Journal of Cheminformatics

EDUCATIONAL Open Access

The rcdk and cluster R packages applied to drug candidate selection

Adrian Voicu[1†], Narcis Duteanu[2*], Mirela Voicu[3†], Daliborca Vlad[4] and Victor Dumitrascu[4]

Abstract

The aim of this article is to show how thevpower of statistics and cheminformatics can be combined, in R, using two packages: *rcdk* and *cluster*.

We describe the role of clustering methods for identifying similar structures in a group of 23 molecules according to their fingerprints. The most commonly used method is to group the molecules using a "score" obtained by measuring the average distance between them. This score reflects the similarity/non-similarity between compounds and helps us identify active or potentially toxic substances through predictive studies.

Clustering is the process by which the common characteristics of a particular class of compounds are identified. For clustering applications, we are generally measure the molecular fingerprint similarity with the Tanimoto coefficient. Based on the molecular fingerprints, we calculated the molecular distances between the methotrexate molecule and the other 23 molecules in the group, and organized them into a matrix. According to the molecular distances and Ward 's method, the molecules were grouped into 3 clusters. We can presume structural similarity between the compounds and their locations in the cluster map. Because only 5 molecules were included in the methotrexate cluster, we considered that they might have similar properties and might be further tested as potential drug candidates.

Keywords: Cytostatic, Molecular fingerprint, Rcdk, Clusters

Introduction

Discovery, synthesis and production of new drugs is still challenging for researchers because of the complex structures of endogenous molecules involved in the pathogenesis of diseases such as AIDS, cancer and autism [16]. Modern drug research is characterized by the growing number of lead molecules and the need to examine and characterize all of these compounds over short periods [14, 39].

Chemical database mining based on the similar compounds search is an in silico method widely used in the drug discovery process [28, 33]. It can be used in the initial stages of drug discovery and speeds up the entire process [10]. The requirement to store, manage and analyse these rapidly growing resources has given rise to a relatively new field known as computer-assisted drug design (CADD) [22, 39, 40].

Computational chemistry is a very effective approach in drug design for the identification of lead compounds. Various virtual screening techniques can be used to reduce the cost and time required to identify a potential drug [2]. As a computational method in drug discovery and virtual screening, clustering of chemical compounds by the similarity of their molecular fingerprints can be used to identify similar structures in a large set of similar data [38, 41]. Their virtual screening performance is comparable to other,more complex, methods. There are many types of fingerprints, each of which represents a different aspect of the molecule [37, 42, 43].

Clustering is an unsupervised machine learning technique that groups data with similar properties. This technique for statistical data analysis is widely used in cheminformatics [19].

*Correspondence: narcis.duteanu@upt.ro
†Adrian Voicu and Mirela Voicu contributed equally to this manuscript.
[2] Dep. CAICAM, Politehnica University of Timisoara, Pirvan Boulevard 6, Timisoara, Romania
Full list of author information is available at the end of the article

A *cluster* is, in this case, a collection of molecules which are organized in groups, according to their molecular fingerprints [3, 17].

Despite the large number of clustering methods, only a few of them are widely used in practice. In this paper, only two of them were used, which proved to be suitable for chemical structure analysis: hierarchical clustering and K-means method. Regardless of the method, the results were the same.

Methods

All the software used for this article can be installed on Windows, Linux or macOS operating systems.

Initially, built as an environment for statistical computing, R, a GNU project, provides a wide variety of packages for cheminformatics that are suitable for calculating molecular fingerprints and clustering [13, 25, 46]. The latest version of R can be downloaded form the CRAN repository. R Studio is considered one of the best IDEs (integrated development environments) for R and was also used for this article. In this paper, R version 3.6.0 and RStudio version 1.2.1335 were used.

Marvin Sketch version 17.3, from ChemAxon, an academic software package, was used to draw, display and characterize the chemical structures [8]. The molecules were imported in Marvin Sketch using their IUPAC names (International Union of Pure and Applied Chemistry) and then saved as SMILES and SDF formats [31]. Then, they were imported and processed in R [12, 28].

R applications for cheminformatics and computational chemistry

Its flexibility and wide application fields have made the R programming environment a popular choice in a large number of areas.

In the field of cheminformatics, R offers several tools that are able to treat a large variety of issues related to the statistical modelling of chemical information. The *rcdk* package, version: 3.4.7.1, used in the present work, provides direct access from the R environment to the CDK (Chemistry Development Kit), a powerful Java framework for cheminformatics [6, 47].

