Systematic Extraction of Analogue Series from Large Compound Collections Using a New Computational Compound–Core Relationship Method

J. Jesús Naveja,†‡§∥ Martin Vogt,†‡∥ Dagmar Stumpfe, † José L. Medina-Franco,§ and Jürgen Bajorath*†

†Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Endenicher Allee 19c, D-53115 Bonn, Germany
‡PECEM, Faculty of Medicine and †Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Mexico City 04510, Mexico

ABSTRACT: Chemical optimization of organic compounds produces a series of analogues. In addition to considering an analogue series (AS) or multiple series on a case-by-case basis, which is often done in the practice of chemistry, the extraction of analogues from compound repositories is of high interest in organic and medicinal chemistry. In organic chemistry, ASs are a source of alternative synthetic routes and also aid in exploring relationships between compounds from different sources including synthetic vs. naturally occurring molecules. In medicinal chemistry, ASs are the major source of structure–activity relationship information and of hits or leads for drug development. ASs might be identified in different ways. For a given reference compound, a substructure search can be carried out using its scaffold. Alternatively, matched molecular pairs can be calculated to retrieve analogues from a compound repository. However, if no query compounds are used, the identification of ASs in databases is a difficult task. Herein, we introduce a computational approach to systematically identify ASs in collections of organic compounds. The approach involves compound decomposition on the basis of well-established retrosynthetic rules, organization of compound–core relationships, and identification of analogues sharing the same core. The method was applied on a large scale to extract ASs from the ChEMBL database, yielding more than 30,000 distinct series.

1. INTRODUCTION

In medicinal chemistry, hit-to-lead and lead optimization campaigns produce a series of analogues. An analogue series (AS) is generally defined as a series of compounds that share the same core structure and carry different R-groups at single or multiple substitution sites.1 ASs are conventionally represented in R-group tables and are the major source of structure–activity relationship (SAR) information.1,2–4 They are usually investigated as individual series in the course of chemical optimization. Computational methods have been introduced to organize large ASs and monitor SAR progression.2–5

Going beyond the analysis of individual ASs, another important task is searching for analogues in compound libraries and databases. If one is interested in identifying analogues of given reference compound(s), substructure search approaches can be applied using the core structure of a reference compound as a query.1,5 For example, this might be attempted in hit expansion when searching for analogues of an interesting active compound. Furthermore, analogues of reference compounds can also be identified without a predefined core structure by searching for matched molecular pairs (MMPs).7 An MMP is defined as a pair of compounds that are only distinguished by a structural modification at a single site.8 This modification can be rationalized as the exchange of a pair of substructures or a chemical transformation.9 To detect analogue relationships via MMPs, chemical transformations are restricted in size to focus on typical R-group replacements.7,10 MMPs can be efficiently generated algorithmically,9 making MMP-based analogue searching generally applicable7 and an attractive alternative to substructure search methods.

A much more difficult task than query-based analogue searching is the identification of ASs in large compound data sets, without prior knowledge. However, this task is highly relevant for knowledge extraction from compounds and activity data. In medicinal chemistry, one would like to identify and extract ASs of any composition from heterogeneous compound sources to maximize SAR information retrieval and provide templates for compound optimization. However, to the best of our knowledge, only one computational method for the systematic identification of ASs has so far been
introduced. This approach is also based upon the MMP formalism. For a given data set, all possible MMPs are generated and organized in an MMP-based network in which nodes represent compounds and edges pairwise MMP relationships. In this network, separate MMP clusters (MMPCs) are formed by individual ASs that can hence be easily identified. Accordingly, this approach is termed herein an MMP cluster (MMPC)-based method. In the simplest case, an AS from a cluster is formed by a matching molecular series (MMS) having a single substitution site. However, separate clusters in the MMP-based network can also be formed by multiple and overlapping MMSs representing ASs with multiple substitution sites. In this case, each participating MMS contributes a unique single site.

