J Comput Aided Mol Des (2013) 27:665–674
DOI 10.1007/s10822-013-9671-5

Systematic mining of analog series with related core structures in multi-target activity space

Disha Gupta-Ostermann · Ye Hu · Jürgen Bajorath

Received: 18 May 2013 / Accepted: 5 August 2013 / Published online: 24 August 2013

Abstract We have aimed to systematically extract analog series with related core structures from multi-target activity space to explore target promiscuity of closely related analogous. Therefore, a previously introduced SAR matrix structure was adapted and further extended for large-scale data mining. These matrices organize analog series with related yet distinct core structures in a consistent manner. High-confidence compound activity data yielded more than 2,300 non-redundant matrices capturing 5,821 analog series that included 4,288 series with multi-target and 735 series with multi-family activities. Many matrices captured more than three analog series with activity against more than five targets. The matrices revealed a variety of promiscuity patterns. Compound series matrices also contain virtual compounds, which provide suggestions for compound design focusing on desired activity profiles.

Keywords Analog series · Core structures · Structural relationships · Compound activity data · Matched molecular pairs · Matching molecular series · Compound series matrix · Biological targets

Introduction

In medicinal chemistry, analog series are usually organized in tables that list R-groups at different sites of a molecular core shared by a given compound series and report corresponding activity values [1]. These tables represent a standard format for SAR analysis. Derivatives of SAR tables have been introduced that organize compound series on the basis of R-group decomposition and display activity data of analogous in a heat map format [2] or in network representations [3]. In addition, approaches that utilize maximum common substructures (MCSs) [4] or scaffold-based compound organization schemes [5–7] are also widely used to represent SAR data. Going beyond the traditional medicinal chemistry focus on single series, methods for the extraction of SAR information from large and heterogeneous compound data sets have been developed in recent years [8]. In this context, SAR matrices have been introduced [9], which capture two or more compound series and display their potency distribution in a colour-coded matrix format. SAR matrices were originally designed to display potency patterns in compound series active against a given target [9]. The SAR matrix data structure utilizes the matched molecular pair formalism [10, 11] for the organization of analog series with related core structures.

There is increasing evidence that many bioactive compounds and drugs specifically interact with multiple targets [12, 13], which extends the traditional single-target focus of medicinal chemistry. Compound promiscuity is intensely studied in pharmaceutical research [14] because it provided the molecular basis of polypharmacology [15]. In recent studies, many compounds with multi-target activities have been identified through data mining [14]. We have been interested in extracting analog series with multi-target activities from currently available active compounds and capturing promiscuity patterns associated with closely related analog series. For systematic mining of such series and the graphical analysis of promiscuity patterns, the SAR matrix data structure [9] has been adapted and further extended.

D. Gupta-Ostermann · Y. Hu · J. Bajorath (✉)
Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstr. 2, 53113 Bonn, Germany
e-mail: bajorath@bit.uni-bonn.de

Materials and methods

Matched molecular pairs

Matched molecular pairs are defined as pairs of compounds that are distinguished by the exchange of a single fragment (substructure) at a specific site [10] *referred to as a chemical transformation* [11]. Distinguishing fragments include R-groups or ring systems and can vary in size. MMPs were systematically calculated using an in-house implementation of the algorithm by Hussain and Rea [11]. All molecules were initially subjected to fragmentation at exocyclic single bonds, so-called single-cuts [11]. The resulting two fragments were stored in an index table as key-value pairs. The larger fragment constituted the key and the corresponding smaller fragment constituted the value. If the two fragments had the same size, each fragment was stored once as a key and the corresponding fragment as a value. If a newly generated key was already contained in the index table, the corresponding value was added to the existing key. For a small compound set shown in Fig. 1a, the generation of the MMP index table is illustrated in Fig. 1b.

Structurally analogous matching molecular series

Matching molecular series (MMS) have previously been introduced as an extension of the MMP concept [16]. *An MMS comprises a series of compounds that share the same core (key) and differ by defined chemical substitutions (values).* In addition, *structurally analogous MMS are defined here as two or more MMS whose core structures (keys) differ at a single site* [16]. For the systematic identification of such structurally related MMS, the keys in the index table were subjected to a second round of fragmentation including all exocyclic single bonds and all combinations of two and three single bonds (so-called dual- and triple-cuts, respectively). Dual-cuts yield three and triple-cuts four fragments. Four-fragment combinations were only retained if they consisted of three fragments with single attachment points and a fragment with three attachment points. Figure 1c illustrates the generation of analogous