CDK is a collection of free Java libraries that supports a wide variety of cheminformatics functionality. This platform allows us to read different molecular formats, calculate molecular descriptors and evaluate molecular fingerprints.

The cluster package, version 2.1.0, can be used to find groups of molecules that share similar chemical properties [2, 23].

The packages can be installed using the function "install.packages()". The general syntax is listed below:

```
install.packages("package_name")
```

To use a package, it must be loaded in the R environment using the function **library().**

In addition to *rcdk*, some other packages were also needed:

```
library(rcdk)
library(chemometrics)
library(rJava)
library(ChemmineR)
library(cluster)
library(rgl)
library(vegan)
library(factoextra)
library(fingerprint)
library(fmcsR)
library(NbClust )
library(iqspr)
library(ggplot2)
library(gridExtra)}
```

Importing and viewing the "drug candidate" molecules in R

In order to manipulate the chemical structures in R, we assigned them a code, starting with CMP1 for methotrexate and ending with CMP24 for the last structure. The Methotrexate molecule (coded as "CMP1") was downloaded from ZINC15, a free database of commercially available compounds, in both SMILE and SDF file format.

In SDF format, the molecule of methotrexate can be imported and visualized in R using the code listed below: [13, 44]

```
CMP1 <- load.molecules( c('CMP1.sdf') )
view.molecule.2d(CMP1[[1]])
```

The result is depicted in Fig. 1:

All the molecules were imported in SDF format and visualized in R as a grid.

```
mols <- load.molecules
(c('CMP1.sdf', 'CMP2.sdf', 'CMP3.sdf',
'CMP4.sdf', 'CMP5.sdf', 'CMP6.sdf',
'CMP7.sdf', 'CMP8.sdf', 'CMP9.sdf',
'CMP10.sdf', 'CMP11.sdf', 'CMP12.sdf',
'CMP13.sdf', 'CMP14.sdf', 'CMP15.sdf',
'CMP16.sdf', 'CMP17.sdf','CMP18.sdf',
'CMP19.sdf','CMP20.sdf','CMP21.sdf',
'CMP22.sdf', 'CMP23.sdf','CMP24.sdf'))
view.molecule.2d
(mols, ncol = 4, width = 200,
 height = 200, depictor = NULL,
 type="isomeric")
view.molecule.2d(mols)
```

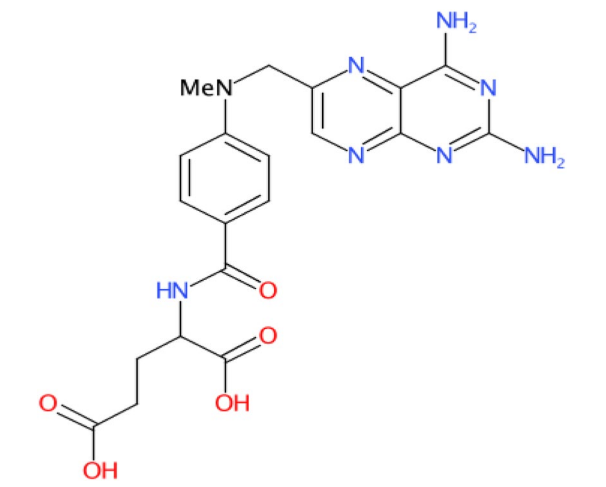

Fig. 1 Methotrexate molecule visualisation in R

The result is depicted in Fig. 2:

Computation of the molecular descriptors (physicochemical properties of the molecules)

The *rcdk* package can also be used to calculate a set of physicochemical properties of the molecules:

The number of atoms:

```
cat(’No. of atoms =’, length(atoms), ’\n’)
No. of atoms = 33
```

The number of chemical bonds:

```
cat(’No. of bonds =’, length(bonds), ’\n’)
No. of bonds = 35
```

The coordinates of the first atom:

```
cat(’No. of bonds =’, length(bonds), ’\n’)
No. of bonds = 35
[1]  2.1434 -4.5375
```

It is also possible to calculate the coordinates for all the atoms present in the molecule:

```
coords <- do.call(’rbind’ ,
lapply(atoms, get.point2d))
coords
```

R can compute a set of molecular descriptors, grouped into 5 different categories:

