Journal of Cheminformatics

POSTER PRESENTATION Open Access

Deterministic clustering of the available chemical space

Philipp Thiel[1,3*], Lisa Peltason[2], Christian Ottmann[1], Oliver Kohlbacher[3]

From 8th German Conference on Chemoinformatics: 26 CIC-Workshop
Goslar, Germany. 11-13 November 2012

Clustering of compound libraries using 2D binary fingerprints is a fundamental task in chemoinformatics and various methods have been described to solve it [1]. These methods can roughly be grouped into deterministic and non-deterministic approaches with two key-characteristics distinguishing them. First, the algorithmic complexity of deterministic approaches is more demanding whereas the non-deterministic methods often try to overcome this drawback by using heuristics to save time and memory. Second, deterministic clustering algorithms, especially agglomerative hierarchical techniques have been shown to yield good results and often perform better than non-deterministic approaches [2]. As a consequence, clustering of small to medium sized libraries with up to 1 million compounds is regularly performed using deterministic techniques whereas libraries comprising millions of compounds are mostly clustered using heuristics like k-means [3].

Here, we present a deterministic approach for clustering huge compound libraries based on all pairwise compound similarities. For this purpose, we use an extremely fast and flexible algorithm for similarity calculations, which we have developed to be purely CPU-based thus having no need for any specialized hardware. Using this similarity method, we implemented a workflow with the following steps. First, we create a set of unique input fingerprints by filtering duplicates that are then stored and finally remapped onto their representative clusters. Second, we calculate all pairwise similarities to construct a similarity network by applying a fixed Tanimoto threshold to select the edges to be inserted into the network. From this similarity network the connected subgraphs are extracted and forwarded to the last step. Finally, connected subgraphs exceeding a predefined size are hierarchically clustered.

As a result, we show that our algorithm for similarity calculation is competitive to recently published CPU-based methods and can perform up to 380 million Tanimoto calculations per second on a current desktop computer. This efficient method allows our workflow to process medium to large libraries on current desktop computers within minutes. To finally demonstrate the power of our clustering workflow, we processed the commercially available chemical space comprising about 17 million compounds [4]. The entire clustering workflow took 63 hours on a compute server using 64 cores and 100 GB main memory to complete.

Author details
[1]Chemical Genomics Centre of the Max Planck Society, Dortmund, 44227, Germany. [2]F. Hoffmann-La Roche AG, CH-4070 Basel, Switzerland. [3]Applied Bioinformatics, Center for Bioinformatics, Quantitative Biology Center and Dept. of Computer Science, University of Tübingen, Tübingen, 72076, Germany.

Published: 22 March 2013

References

1. Olah MM, Bologa CG, Oprea TI: **Strategies for compound selection.** *Curr Drug Discovery Technol* 2004, **3**:211-220.
2. Downs G, Willett P, Fisanick W: **Similarity Searching and Clustering of Chemical-Structure Databases Using Molecular Property Data.** *J Chem Inf Model* 1994, **34**:1094-1102.
3. Boecker A, Derksen S, Schmidt E, Teckentrup A, Schneider G: **A hierarchical clustering approach for large compound libraries.** *J Chem Inf Model* 2005, **45**:807-815.
4. Irwin JJ, Sterling T, Mysinger MM, Bolstad E, Coleman RG: **ZINC: A Free Tool to Discover Chemistry for Biology.** *J Chem Inf Model* 2012, **52**:1757-1768.

doi:10.1186/1758-2946-5-S1-P53
Cite this article as: Thiel *et al.*: Deterministic clustering of the available chemical space. *Journal of Cheminformatics* 2013 **5**(Suppl 1):P53.

* Correspondence: philipp.thiel@cgc.mpg.de
[1]Chemical Genomics Centre of the Max Planck Society, Dortmund, 44227, Germany
Full list of author information is available at the end of the article