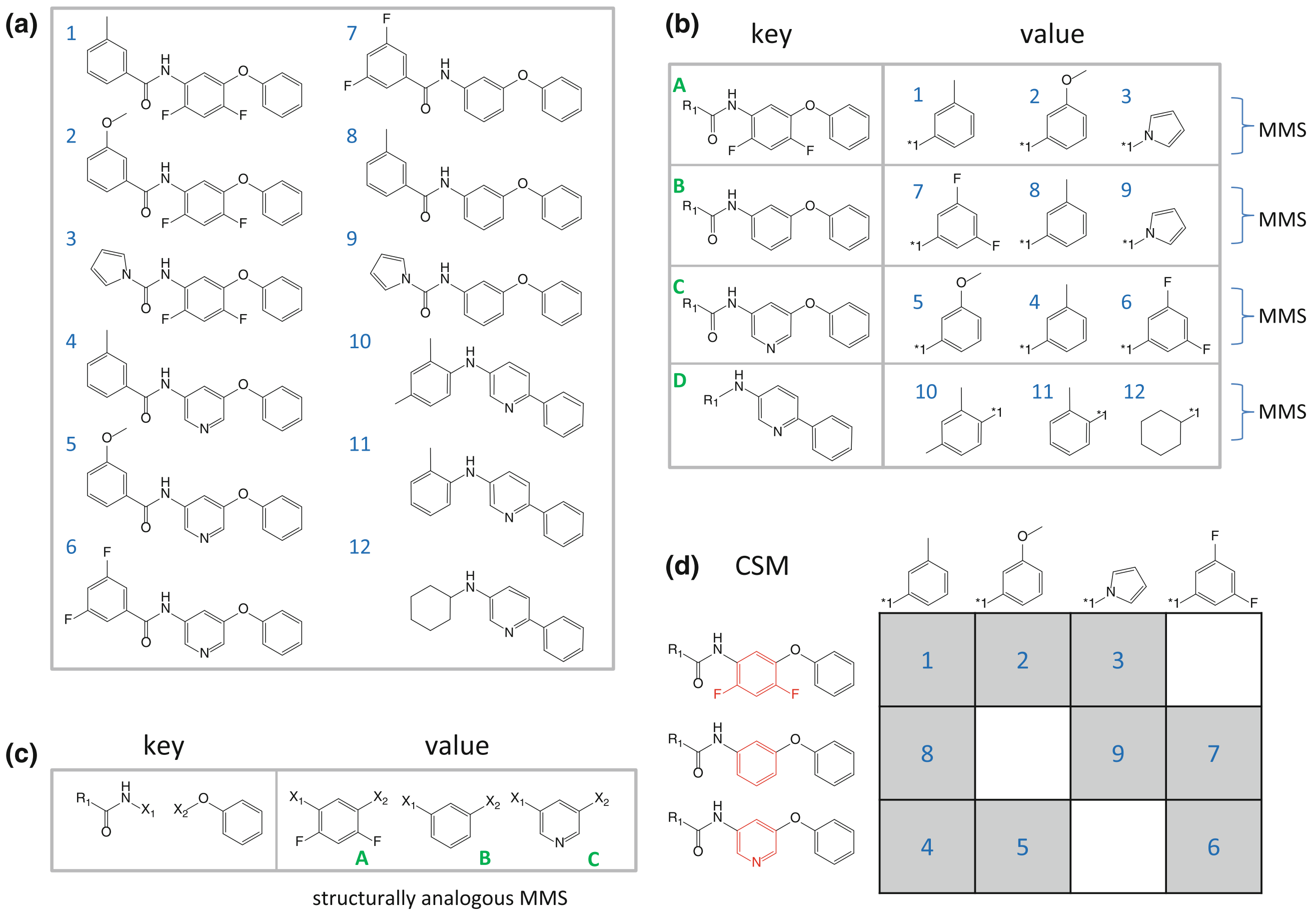

Fig. 1 MMP and matrix generation. **a** A set of 12 test compounds is shown. **b** MMP index table derived from these compounds form four MMS (*A*, *B*, *C* and *D*). **c** Three of these MMS (*A*, *B*, and *C*) that are structurally analogous are revealed by the index table. **d** CSM representing the three structurally analogous MMS. Distinguishing fragments in the structurally related cores are highlighted in *red*

MMS from keys in the original index table (Fig. 1b) through double-cut fragmentation.

Figure 2 shows three exemplary MMS that are further transformed into Bemis–Murcko (BM) scaffolds [17] by removing R-groups from compounds as well as cyclic skeletons (CSKs) [18] by setting all bond orders in BM scaffolds to one and converting all heteroatoms to carbon. Structurally analogous MMS are generally difficult to identify because maximum common substructure methods are not feasible in this case to capture related series. However, they can be systematically identified by applying the MMP-based double-fragmentation procedure described above.

Compound series matrix

MMS that are structurally analogous are combined in a matrix in which rows are formed by keys and columns by corresponding values, as illustrated in Fig. 1d. Each combination of a key and value fragment defines an individual compound. Compounds present in the same row form an individual MMS. The matrix format was adapted from SAR matrices [9] that were designed to analyze SAR information contained in analog series focusing on compound potency, as illustrated in Fig. 3 (left). Here, the matrix structure was utilized for data mining and used to systematically extract all analog series with multi-target activities from currently available active compounds. Accordingly, the matrix structure is termed compound series matrix (CSM). For matrix annotation, all target annotations were collected for available active compounds. Cells in a matrix were color-coded according to the number of targets that compounds were active against, as shown in Fig. 3 (right).

As a consequence of the systematic fragmentation procedure, it is possible that a compound is represented multiple times in matrices as distinct key-value combinations. Therefore, compound redundancy in matrices is minimized as follows:

1. If keys of key-value combinations representing the same compound form substructure relationships, only the larger key fragment is retained.
2. If two MMS sharing the same value fragments yield identical compounds, one of these series is randomly selected and removed.
3. If two different value fragments in an MMS yield the same compounds, the value fragment associated with the smaller number of compounds is removed.