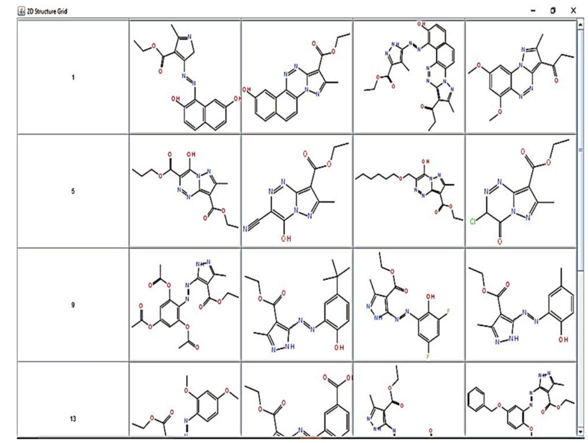

Fig. 2 Molecule set visualization

```
dc <- get.desc.categories()
dc
[1] "hybrid" "constitutional" "topological"
[4] "electronic" "geometrical"
```

Category 2 (constitutional), important in QSAR, contains 15 descriptors, which are listed below: [14].

```
dn <- get.desc.names(dc[2])
 dn
XlogP, Weight, RuleOfFiveDescriptor,
RotatableBondsCount, MannholdLogP,
LongestAliphaticChain, LargestPiSystem,
LargestChain, BondCount,
BasicGroupCount, AtomCount,
AromaticBondsCount, AromaticAtomsCount,
cdk.qsar.descriptors, ALOGP, AcidicGroup
```

Regarding drug design, the evaluation of AlogP is given a higher importance than that of other descriptors:

```
aDesc <- eval.desc(meth, dn[14])
aDesc
ALogP  ALogp2      AMR
1 -3.4898 12.1787 113.7535
allDescs <- eval.desc(mol, dn)
allDescs
XLogP       MW       LipinskiFailures
2.955      339.1219         0
nRotB MLogP  nAtomLAC  nAtomP
1      5    2.67      0        21
nAtomLC  nB  nBase  nAtom
5        27     0    42
nAromBond  naAromAtom   ALogP
1   11          10         0.1535
AMR  nAcid   ALogp2
1   92.9528     0 0.02356225
```

Computation of the molecular fingerprints

Molecular fingerprints can be computed by several methods, but in the case of aromatic compounds the "extended" method is preferred. "Extended" fingerprints have a length (the number of bits) of 1024, compared to 166 for "maccs "type fingerprints [11, 24, 36].

```
fps <- get.fingerprint
(CMP1, type=’extended’)
fps
 Fingerprint object
 name =
 length =  1024
 folded =  FALSE
 source =  CDK
 bits on = 13 15 26 27 33 37 42 47 53 54 55
 56 57 62 63 65 66 69 71 76 79 84 86 87 90
 103 117 119 123 147 151 154 155 157 169 174
 184 188 202 210 212 217 220 223 227 228 245
 252 257 260 266 272 275 282 290 296 303 311
 318 324 326 339 350 353 355 367 382 396 397
 402 404 419 422 439 446 447 451 452 454 465
 468 478 482 497 505 517 518 519 520 524 529
 530 535 542 547 561 565 582 587 597 606 607
 609 613 617 618 622 623 629 633 647 680 689
 697 699 705 711 713 715 718 729 742 750 753
 754 779 785 787 788 791 801 813 814 824 831
 833 841 851 852 858 864 878 885 886 892 897
 902 908 915 917 921 922 924 925 926 927 934
 937 943 950 953 964 973 977 980 981 987 991
 993 996 1010 1011 1012 1014 1015
```

Similarly we computed the molecular fingerprints for the entire set of molecules:

```
fps <- lapply(mols, get.fingerprint,
type=’extended’)
fps
```

Computation of the intermolecular distances by the Tanimoto index

The Tanimoto coefficient can be expressed as:

$$S_{A,B} = c/[a + b - c]$$

where S is the similarity, a is the number of on bits in molecule A, b is number of on bits in molecule B, and c is the number of on bits in both molecules [49].

