Li, Weizhong
Journal of chemical information and modeling ISSN 1549-9596 2006, vol. 46, no5, pp. 1919-1923 [5 page(s) (article)]

Abstract

As a result of the recent developments of high-throughput screening in drug discovery, the number of available screening compounds has been growing rapidly. Chemical vendors provide millions of compounds; however, these compounds are highly redundant. Clustering analysis, a technique that groups similar compounds into families, can be used to analyze such redundancy. Many available clustering methods focus on accurate classification of compounds; they are slow and are not suitable for very large compound libraries. Here is described a fast clustering method based on an incremental clustering algorithm and the 2D fingerprints of compounds. This method can cluster a very large data set with millions of compounds in hours on a single computer. A program implemented with this method, called cd-hit-fp, is available from http://chemspace.org.