



BRIEF REPORTS

Open Access

RMol: a toolset for transforming SD/Molfile structure information into R objects

Martin Grabner¹, Kurt Varmuza² and Matthias Dehmer^{1*}

Abstract

Background: The graph-theoretical analysis of molecular networks has a long tradition in chemoinformatics. As demonstrated frequently, a well designed format to encode chemical structures and structure-related information of organic compounds is the Molfile format. But when it comes to use modern programming languages for statistical data analysis in Bio- and Chemoinformatics, R as one of the most powerful free languages lacks tools to process Molfile data collections and import molecular network data into R.

Results: We design an R object which allows a lossless information mapping of structural information from Molfiles into R objects. This provides the basis to use the `RMol` object as an anchor for connecting Molfile data collections with R libraries for analyzing graphs. Associated with the `RMol` objects, a set of R functions completes the toolset to organize, describe and manipulate the converted data sets. Further, we bypass R-typical limits for manipulating large data sets by storing R objects in `bz`-compressed serialized files instead of employing `RData` files.

Conclusions: By design, `RMol` is a R toolset without dependencies to other libraries or programming languages. It is useful to integrate into pipelines for serialized batch analysis by using network data and, therefore, helps to process `sdf`-data sets in R efficiently. It is freely available under the BSD licence. The script source can be downloaded from <http://sourceforge.net/p/rmol-toolset>.

Background

To solve many tasks in Bio- and Chemoinformatics, the analysis of chemical and biological structures represented by networks has been proven powerful [1,2]. A typical problem in this area is to characterize the structure of molecular networks quantitatively by using graph measures [3-7] or to predict physicochemical properties of the molecules by taking structural features into account [8].

For quantifying structural information of molecular networks, one often needs quantitative or comparative network measures to analyze the structure of the underlying networks [1,2]. For instance, `Dragon` [9] is a commercial and well-known software to calculate so-called molecular descriptors from SD/Molfile data [10] and other data formats specializing in chemical structures. But when using the programming environment R [11], there is yet no

interface to employ structural information of molecular networks encoded by SD/Molfile data.

To tackle this problem and, hence, to spread out the usage of R in chemically and biologically-driven disciplines, we develop an R toolset for transforming SD/Molfile structure information into R objects. As structural information of the networks is now available in R, we hope that our tool may stimulate the Bioinformatics community to explore problems centered around chemical and molecular networks by using existing R packages.

Tools for graph analysis

In this section, we briefly sketch some tools for analyzing graphs by using R. An extensive review of such tools can be found in [12,13].

Among other environments suitable for graph analysis [14], the script language R has gained much importance. Not only because basic functions are allocated by R packages such as the packages `graph` [15] from `Bioconductor` [16] and `igraph` [17] but also because of packages such as `QuACN` [18]. The latter contributed extensively to analyze networks quantitatively [18] with

*Correspondence: matthias.dehmer@umit.at

¹Department of Biomedical Sciences and Engineering, Institute for Bioinformatics and Translational Research, University for Health Sciences, Medical Informatics and Technology (UMIT), Eduard Wallnöfer Zentrum 1, A-6060, Hall in Tyrol, Austria