Based on molecular fingerprints calculated using the Tanimoto method, the molecular distances between the methotrexate molecule and the other 23 molecules in the group can be evaluated: [24, 30].

```
query.mol<-load.molecules( c(’meth.sdf’) )
target.mols<-mols
fp.sim<-fp.sim.matrix(fps,method=’tanimoto’)
fp.dist <- 1 - fp.sim
fp.dist
```

Using this method, a complete set of distances, in matrix form, between each of the 23 molecules of interest was obtained. By analysing these results, it is possible to identify all the molecules located at a certain distance from the target molecule (0.5 in our example): [48].

```
query.fp<-get.fingerprint(CMP1[[1]]
type = ’maccs’)
target.mols <-mols
target.fps <- lapply(target.mols,
get.fingerprint, type=’maccs’)
target.fps
sims <- data.frame(sim=do.call
(rbind, lapply(target.fps,
   fingerprint::distance,
   fp2=query.fp, method=’tanimoto’)))
subset(sims, sim >= 0.5)
hits <- which(sims >= 0.5)
hits
[1] 0.3809524 0.4285714 0.5000000 0.3974359
    0.5128205 0.4473684
[7] 0.5121951 0.4430380
> hits <- which(sims > 0.5)
> hits
[1] 5 7
```

From the data presented above we can conclude that only molecules 5 and 7 meet our criteria. This method is the basis for fingerprint-based clustering.

Results and Discussion

In the present study, we used a group of 23 newly synthesized molecules. All of them share the following characteristics: they are pyrazole derivatives, that have never been synthesized, there is no data about them in the literature or in chemical databases, and they have the potential to be drug candidates, such as purine derivatives. Our intention was to check whether the studied chemical compounds can be considered as possible lead molecules [26]. Because the costs of clinical trials are high, even in the preclinical phase, pre-sorting these candidates by computational chemistry and cheminformatics methods would be beneficial [2, 45]. According to the similarity property principle (SPP), which says that drugs with similar molecular structures are likely to have the same properties, a new drug candidate can be identified upon its similarity with another known drug, regardless of how the similarity is evaluated [5]. As a screening criterion, we used the comparison with the traditional methotrexate molecule [32].

Methotrexate is a cytotoxic substance widely used in cancer therapy. It was one of the first purine-inhibiting antimetabolites on the market, and it interferes with the growth of different molecules present in human body, such as like highly reproductive cancer cells. Even though this molecule cannot be considered a "gold standard" for this compound class, the above arguments have contributed to this choice.

Clustering the dataset of molecules

Different types of methods can be used for clustering, including partitioning methods (K-means), hierarchical clustering, fuzzy clustering, density-based clustering and model-based clustering. The K-means and hierarchical clustering were chosen because they are suitable for our goal [29].

The number of clusters

The optimal number of clusters can be estimated using the NbClust package. The function fviz.nbclust is used for visualizing the result [1]. The R code for the elbow method is presented below:

```
fviz_nbclust(fp.dist,kmeans, method = "wss")
 +geom_vline(xintercept = 3, linetype = 2)
```

The result is depicted in Fig. 3:

The optimal number of clusters is 3.

All the considered molecules were then grouped into clusters by taking into account the calculated intermolecular distances.

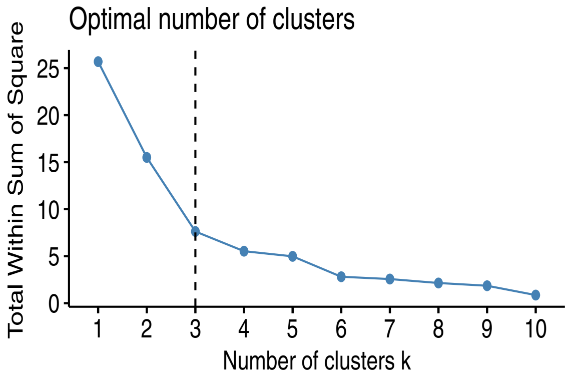

Fig. 3 Estimation of the optimal number of clusters

Hierarchical clustering with the hclust package

The hierarchical clustering algorithm creates clusters with sets of data that are similar internally but different from each other externally [30]. The most common and useful graphical representation of molecular clusters is hierarchical clustering (dendrogram). We performed hierarchical clusterization using Ward's method [30, 37].

Ward's method is based on an ANOVA approach and its goal is to maximize the r^2 value.

To obtain this dendogram, we used the following R code:

```
d <- dist(fp.dist, method = "euclidean")
res.hc <- hclust(d, method = "ward.D2" )
grp <- cutree(res.hc, k = 3)
plot(res.hc, cex = 0.6) # plot tree
rect.hclust(res.hc, k = 3, border = 2:5)
```

The graphical representation of the dendrogram is depicted in Fig. 4.

