

ChemmineR-V2: Analysis of Small Molecule and Screening Data

Yiqun Cao, Tyler Backman, Yan Wang, Thomas Girke
Email contact: thomas.girke@ucr.edu

February 21, 2011

1 Introduction

ChemmineR is an R package for analyzing small molecule and screening data. Its new version ChemmineR-V2 contains functions for processing SDFs (structure data files), molecule depictions, structural similarity searching, clustering/diversity analyses of compound libraries with a wide spectrum of algorithms and utilities for managing complex data sets from high-throughput compound bio-assays.

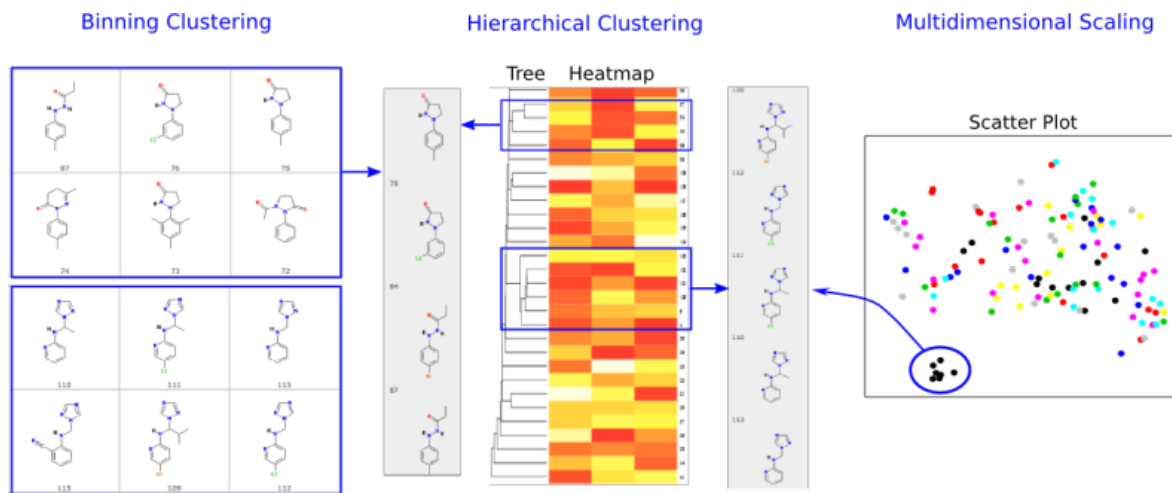


Figure 1: Selected functionalities provided by the *ChemmineR* package.

In addition, *ChemmineR* offers visualization functions for compound clustering results and chemical structures. The integration of chemoinformatic tools with the R programming environment has many advantages, such as easy access to a wide spectrum of statistical methods, machine learning algorithms and graphic utilities. The first version of this package was published in Cao et al. (2008). Since then many additional utilities have been added to the package and many more are under development for future releases.

2 Getting Started

2.1 Installation

The R software for running ChemmineR can be downloaded from CRAN (<http://cran.at.r-project.org/>). The ChemmineR package can be installed from R using the `biocLite` install command.

```
> source("http://bioconductor.org/biocLite.R")
> biocLite("ChemmineR")
```

2.2 Loading the Package and Documentation

```
> library("ChemmineR") # Loads the package

> library(help="ChemmineR") # Lists all functions and classes
> vignette("ChemmineR") # Opens this PDF manual from R
```

2.3 Five Minute Tutorial

The following code gives an overview of the most important functionalities provided by *ChemmineR*. Copy and paste of the commands into the R console will demonstrate these utilities.

Create Instances of *SDFset* class:

```
> data(sdfsampl)
> sdfset <- sdfsampl
> sdfset # Returns summary of SDFset
```

An instance of "SDFset" with 100 molecules

```
> sdfset[1:4] # Subsetting of object
```

An instance of "SDFset" with 4 molecules

```
> sdfset[[1]] # Returns summarized content of one SDF
```

An instance of "SDF"

```
<<header>>
```

```

Molecule_Name
"650001"
Source
" -OEChem-07071010512D"
Comment
""
Counts_Line
" 61 64 0 0 0 0 0 0 0999 V2000"
```

```
<<atomblock>>
```

	C1	C2	C3	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16
O_1	7.0468	0.0839	0	0	0	0	0	0	0	0	0	0	0	0	0
O_2	12.2708	1.0492	0	0	0	0	0	0	0	0	0	0	0	0	0
...
H_60	1.8411	-1.5985	0	0	0	0	0	0	0	0	0	0	0	0	0
H_61	2.6597	-1.2843	0	0	0	0	0	0	0	0	0	0	0	0	0

```
<<bondblock>>
```