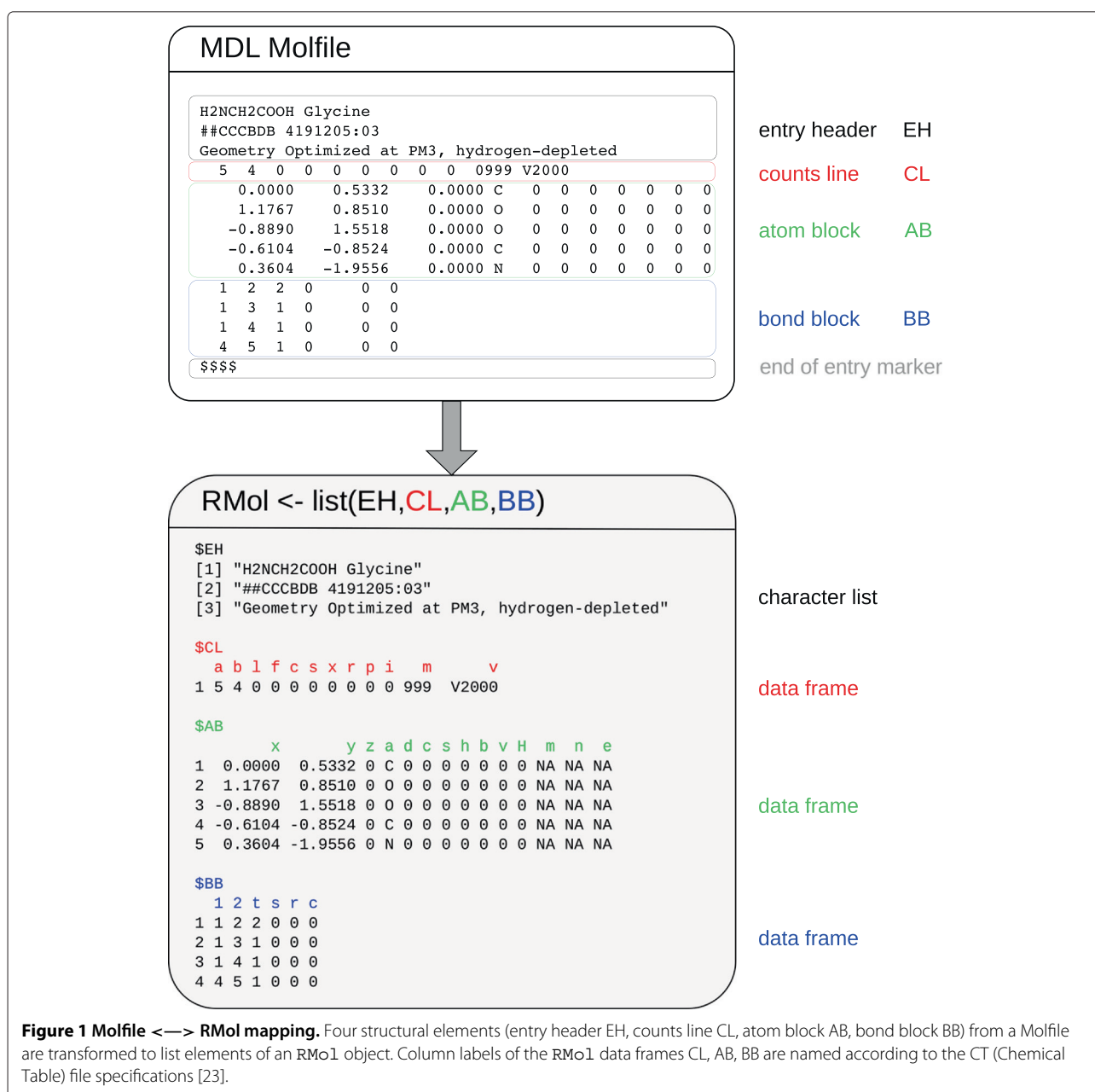
Full list of author information is available at the end of the article

R. Note that QuACN is an R-tool for calculating ca. 150 quantitative network measures which can be mostly interpreted as complexity indices [19].

Also, Guha [20] developed a set of wrapper functions providing R user access to the functions and objects of the CDK [21] representing a Java framework for cheminformatics. The cheminformatics package ChemMineR [22] written in R includes in updated versions functions which are capable of reading and extracting structural information from different data formats including SD-files. But we emphasize that both packages are conceptionally focused on the inspection of single

networks or comparison of limited data sets, in contrast to the RMol script collection which fits well into workflows with serialized pipelines. To reduce dependencies, we extract SD-file information with our own R parser. Also, to avoid repetitive transformation processes we store the chemical and molecular network information in R objects.

In this report, we present an R toolset for linking available R functions with existing data collections representing chemical structures by mapping structural information of MDL-Molfiles to an R object called RMol. Figure 1 illustrates the lossless mapping from Molfile information



to a list consisting of four elements that represents an `RMol` object.

Results and discussion

Besides the definition of the novel `RMol` object for encoding chemical structure information by using R, we develop a set of functions for the programming environment R to accomplish and facilitate efficient graph-based analysis of the underlying molecular networks, e.g., chemical structures. More precisely, this tool covers the following functionalities:

- Importing chemical structure data from an SD-file (Molfile format) into R.
- Handling of `RMol` data sets as serialized bz-compressed files (to bypass memory limits).
- Providing simple statistics of chemical structures or structure data sets in `RMol` format.
- A filter for selecting chemical structures and reorganizing data collections in `RMol` format.
- Generating adjacency matrices or connection tables from chemical structures in `RMol` format.
- Converting `RMol` objects into attribute-extended `graphNEL` objects. By doing so, this links directly to R packages for graph analysis (e.g., see `graph`, `igraph`, `QuACN`).

