Journal of
CHEMICAL INFORMATION
AND MODELING

Article

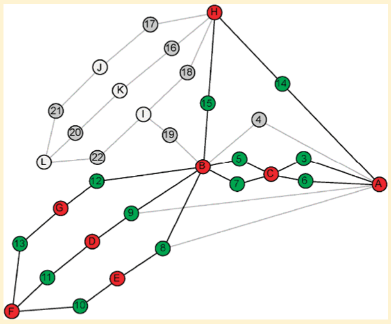
pubs.acs.org/jcim

Graph Mining for SAR Transfer Series

Disha Gupta-Ostermann,[†] Mathias Wawer,[†,#] Anne Mai Wassermann,[†] and Jürgen Bajorath*[,†]

[†]Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstrasse 2, D-53113 Bonn, Germany

ABSTRACT: The transfer of SAR information from one analog series to another is a difficult, yet highly attractive task in medicinal chemistry. At present, the evaluation of SAR transfer potential from a data mining perspective is still in its infancy. Only recently, a first computational approach has been introduced to evaluate SAR transfer events. Here, a substructure relationship-based molecular network representation has been used as a starting point to systematically identify SAR transfer series in large compound data sets. For this purpose, a methodology is introduced that consists of two stages. For graph mining, an algorithm has been designed that extracts all parallel series from compound data sets. A parallel series is formed by two series of analogs with different core structures but pairwise corresponding substitution patterns. The SAR transfer potential of identified parallel series is then evaluated using a scoring function that emphasizes corresponding potency progression over many analog pairs and large potency ranges. The substructure relationship-based molecular network in combination with the graph mining algorithm currently represents the only generally applicable approach to systematically detect SAR transfer events in large compound data sets. The combined approach has been evaluated on a large number of compound data sets and shown to systematically identify SAR transfer series.



INTRODUCTION

SAR exploration is one of the central tasks in medicinal chemistry.[1,2] Analysis efforts typically focus on both individual compound series to aid in the design of analogs[1] and on entire compound data sets.[3] While increasingly large sets of biological screening and compound optimization data become available for current pharmaceutical targets, computational methods are employed to extract available SAR information from such compound sources.[4] Hence, in addition to classical SAR studies on individual compound series, SAR data mining is a topic of growing interest. For large-scale SAR exploration, graphical analysis methods are of particular interest.[3,4] In combination with numerical analysis functions, SAR visualization methods often provide a direct access to SAR information contained in compound data sets from various sources.[3,4] In particular, molecular networks have become increasingly popular for SAR visualization.[4] For example, in similarity-based molecular networks, nodes represent active compounds and edges similarity relationships.[4] Such networks are then annotated with compound potency and other SAR-relevant information. Recently, a first SAR network design has been introduced where calculated pairwise molecular similarity values are replaced with well-defined substructure relationships, the so-called bipartite matching molecular series graph (BMMSG).[5] The substructure-based representation was designed to further improve the chemical interpretability of SAR networks.[5] In this type of network, edges represent specific fragments that distinguish compounds with otherwise identical structures. This makes it possible to interactively navigate the network by directly following structural (and potency) relationships between active compounds.

One of the tasks in SAR analysis where the exploration of individual compound series and data mining approaches meet is the evaluation of SAR transfer. In many instances, lead series yielding promising SARs cannot be further developed because of therapeutic liabilities associated with the given chemotype. In medicinal chemistry, it is then of considerable interest to explore alternative chemotypes that might yield equally promising SARs. In an ideal scenario, one would like to replace an undesired core structure with another for which corresponding analogs display comparable (and predictable) SAR characteristics and potency progression, yielding a series that can be further developed. The successful identification of such alternative series then constitutes an SAR transfer event. Recently, the only currently available computational methodology for the detection of SAR transfer events has been introduced.[6] This approach specifically focuses on single-ring replacements in molecular scaffolds to identify SAR transfer series and hence is not generally applicable.