	C1	C2	C3	C4	C5	C6	C7
1	1	16	2	0	0	0	0
2	2	23	1	0	0	0	0
...
63	33	60	1	0	0	0	0
64	33	61	1	0	0	0	0

```
<<datablock>> (33 data items)
```

PUBCHEM_COMPOUND_CID	PUBCHEM_COMPOUND_CANONICALIZED
"650001"	"1"
PUBCHEM_CACTVS_COMPLEXITY	PUBCHEM_CACTVS_HBOND_ACCEPTOR
"700"	"7"
"..."	

```
> view(sdfset[1:4]) # Returns summarized content of many SDFs, not printed here
> as(sdfset[1:4], "list") # Returns complete content of many SDFs, not printed here
```

An *SDFset* is created during the import of an SD file:

```
> sdfset <- read.SDFset("http://faculty.ucr.edu/~tgirke/Documents/
+ R_BioCond/Samples/sdfsamples.sdf")
```

Miscellaneous accessor methods for *SDFset* container:

```
> header(sdfset[1:4]) # Not printed here
```

```
> header(sdfset[[1]])
```

Molecule_Name
"650001"
Source
" -OEChem-07071010512D"
Comment
" "
Counts_Line
" 61 64 0 0 0 0 0 0 0999 V2000"

```
> atomblock(sdfset[1:4]) # Not printed here
```

```
> atomblock(sdfset[[1]])[1:4,]
```

	C1	C2	C3	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16
0_1	7.0468	0.0839	0	0	0	0	0	0	0	0	0	0	0	0	0
0_2	12.2708	1.0492	0	0	0	0	0	0	0	0	0	0	0	0	0
0_3	12.2708	3.1186	0	0	0	0	0	0	0	0	0	0	0	0	0
0_4	7.9128	2.5839	0	0	0	0	0	0	0	0	0	0	0	0	0

```
> bondblock(sdfset[1:4]) # Not printed here
```

```
> bondblock(sdfset[[1]])[1:4,]
```

	C1	C2	C3	C4	C5	C6	C7
1	1	16	2	0	0	0	0
2	2	23	1	0	0	0	0
3	2	27	1	0	0	0	0
4	3	25	1	0	0	0	0

```
> datablock(sdfset[1:4]) # Not printed here
```

```
> datablock(sdfset[[1]])[1:4]
```

PUBCHEM_COMPOUND_CID	PUBCHEM_COMPOUND_CANONICALIZED
"650001"	"1"
PUBCHEM_CACTVS_COMPLEXITY	PUBCHEM_CACTVS_HBOND_ACCEPTOR
"700"	"7"

Assigning compound IDs and keeping them unique:

```
> cid(sdfset)[1:4] # Returns IDs from SDFset object
```

```
[1] "CMP1" "CMP2" "CMP3" "CMP4"
```

```
> sdfid(sdfset)[1:4] # Returns IDs from SD file header block
```

```
[1] "650001" "650002" "650003" "650004"
```

```
> unique_ids <- makeUnique(sdfid(sdfset))
```

```
[1] "No duplicates detected!"
```

```
> cid(sdfset) <- unique_ids
```

Converting the data blocks in an *SDFset* to a matrix:

```
> blockmatrix <- datablock2ma(datablocklist=datablock(sdfset))
```

```
> # Converts data block to matrix
```

```
> numchar <- splitNumChar(blockmatrix=blockmatrix)
```

```
> # Splits to numeric and character matrix
```

```
> numchar[[1]][1:2,1:2] # Slice of numeric matrix
```

```

      PUBCHEM_COMPOUND_CID  PUBCHEM_COMPOUND_CANONICALIZED
650001          650001                      1
650002          650002                      1

```

```
> numchar[[2]][1:2,10:11] # Slice of character matrix
```

```

      PUBCHEM_MOLECULAR_FORMULA
650001 "C23H28N4O6"
650002 "C18H23N5O3"
      PUBCHEM_OPENEYE_CAN_SMILES
650001 "CC1=CC(=NO1)NC(=O)CCC(=O)N(CC(=O)NC2CCCC2)C3=CC4=C(C=C3)OCCO4"
650002 "CN1C2=C(C(=O)NC1=O)N(C(=N2)NCCCC)CCCC3=CC=CC=C3"

```

Compute atom frequency matrix, molecular weight and formula:

```
> propma <- data.frame(MF=MF(sdfset), MW=MW(sdfset), atomcountMA(sdfset))
> propma[1:4, ]
```

```

      MF      MW  C  H N O S F Cl
650001 C23H28N4O6 456.4916 23 28 4 6 0 0 0
650002 C18H23N5O3 357.4069 18 23 5 3 0 0 0
650003 C18H18N4O3S 370.4255 18 18 4 3 1 0 0
650004 C21H27N5O5S 461.5346 21 27 5 5 1 0 0

