Parallel Structural Graph Clustering

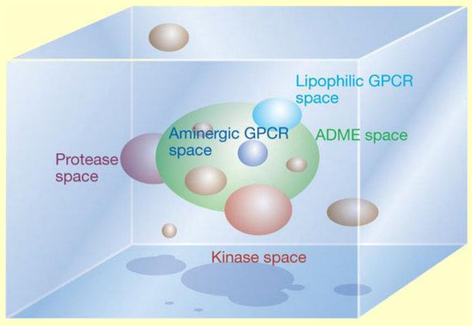
Madeleine Seeland[1], Simon A. Berger[2], Tobias Girschick[1], Alexandros Stamatakis[2] and Stefan Kramer[1]
[1] Institut für Informatik/I12, Technische Universität München
[2] Heidelberg Institute for Theoretical Studies

Technische Universität München

The goal of clustering a graph database of molecular structures is to identify groups of similar structures, such that intra-group similarity is high and inter-group similarity is low. This can serve to structure the chemical space and to improve the understanding of the data.
In previous work we proposed a graph-based clustering approach based on the frequent graph mining algorithm gSpan. In the proposed approach, clusters encompass all molecules that share a sufficiently large common substructure. The size of the common substructure of a compound in a cluster has to take at least a user-specified fraction of its overall size. The algorithm processes the instances in one defined order, one after the other, and produces overlapping (non-disjoint) and non-exhaustive clusters.
Several experiments were designed to evaluate the effectiveness and efficiency of the structural clustering algorithm on various real-world data sets of molecular graphs. We showed that the approach is able to rediscover known structure classes in the NCI standard anti-cancer agents. Moreover a baseline comparison with a PubChem Tanimoto fingerprint-based clustering was presented.
In recent work, we addressed the problem of clustering large graph databases of molecular structures. We parallelized the structural clustering algorithm to take advantage of high-performance parallel hardware and further improved the algorithm in three ways: a refined cluster membership test based on a set abstraction of graphs, sorting graphs according to size, to avoid cluster membership tests in the first place, and the definition of a cluster representative once the cluster scaffold is unique, to avoid cluster comparisons with all cluster members. In experiments on a large database of chemical structures, we showed that running times can be reduced by a large factor for one parameter setting used in previous work. For harder parameter settings, it was possible to obtain results within reasonable time for 300,000 structures, compared to 10,000 structures in previous work. This shows that structural, scaffold-based clustering of smaller libraries for virtual screening is already feasible.

Motivation

- Building structural categories is an important step in structuring the chemical space for (semi-) automatic methods for toxicity prediction
- PSCG (Parallel Structural Clustering of Graphs) [1,2] is able to discover groups of structurally similar/dissimilar graphs
- PSCG can be useful for:
  - Prestructuring the chemical space, e.g. for virtual screening
  - Descriptor calculation (e.g., for QSAR studies)
  - Computing local models for classification or regression
  - Calculation of the applicability domain of models



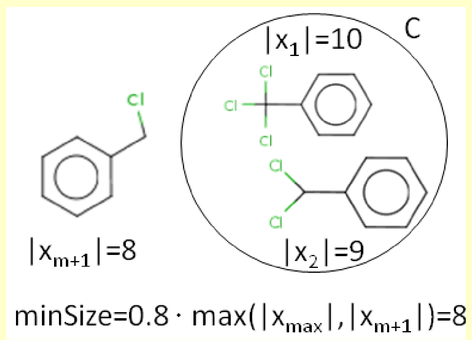
PSCG – Problem Formulation

- Structural clustering is the problem of finding groups of graphs according to scaffolds, i.e., large structural overlaps that are shared among all cluster members
- Input: Set of graph objects $X = \{x_1,..., x_n\}$ and user-defined parameters θ and *minGraphSize*
- Output: Set of clusters with maximum size, s.t. cluster members share at least one common subgraph that covers a specific fraction of the graphs in the cluster

$$\exists\, s \in cs(\{x_1,..., x_n\}) \forall\, x_i \in C : |s| \geq \theta\, |x_i|$$

- Minimum threshold for the size of the common subgraphs shared by the query graph x_{m+1} and the graphs in the cluster:

$$minSize = \theta \max(|x_{max}|, |x_{m+1}|)$$



- Overlapping and non-exhaustive clustering

PSCG – Technical Details

- Modified version of gSpan [3], gSpan'', to compute a sufficiently large common subgraph
  - Search for common subgraphs terminates as soon as a subgraph of size *minSize* is found
- Parallelization of the structural clustering algorithm based on master-worker paradigm
  - Clusters are distributed among a set of workers
  - Maintenance of cluster membership information for each graph → indication whether a new cluster needs to be created
- Refined cluster membership tests to reduce the number of expensive subgraph search computations:
  - Set abstraction of graphs which serves as an upper bound for the size of the MCS
  - Size-based clustering criterion which constrains the set of graphs being considered for clustering
  - Definition of a cluster representative for each cluster once a unique cluster scaffold is found