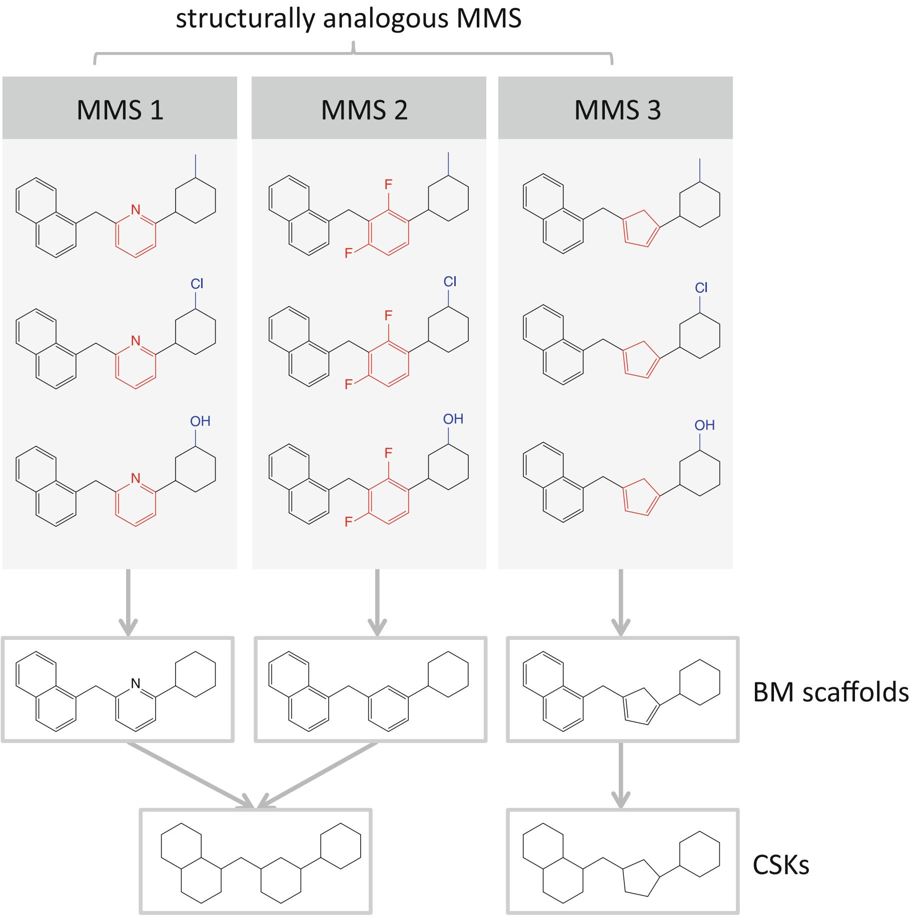

Fig. 2 Structurally analogous matching molecular series. Shown are three exemplary analog series (MMS) with structurally related yet distinct cores. Corresponding substituents are highlighted in *blue* and modifications that distinguish related core structures in *red*. The generation of BM scaffolds and CSKs from analogs is shown at the *bottom*

Matrix coverage

Matrix coverage C defines the proportion of CSM cells that are populated with known active ("real") compounds.

$$C = n/(rows * columns)$$

Here, n gives the number of populated matrix cells and *rows* and *columns* refer to the numbers of rows (keys) and columns (values) forming the matrix. Coverage values range from 0 to 1 and reflect matrix population density for known compounds.

Importantly, by design CSMs consist of known active compounds and virtual compounds that extend/complement structurally related series but are not yet available. As further discussed below, virtual compounds are highly relevant for CSM analysis because they provide compound design suggestions. Accordingly, the matrix coverage parameter provides a measure for virtual compound content of CSMs; the larger C, the larger the population density for real compounds in a given matrix; the smaller C, the sparser the matrix and the larger the population of virtual compounds. Thus, CSMs can be ranked on the basis of matrix coverage. For example, CSMs with larger coverage are prioritized if one searches for extensively explored core structures and/or chemical substitutions. Alternatively, CSMs with smaller coverage are preferred if one searches for opportunities to design new analogs of active compounds or series of interest.

Matrix generation

Keys from original set of MMS comprising of at least three compounds were subjected to a second round of fragmentation to generate CSMs formed by structurally analogous MMS. A CSM was required to contain a minimum of two MMS with at least three compounds each. CSMs comprising compounds that were subsets of other larger matrices were discarded to minimize matrix redundancy. If CSMs consisted of the same set of compounds, the matrix with larger coverage and larger average key size was retained to prioritize analogs with smaller substituent exchanges.

Implementation

Routines required to build the index tables, identify analogous MMS, generate CSMs, and visualize matrices were implemented in Java using OpenEye tool kits [19, 20].