```

Assign matrix data to data block:

```
> datablock(sdfset) <- propma
> datablock(sdfset[1])
```

```

$`650001`
      MF      MW      C      H      N      O
"C23H28N4O6" "456.4916" "23"    "28"    "4"    "6"
      S      F      Cl
      "0"    "0"    "0"

```

String searching in *SDFset* ():

```
> grepSDFset("650001", sdfset, field="datablock", mode="subset")
> # Returns summary view of matches. Not printed here.
> .
```

```
> grepSDFset("650001", sdfset, field="datablock", mode="index")
```

```

1 1 1 1 1 1 1 1 1
1 2 3 4 5 6 7 8 9

```

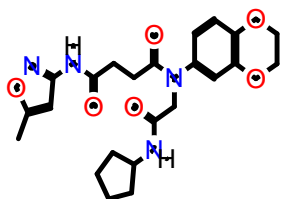
Export SDFset to SD file:

```
> write.SDF(sdfset[1:4], file="sub.sdf", sig=TRUE)
```

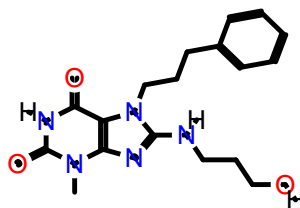
Plot molecule structure of one or many SDFs:

```
> plot(sdfset[1:4], print=FALSE) # Plots structures to R graphics device
```

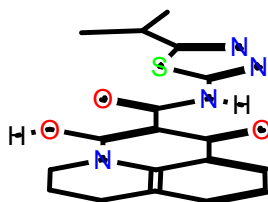
650001



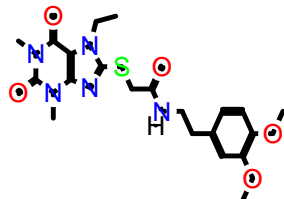
650002



650003



650004



```
> sdf.visualize(sdfset[1:4]) # Compound viewing in web browser
```

Structure similarity searching and clustering:

```
> apset <- sdf2ap(sdfset)
> # Generate atom pair descriptor database for searching
> cmp.search(apset, apset[1], type=3, cutoff = 0.3, quiet=TRUE)
```

	index	cid	scores
1	1	650001	1.0000000
2	96	650102	0.3516643
3	67	650072	0.3117569
4	88	650094	0.3094629
5	15	650015	0.3010753

```
> # Search apset database with single compound.
> cmp.cluster(db=apset, cutoff = c(0.65, 0.5), quiet=TRUE)[1:4,]
```

UCR :: IIGB :: CEPCEB

ChemMine | **ChemmineR**

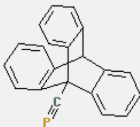
Systemics Network GCD Expression POND CWN BAP DB ChemMine Links

Home
Readme
2010 Project
Protocols
CMP Sources
Search Database
Annotation
Structure
Screen Data
Workbench
Manage CMPs
Descriptors
Clustering
Clusters
Software
ChemmineR
Links
Login

[View Previously Accessed Compounds >>>](#) Width of information table:

Reference Compound (ka-01834)

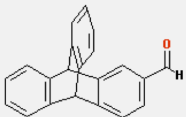
[View SDF](#) [Structure Search](#) [Add to Selection](#)



Similarities With All	
ids	scores
3	0.550335570
43	0.484662577
42	0.484662577
1	0.484662577
4	0.480122324
2	0.480122324
44	0.356097561
46	0.312500000
11	0.311653117
35	0.287719298

(ChemmineR_Unnamed_Compound_3)

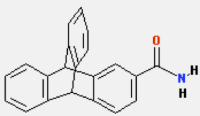
[View SDF](#) [Structure Search](#) [Add to Selection](#)



Similarities With All	
ids	scores
3	1.000000000
43	0.785977860
42	0.785977860
1	0.785977860
2	0.766423358
4	0.646258503
44	0.561797753
11	0.415204678
35	0.390151515
46	0.368539326

(ChemmineR_Unnamed_Compound_43)

[View SDF](#) [Structure Search](#) [Add to Selection](#)



Similarities With All	
ids	scores
43	1.000000000
42	0.840000000
1	0.840000000
3	0.785977860
2	0.709459459
4	0.611464968
44	0.601108033
11	0.390109890
46	0.339702760
35	0.223128352

Figure 2: Visualization webpage created by calling `sdf.visualize`.

```
sorting result...
      ids CLSZ_0.65 CLID_0.65 CLSZ_0.5 CLID_0.5
48 650049         2        48         2        48
49 650050         2        48         2        48
54 650059         2        54         2        54
55 650060         2        54         2        54
```

3 Overview of Classes and Functions

The following list gives an overview of the most important S4 classes, methods and functions available in the ChemmineR package. The help documents of the package provide much more detailed information on each utility. The standard R help documents for these utilities can be accessed with this syntax: `?function_name` (e.g. `?cid`) and `?class_name-class` (e.g. `?SDFset-class`).