K-means Clustering

K-means clustering is one of the most commonly used clustering algorithms because it is easy to code and implement. Each cluster has a centre, which is called a centroid. The algorithm combines the distances between points and centroids [27]. The R code for K-means clustering is shown below.

```
> fviz_nbclust(fp.dist,
   method = "gap_stat")
> km.res <- kmeans
   (fp.dist, 3, nstart = 10)
> km.res
K-means clustering
with 3 clusters of sizes
 6, 7, 11
Clustering vector:
 [1] 1 1 3 3 1 1 1 1
      2 3 3 3 3 3 3 3
      3 2 2 2
      3 2 2 2
Within cluster sum of squares by cluster:
[1] 3.457616 1.402913 2.765955
 (between_SS / total_SS =  70.3 %)
```

The result is presented in Fig. 5:

The molecules included in cluster 1, containing methotrexate (CMP1), were visualized using the following R code:

```
sdfset <- read.SDFset("mysdf.sdf")
sdfset
cid(sdfset)[1:6]
plot(sdfset[1:6], print=FALSE)
 [1] "CMP1" "CMP2" "CMP3"
      "CMP4" "CMP5" "CMP6"
```

The result are depicted in Fig. 6:

Clustering validation

Statistics for K-means clustering

```
silinfo <- km.res$silinfo
names(silinfo)
km_stats <- cluster.stats
(fp.dist,  km.res$cluster)
km_stats
```

The R code for the cluster statistics is listed below:

The most important information for the cluster analysis provided by this function can be considered the silhouette index and Dunn index: [7, 34].

```
$dunn
[1] 0.5443968
$sindex
[1] 0.4565217
```

The Dunn index is equal to the ratio of the smallest inter-cluster distance divided by the largest intra-cluster distance [20].

It takes a value between zero and infinity, and a higher DI means that clusters are compact and well separated. A larger distance between clusters means a better separation, and smaller cluster sizes lead to a higher Dunn Index [1, 21].

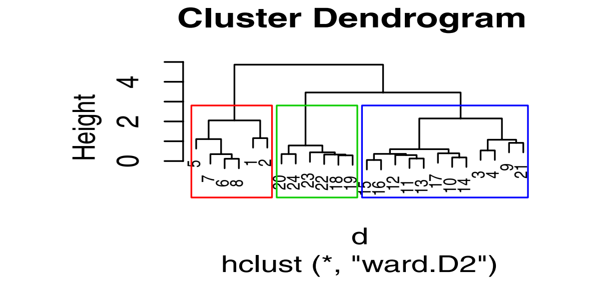

Fig. 4 Dendrogram—hierarchical clustering using Ward's method

The Silhouette coefficient is a method of cluster validation that combines both cohesion and separation [35]. It measures, for each point M_i, the mean distance to each cluster, and the mean distance to the other points in its cluster. Silhouette values range between −1 and 1 [4]. A Silhouette coefficient with a value near +1 indicates that the point is far from its neighbouring cluster and very close to the cluster to which it is assigned. These values are preferred.

The R code for visualizing a Silhouette plot for K-means clustering:

```
library("cluster")
sil <- silhouette(km.res$cluster,
dist(fp.dist))
head(sil[, 1:3], 10)
plot(sil, main ="Silhouette plot - K-means")
```

The plot is visualized in Fig. 7:

The Silhouette plot for K-means clustering reveals a coefficient of 0.4 for the first cluster and a mean value of 0.49 for all the other clusters. These values can be considered acceptable.

Conclusions

Cheminformatics is a dynamic and powerful field that is considered the heart of modern drug design [9, 15]. It plays an important role in collecting, storing and analysing chemical data [18]. It is also an emerging interdisciplinary field that aims to discover new chemical entities that ultimately result in the design of a new active molecule (chemical data) [14, 22].

This work focused on cheminformatics and its application in the discovery and testing of new active molecules.