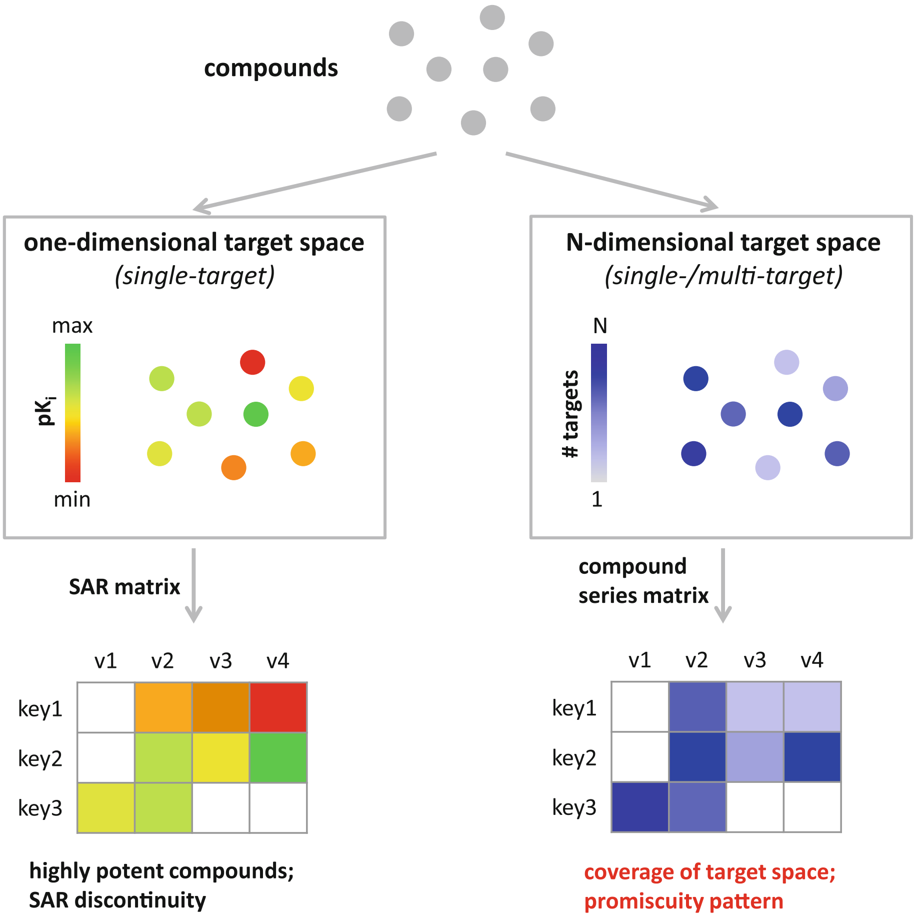

Fig. 3 Matrices of different design concept. A schematic comparison of the SAR matrix method and its CSM extension for mining of multi-target activity space is shown

Data sets

Compounds containing rings and a maximum of 45 acyclic single bonds with activity against human targets were extracted from ChEMBL (release 15) [21]. Compounds with direct target interactions, available K_i values, and a potency of at least 10 μM were selected and their K_i measurement-based target annotations were compiled.

Results and discussion

Study concept

This matrix format was originally introduced as the so-called SAR matrix [9] to monitor potency distributions of analogs active against a given target in series using potency-based color-coding of matrix cells, as illustrated on the left in Fig. 3. SAR matrices were prioritized if they contained many highly potent compounds or displayed SAR discontinuity [9]. Here, we do not utilize the matrix format for SAR analysis but rather adapt the data structure for systematic mining of related compound series in high-dimensional target space, as illustrated on the right in Fig. 3. Transitioning from SAR analysis of single-target compound sets to systematic mining of compound activity data in multi-target space required methodological extensions. In CSMs, an alternative color code was introduced focusing on multi-target activities and efficient compound and matrix redundancy tests were implemented. CSMs were primarily designed to map structurally related analog series in target space, capture multi-target activities associated with closely related compounds, and reveal (visualize) promiscuity patterns. Furthermore, we emphasize another previously unconsidered aspect of matrix mining. Analysis of CSMs makes it possible to bridge between data mining and compound design by focusing on virtual compounds contained in matrices that complement existing series. Depending on their matrix environment and multi-target activity patterns of neighboring compounds, virtual CSM compounds can be prioritized that are likely to display desired activity against selected targets, as further discussed below.

Analog series with related core structures

In Fig. 2, exemplary MMS are shown that are by definition characterized by the presence of closely related yet distinct core structures and corresponding substituents. In this study, we have aimed to systematically identify and characterize such series, which are of particular interest for a comparative SAR analysis and practical medicinal chemistry applications, yet difficult to extract from databases. To these ends, we have adapted an MMP-based dual-fragmentation approach to identify structurally related MMS and organize them in CSMs, as detailed in Methods section.

Mining compound series matrices

On the basis of our selection criteria, 37,850 unique compounds were obtained having a total of 62,784 target annotations. The number of target annotations per compound ranged from one to 35 and the compounds were active against a total of 342 targets.

From the pool of 37,850 compounds, CSMs were systematically generated and a total of 2,337 non-redundant

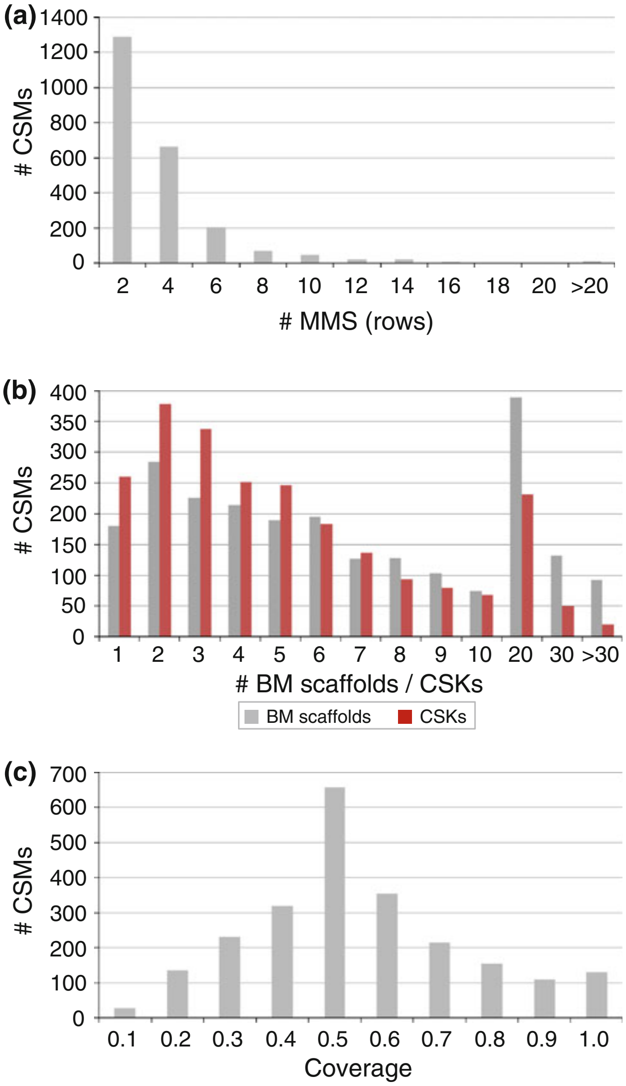

Fig. 4 Compound series matrix content. **a** The number of CSMs containing increasing numbers of MMS (*rows*) is reported in a histogram. **b** The distribution of BM scaffolds (*gray*) and CSKs (*red*) over CSMs is reported. **c** For all CSMs, matrix coverage is reported