A prerequisite for the assessment of SAR transfer potential is the identification of parallel series of analogs with corresponding substitutions. Once parallel series have been identified, potency values can be taken into account and potency progression assessed. In substructure relationship-based compound networks, parallel series were previously observed to form characteristic subgraph patterns, which makes these networks an attractive source for the identification of such series and the assessment of SAR transfer potential.[5] However, such subgraphs of varying topological complexity and appearance are

Received: February 6, 2012
Published: March 21, 2012

dx.doi.org/10.1021/ci300071y | *J. Chem. Inf. Model.* 2012, 52, 935−942

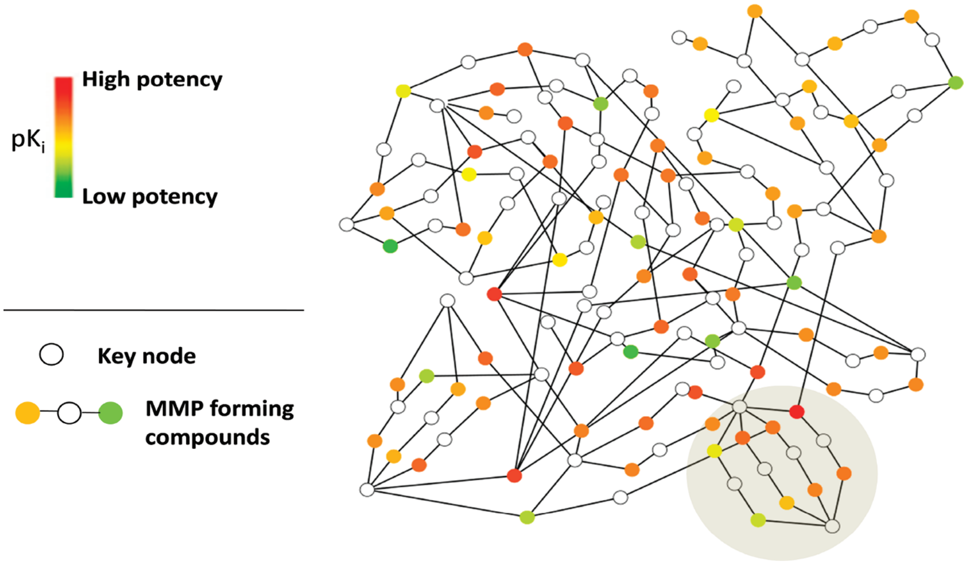


Figure 1. BMMSG representation. A prototypic BMMSG is shown. Key nodes are displayed in white and molecule nodes are colored by compound potency using a continuous color code from green (lowest potency) via yellow to red (highest potency in a data set). An MMP is represented by two molecule nodes connected to the same key. Edges represent transformations, i.e., distinguishing substructures. At the lower right of the graph, a spindle structure is highlighted.

often difficult to identify in graphs of large compound data sets. This has motivated us to develop a graph mining algorithm for the extraction of all possible parallel series from such network representations. The SAR transfer character of candidate series is then assessed using a newly introduced scoring scheme. On the basis of these scores, parallel series can be ranked according to their SAR transfer potential. When used in combination, the graph mining algorithm and scoring function make it possible to automatically extract available SAR transfer series from large compound data sets.

■ CONCEPTS, METHODS, AND MATERIALS

Matched Molecular Pairs. Matched molecular pairs (MMPs)[7,8] are defined as pairs of compounds that are distinguished by a single fragment (substructure) at a specific site. Distinguishing fragments might vary in size and include, for example, R-groups or ring systems, which might be substituents or part of the core structure. The exchange of the distinguishing substructures is referred to as a molecular transformation.[9] The MMP concept provides one of the foundations of the BMMSG design, as further detailed below.

MMP Identification. MMPs were detected using an in-house implementation of the MMP search algorithm of Hussain and Rea.[9] First, compounds are fragmented by systematically deleting all nonring single bonds between two non-hydrogen atoms. A single cut results into two fragments that are added to an index table: the larger fragment constitutes the key and the smaller fragment the value that is linked to the source compound. If fragments having the same size are generated, each is stored once as key and the other as the corresponding value. Furthermore, the deletion of combinations of two or three bonds generates a core fragment and two or three terminal fragments, respectively. If the core contains less heavy atoms than are contained in the terminal substructures, the fragments are added to the index table, with the set of terminal fragments indexed as the key and the core fragment as the corresponding value. MMPs are extracted from the index table by searching for keys having more than one value. It should also be noted that stereochemical information is often retained in MMPs if bonds carrying stereochemical information have not been deleted. Because compounds are systematically fragmented during MMP generation, keys and/or values fragments containing stereochemical information are retained.

BMMSG Generation. The BMMSG data structure[5] is constructed on the basis of the MMP index table. The graph contains two types of nodes and is hence bipartite: key nodes and molecule nodes.[5] All keys that are associated with more than one value in the index table are displayed. For each key, molecules containing this substructure are derived from corresponding value fragments and connected to the key via edges. Accordingly, edges correspond to values (and are also graphically associated with value fragments). All molecule nodes that share the same key represent a so-called matching molecular series (MMS). A compound can belong to several MMS. An exemplary BMMSG is shown in Figure 1. Key nodes are displayed in white, and all molecule nodes are colored according to the compound potency range in the data set using a color spectrum illustrated in Figure 1.

