BIOINFORMATICS APPLICATIONS NOTE
Vol. 24 no. 15 2008, pages 1733–1734
doi:10.1093/bioinformatics/btn307

Structural bioinformatics

ChemmineR: a compound mining framework for R

Yiqun Cao[1], Anna Charisi[2], Li-Chang Cheng[1], Tao Jiang[1] and Thomas Girke[2,*]

[1]Department of Computer Science and Engineering and [2]Department of Botany and Plant Sciences, University of California, Riverside, California, USA

Received and revised on April 21, 2008; accepted on June 10, 2008

Advance Access publication July 2, 2008

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Software applications for structural similarity searching and clustering of small molecules play an important role in drug discovery and chemical genomics. Here, we present the first open-source compound mining framework for the popular statistical programming environment R. The integration with a powerful statistical environment maximizes the flexibility, expandability and programmability of the provided analysis functions.

Results: We discuss the algorithms and compound mining utilities provided by the R package *ChemmineR*. It contains functions for structural similarity searching, clustering of compound libraries with a wide spectrum of classification algorithms and various utilities for managing complex compound data. It also offers a wide range of visualization functions for compound clusters and chemical structures. The package is well integrated with the online ChemMine environment and allows bidirectional communications between the two services.

Availability: *ChemmineR* is freely available as an R package from the ChemMine project site: http://bioweb.ucr.edu/ChemMineV2/chemminer

Contact: thomas.girke@ucr.edu

1 INTRODUCTION

High-throughput screens (HTS) of small molecules for drug discovery and chemical genomics have evolved into routine technologies for analyzing protein functions and cellular networks on a systems biology level. At the same time, millions of drug-like molecule structures have become freely available to the public, mainly through online compound database projects, such as PubChem, ChemBank, Zinc, ChemMine, ChemDB and others (Chen *et al.*, 2005; Girke *et al.*, 2005; Irwin and Shoichet, 2005; Seiler *et al.*, 2008). To search and analyze the vast amounts of available compound and screening information, and to assemble diverse screening libraries, efficient compound analysis tools are a critical enabling resource. Unfortunately, most of the available software in this area is only commercially available, and open-source approaches are still the exception (e.g. OpenBabel, JOELib, Guha *et al.*, 2006). The long-term goal of the *ChemmineR* project is to narrow this resource gap by providing free access to a flexible and expandable open-source framework for the analysis of small molecule data from chemical genomics, agrochemical and drug discovery screens.

2 APPROACH

2.1 Overview

The development of compound analysis software for the statistical environment and programming language R has many obvious advantages (R Development Core Team, 2008). To name just a few: (1) R is one of the most widely used data mining environments used in bioinformatics. (2) The software and its associated packages are available for all common operating systems. (3) CPU and memory intensive calculations can be computed in high-performance languages, like C. (4) The data objects and base functions available in R are extremely efficient for typical compound and screening data mining routines. (5) Finally, an unmatched spectrum of data mining resources is available in R, such as extensive graphics utilities, powerful statistical functions, and a wide variety of algorithms for clustering and machine learning tasks.

The current release of the *ChemmineR* package contains functions for 2D structural similarity comparisons between compounds and similarity searching against compound databases. Both methods use the highly accurate atom pair approach for scoring structural similarities (Carhart *et al.*, 1985; Chen and Reynolds, 2002). In addition, the package provides various functions for clustering entire compound libraries and visualizing clustering results and compound structures (Fig. 1). All functions and data objects are well integrated into the existing infrastructure of the R environment. An overview of the *ChemmineR*-specific functions is provided in the online instructions and the PDF manual of this project.

2.2 Compound import and descriptor calculation

Compound structures are imported into *ChemmineR* in the generic structure definition file (SDF) format. Single compounds are imported by providing an SDF with one compound structure, whereas entire compound databases are imported by providing all SDF formatted structures concatenated in one batch file. The atom pair descriptors are calculated during the SDF import and stored in a searchable descriptor database as a list object (Chen and Reynolds, 2002). Because the calculation of descriptors for thousands of compounds can be time consuming, functions are provided to efficiently store and reload existing descriptor databases as binary files. Custom compound databases can be generated with a SDF subsetting function.

2.3 Compound searching, clustering and viewing

A search function is available to perform structural similarity searches against the generated atom pair descriptor databases. The default setting uses the Tanimoto coefficient as similarity measure (Holliday *et al.*, 2003). The search function can return all entries in a compound database sorted by similarity score. Alternatively, the search results can be limited by a similarity threshold or a desired number of similar compounds. To view the compounds structures of search results, the function can automatically upload the returned compounds to the online ChemMine service where they are rendered into chemical structure images (see below).

[*] To whom correspondence should be addressed.

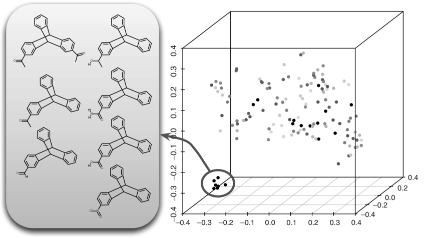

Fig. 1. Sample 3D scatter plot of a clustering result and online visualization of the corresponding compound structures using *ChemmineR*.

The compound search function calls internally a generic function for calculating atom pair-based similarities between compound structures (Chen and Reynolds, 2002). This similarity function can be used to calculate pairwise compound similarities or to design custom subroutines for similarity scoring and searching.