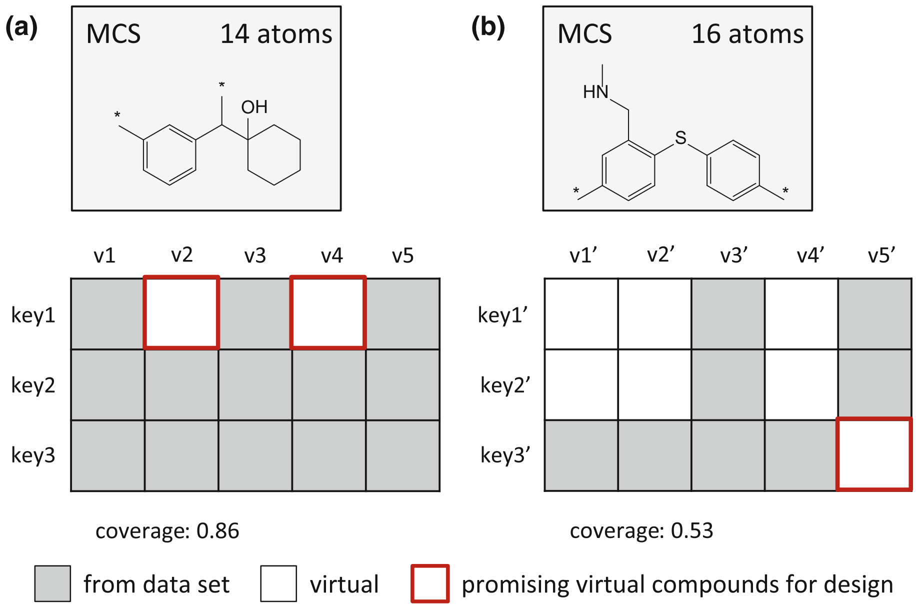

Fig. 5 Compound series matrices with different coverage. Two model CSMs are shown containing the same number of series and substituents but having different matrix coverage. For both CSMs, the maximum common substructure (MCS) is given. Cells are shown in *gray* if they represent data set compounds and *white* if they represent virtual compounds. Cells yielding promising compound design suggestions are highlighted in *red*

matrices were obtained. The structurally related compound series forming each of these matrices are made freely available via the following URL: http://www.lifescienceinformatics.uni-bonn.de.

Matrix composition

Figure 4a reports the distribution of MMS (rows) over CSMs. The majority of CSMs, i.e., more than 1,300, contained two MMS. In addition, more than 600 and 200 CSMs consisted of three to four and five to six MMS, respectively. Matrices with seven to 14 MMS were also frequently obtained and individual CSMs with more than 20 series were identified. In total, the CSMs represented 5,821 unique MMS. Among these MMS, there were 4,288 series with multi-target activities, 735 of which were active against targets from different families.

Figure 4b reports the distribution of BM scaffolds and CSKs for all CSMs. Among 2,337 CSMs, a total of 180 and 260 CSMs contained compounds that were represented by the same BM scaffolds and CSKs, respectively. The remaining ~89–92 % of CSMs contained multiple core structures yielding different scaffolds and CSKs. Moreover, 689 and 369 CSMs were found to represent more than nine BM scaffolds and CSKs, respectively. On average, CSMs contained compounds with nine scaffolds and six CSKs, respectively.

Figure 4c reports the coverage of CSMs. The distribution displays a peak at a coverage value of 0.5. Hence, in these matrices, 50 % of all theoretically possible analogs were present, providing opportunities for the exploration of additional analogs. CSMs with higher coverage were also frequently observed. Figure 5 shows two model CSMs that yielded MCSs of comparable size and contained the same number of MMS and value fragments. However, these CSMs were distinguished by different degrees of coverage, i.e., 0.86 versus 0.53. White cells represent virtual compounds combining key and value fragments that occurred in other compounds. Virtual compounds that are adjacent

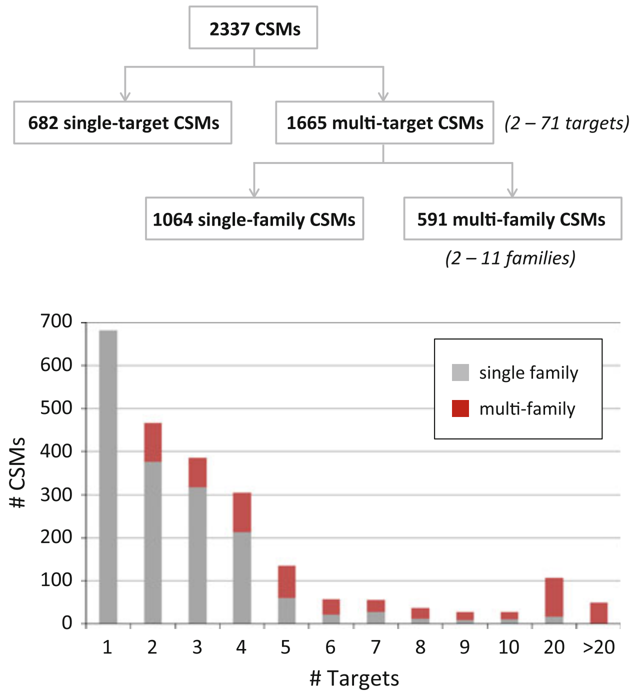

Fig. 6 Mining multi-target activity space. At the top, all generated CSMs are classified according to single- and multi-target activities. At the bottom, the distribution of CSMs annotated with increasing numbers of targets is reported distinguishing between single- (*gray*) and multi-family (*red*) activity

in the matrix to data set compounds with desired (multi-target) activity (highlighted in Fig. 5) provide promising compound design suggestions.