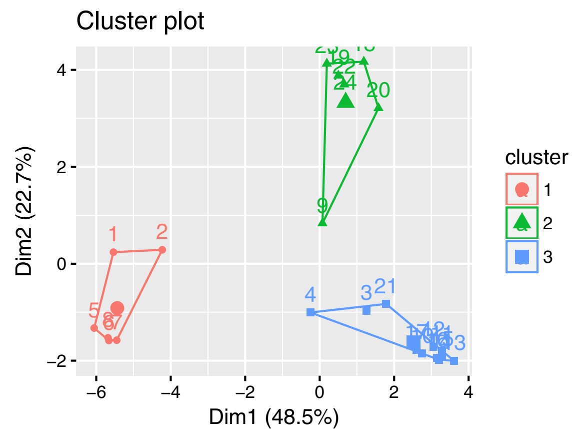

Fig. 5 Polygonal clusters

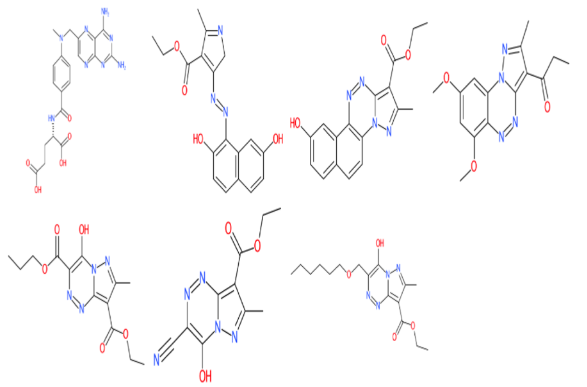

Fig. 6 Molecules similar to methotrexate

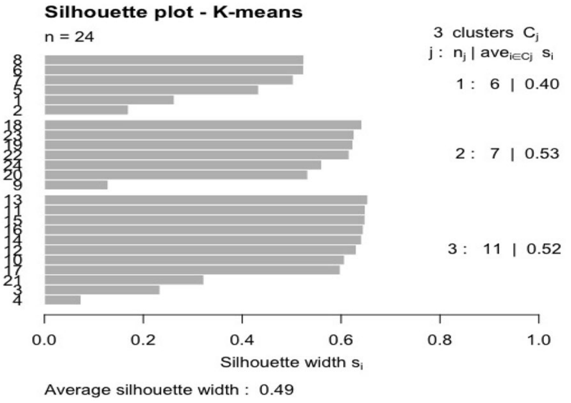

Fig. 7 Silhouette plot for K-means clustering

In addition we focused on modern data mining techniques that help chemists and medical researchers to discover, produce and test new active molecules for the treatment of certain diseases.

Our goal was to test in vitro a set of 23 newly synthesized molecules, about which we do not have enough experimental data. The lack of information about the physicochemical properties of the respective molecules, especially those related to QSAR, was supplemented by the computational methods offered by the *rcdk* software package [12, 25].

Because we studied 23 compounds from the pyrazole class, we expected that at least some of them would behave similar to the like cytotoxic antimetabolite class (purine inhibitors). The clusters were obtained using hierarchical and K-means clustering methods. The results of clustering were confirmed using the Dunn index and Silhouette coefficient.

To avoid the additional costs that pre-clinical and clinical trials for all these compounds would have involved, we tried to reduce the number of "candidate drugs" by computational methods. This reduction was accomplished by calculating the molecular fingerprints of all the studied molecules and then comparing them with the molecular marker methotrexate, which still has a wide use. As a result of this comparison and after the clusterization of the molecules according to the Tanimoto distances, an optimal number of 3 clusters was obtained. In the cluster containing the methotrexate molecule of, marked with a 1, we can also find the molecules marked with a 2, 5, 6 and 7. The remaining 17 molecules are part of the other two clusters [30]. Therefore, starting from the assumption that "similar chemical structures have similar biological properties and actions", the number of compounds worth considering for further studies has been significantly reduced, from 23 to 4, which will lead to a significant decrease in all future costs [9].

Acknowledgements
We would like to thank Dr. Sabina Nitu who kindly provide us the structures and additional information regarding the studied compounds.

Authors' contributions
AV—clustering and computing in R and Chem Axon—MS, MV, DV and VD—documenting and selection of possible active molecules AV, MV, ND—wrote the first draft of the manuscript. All authors contributed to the interpretation of experimental data, discussion and the preparation of the final manuscript. All authors read and approved the final manuscript.

Data and material availability
Data and materials are available on GitHub:
The R code used for this paper:
clusterch.R
https://github.com/voicuadr/RClusters/blob/master/clusterch.R
The cemical structures in sdf file format:
CMP.sdf
https://github.com/voicuadr/RClusters/blob/master/CMP.sdf

Competing interests
The authors declare that they have no competing interests.

Author details
[1] Department of Medical Informatics and Biostatistics, Victor Babes University of Medicine and Pharmacy, E. Murgu 2, 300041 Timisoara, Romania. [2] Dep. CAICAM, Politehnica University of Timisoara, Pirvan Boulevard 6, Timisoara, Romania. [3] Department of Pharmacology-Clinical Pharmacy, Victor Babes University of Medicine and Pharmacy, E. Murgu 2, 300041 Timisoara, Romania. [4] Department of Pharmacology, Victor Babes University of Medicine and Pharmacy, E. Murgu 2, 300041 Timisoara, Romania.

