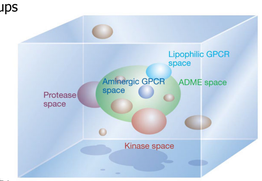
- Embedding algorithms assign 3D coordinates to each compound or cluster, according to the feature values of the compounds
- Different approaches are provided:
  - Principal Component Analysis (PCA)
  - Multidimensional Scaling Using Majorization (SMACOF)
  - T-distributed Stochastic Neighbor Embedding (tSNE)
- Developers can easily plug in their own/preferred 3D Embedding algorithm

Features	LogP	Weight	W. path number	W. polarity number	..
C1(C=CC=CC=1C(=C(Cl)Cl)C2=CC=C(C=C2)Cl)C1	4.27	315.938	581.0	27.0	
C(C(C1C=CC(=CC=1)O)C2C=CC(=CC=2)O)(Cl)(Cl)Cl	2.74	315.982	678.0	28.0	
...					

3D	X	Y	Z
C1(C=CC=CC=1C(=C(Cl)Cl)C2=CC=C(C=C2)Cl)C1	-45.16	45.81	-30.53
C(C(C1C=CC(=CC=1)O)C2C=CC(=CC=2)O)(Cl)(Cl)Cl	-2.68	-43.04	-91.27
...			

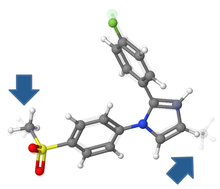
Cluster Compounds

- Compounds in the dataset are assigned to subgroups according to their similarity
- Supported cluster algorithms:
  - k-Means Clustering
    - Fixed number of k clusters
    - Random initialization, iterative update of clusters and cluster centroids
  - Hierarchical Clustering
    - Each compound is single cluster
    - Sequentially merge similar clusters
  - Structural Clustering
    - Finds groups that share structural similarity
    - Compounds are assigned to clusters when there exists a common subgraph of sufficient size
- Developers can plug in new cluster algorithms



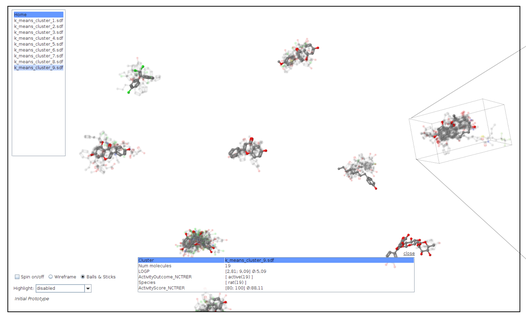
3D Alignment of Compounds

- Compounds in a cluster are likely to share common subgraphs:
  - This subgraph is already available if structural clustering is performed
  - Alternatively, the maximum common subgraph can be computed within each cluster
- The compounds within a cluster can be superimposed/aligned according to this subgraph:
  This shows differences between compounds



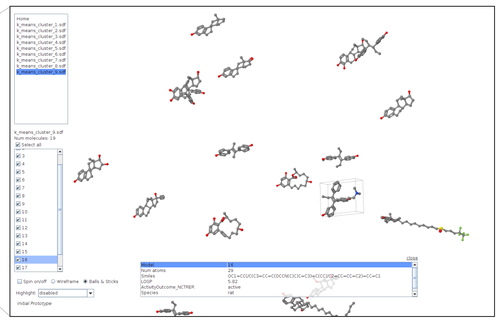
3D Visualization

Dataset Overview -- Clusters



- Datasets are separated into clusters, arranged in 3D space
- The intuitive interface of the 3D viewer allows to:
  - Zoom/rotate the clusters
  - Get valuable information on clusters via mouse over
  - Examine a cluster by clicking on it
- The embedding into 3D space (positions/distance between clusters) reflects the similarity between clusters
- Cluster can be removed from the dataset

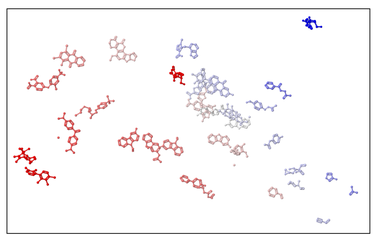
Inside Cluster View



- By selecting a cluster, the view zooms into the cluster and displays only the compounds included
- Details for each compound are available via mouse over
- Like the clusters, the compounds are embedded into the 3D space as well: the position/ distance between compounds within the cluster reflects the similarity between compounds
- Compounds can be removed from the dataset

Highlight Features and Endpoints

- All compound properties can be highlighted:
  The compounds are colored according to the numeric value, a high value is indicated by red, a low value is indicated by blue
- Also available for the cluster overview
- Gives an intuitive explanation towards the quality of the clustering approach:
  'Does the clustering algorithm separate active from inactive compounds?'



Open-Source Webstart Application

- Java program that comes in two variants:
  - Java Web Start application (can directly started from a web browser)
  - Local installation that makes use of non-java libraries
- ChesS-Mapper is available at http://opentox.informatik.uni-freiburg.de/ches-mapper

Powered by:

References

[1] Seeland, M, Girschick, T, Buchwald, F, Kramer, Online Structural Graph Clustering Using Frequent Subgraph Mining, 2010, Machine Learning and Knowledge Discovery in Databases, 213--228, Springer
[2] Jmol: an open-source Java viewer for chemical structures in 3D. http://www.jmol.org

Acknowledgements

This work has been supported by the EU FP7 project (HEALTH-F5-2008-200787) OpenTox (http://www.opentox.org).