Target distribution

Major goals of this study have been to mine multi-target activity space using CSMs and identify structurally related analog series with activity against increasing numbers of targets. As reported in Fig. 6, ~29 % of CSMs (682) were composed of compounds with single-target activity. By contrast, a total of 1,064 CSMs identified compounds with activity against multiple targets from the same family. Unexpectedly, 591 CSMs were also found to contain compounds active against targets from two to 11 different families. The target distribution is reported at the bottom of Fig. 6. CSMs containing compounds with reported activity against two to five targets were frequently observed and smaller

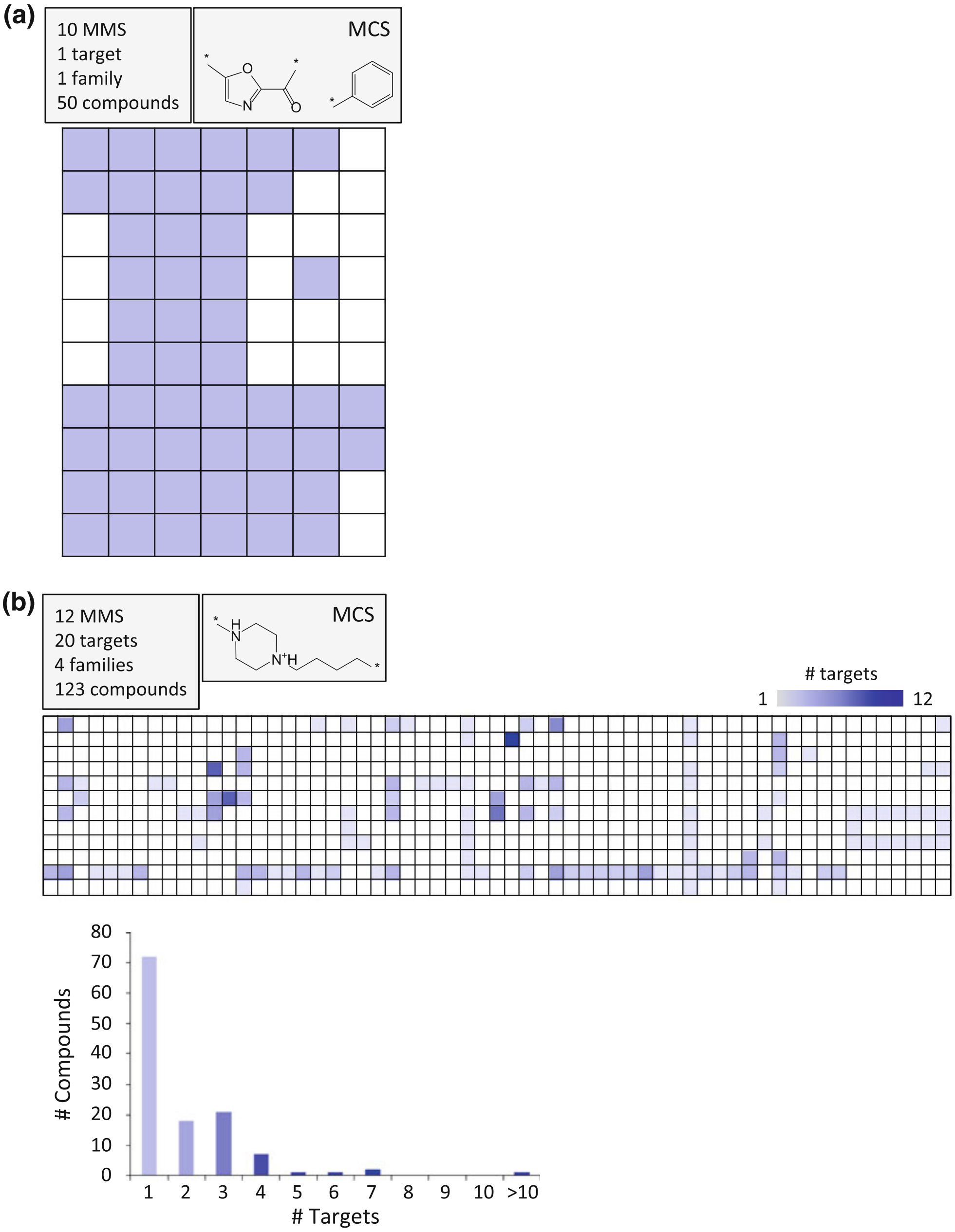

Fig. 7 Single- and multi-target compound series matrices. **a** Single-target CSM with analogs active against anandamide amidohydrolase. The MCS of all series is shown. The CSM label (*top left*) reports the number of MMS, targets the matrix compounds are active against, families these targets belong to, and the total number of analogs. **b** Multi-target CSM with analogs active against 20 targets belonging to four families. Target coverage of analogs is reflected by increasingly *dark blue* shading of cells. The *histogram at the bottom* reports the number of matrix compounds with activity against increasing numbers of targets

numbers of CSMs covered a wide range of up to more than 20 targets. On average, CSMs displayed activity against four targets. In general, the proportion of multi-family CSMs increased with the increasing numbers of targets.