Herein, we introduce another computational methodology for the systematic identification of ASs in repositories of organic compounds, which does not rely on the MMP formalism. Rather, it is based upon the decomposition of single compounds according to well-established retrosynthetic rules and subsequent organization of compound–core relationships (CCRs). In a large-scale application, this new compound–core relationship (CCR) method was applied to systematically extract ASs from the ChEMBL database and compared with the MMPC approach.

2. RESULTS AND DISCUSSION

Extracting ASs from compound repositories without prior knowledge or query compounds is a difficult task. The CCR method introduced herein for systematically identifying ASs in databases of any composition is conceptually simple and generally applicable. The method comprises three sequential steps, including the generation of cores, exploration of compound–core relationships, and identification of analogue series, which are discussed in the following sections.

2.1. Methodological Concept. 2.1.1. Generation of Cores. The primary goal of the method is the identification of core structures and corresponding analogues such that the compounds can be readily reconstructed from the cores by substitutions at one or more sites and organized into ASs. The basis for reconstruction is provided by systematically applying combinations of possible bond deletions in each compound using retrosynthetic rules. Specifically, for each database compound, all possible combinations of one to five (or any other predefined number of) bonds are systematically subjected to retrosynthetic cleavage. Hence, a maximum number of five substitution sites per AS are covered. Each combination of applicable retrosynthetic rules leading to the corresponding elimination of single or multiple bonds yields a potential core. The core is considered valid if it consists of a single substructure containing an individual end point (substitution site) for each cleaved bond. Figure 1 illustrates the generation of cores for two analogues having two retrosynthetic cleavage sites. In addition to the three cores obtained from each analogue through retrosynthetic modification, each original compound is recorded as a core with no cleavage sites. Substitution sites in cores are recorded. Furthermore, it is required that the core and eliminated fragments (substituents) meet a predefined size
ratio. In our proof-of-concept study presented herein, we applied the rule that the core must contain at least two-thirds of the heavy atoms comprising the original compound. In other words, the ratio of the number of heavy atoms in the core to the sum of the total number of heavy atoms in all substituents must be at least 2:1. If these requirements are met, a core is accepted for further analysis. For a given database, all possible cores are generated and then "generalized". During the generalization step, all substitution sites are disregarded by introducing hydrogen atom substitutions at each site such that different cores become identical if they only differ in the position of substitution sites. In Figure 1, two identical cores resulting from generalization are highlighted.

2.1.2. Exploration of Compound–Core Relationships. Original database compounds that are identical to hydrogen-substituted cores are assigned to the corresponding cores as the smallest possible analogues. Generalization of cores is followed by reconstruction of recorded substitution sites and the assignment of additional database compounds to cores that differ at given substitution sites. The generalization and reconstruction steps ensure that compounds with all possible substitutions are assigned to corresponding cores, for example, analogues with ortho-, meta-, and/or para-substitution at one or more rings. We note that this cannot be accomplished on the basis of MMPs. The assignment of compounds to cores with reconstructed substitution sites yields all possible compound–core relationships in an organized form. Figure 1 illustrates the reconstruction of a single core with two substitution sites representing two exemplary analogues.

2.1.3. Identification of Analogue Series. An AS is formed if at least two compounds are associated with a core. Because all possible cores meeting the acceptance criteria are involved in CCRs, analogues forming an AS are often associated with multiple cores. ASs might consist of distinct sets of analogues, i.e., analogues belonging to one and only one AS, or overlapping sets of compounds. In addition, an AS might be fully contained as a subset in another series. The latter case is disambiguated by removal of ASs forming a subset of another. In addition, if two ASs contain exactly the same analogues, the one associated with the largest core is retained.

Figure 2. Compound–core relationships and identification of analogue series. (a) AS associated with three retrosynthetic cores. The core at the top represents all analogues (depicted on a purple background), whereas the two remaining cores represent two analogues each (encircled in green and red, respectively). (b) Two overlapping ASs are shown, each of which is associated with an individual core. The core at the top represents four analogues (depicted on a purple background) and the core at the bottom three (encircled in red). One of the analogues is shared by both series.
Other possible cases must also be taken into consideration. Figure 2a illustrates the frequently observed situation that an AS is associated with multiple cores. One of the cores might represent the entire AS and others subsets of analogues comprising the series. In this case, the core associated with the entire AS is retained to represent the series.