3.1 Molecular Structure Data

Classes

- *SDFstr*: intermediate string class to facilitate SD file import; not important for end user
- *SDF*: container for single molecule imported from an SD file
- *SDFset*: container for many SDF objects; most important structure container for end user

Functions/Methods

- Accessor methods for *SDF/SDFset*
 - Object slots: `cid`, `header`, `atomblock`, `bondblock`, `datablock` (`sd fid`, `data blocktag`)
 - Summary of *SDFset*: `view`
 - Matrix conversion of data block: `datablock2ma`, `splitNumChar`
 - String search in *SDFset*: `grepSDFset`
- Coerce one class to another
 - Standard syntax `as(..., "...")` works in most cases. For details see R help with `?SDFset-class`.
- Utilities
 - Atom frequencies: `atomcountMA`, `atomcount`
 - Molecular weight: `MW`
 - Molecular formula: `MF`
- Compound structure depictions
 - R graphics device: `plot`, `plotStruc`
 - Online: `cmp.visualize`

3.2 Structure Descriptor Data

Classes

- *AP*: container for atom pair descriptors of a single molecule
- *APset*: container for many AP objects; most important structure descriptor container for end user

Functions/Methods

- Create *AP/APset* instances
 - From *SDFset*: `sdf2ap`
 - From SD file: `cmp.parse`
 - Summary of *AP/APset*: `view`, `db.explain`
- Accessor methods for AP/APset
 - Object slots: `ap`, `cid`
- Coerce one class to another
 - Standard syntax `as(..., "...")` works in most cases. For details see R help with `?“APset-class”`.
- Structure Similarity comparisons and Searching
 - Compute pairwise similarities : `cmp.similarity`
 - Search APset database: `cmp.search`
 - Compute pairwise similarities : `cmp.similarity`
- AP-based Structure Similarity Clustering
 - Single-linkage binning clustering: `cmp.cluster`
 - Visualize clustering result with MDS: `cluster.visualize`
 - Size distribution of clusters: `cluster.sizestat`

4 Importing Compounds

The following code gives an overview of the most important import/export functionalities provided by *ChemmineR*. The example creates an instance of the *SDFset* class using as sample data set the first 100 compounds from this PubChem SD file (SDF): `Compound_00650001_00675000.sdf.gz` (`ftp://ftp.ncbi.nih.gov/pubchem/Compound/CURRENT-Full/SDF/`).

SDFs can be imported with the `read.SDFset` function:

```
> sdfset <- read.SDFset("http://faculty.ucr.edu/~tgirke/Documents/  
+ R_BioCond/Samples/sdfsamples.sdf")
```

```
> data(sdfsample) # Loads the same SDFset provided by the library
> sdfset <- sdfsample
> valid <- validSDF(sdfset) # Identifies invalid SDFs in SDFset objects
> sdfset <- sdfset[valid] # Removes invalid SDFs, if there are any
```

Import SD file into *SDFstr* container:

```
> sdfstr <- read.SDFstr("http://faculty.ucr.edu/~tgirke/Documents/
+                      R_BioCond/Samples/sdfsample.sdf")
```

Create *SDFset* from *SDFstr* class:

```
> sdfstr <- as(sdfset, "SDFstr")
> sdfstr
```

An instance of "SDFstr" with 100 molecules

```
> as(sdfstr, "SDFset")
```

An instance of "SDFset" with 100 molecules

5 Export of Compounds

Write objects of classes *SDFset*/*SDFstr*/*SDF* to SD file:

```
> write.SDF(sdfset[1:4], file="sub.sdf")
```

Writing customized *SDFset* to file containing *ChemmineR* signature, IDs from *SDFset* and no data block:

```
> write.SDF(sdfset[1:4], file="sub.sdf", sig=TRUE, cid=TRUE, db=NULL)
```

Example for injecting a custom matrix/data frame into the data block of an *SDFset* and then writing it to an SD file:

```
> props <- data.frame(MF=MF(sdfset), MW=MW(sdfset), atomcountMA(sdfset))
> datablock(sdfset) <- props
> write.SDF(sdfset[1:4], file="sub.sdf", sig=TRUE, cid=TRUE)
```

Indirect export via *SDFstr* object:

```
> sdf2str(sdf=sdfset[[1]], sig=TRUE, cid=TRUE)
> # Uses default components
> sdf2str(sdf=sdfset[[1]], head=letters[1:4], db=NULL)
```

