Data Article

Searchable database of frequent R-groups in medicinal chemistry and their preferred replacements

Kosuke Takeuchi, Ryo Kunimoto, Jürgen Bajorath*

Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Friedrich-Hirzebruch-Allee 6, Bonn D-53115, Germany

ARTICLE INFO

Article history:
Received 24 August 2021
Revised 29 September 2021
Accepted 4 October 2021
Available online 8 October 2021

Keywords:
Medicinal chemistry
Compound optimization
Analog series
R-groups
Network analysis
R-group replacement database

ABSTRACT

In compound optimization, analogue series (ASs) are generated by introducing different R-groups (substituents, functional groups) at specific substitution sites. Systematic investigations of R-groups in medicinal chemistry have so far been rare. We have carried out a large-scale computational analysis of R-groups on the basis of ASs covering currently available bioactive compounds (Takeuchi et al., 2021). With the aid of a network data structure, frequently used R-groups and preferred replacements were identified. On the basis of these data, R-group replacement hierarchies were derived and organized in a searchable database that is made freely available. This contribution complements our systematic analysis (Takeuchi et al., 2021) by specifying the data we have generated and detailing their open access deposition.

© 2021 The Author(s). Published by Elsevier Inc.
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

DOI of original article: 10.1016/j.ejmech.2021.113771
* Corresponding author.
E-mail address: bajorath@bit.uni-bonn.de (J. Bajorath).
### Specifications Table

| Subject | Medicinal chemistry, drug discovery |
|---------|-------------------------------------|
| Specific subject area | Computational analysis of active compounds, extraction of R-groups from analogue series, identification of frequently used R-groups and preferred replacements, design of a database system. |
| Type of data | Tables |
| How data were acquired | Qualifying active compounds were systematically curated from ChEMBL [2]. All R-group data were generated in-house. |
| Data format | Secondary data |
| Parameters for data collection | Tables in text files (consistently formatted, searchable) |
| Description of data collection | The network data structure for R-group replacements is made available as a CSV file. For the 500 most frequently used R-groups, systematically generated replacement sequences are provided in a searchable XML file. R-group structures are reported as SMILES strings [3]. |
| Data source location | Department of Life Science Informatics and Data Science, B-IT, University of Bonn, Friedrich-Hirzebruch-Allee 6, D-53115 Bonn, Germany. |
| Data accessibility | The data sets are freely available in a deposition on the Zenodo open access platform that is accessible via the following link: https://doi.org/10.5281/zenodo.4741973. In addition, the data sets are deposited on Mendeley Data and can be accessed via the following link: https://doi.org/10.17632/pz7kkhk2bv.1. |
| Related research article | K. Takeuchi, R. Kunimoto, J. Bajorath, Systematic Mapping of R-group Space Enables the Generation of an R-group Replacement System for Medicinal Chemistry, Eur. J. Med. Chem. 2021, 225, 113771. https://doi.org/10.1016/j.ejmech.2021.113771 [1]. |

### Value of the Data

- The collection of frequently used R-groups and associated replacement hierarchies provides a large knowledge base for compound optimization. R-groups of interest can be searched using the new database, the first of its kind, and suitable replacements across bioactive compounds can be identified, hence providing practical decision support for medicinal chemistry which analogues to consider next.
- Since the computational approach for R-group identification was generalized (see below), we have obtained the -to our knowledge- by far largest collection of formally defined R-groups available thus far, with more than 50,000 candidate groups.
- In addition to the R-group replacement database, we also provide the original network data structure from which frequent R-groups and their replacements were obtained. Thus, the R-group database and replacement system can be modified or extended according to user preferences.

### 1. Data Description

From 343,373 qualifying bioactive compounds obtained from ChEMBL [2], a total 17,254 ASs with at least three analogues (comprising 314,525 unique compounds) were extracted using an algorithmic variant [1] of the compound-core relationship (CCR) method [4]. From each AS, the core structure was isolated and substitution sites were indexed, yielding a total of 61,312 sites. For each substitution site, all R-groups from the corresponding AS were sampled, leading to
the identification of a total of 50,545 unique R-groups. For each site, all possible pairwise R-group replacements were enumerated. Then, replacements for all 61,312 substitution sites were combined.

On the basis of these data, a global R-group network was constructed in which nodes represented individual R-groups and edges pairwise replacements. Two network parameters were calculated including the 'node degree' (ND; defined as the number of replacements/edges per node) and the 'edge weight' (EW; defined as the frequency with which each unique replacement occurred over all substitution sites). Most frequently used R-groups were identified on the basis of ND values and their preferred replacements on the basis of EW values.