Figure 2b shows an example of overlapping ASs having different cores. The series share one analogue that is associated with both cores. This example also illustrates the rationale for consistently applying core/substituents size ratio restrictions. We note that the smaller core at the bottom in Figure 2b is a substructure of the larger one at the top. Due to the applied 2:1 size ratio restriction, the three analogues at the top are not presented by the small core at the bottom. This provides a basis for separating the series into two smaller ASs. The confined set of six analogues in this example could have been easily combined into a single AS by assigning two cores to the series. However, application of the size ratio restriction as a criterion for separating overlapping series generally avoids the situation that increasingly large compounds associated with cores that are substructures of each other form elongated “pseudo-AS” that might be artificial in nature and not meaningful chemically. Albeit rarely observed (see Section 2.2), this possible complication should strictly be avoided to ensure chemical relevance of computed ASs. Therefore, in overlapping ASs, each analogue is assigned to the largest AS it belongs to and removed from others. If the number of compounds in alternative ASs is the same, the AS associated with the larger core is selected. Furthermore, if the cores have an identical size, preference is given to the one with fewer substitution sites. Application of these criteria ensures that nearly all overlapping series are disambiguated, as further discussed below. The protocol outlined above guarantees that each AS is ultimately associated with a single core and each compound is associated with no more than one AS. Distinguishing between different CCRs is also of practical relevance. The consistent association of analogues and cores on the basis of size ratio restrictions and the selection of largest possible cores ensures that newly identified ASs are well-defined and can be easily represented in standard R-group tables, as illustrated in Figure 3. Hence, such ASs are readily available for follow-up analysis in medicinal chemistry.

2.2. Evaluation. 2.2.1. Large-Scale Search Application. In a proof-of-concept application, the CCR method was applied to systematically search for ASs in 244,704 active compounds from the ChEMBL database (for details, see the Materials and Methods section). A total of 30,431 ASs containing 145,269 compounds were identified, 8,359 of which contained cores with multiple substitution sites. Table 1 reports the size distribution of these ASs, 90% containing between two and nine analogues, 7.5% containing between 10 and 19, and 2.5% containing more than 19 analogues. Furthermore, with increasing size, the proportion of ASs with multiple substitution sites and the average number of substitution sites per AS also increased. For example, the 768 ASs containing at least 20 analogues included 380 series with multiple substitution sites and had on average close to two substitution sites per AS (with a maximum of sites).

Importantly, 18,606 (61%) of the identified ASs containing 64,323 compounds were nonoverlapping and associated with a single core representing the entire series, corresponding to the example shown in Figure 2a. Furthermore, 11,825 ASs (39%) containing 80,946 compounds were obtained from a set of 24 202 initially overlapping AS, as illustrated in Figure 2b. Most of the overlapping ASs were separated into well-defined series by uniquely assigning each compound to a single core. Disambiguation (as detailed above) was not possible for a very small subset of 96 overlapping ASs, 82 of which contained less than five compounds.

Thus, taken together, the results of systematic search calculations using the CCR method revealed that the majority of newly identified ASs was distinct from others. In cases where series overlap was detected, separation into nonoverlapping ASs was mostly unambiguous. Pseudo-ASs were not detected.