In the following, we explain some items of `RMol` in more detail. The function `Sdf2RMol` has been developed to process SD-files and convert chemical structure information from the Molfile portions into `RMol` objects.

Concretely `Sdf2RMol` represents a working script, which combines an entry picking routine (`pickSdfEntry`) with an `RMol` specific parser (`parseSdfEntry`) using regular expressions to scan the Molfile sections of SD-files according to the CT-file format specifications [23]. Moreover `Sdf2RMol` completes the conversion pipeline with error logging and internal routines for checking feeded entries for consistency and plausibility. Finally the resulting R objects are streamed as data sets into serialized bz-compressed files.

These files are denoted with the file ending `.Rbz` and referred to as "Rbz-files". Rbz-files help to bypass R-typical memory limits for huge data collections and are useful storage containers for any R object. By design, Rbz-files contain the R objects as serialized list elements `S[i]`, where `S[i]=list(objectname[i],objectcontent[i])`. They are a useful data source for any R driven process pipeline.

The functions `RData2Rbz` and `Rbz2RData` allow the transformation of the serial Rbz-format to standard `RData`-format and vice versa. For users who are not familiar with connection manipulation in R, we also include the functions `RbzOpen`, `NextRbzObject`, `RbzClose` to alleviate the handling of Rbz-files.

To extract and summarize properties of Rbz-packed `RMol` data sets, the functions `RbzSummary` and `RbzSummaryReport` are useful. To manipulate and split these data sets `RbzFilter` is available.

The raw `graphNEL` class, as defined in the R package `graph` is sufficient to build representations for graphs without vertex and edge labels. However, to perform the analysis of labeled graphs (e.g., graphs representing chemical structures with hetero atoms and bond types) by using the `QuACN` package, `graphNEL` needs to be extended with bond and atom attributes. `RMol` contains the function `RMol2QuACNgN` to pack the relevant information into these attribute-extended `graphNEL` objects. All `RMol` functions are put together in one R script. After sourcing this script all functions will be available to support in preparing chemical structure data for analyzing molecular networks.

Conclusions

In this report, we presented an R toolset to convert the structure information of molecular graphs encoded by SD/Molfiles into R objects. It complements existing packages capable of reading SD-file information by easing batch processing and using pure R scripts without dependencies. In combination with R packages designed for analysing graph and network properties it represents a connector module for R workflows which process structure information from SD-files. This toolset can also support other R packages for analyzing networks structurally and, thus, makes a further contribution towards demonstrating the power of R for network analysis in Chemo- and Bioinformatics.

So far, it was not common to investigate SD-file data collections by using R and packages thereof. The new toolset `RMol` may encourage the community to spread out the usage of R for chemically and biologically driven areas.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

MG developed the R package. KV also participated in this process. MD, MG and KV wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Funding

Martin Grabner, Kurt Varmuza and Matthias Dehmer thank the Austrian Science Funds for supporting this work (project P22029-N13).

Author details

¹Department of Biomedical Sciences and Engineering, Institute for Bioinformatics and Translational Research, University for Health Sciences, Medical Informatics and Technology (UMIT), Eduard Wallnöfer Zentrum 1, A-6060, Hall in Tyrol, Austria. ²Laboratory for Chemometrics, Institute of Chemical Engineering, Vienna University of Technology, Getreidemarkt 9/166, A-1060 Vienna, Austria.