The initially obtained global network structure was analyzed and many replacements of frequent R-groups were found to be dominated by the H atom, methyl group, and phenyl ring. Among these, the H atom was a special case because it was always present at a substitution site before a replacement with another R-group occurred and involved in a replacement when another group was introduced. To avoid domination of preferred replacements by these three groups, we further refined the network structure by omitting them, yielding the final network containing a total of 50,311 unique R-groups and 2,937,991 pairwise replacements.

The final network was used to identify preferred replacements of frequent R-groups having largest EW values. Therefore, we determined ‘first layer’ and ‘second layer’ R-group replacements. First layer replacements were formed by immediate network neighbors of a prioritized R-group and second layer replacements by nearest neighbors of the first layer groups. We then systematically generated ‘5 × 2’ first and second layer replacements. Hence, for a selected R-group, this data structure combined the top five most frequent first layer replacements with the top two second layer replacements (if available).

Fig. 1 shows an exemplary 5 × 2 replacement hierarchy. For a given R-group, this data structure formally defines 10 first and second layer replacement sequences.

The 500 most frequently used R-groups were determined and for each of these groups, the 5 × 2 replacement hierarchy was generated.

On the basis of the network data, the composition of replacement hierarchies can be modified at will. Since the dominant methyl group and phenyl ring were omitted from the final network (in addition to the H atom), these R-groups can be considered as additional generic replacements in all 5 × 2 hierarchies.

The 50,311 unique R-groups and 2,937,991 pairwise replacements forming the final network are provided as a CSV file containing three columns (source, target, and weight) defining weighted edges and participating nodes. In addition, the database containing the 500 most frequently used R-groups from the final network and the 5 × 2 replacement hierarchies is made available as a searchable XML file. Both files are contained in our open access depositions specified above.
2. Experimental Design, Materials and Methods

The analysis involved the following steps: (1) Compound selection; (2) Systematic identification of ASs; (3) Core structure isolation and indexing of substitution sites; (4) Site-specific sampling of R-groups; (5) Generation of an R-group network data structure revealing frequent R-groups and replacements; and (6) Design of R-group replacement hierarchies. All compounds and ASs were recorded and processed as SMILES strings [3]. The CCR method [4] systematically decomposes compounds and generates ASs with indexed substitution sites. For our analysis, the approach was generalized such that decomposition based on pre-defined chemical rules was replaced by systematic fragmentation of exocyclic single bonds, thereby substantially increasing fragment space [1]. Steps (4) and (6) were carried out with in-house scripts. The network data structure was generated and analyzed using Cytoscape [5].
Ethics Statement

This is a secondary data set and thus did not involve any human or animal testing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have, or could be perceived to have, influenced the work reported in this article.

CRediT Author Statement

Kosuke Takeuchi: Data curation, Formal analysis, Writing – review & editing; Ryo Kunimoto: Conceptualization, Formal analysis, Writing – review & editing; Jürgen Bajorath: Conceptualization, Supervision, Formal analysis, Writing – original draft, Writing – review & editing.

Acknowledgments

We thank Huabin Hu for help with the open access data deposition.

References

[1] K. Takeuchi, R. Kunimoto, J. Bajorath, Systematic mapping of R-group space enables the generation of an R-group replacement system for medicinal chemistry, Eur. J. Med. Chem. 225 (2021) 113771.
[2] A. Gaulton, A. Hersey, M. Nowotka, A.P. Bento, J. Chambers, D. Mendez, P. Mutowó, F. Atkinson, L.J. Bellis, E. Cibirán-Uhalte, M. Davies, N. Dedman, A. Karlsson, M.P. Magariños, J.P. Overington, G. Papadatos, I. Smit, A.R. Leach, The ChEMBL database in 2017, Nucleic Acids Res. 45 (2017) D945–D954.
[3] D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, J. Chem. Inf. Comput. Sci. 28 (1988) 31–36.
[4] J.J. Naveja, M. Vogt, D. Stumpfe, J.L. Medina-Franco, J. Bajorath, Systematic extraction of analogue series from large compound collections using a new computational compound–core relationship method, ACS Omega 4 (2019) 1027–1032.
[5] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: a software environment for integrated models of biomolecular interaction networks, Genome Res. 13 (2003) 2498–2504.