2.2.2. Method Comparison. For comparison, search calculations on the basis of 244,704 ChEMBL compounds were repeated using the MMPC approach,11 the only other computational methodology available to date for systematically identifying ASs. The results are reported in Table 2. MMPC calculations identified 22,111 ASs that covered a total of 103,154 ChEMBL compounds. These series included 3,509 ASs (15.9%) with multiple substitution sites. In contrast, the CCR search calculations detected 30,431 ASs that covered a total of 145,326 compounds and included 8,359 ASs (27.5%) with multiple substitution sites. Most of the ASs obtained by MMPC were also detected using the CCR method, with some variation in the composition of individual (especially larger) series. Moreover, nearly all analogues (97%) obtained by MMPC were identified using the CCR approach, which yielded 45,508 additional analogues. MMPC calculations yielded 2,191 ASs comprising 10 or more analogues. Of these series, 1986 ASs (91%) having more than 50% compound overlap were also identified by CCR including 1406 ASs with at least 80% compound overlap and 730 identical ASs. The...
overlap was calculated as the Jaccard index, i.e., the ratio of the number of shared analogues to the total number of unique analogues in a pair of corresponding series. CCR calculations identified a total of 3040 ASs with 10 or more analogues including 1352 ASs that were not detected using MMPC.

The MMPC/CCR comparison showed that the CCR method identified a significantly larger number of ASs, with a larger proportion of series having multiple substitution sites, and achieved a larger global compound coverage.

2.4. Conclusions. The identification of ASs in compound repositories without prior knowledge is of considerable relevance for the practice of organic and medicinal chemistry. ASs and the associated activity information can be used to rationalize and/or guide chemical synthesis and optimization efforts. However, only little has been done so far to automatically identify and extract ASs from databases, leaving much room for further developments. Herein, we have introduced a new computational approach to systematically search for ASs. The CCR method relies on the decomposition of single compounds on the basis of retrosynthetic rules, systematic generation of cores and compound-core relationships, and identification of ASs on the basis of organized and prioritized relationships. By design, the methodology is conceptually simple yet generally applicable. As such, it is thought to represent an attractive addition to the current repertoire of computational methods with utility for organic and medicinal chemistry. In our proof-of-concept investigation, a systematic search for ASs in ChEMBL identified a large number of ASs. The majority of ASs were nonoverlapping and distinct from others and associated with an individual core representing the entire AS. Such series should be of considerable interest for further SAR analysis and the identification of target-selective or promiscuous compounds. In summary, the CCR method introduced herein represents a new and general approach for systematically identifying ASs. It should be of interest to computational as well as organic and medicinal chemists including investigators aiming to explore relationships between compounds from different sources such as natural products and synthetic compounds. Such analyses will provide interesting topics for future application-oriented research.

3. MATERIALS AND METHODS

3.1. Retrosynthetic Rules. As retrosynthetic rules for compound decomposition, a well-established set of 13 retrosynthetic combinatorial analysis procedure (RECAP) rules was applied.14 We emphasize that the CCR methodology does not depend on a given set of rules. Depending on individual preferences or project requirements, any chosen set of reaction/retrosynthetic rules can be used. This is particularly relevant for applications in organic chemistry when new synthesis schemes are explored and compared with others.

3.2. Core Generation Details. The systematic generation of cores is among the three central components of the CCR method. Further details are provided. Bonds in compounds are cleaved according to RECAP rules and respective substituents are removed. If multiple RECAP rules are applicable to a given compound, all possible combinations are explored to generate cores. For example, if three rules A, B, and C apply, seven cores are obtained, including three with single cleavage sites (A, B, and C), three with dual sites (A/B, A/C, and B/C), and one with three cleavage sites (A/B/C). However, cores are only accepted to establish compound—core relationships if the ratio of the number of heavy atoms forming the core to the number of heavy atoms of all eliminated substituents is at least 2:1. The number of bonds in a compound to which RECAP rules applied was limited (and rarely larger than 20). Consequently, the exhaustive exploration of all possible combinations and resulting cores did not pose a combinatorial problem in most cases. In addition, the 2:1 size ratio restriction further reduced the number of cores for analyzing compound—core relationships. Nonetheless, a computational time restriction of 100 s per compound was implemented for core generation. However, due to this constraint, only 629 of 244 704 ChEMBL compounds failed to produce cores. The protocol for compound decomposition according to retrosynthetic rules was implemented in Java with the aid of the OEChem toolkit.15

3.3. Implementation of the CCR Algorithm. The CCR algorithm for systematically identifying ASs, as detailed in the Results and Discussion section, was implemented in Python.