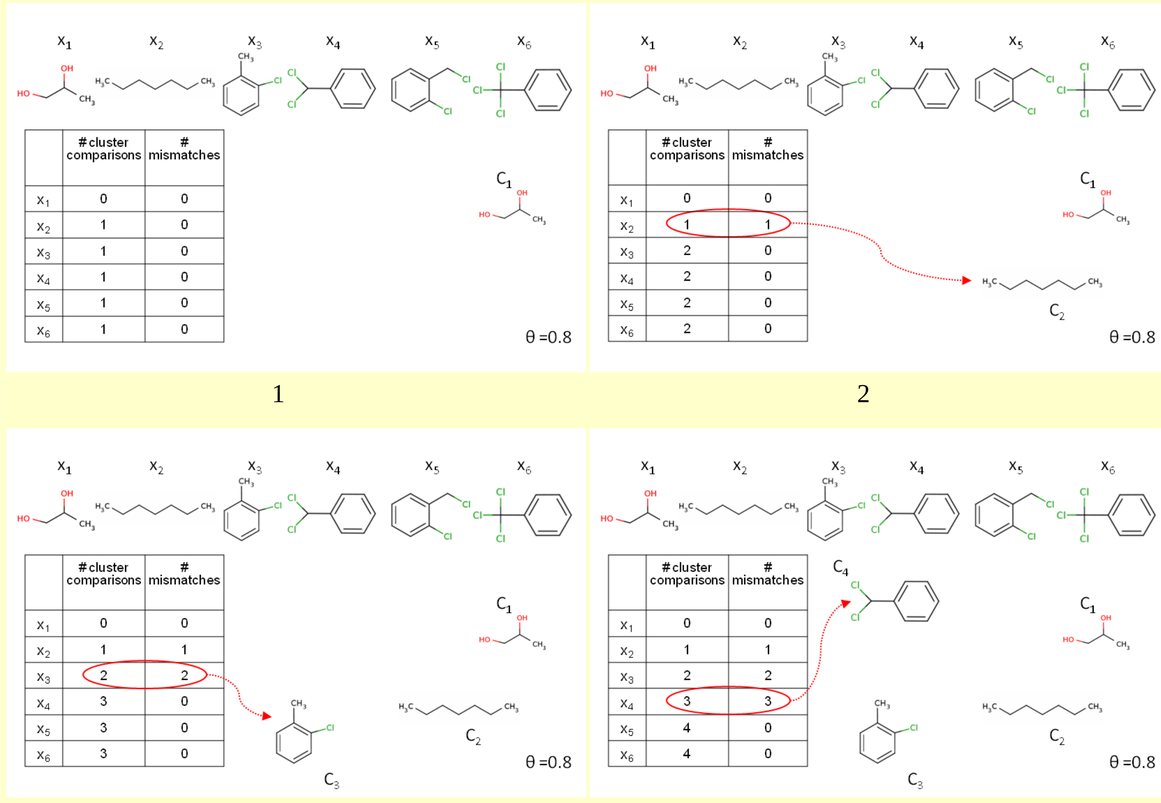


Fig. 1. Example sequence of steps of PSCG

PSCG – Example Output

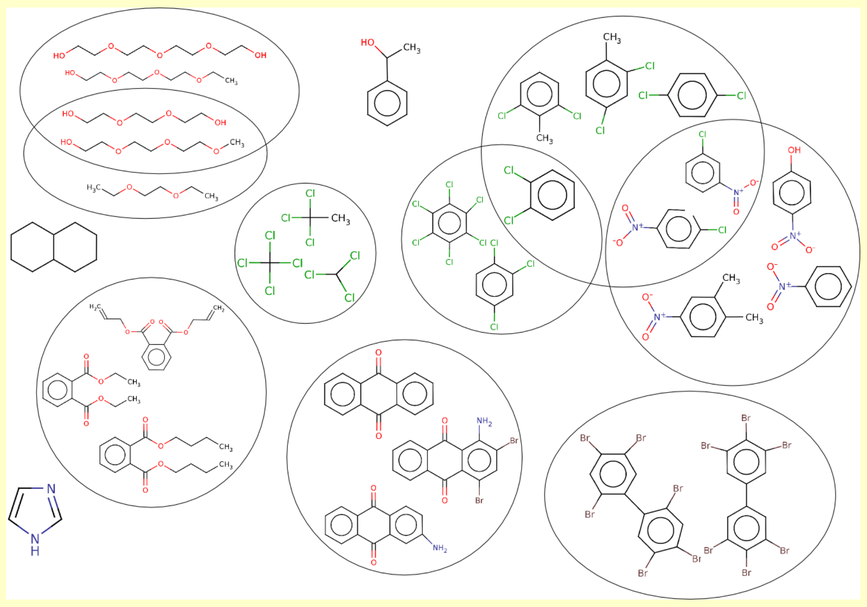


Fig. 2. Example output of PSCG on a subset of the RepDose data set for θ = 0.7

Results

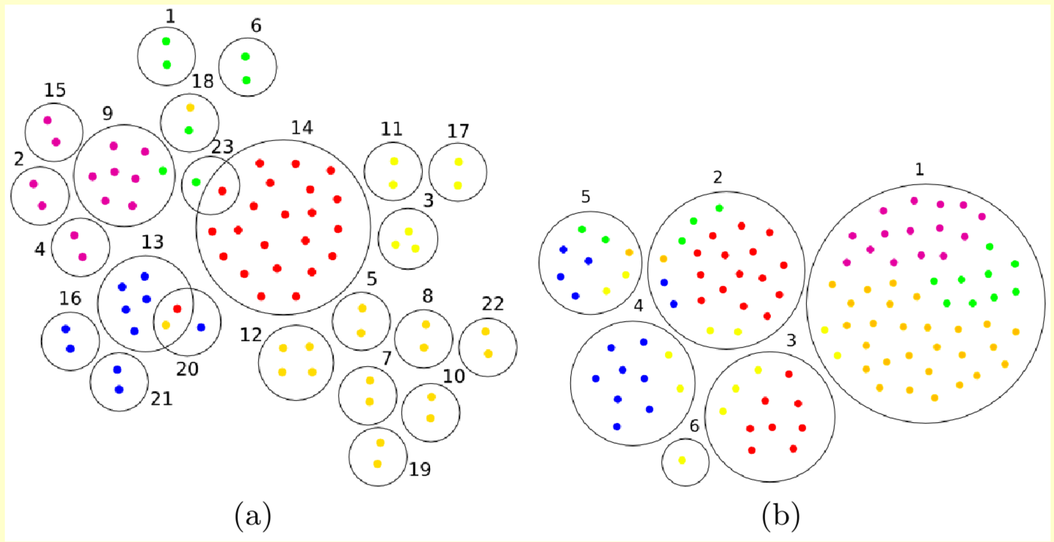


Fig. 3. Output of (a) PSCG for θ = 0.6 and (b) a graph-based clustering based on variational Dirichlet process (DP) mixture models for α = 0.1 and m = 1000 on the standard anti-cancer agents (SACA) data set. The cluster instances are colored according to the six SACA classes.

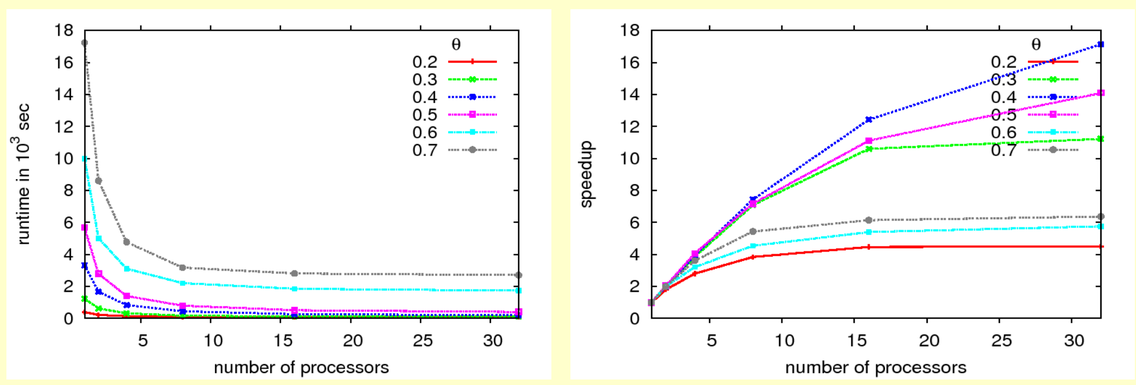


Fig. 4. Runtime performance and speedup of PSCG on the first 10,000 graphs of the NCI anti-HIV data set.

Table 1. Runtime (in sec) of the sequential clustering version vs. PSCG on the first 10,000 graphs of the NCI anti-HIV data set for different values of θ.

θ	0.2	0.3	0.4	0.5
t_{seq}	747,000	1,068,420	1,434,780	2,087,280
t_{par}	396	1,244	3,394	6,235

Table 2. Runtime (in sec) and number of cluster of PSCG on three data sets sampled from the ChemDB data set for θ=0.4 and θ=0.6.

$\|D\|$	$\theta = 0.4$	$\theta = 0.6$
100,000	31,103 ●	67,563 ●
200,000	122,204 ●	349,568 ●
300,000	610,577 ○	1,163,761 ⋆

$\|D\|$	$\theta = 0.4$	$\theta = 0.6$
100,000	4,112	16,295
200,000	6,096	25,685
300,000	9,811	38,775

●: 32 processors ○: 96 processors ⋆: first half: 96 processors, second half: 48 processors

References

[1] M. Seeland, T. Girschick, F. Buchwald, and S. Kramer. Online structural graph clustering using frequent subgraph mining. In: *Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases*, vol. 3, pages 213–228, 2010.
[2] M. Seeland, S. A. Berger, A. Stamatakis, and S. Kramer. Parallel Structural Graph Clustering In: *Proceedings of the 2011 European Conference of Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, vol. 3, pages 256-272, 2011.
[3] X. Yan, and J. Han. gSpan: Graph-based substructure pattern mining. In: *Proceedings of the 2002 IEEE International Conference on Data Mining*, pages 721–724, 2002.