Write *SDF*, *SDFset* or *SDFstr* classes to file:

```
> write.SDF(sdfset[1:4], file="sub.sdf", sig=TRUE, cid=TRUE, db=NULL)
> write.SDF(sdfstr[1:4], file="sub.sdf")
> cat(unlist(as(sdfstr[1:4], "list")), file="sub.sdf", sep="\n")
```

6 Working with SDF/SDFset Classes

Several methods are available to return the different data components of *SDF/SDFset* containers in batches. The following examples list the most important ones. To save space their content is not printed in the manual.

```
> view(sdfset[1:4]) # Summary view of several molecules
> length(sdfset) # Returns number of molecules
> sdfset[[1]] # Returns single molecule from SDFset as SDF object
> sdfset[[1]][[2]] # Returns atom block from first compound as matrix
> sdfset[[1]][[2]][1:4,]
> c(sdfset[1:4], sdfset[5:8]) # Concatenation of several SDFsets
```

The `grepSDFset` function allows string matching/searching on the different data components in *SDFset*. By default the function returns a SDF summary of the matching entries. Alternatively, an index of the matches can be returned with the setting `mode="index"`.

```
> grepSDFset("650001", sdfset, field="datablock", mode="subset")
> # To return index, set mode="index")
> .
```

Utilities to maintain unique compound IDs:

```
> sdfid(sdfset[1:4])
> # Retrieves CMP IDs from Molecule Name field in header block.
> cid(sdfset[1:4])
> # Retrieves CMP IDs from ID slot in SDFset.
> unique_ids <- makeUnique(sdfid(sdfset))
> # Creates unique IDs by appending a counter to duplicates.
> cid(sdfset) <- unique_ids # Assigns uniquified IDs to ID slot
```

Subsetting by character, index and logical vectors:

```
> view(sdfset[c("650001", "650012")])
> view(sdfset[4:1])
> mylog <- cid(sdfset) %in% c("650001", "650012")
> view(sdfset[mylog])
```

Accessing *SDF/SDFset* components: header, atom, bond and data blocks

```
> atomblock(sdf); sdf[[2]]; sdf[["atomblock"]]
> # All three methods return the same component
> header(sdfset[1:4])
> atomblock(sdfset[1:4])
> bondblock(sdfset[1:4])
> datablock(sdfset[1:4])
> header(sdfset[[1]])
> atomblock(sdfset[[1]])
> bondblock(sdfset[[1]])
> datablock(sdfset[[1]])
```

Replacement Methods:

```
> sdfset[[1]][[2]][1,1] <- 999
> atomblock(sdfset)[1] <- atomblock(sdfset)[2]
> datablock(sdfset)[1] <- datablock(sdfset)[2]
```

Assign matrix data to data block:

```
> datablock(sdfset) <- as.matrix(iris[1:100,])
> view(sdfset[1:4])
```

Class coercions from *SDFstr* to *list*, *SDF* and *SDFset*:

```
> as(sdfstr[1:2], "list")
> as(sdfstr[[1]], "SDF")
> as(sdfstr[1:2], "SDFset")
```

Class coercions from *SDF* to *SDFstr*, *SDFset*, list with SDF sub-components:

```
> sdfcomplist <- as(sdf, "list")
> sdfcomplist <- as(sdfset[1:4], "list"); as(sdfcomplist[[1]], "SDF")
> sdflist <- as(sdfset[1:4], "SDF"); as(sdflist, "SDFset")
> as(sdfset[[1]], "SDFstr")
> as(sdfset[[1]], "SDFset")
```

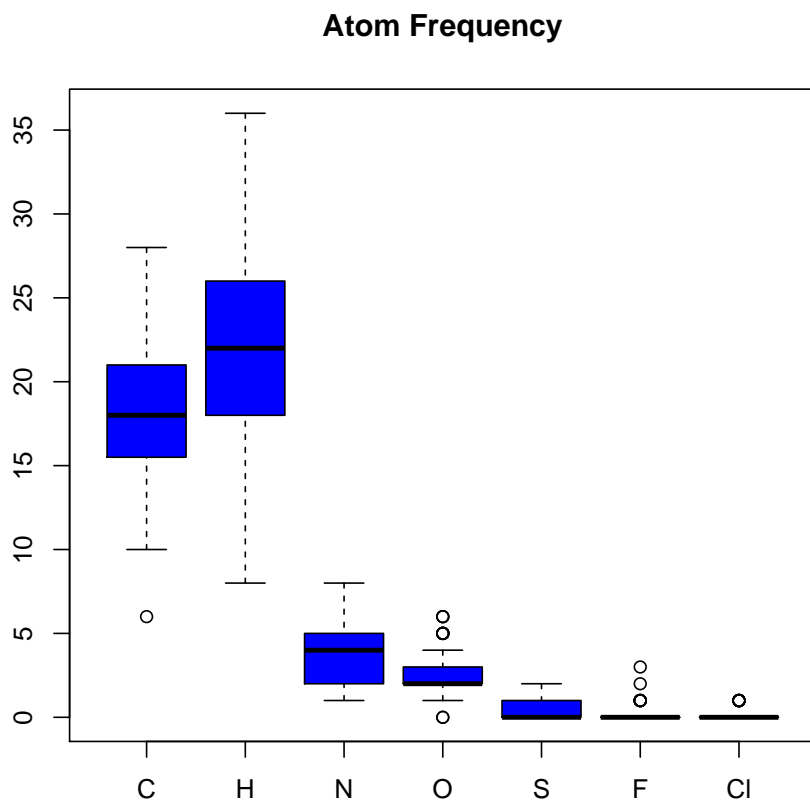
Class coercions from *SDFset* to lists with components consisting of SDF or sub-components:

```
> as(sdfset[1:4], "SDF")
> as(sdfset[1:4], "list")
> as(sdfset[1:4], "SDFstr")
```