Received: 8 January 2019 Accepted: 20 December 2019
Published online: 20 January 2020

References

1. Arbelaitz O, Gurrutxaga I, Muguerza J, PéRez JM, Perona I (2013) An extensive comparative study of cluster validity indices. Pattern Recognit 46(1):243–256
2. Backman Tyler WH, Yiqun C, Thomas G (2011) Chemmine tools: an online service for analyzing and clustering small molecules. Nucleic Acids Res 39(suppl–2):W486–W491
3. Bajusz D, Rácz A, Héberger K (2015) Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? J Cheminform 7(1):20
4. Baridam BB (2012) More work on k-means clustering algorithm: the dimensionality problem. Int J Comput Appl 44(2):23–30

5. Begam BF, Kumar JS (2012) A study on cheminformatics and its applications on modern drug discovery. Procedia Eng 38:1264–1275
6. Beisken S, Meinl T, Wiswedel B, de Figueiredo LF, Berthold M, Steinbeck C (2013) Knime-cdk: workflow-driven cheminformatics. BMC Bioinform 14(1):257
7. Brock G, Pihur V, Datta S, Datta S et al. (2008) clValid, an R package for cluster validation. J Stat Softw 25(4):1–22
8. ChemAxon L (2013) Marvinsketch. https://chemaxon.com/products/marvin
9. DiMasi JA, Hansen RW, Grabowski HG (2003) The price of innovation: new estimates of drug development costs. J Health Econ 22(2):151–185
10. Georgiou KR, Scherer MA, Fan CM, Cool JC, King TJ, Foster BK, Xian CJ (2012) Methotrexate chemotherapy reduces osteogenesis but increases adipogenic potential in the bone marrow. J Cell Physiol 227(3):909–918
11. Godden JW, Stahura FL, Bajorath J (2005) Anatomy of fingerprint search calculations on structurally diverse sets of active compounds. J Chem Inform Model 45(6):1812–1819
12. Guha R, Cherto MR (2017) Integrating the CDK with R. Chemical informatics functionality in R, pp 1–17
13. Guha R et al (2007) Chemical informatics functionality in r. J Stat Softw 18(5):1–16
14. Guha R, Gilbert K, Fox G, Pierce M, Wild D, Yuan H (2010) Advances in cheminformatics methodologies and infrastructure to support the data mining of large, heterogeneous chemical datasets. Curr Comput Aided Drug Design 6(1):50–67
15. Hassan Baig M, Ahmad K, Roy S, Mohammad Ashraf J, Adil M, Haris Siddiqui M, Khan S, Amjad Kamal M, Provazník I, Choi I (2016) Computer aided drug design: success and limitations. Curr Pharma Design 22(5):572–581
16. Hughes JP, Rees S, Kalindjian SB, Philpott KL (2011) Principles of early drug discovery. Br J Pharmacol 162(6):1239–1249
17. Jacques Julien, Preda Cristian (2014) Functional data clustering: a survey. Adv Data Anal Classif 8(3):231–255
18. Karthikeyan M, Vyas R (2014) Machine learning methods in chemoinformatics for drug discovery. In: Karthikeyan M, Vyas R (eds) Practical chemoinformatics. Springer, New Delhi, pp 133–194
19. Kovács F, Legány C, Babos A (2005) Cluster validity measurement techniques. In: 6th International symposium of hungarian researchers on computational intelligence, p 35. Citeseer
20. Kryszczuk K, Hurley P (2010) Estimation of the number of clusters using multiple clustering validity indices. In: International workshop on multiple classifier systems. Springer, pp 114–123
21. Liu Y, Li Z, Xiong H, Gao X, Wu J (2010) Understanding of internal clustering validation measures. In: 2010 IEEE international conference on data mining. IEEE, pp 911–916
22. Macalino SJY, Gosu V, Hong S, Choi S (2015) Role of computer-aided drug design in modern drug discovery. Arch Pharm Res 38(9):1686–1701
23. MacCuish JD, MacCuish NE (2014) Chemoinformatics applications of cluster analysis. Wiley Interdiscip Rev Comput Mol Sci 4(1):34–48
24. Martin E, Cao E (2015) Euclidean chemical spaces from molecular fingerprints: hamming distance. J Comput Aided Mol Design 29(5):387–395
25. Mente S, Kuhn M (2012) The use of the r language for medicinal chemistry applications. Curr Topics Med Chem 12(18):1957–1964