Thus, a large number of multi-target CSMs was identified that captured activity of closely related analog series against targets from different families. The CSM-based identification and structural organization of these series

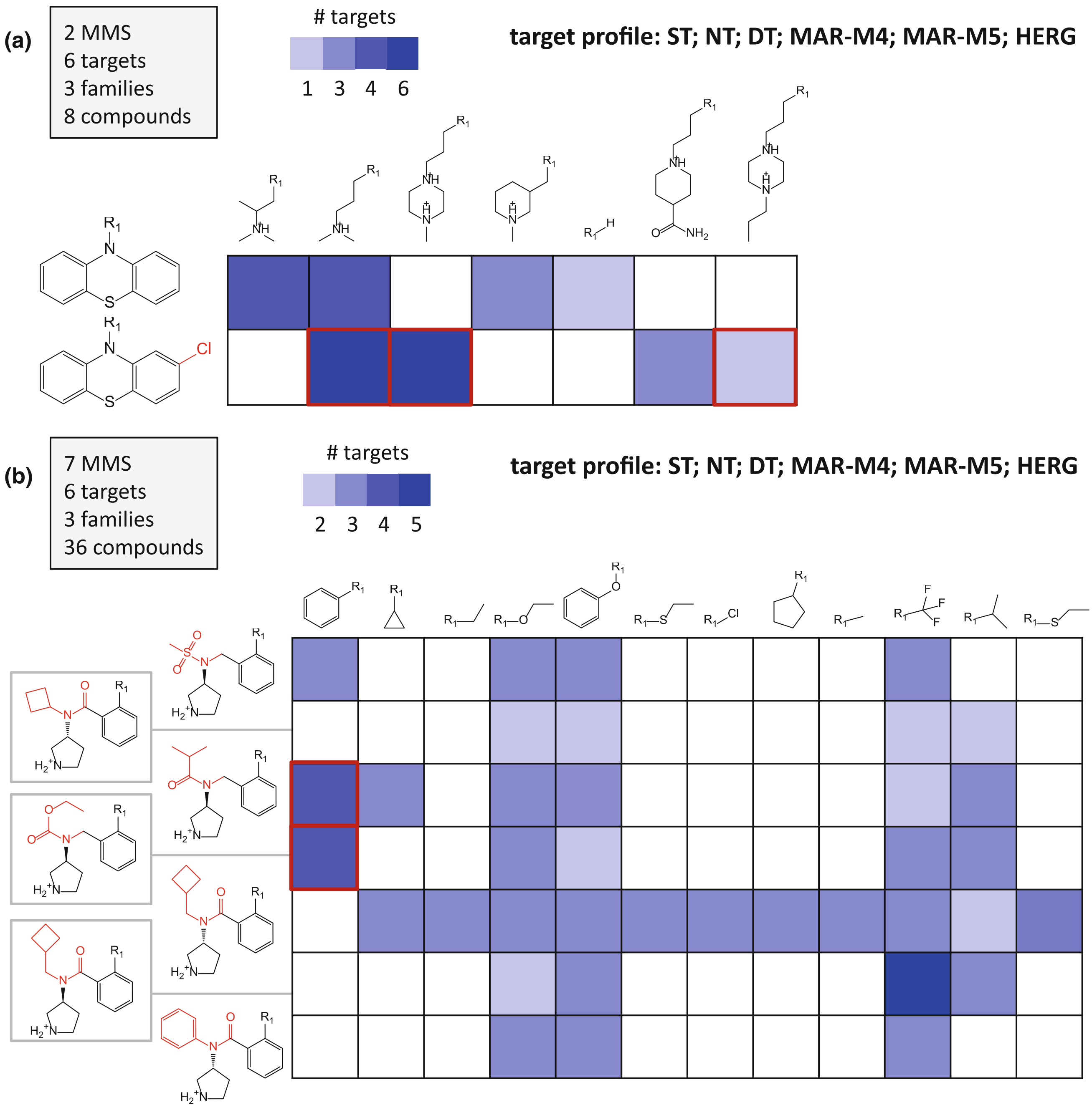

Fig. 8 Compound series matrices representing the same target profile. In **a** and **b**, two exemplary CSMs cover structurally distinct series with activity against the same targets. **a** CSM containing two MMS with eight compounds. **b** CSM containing seven MMS with 36 compounds with activities against 2–5 targets. The representation is according to Fig. 7. Substructures distinguishing the core fragments are highlighted in *red*. The series in these two matrices display different promiscuity patterns. Target abbreviations: *ST* serotonin transporter, *NT* norepinephrine transporter, *DT* dopamine transporter, *MAR-M4* muscarinic acetylcholine receptor M4, *MAR-M5* muscarinic acetylcholine receptor M5, *HERG* HERG ion channel. Cells in the two matrices with reported activity against the HERG anti-target are highlighted in *red*

made it possible to study compound promiscuity patterns in detail.

Promiscuity patterns

In Fig. 7, two exemplary CSMs capturing analog series with single-target activity or target promiscuity are shown. In Fig. 7a, the single-target CSM contained 10 MMS and 50 compounds that were active against anandamide amidohydrolase. By contrast, the larger CSM in Fig. 7b contained 12 series with 123 analogs active against one to 12 targets. The histogram at the bottom of Fig. 7b reports the distribution of target annotations for all analogs, revealing a subset of 51 promiscuous compounds with activity against increasing numbers of targets. In total, the compounds were active against 20 unique targets belonging to four different families. Hence, this example illustrates that CSMs enable the detection of progressive target promiscuity patterns among analogs belonging to closely related series.