7 Molecular Property Functions

Several methods and functions are available to compute basic compound descriptors, such as molecular formula (MF), molecular weight (MW) and atom frequencies.

```
> propma <- atomcountMA(sdfset)
> boxplot(propma, col="blue", main="Atom Frequency")
```



```
> boxplot(rowSums(propma), main="All Atom Frequency")
```

Compute MW and formula:

```
> MW(sdfset[1:4]); MF(sdfset[1:4])
> propma <- data.frame(MF=MF(sdfset), MW=MW(sdfset), atomcountMA(sdfset)); propma[1:4,]
> datablock(sdfset) <- propma # Works with all SDF components
> test <- apply(propma[1:4,], 1, function(x) data.frame(col=colnames(propma), value=x))
> sdf.visualize(sdfset[1:4], extra = test)
```

The following shows an example for assigning the values stored in a matrix (*e.g.* property descriptors) to the data block components in an *SDFset*. Each matrix row will be assigned to the corresponding slot position in the *SDFset*.

```
> datablock(sdfset) <- propma
> datablock(sdfset)[1:4]
```

\$CMP1

C	H	N	O	S	F	Cl
"23"	"28"	"4"	"6"	"0"	"0"	"0"

\$CMP2

```
  C    H    N    O    S    F    Cl
"18" "23" "5"  "3"  "0"  "0"  "0"
```

```
$CMP3
```

```
  C    H    N    O    S    F    Cl
"18" "18" "4"  "3"  "1"  "0"  "0"
```

```
$CMP4
```

```
  C    H    N    O    S    F    Cl
"21" "27" "5"  "5"  "1"  "0"  "0"
```

The data blocks in SDFs contain often important annotation information about compounds. The `datablock2ma` function returns this information as matrix for all compounds stored in an *SDFset* container. The `splitNumChar` function can then be used to organize all numeric columns in a *numeric matrix* and the character columns in a *character matrix* as components of a *list* object.

```
> datablocktag(sdfset, tag="PUBCHEM_NIST_INCHI")
> datablocktag(sdfset, tag="PUBCHEM_OPENEYE_CAN_SMILES")
```

Convert entire data block to matrix:

```
> blockmatrix <- datablock2ma(datablocklist=datablock(sdfset))
> # Converts data block to matrix
> numchar <- splitNumChar(blockmatrix=blockmatrix)
> # Splits matrix to numeric matrix and character matrix
> numchar[[1]][1:4,]; numchar[[2]][1:4,]
> # Splits matrix to numeric matrix and character matrix
> .
```

8 Rendering Chemical Structure Images

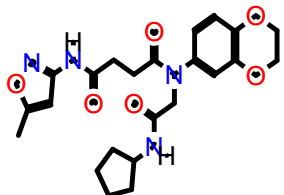
8.1 R Graphics Device

A new plotting function for compound structures has been added to the package recently. This function uses the native R graphics device for generating compound depictions. At this point this function is still in an experimental developmental stage but should become stable soon.