26. Mioc M, Avram S, Tomescu AB, Chiriac DV, Heghes A, Voicu M, Voicu A, Citu C, Kurunczi L (2017) Docking study of 3-mercapto-1, 2, 4-triazole derivatives as inhibitors for vegfr and egfr. Rev Chim 68(3):500–503
27. Morissette L, Chartier S (2013) The k-means clustering technique: general considerations and implementation in mathematica. Tutor Quant Methods Psychol 9(1):15–24
28. Muchmore SW, Edmunds JJ, Stewart KD, Hajduk PJ (2010) Cheminformatic tools for medicinal chemists. J Med Chem 53(13):4830–4841
29. Murtagh F, Contreras P (2012) Algorithms for hierarchical clustering: an overview. Wiley Interdiscip Rev Data Min Knowl Discov 2(1):86–97
30. Murtagh F, Legendre P (2014) Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? J Classif 31(3):274–295
31. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR (2011) Open babel: an open chemical toolbox. J Cheminform 3(1):33
32. OBoyle NM (2012) Towards a universal smiles representation-a standard method to generate canonical smiles based on the inchi. J Cheminform 4(1):22
33. Prakash N, Gareja DA (2010) Cheminformatics. J Proteomics Bioinform 3:249–252
34. Rendón E, Abundez I, Arizmendi A, Quiroz EM (2011) Internal versus external cluster validation indexes. Int J Comput Commun 5(1):27–34
35. Rendón E, Abundez IM, Gutierrez C, Zagal SD, Arizmendi A, Quiroz EM, Arzate HE (2011) A comparison of internal and external cluster validation indexes. In: Proceedings of the 5th WSEAS international conference on computer engineering and applications, pp 158–163
36. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. J Chem Inform Model 50(5):742–754
37. Saeed F, Salim N, Abdo A (2012) Voting-based consensus clustering for combining multiple clusterings of chemical structures. J Cheminform 4(1):37
38. Sliwoski G, Kothiwale S, Meiler J, Lowe EW (2014) Computational methods in drug discovery. Pharmacol Rev 66(1):334–395
39. Szymański P, Markowicz M, Mikiciuk-Olasik E (2012) Adaptation of high-throughput screening in drug discovery–toxicological screening tests. Int J Mol Sci 13(1):427–452
40. Taft CA, Da Silva VB et al (2008) Current topics in computer-aided drug design. J Pharm Sci 97(3):1089–1098
41. Taguchi Y-H (2017) Identification of candidate drugs using tensor-decomposition-based unsupervised feature extraction in integrated analysis of gene expression between diseases and drugmatrix datasets. Sci Rep 7(1):13733
42. Vogt M, Stumpfe D, Geppert H, Bajorath J (2010) Scaffold hopping using two-dimensional fingerprints: true potential, black magic, or a hopeless endeavor? Guidelines for virtual screening. J Med Chem 53(15):5707–5715
43. Wagener M, van Geerestein VJ (2000) Potential drugs and nondrugs: prediction and identification of important structural features. J Chem Inf Comput Sci 40(2):280–292
44. Warr WA (2011) Representation of chemical structures. Wiley Interdiscip Rev Comput Mol Sci 1(4):557–579
45. Willett P (2009) Similarity methods in chemoinformatics. Annu Rev Inform Sci Technol 43:3–71
46. Willett Peter (2010) Similarity searching using 2d structural fingerprints. In: Chemoinformatics and computational chemical biology. Springer, pp 133–158
47. Willighagen EL, Mayfield JW, Alvarsson J, Berg A, Carlsson L, Jeliazkova N, Kuhn S, Pluskal T, Rojas-Chertó M, Spjuth O et al (2017) The chemistry development kit (cdk) v2. 0: atom typing, depiction, molecular formulas, and substructure searching. J Cheminform 9(1):33
48. Zhang B, Vogt M, Maggiora GM, Bajorath J (2015) Design of chemical space networks using a tanimoto similarity variant based upon maximum common substructures. J Comput Aided Mol design 29(10):937–950
49. Zhang C, Idelbayev Y, Roberts N, Tao Y, Nannapaneni Y, Duggan BM, Min J, Lin EC, Gerwick EC, Cottrell GW et al (2017) Small molecule accurate recognition technology (smart) to enhance natural products research. Sci Rep 7(1):14243

Publisher's Note