3.4. Searching for Analogue Series. Systematic search calculations using the CCR and MMPC reference methods were carried out in a curated version of ChEMBL release 23.13 Only compounds with direct interactions (target relationship type “D”) with human targets at the highest confidence level (target confidence score 9) and available Ki or IC50 values were selected, yielding a total of 244 704 active compounds. The application of these selection criteria was not essential for the analysis but ensured that detected ASs exclusively consisted of compounds for which meaningful activity data were available.

AUTHOR INFORMATION

Corresponding Author

*E-mail: bajorath@bit.uni-bonn.de. Phone: 49-228-7369-100.
Author Contributions

J.J.N. and M.V. contributed equally to this work.

Author Contributions

The study was carried out and the manuscript was written with contributions of all authors. All authors have approved the final version of the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

J.J.N. is grateful to Consejo Nacional de Tecnología, Mexico (CONACyT) for a scholarship (grant no. 622969) and to the German Academic Exchange Service (DAAD) for a short-term research grant (program no. 53378443). The authors thank OpenEye Scientific Software for an academic software license.

REFERENCES

(1) The Practice of Medicinal Chemistry; 3rd ed.; Wermuth, C. G., Ed.; Academic Press-Elsevier: Burlington, San Diego, London, U.K., 2008.
(2) Agrafiotis, D. K.; Shemanarev, K.; Connolly, P. J.; Farnum, M.; Lobanov, V. S. SAR Maps: A New SAR Visualization Technique for Medicinal Chemists. J. Med. Chem. 2007, 50, 5926−5937.
(3) Zhang, B.; Hu, Y.; Bajorath, J. AnalogExplorer: A New Method for Graphical Analysis of Analog Series and Associated Structure-Activity Relationship Information. J. Med. Chem. 2014, 57, 9184−9194.
(4) Maynard, A. T.; Roberts, C. D. Quantifying, Visualizing, and Monitoring Lead Optimization. J. Med. Chem. 2016, 59, 4189−4201.
(5) Shanmugasundaram, V.; Zhang, L.; Kayastha, S.; de la Vega de León, A.; Dimova, D.; Bajorath, J. Monitoring the Progression of Structure-Activity Relationship Information During Lead Optimization. J. Med. Chem. 2016, 59, 4235−4244.
(6) Barnard, J. M. Substructure Searching Methods: Old and New. J. Chem. Inf. Comput. Sci. 1993, 33, 532−538.
(7) Dimova, D.; Stumpfe, D.; Bajorath, J. Systematic Assessment of Analog Relationships between Bioactive Compounds and Promiscuity of Analog Sets. Med. Chem. Commun. 2016, 7, 230−236.
(8) Griffen, E.; Leach, A. G.; Robb, G. R.; Warner, D. J. Matched Molecular Pairs as a Medicinal Chemistry Tool. J. Med. Chem. 2011, 54, 7739−7750.
(9) Hussain, J.; Rea, C. Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets. J. Chem. Inf. Model. 2010, 50, 339−348.
(10) Hu, X.; Hu, Y.; Vogt, M.; Stumpfe, D.; Bajorath, J. MMP-cliffs: Systematic Identification of Activity Cliffs on the Basis of Matched Molecular Pairs. J. Chem. Inf. Model. 2012, 52, 1138−1145.
(11) Stumpfe, D.; Dimova, D.; Bajorath, J. Computational Method for Systematic Identification of Analog Series and Key Compounds Representing Series and Their Biological Activity Profiles. J. Med. Chem. 2016, 59, 7667−7676.
(12) Wawer, M.; Bajorath, J. Local Structural Changes, Global Data Views: Graphical Substructure-Activity Relationship Trailing. J. Med. Chem. 2011, 54, 2944−2951.
(13) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. Nucleic Acids Res. 2012, 40, D1100−D1107.
(14) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP – Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. J. Chem. Inf. Comput. Sci. 1998, 38, 511−522.