Matrices representing the same target profile

As described above, multi-target CSMs were frequently identified. In Fig. 8, two exemplary CSMs with different core structures are shown that contain two and seven MMS, respectively. These series consisted of different numbers of analogs with activity against varying numbers of six targets from three different families. For the majority of these series, closely related analogs were found to be active against overlapping yet distinct targets. These CSMs had very different matrix coverage, i.e., 0.57 (Fig. 8a) versus 0.29 (Fig. 8b). Thus, the CSM in Fig. 8b provided more opportunities to design analogs and “fill” the matrix. However, compounds with HERG anti-target activity were also found in these CSMs (highlighted in Fig. 8). Hence, these compounds point at likely liabilities associated with individual series, which would suggest to carefully investigate potential anti-target activities of closely related analogs captured in these matrices.

Mapping of virtual matrix compounds to drugs

From the 2,337 CSMs reported herein, all virtual compounds were extracted and mapped to 6,081 approved and experimental drugs assembled from DrugBank [22]. A total of 48 drugs were found to match virtual compounds derived from 25 different matrices. Most of these drugs were annotated with targets that overlapped with the target profiles of the corresponding matrices. Hence, virtual matrix compounds are a potential source of interesting drug (-like) molecules.

Conclusions

In this study, we have searched for closely related analog series with multi-target activities by applying the CSM concept and described promiscuity patterns emerging from these series. CSMs represent multiple structurally analogous series and closely related virtual compounds in a well-defined manner. Moreover, the matrix data structure was used here for systematic compound data mining and the exploration of multi-target activity space on the basis of currently available bioactivity data. Virtual analogs adjacent to compounds with desired activities yield design suggestions. We have identified 4,288 series with diverse multi-target and 735 series with multi-family activities. The identification of these series in CSMs and their structural organization makes it possible to analyze promiscuity patterns at the level of structurally related analogs. All CSMs are made freely available to provide a basis for further analysis of analog series with multi-target activities.

References

1. Wermuth CG (ed) (2008) The practice of medicinal chemistry, 3rd edn. Academic Press, San Diego
2. Agrafiotis DK, Shemanarev M, Connolly PJ, Farnum M, Lobanov VS (2007) SAR maps: a new SAR visualization technique for medicinal chemists. J Med Chem 50(24):5926–5937
3. Wassermann AM, Bajorath J (2012) Directed R-group combination graph: a methodology to uncover structure–activity relationship patterns in series of analogs. J Med Chem 55(3):1215–1226
4. Cho SJ, Sun Y (2008) Visual exploration of structure–activity relationship using maximum common framework. J Comput Aided Mol Des 22(8):571–578
5. Schuffenhauer A, Ertl P, Roggo S, Wetzel S, Koch MA, Waldmann H (2007) The scaffold tree—visualization of the scaffold universe by hierarchical scaffold classification. J Chem Inf Model 47(1):47–58
6. Agrafiotis DK, Wiener JJ (2010) Scaffold explorer: an interactive tool for organizing and mining structure–activity data spanning multiple chemotypes. J Med Chem 53(13):5002–5011
7. Gupta-Ostermann D, Hu Y, Bajorath J (2012) Introducing the LASSO graph for compound data set representation and structure–activity relationship analysis. J Med Chem 55(11):5546–5553
8. Wawer M, Lounkine E, Wassermann AM, Bajorath J (2010) Data structures and computational tools for the extraction of SAR information from large compound sets. Drug Discov Today 15(15–16):631–639
9. Wassermann AM, Haebel P, Weskamp N, Bajorath J (2012) SAR matrices: automated extraction of information-rich SAR tables from large compound data sets. J Chem Inf Model 52(7):1769–1776
10. Kenny PW (2005) Sadowski J (2005) Structure modification in chemical databases. In: Oprea TI (ed) Chemoinformatics in drug discovery. Wiley-VCH, Weinheim, Germany, pp 271–285
11. Hussain J, Rea C (2010) Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. J Chem Inf Model 50(3):339–348

12. Knight ZA, Lin H, Shokat KM (2010) Targeting the cancer kinome through polypharmacology. Nat Rev Cancer 10(2):130–137
13. Paolini GV, Shapland RH, van Hoorn WP, Mason JS, Hopkins AL (2006) Global mapping of pharmacological space. Nat Biotechnol 24(7):805–815
14. Hu Y, Bajorath J (2013) Compound promiscuity: what can we learn from current data? Drug Discov Today 18(13–14):644–650
15. Boran AD, Iyengar R (2010) Systems approaches to polypharmacology and drug discovery. Curr Opin Drug Discov Dev 13(3):297–309
16. Wawer M, Bajorath J (2011) Local structural changes, global data views: graphical substructure–activity relationship trailing. J Med Chem 54(8):2944–2951
17. Bemis GW, Murcko MA (1996) The properties of known drugs. 1. Molecular frameworks. J Med Chem 39(15):2887–2893
18. Xu YJ, Johnson M (2002) Using molecular equivalence numbers to visually explore structure features that distinguish chemical libraries. J Chem Inf Comput Sci 42(4):912–926
19. OEChem TKV (2013) April, Open Eye Scientific Software Inc, Santa Fe, New Mexico
20. OEDepict TKV (2013) April, Open Eye Scientific Software Inc, Santa Fe, New Mexico
21. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP (2012) ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 40:D1100–D1107
22. Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V, Djoumbou Y, Eisner R, Guo AC, Wishart DS (2011) DrugBank 3.0: a comprehensive resource for ‘omics’ research on drugs. Nucleic Acids Res 40:D1035–D1041