Structure-based clustering is required for many analysis steps of compound libraries and HTS datasets. *ChemmineR* provides a novel binning clustering method that is optimized for these compound analysis tasks. The algorithm uses single-linkage clustering to join compounds into similarity groups, where every member in a cluster shares with at least one another member a similarity value above a user-specified threshold. The algorithm is optimized for speed and memory efficiency by avoiding the calculation of an all-against-all distance matrix. This is achieved by calculating on-the-fly only the distance values that are required in each clustering step. Because an optimum similarity threshold is often unknown, a series of binning clustering result can be calculated simultaneously for several user-specified thresholds. Cluster results for several thresholds can be calculated almost with the same speed as for a single threshold by issuing multiple clustering processes simultaneously, but calculating the required distances only once.

If desired by the user, then the binning clustering function can generate an all-against-all distance matrix for clustering compound sets with many other classification algorithms available in R, such as hierarchical clustering or *K*-means. In addition, *ChemmineR* provides an interactive wrapper function for multidimensional scaling (MDS) clustering. The online instructions provide several examples on how to cluster compound sets in *ChemmineR* with external clustering utilities. These include examples for using the fully interactive visual data mining tool RGGobi (Lang *et al.*, 2007).

The *ChemmineR* package provides bidirectional communications with online tools and databases available on the ChemMine portal (Girke *et al.*, 2005). The service allows users to view and compare any combination of compound structure images in large batches via a standard internet browser along with extensive compound annotation information and custom data tables for basic QSAR analyses (Gedeck *et al.*, 2006). This includes structure viewing of extensive similarity search results generated by *ChemmineR*. All online viewing utilities can be accessed directly from R simply by selecting the online viewing argument in various *ChemmineR* functions or issuing a dedicated data exchange function. Uploading compound data to the ChemMine interface gives the user access to many additional tools available on ChemMine's online compound analysis WorkBench. This includes the calculation of physicochemical property descriptors (Guha *et al.*, 2006), inter-conversions between different structure formats (e.g. SMILES and SDF), searching of the millions of drug-like compounds available in ChemMine, and easy access to published bioactivity and target protein information.

The *ChemmineR* framework will be expanded in the future by adding many more useful compound and screening data analysis functions. These include functions for (1) calculating physicochemical properties of compounds directly in R, (2) local similarity searching based on most common substructures (MCS, Raymond *et al.*, 2002), (3) various utilities for QSAR modeling (Gedeck *et al.*, 2006) and (4) wrapper functions for interfacing directly with other open-source small molecule analysis projects, such as OpenBabel and JOELib (Guha *et al.*, 2006; O'Boyle *et al.*, 2008). Extensive user tutorials and download options of different package versions will be available from the *ChemmineR* project site and from the BioConductor site (Gentleman *et al.*, 2005).

3 DISCUSSION

ChemmineR is the first open-access compound mining toolkit for the popular statistical environment R. The package provides flexible functions for powerful structural similarity searches, compound clustering, screening library management and online batch viewing of chemical structures. Users with a basic understanding of the R environment can easily customize the provided functions and design sophisticated compound library analysis pipelines that utilize the extensive statistical and machine learning resources available in R.

ACKNOWLEDGEMENTS

Funding: This work was supported by NSF grants: IOB-0420033, IOB-0420152, DGE-0504249, and IIS-0711129.

Conflict of Interest: none declared.

REFERENCES

Carhart,R. *et al.* (1985) Atom pairs as molecular features in structure-activity studies: definition and applications. *J. Chem. Inf. Comput. Sci.*, **25**, 64–73.

Chen,J. *et al.* (2005) ChemDB: a public database of small molecules and related chemoinformatics resources. *Bioinformatics*, **21**, 4133–4139.

Chen,X. and Reynolds,C. (2002) Performance of similarity measures in 2D fragment-based similarity searching: comparison of structural descriptors and similarity coefficients. *J. Chem. Inf. Comput. Sci.*, **42**, 1407–1414.

Gedeck,P. *et al.* (2006) QSAR–how good is it in practice? Comparison of descriptor sets on an unbiased cross section of corporate data sets. *J. Chem. Inf. Model*, **46**, 1924–1936.

Gentleman,R. *et al.* (2005) *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, New York.

Girke,T. *et al.* (2005) ChemMine. A compound mining database for chemical genomics. *Plant Physiol.*, **138**, 573–577.

Guha,R. *et al.* (2006) The Blue obelisk-interoperability in chemical informatics. *J. Chem. Inf. Model*, **46**, 991–998.

Holliday,J.D. *et al.* (2003) Analysis and display of the size dependence of chemical similarity coefficients. *J. Chem. Inf. Comput. Sci.*, **43**, 819–828.

Irwin,J.J. and Shoichet,B.K. (2005) ZINC–a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model*, **45**, 177–182.

Lang,D.T. *et al.* (2007) rggobi: interface between R and GGobi. R package version 2.1.7.

O'Boyle,N.M. *et al.* (2008) Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. *Chem. Cent. J.*, **2**, 1–7.

R Development Core Team (2008). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Raymond,J. *et al.* (2002) Heuristics for similarity searching of chemical graphs using a maximum common edge subgraph algorithm. *J. Chem. Inf. Comput. Sci.*, **42**, 305–316.

Seiler,K.P. *et al.* (2008) ChemBank: a small-molecule screening and cheminformatics resource database. *Nucleic Acids Res.*, **36** (Database issue), 351–359.