Received: 27 September 2012 Accepted: 30 October 2012
Published: 14 November 2012

References

1. Emmert-Streib F, Dehmer M: **Networks for systems biology: conceptual connection of data and function.** *IET Syst Biol* 2011, **5**:185–207.
2. Gasteiger J, Engel T: *Chemoinformatics - A Textbook*. Weinheim, Germany: Wiley VCH; 2003.
3. Bonchev D: *Information Theoretic Indices for Characterization of Chemical Structures*. Chichester: Research Studies Press; 1983.
4. Dehmer M, Varmuza K, Borgert S, Emmert-Streib F: **On entropy-based molecular descriptors: statistical analysis of real and synthetic chemical structures.** *J Chem Inf Model* 2009, **49**:1655–1663.
5. Dehmer M, Barbarini N, Varmuza K, Graber A: **A large scale analysis of information-theoretic network complexity measures using chemical structures.** *PLoS ONE* 2009, **4**(12):e8057.
6. Dehmer M, Barbarini N, Varmuza K, Graber A: **Novel topological descriptors for analyzing biological networks.** *BMC Struct Biol* 2010, **10**:18.
7. Todeschini R, Consonni V, Mannhold R: *Handbook of Molecular Descriptors*. Weinheim, Germany: Wiley-VCH; 2002.
8. Devillers J, Balaban AT: *Topological Indices and Related Descriptors in QSAR and QSPR*. Amsterdam, The Netherlands: Gordon and Breach Science Publishers; 1999.
9. Todeschini R, Consonni V, Mauri A, Pavan M: *Dragon, software for calculation of molecular descriptors*. Milano, Italy: Talet srl; 2004.
10. Dalby A, Nourse JG, Hounshell WD, Gushurst AKI, Grier DL, Leland BA, Laufer J: **Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited.** *J Chem Inf Comput Sci* 1992, **32**(3):244–255.
11. R Development Core Team: *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2008. [<http://www.R-project.org>.] [ISBN 3-900051-07-0].
12. Meur NL, Gentleman R: **Analyzing biological data using R: methods for graphs and networks.** *Methods Mol Biol* 2012, **804**:343–373.
13. Müller L, Dehmer M, Emmert-Streib F: **Network-based Methods for Computational Diagnostics by Means of R.** In *Computational Medicine*. Edited by Trajanoski Z. Vienna, Austria: Springer; 2012:185–197.
14. Hagberg AA, Schult DA, Swart PJ: **Exploring network structure, dynamics, and function using NetworkX.** In *Proceedings of the 7th Python in Science Conference, SciPy2008*; 2008:11–15.
15. Gentleman R, Whalen E, Falcon S: *graph: A package to handle graph data structures*; 2010. [<http://CRAN.R-project.org/package=graph>]. [R package version 1.26.0].
16. Gentleman R, Carey V, Bates D, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini A, Sawitzki G, Smith C, Smyth G, Tierney L, Yang J, Zhang J: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biol* 2004, **5**(10):R80. [<http://genomebiology.com/2004/5/10/R80>].
17. Csardi G, Nepusz T: **The igraph software package for complex network research.** *InterJournal* 2006. **Complex Systems**: 1695, [<http://igraph.sf.net>].
18. Müller LAJ, Kugler KG, Dander A, Graber A, Dehmer M: **QuACN - An R package for analyzing complex biological networks quantitatively.** *Bioinformatics* 2011, **27**(1):140–141.
19. Dehmer M, Emmert-Streib F, Tsoy Y, Varmuza K: **Quantifying structural complexity of graphs: information measures in mathematical chemistry.** In *Quantum Frontiers of Atoms and Molecules*. Edited by Putz M: Nova Publishing; 2011:479–498.
20. Guha R: **Chemical informatics functionality in R.** *J Stat Software* 2007, **18**(6):1–16.
21. Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, Willighagen E: **The Chemistry Development Kit (CDK): an open-source Java Library for Chemo- and Bioinformatics.** *J Chem Inf Comput Sci* 2003, **43**(2):493–500. [<http://pubs.acs.org/doi/abs/10.1021/ci025584y>]. [PMID: 12653513].
22. Cao Y, Charisi A, Cheng CL, Jiang T, Girke T: **ChemmineR: a compound mining framework for R.** *Bioinformatics* 2008, **24**(15):1733–1734.
23. Symyx Solutions Inc: 2010, [<http://accelrys.com/products/informatics/cheminformatics/ctfile-formats/no-fee.php>]. [CT File Formats, Symyx Solutions, Inc].

doi:10.1186/1751-0473-7-12

Cite this article as: Grabner *et al.*: Rmol: a toolset for transforming SD/Molfile structure information into R objects. *Source Code for Biology and Medicine* 2012 **7**:12.

Submit your next manuscript to BioMed Central
and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit