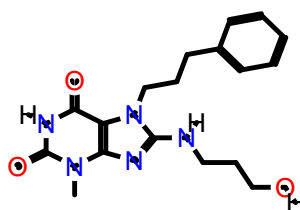
Plot Compound Structures with R's graphics device

```
> data(sdfsample); sdfset <- sdfsample
> plot(sdfset[1:4], print=FALSE) # 'print=TRUE' returns SDF summaries
```

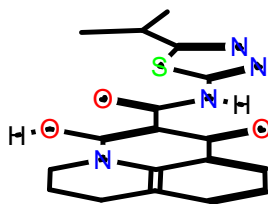
CMP1



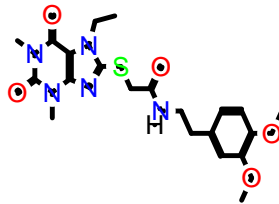
CMP2



CMP3



CMP4



Customize plot:

```
> plot(sdfset[1:4], griddim=c(2,2), print_cid=letters[1:4], print=FALSE,  
+       noHbonds=FALSE)
```

8.2 Online with ChemMine Tools

Alternatively, one can visualize compound structures with a standard web browser using the online ChemMine Tools service. The service allows to display other information next to the structures using the extra argument of the `sdf.visualize` function. The following examples demonstrate, how one can plot and annotate structures by passing on extra data as vector of character strings, matrices or lists.

Plot structures using web service ChemMine Tools:

```
> sdf.visualize(sdfset[1:4])
```

Add extra annotation as *vector*:

```
> sdf.visualize(sdfset[1:4], extra=month.name[1:4])
```

Add extra annotation as *matrix*:

UCR :: IIGB :: CEPCEB

ChemMine **ChemmineR**

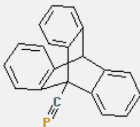
Systemics Network GCD Expression POND CWN BAP DB ChemMine Links

Home
Readme
2010 Project
Protocols
CMP Sources
Search Database
Annotation
Structure
Screen Data
Workbench
Manage CMPs
Descriptors
Clustering
Clusters
Software
ChemmineR
Links
Login

[View Previously Accessed Compounds >>>](#) Width of information table:

Reference Compound (ka-01834)

[View SDF](#) [Structure Search](#) [Add to Selection](#)

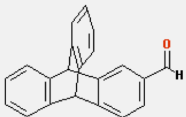


Similarities With All

ids	scores
3	0.550335570
43	0.484662577
42	0.484662577
1	0.484662577
4	0.480122324
2	0.480122324
44	0.356097561
46	0.312500000
11	0.311653117
35	0.287719298

(ChemmineR_Unnamed_Compound_3)

[View SDF](#) [Structure Search](#) [Add to Selection](#)

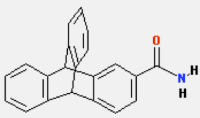


Similarities With All

ids	scores
3	1.000000000
43	0.785977860
42	0.785977860
1	0.785977860
2	0.766423358
4	0.646258503
44	0.561797753
11	0.415204678
35	0.390151515
46	0.368539326

(ChemmineR_Unnamed_Compound_43)

[View SDF](#) [Structure Search](#) [Add to Selection](#)



Similarities With All

ids	scores
43	1.000000000
42	0.840000000
1	0.840000000
3	0.785977860
2	0.709459459
4	0.611464968
44	0.601108033
11	0.390109890
46	0.339702760
35	0.223128352

Figure 3: Visualization webpage created by calling `sdf.visualize`.


```
> extra <- apply(propma[1:4,], 1, function(x)
+               data.frame(Property=colnames(propma), Value=x))
> sdf.visualize(sdfset[1:4], extra=extra)
```

Add extra annotation as *list*:

```
> sdf.visualize(sdfset[1:4], extra=bondblock(sdfset[1:4]))
```

9 Similarity Comparisons and Searching

9.1 AP/APset Classes for Storing Atom Pair Descriptors

The function `sdf2ap` computes atom pair descriptors for one or many compounds (Carhart et al., 1985; Chen and Reynolds, 2002). It returns a searchable atom pair database stored in a container of class *APset*, which can be used for structural similarity searching and clustering. As similarity measure, the Tanimoto coefficient or related coefficients can be used. An *APset* object consists of one or many *AP* entries each storing the atom pairs of a single compound. Note: the deprecated `cmp.parse` function is still available which also generates atom pair descriptor databases, but directly from an SD file. Since the latter function is less flexible it may be discontinued in the future.

Generate atom pair descriptor database for searching:

```
> ap <- sdf2ap(sdfset[[1]]) # For single compound
> ap
```

An instance of "AP"

```
<<atom pairs>>
```

```
53688190976 53688190977 53688190978 53688190979 53688190980 ... length: 528
```

```
> apset <- sdf2ap(sdfset) # For many compounds.
```

```
> view(apset[1:4])
```

```
$`650001`
```

An instance of "AP"

```
<<atom pairs>>
```

```
53688190976 53688190977 53688190978 53688190979 53688190980 ... length: 528
```

```
$`650002`
```

An instance of "AP"

```
<<atom pairs>>
```

```
53688190976 53688190977 53688190978 53688190979 53689239552 ... length: 325
```

```
$`650003`
```

An instance of "AP"

```
<<atom pairs>>
```

```
52615496704 53688190976 53688190977 53689239552 53697627136 ... length: 325
```

```
$`650004`  
An instance of "AP"  
<<atom pairs>>  
52617593856 52618642432 52619691008 52619691009 52628079616 ... length: 496
```

9.2 Large SDF and Atom Pair Databases

When working with large data sets it is often desirable to save the *SDFset* and *APset* containers as binary R objects to files for later use. This way they can be loaded very quickly into a new R session without recreating them every time from scratch.

Save and load of *SDFset* and *APset* containers:

```
> save(sdfset, file = "sdfset.rda", compress = TRUE)  
> load("sdfset.rda")  
> save(apset, file = "apset.rda", compress = TRUE)  
> load("apset.rda")
```

9.3 Pairwise Compound Comparisons

The `cmp.similarity` function computes the atom pair similarity between two compounds using the Tanimoto coefficient as similarity measure.

```
> cmp.similarity(apset[1], apset[2])
```

```
[1] 0.2637037
```

```
> cmp.similarity(apset[1], apset[1])
```

```
[1] 1
```

9.4 Similarity Searching

The `cmp.search` function searches an atom pair database for compounds that are similar to a query compound. The following example returns a data frame where the rows are sorted by the Tanimoto similarity score (best to worst). The first column contains the indices of the matching compounds in the database. The argument `cutoff` can be a similarity cutoff, meaning only compounds with a similarity value larger than this cutoff will be returned; or it can be an integer value restricting how many compounds will be returned. When supplying a cutoff of 0, the function will return the similarity values for every compound in the database.

```
> cmp.search(apset, apset[1], type=3, cutoff = 0.2, quiet=TRUE)
```

	index	cid	scores
1	1	650001	1.0000000
2	96	650102	0.3516643
3	67	650072	0.3117569
4	88	650094	0.3094629
5	15	650015	0.3010753

6	77	650082	0.2960969
7	31	650032	0.2848181
8	98	650104	0.2777778
9	86	650092	0.2739274
10	83	650089	0.2738462
11	64	650069	0.2736842
12	85	650091	0.2724796
13	72	650077	0.2674591
14	4	650004	0.2641975
15	2	650002	0.2637037
16	51	650054	0.2633411
17	23	650024	0.2581121
18	74	650079	0.2575107
19	11	650011	0.2559653
20	38	650039	0.2539062
21	79	650085	0.2518337
22	70	650075	0.2506297
23	75	650080	0.2496552
24	25	650026	0.2485795
25	93	650099	0.2438163
26	32	650033	0.2410959
27	69	650074	0.2408840
28	52	650056	0.2330346
29	43	650044	0.2322503
30	63	650068	0.2321900
31	47	650048	0.2320099
32	66	650071	0.2301459
33	91	650097	0.2251908
34	78	650083	0.2225313
35	94	650100	0.2208333
36	3	650003	0.2185714
37	16	650016	0.2176471
38	18	650019	0.2163389
39	99	650105	0.2159091
40	39	650040	0.2127329
41	68	650073	0.2124601
42	45	650046	0.2112971
43	71	650076	0.2107438
44	20	650021	0.2099291
45	22	650023	0.2098361
46	9	650009	0.2098361
47	12	650012	0.2082153
48	92	650098	0.2071097
49	61	650066	0.2065064
50	60	650065	0.2065064
51	19	650020	0.2034884
52	40	650041	0.2019544

Return main components of APset objects:

```
> cid(apset[1:4]) # Compound IDs
> ap(apset[1:4]) # Atom pair descriptors
> db.explain(apset[1]) # Return atom pairs in human readable format
```

Coerce APset to other objects:

```
> apset2descdb(apset) # Returns old list-style AP database
> tmp <- as(apset, "list") # Returns list
> as(tmp, "APset") # Converts list back to APset
```

10 Clustering Compound Structures

```
> cmp.cluster(db=apset, cutoff = c(0.65, 0.5))[1:4,] # Binning clustering using variable simi
```

References

- Y Cao, A Charisi, L C Cheng, T Jiang, and T Girke. ChemmineR: a compound mining framework for R. *Bioinformatics*, 24(15):1733–1734, Aug 2008. doi: 10.1093/bioinformatics/btn307. URL <http://www.hubmed.org/display.cgi?uids=18596077>.
- R.E. Carhart, D.H. Smith, and R. Venkataraghavan. Atom pairs as molecular features in structure-activity studies: definition and applications. *Journal of Chemical Information and Computer Sciences*, 25(2):64–73, 1985.
- X. Chen and C.H. Reynolds. Performance of Similarity Measures in 2D Fragment-Based Similarity Searching: Comparison of Structural Descriptors and Similarity Coefficients. *Journal of Chemical Information and Computer Sciences*, 42(6):1407–1414, 2002.