Parallel and SAR Transfer Series. A parallel series is defined as two series of analogs that are based upon different core structures and that can be organized into compound pairs with the same substitutions, as illustrated in Figure 2a. SAR transfer occurs in parallel series when pairs of compounds with corresponding substitutions display comparable potency progression. In SAR transfer series, the absolute potency values of individual series might differ, for example, one series might be active at the micromolar and the other at the nanomolar level. However, in SAR transfer events, corresponding substitutions should lead to comparable potency differences between corresponding analogs in the two series. This is evaluated by an alignment of compound pairs with corresponding substitutions in both series in the order of increasing potency,[6] as also illustrated in Figure 2a. Hence, not all parallel series represent SAR transfer events, and, in addition, SAR transfer might not be complete along the alignment. Therefore, following the

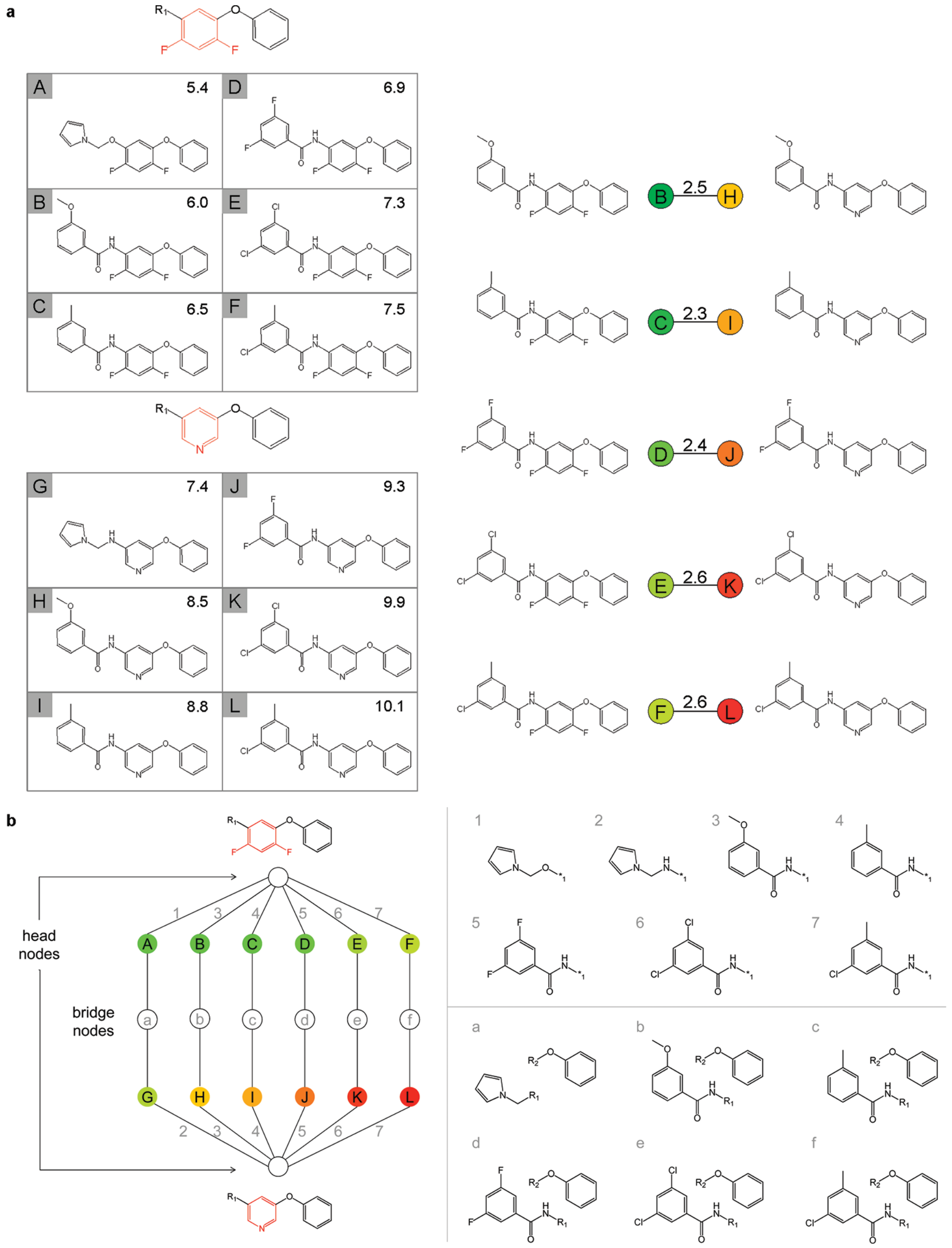


Figure 2. SAR transfer series. (a) Shown are two structurally related analog series containing six compounds each. The distinguishing substructures are colored in red. Compounds (A-L) are annotated with their pK_i values. Five analogs in each series form a parallel series, as illustrated by the compound alignment on the right. Molecule nodes are numbered and color-coded according to potency, and edges between nodes are annotated with the logarithmic potency difference of analogs forming a pair. Because corresponding structural modifications are accompanied by similar changes in potency in both series (i.e., potency differences between corresponding analogs remain largely constant along the pairwise alignment), the two series represent an SAR transfer event. (b) The BMMSG spindle structure capturing the SAR transfer series in (a) is shown. The subgraph contains the key nodes (white) that represent the core structures of each series and the molecule nodes A-L colored by potency. The corresponding substitutions 1−7 are provided that are represented by the edges connecting molecule and key nodes. Because each corresponding substitution generates a larger substructure shared by a pair of analogs, an additional layer of key nodes a-f is generated that connects compounds belonging to each analog pair, giving rise to the spindle architecture. Following the generalizations of its topology, as described in the text, the two key nodes at the top and at the bottom are head nodes and the intermittent layer of keys represents bridge nodes.

detection of parallel series, SAR transfer potential must be subsequently evaluated.

Parallel Series in BMMSGs. In BMMSGs, parallel series are typically detected as subgraphs having a characteristic topology reminiscent of a spindle.[5] An exemplary spindle structure is highlighted in the BMMSG in Figure 1, and other examples are provided in Figures 2b and 3. However, as will be described in more detail below, the spindle topology might be distorted and/or appear in complex subgraph environments and might hence often be difficult to recognize in BMMSGs on the basis of visual inspection. For automated detection, it is important to generalize the topological characteristics of spindle-type subgraphs representing parallel series. The following generalizations are introduced:

(1) In a BMMSG, two parallel series form several distinct paths of length five that share a common start and end point, i.e., key nodes we here refer to as “head nodes”.

(2) Each pair of such parallel paths forms a cycle of eight nodes; each head node is connected to a set of molecule nodes constituting an MMS.

(3) Compounds that belong to different series but display corresponding substitutions are connected through another key node termed “bridge node”; each pair of corresponding compounds differs by the substructure exchange defined by the connecting edges to the bridge node. A bridge node might be connected to multiple analogs of a series.

Head and bridge nodes are indicated in Figure 2b.

Graph Mining for Parallel Series. Given the above generalizations, parallel series yield cyclic subgraphs in BMMSGs. Accordingly, without loss of parallel series, a BMMSG can be simplified by iteratively removing all terminal nodes (with degree one) until no node having only a single neighbor remains. This reduces the number of search steps during the graph traversal process to identify circular subgraphs, as described below. Parallel series are identified using a “breadth-first” search. The algorithm is described in detail in the Results section.

Identification of SAR Transfer Series. The graph mining algorithm identifies parallel series. If parallel series represent an SAR transfer event, corresponding chemical modifications in both series must be accompanied by similar changes in potency. This is assessed by calculating logarithmic potency differences between corresponding compound pairs. If potency differences are comparable for all pairs of compounds, SAR transfer is observed. By contrast, if parallel series display large variations in potency differences, potency progression within the two series is different and the SAR of one series is not transferable to the other. To prioritize parallel series with similar potency progression and identify SAR transfer series, the following ranking function is applied to the alignment of parallel series (according to Figure 2a)

$$\text{score}(\text{series}_{i,j}) = \frac{\max\limits_{s=i,j}(\text{range}^{s})}{\text{sd}(d)} \cdot \log(n)$$

with

$$\text{range}^{s} = \max_{k=1,\ldots,n}(\text{pot}_{k}^{s}) - \min_{k=1,\ldots,n}(\text{pot}_{k}^{s})$$

and

$$d_k = \text{pot}_{k}^{i} - \text{pot}_{k}^{j}$$

Here, n corresponds to the number of pairs of corresponding compounds in the parallel series, pot_k^s denotes the logarithmic potency of the kth compound in series s, range^s gives the potency range spanned by series s, and $\text{sd}(d)$ is the standard deviation of logarithmic potency differences between corresponding compounds. $\log(n)$ is calculated to balance the influence of the number of compound pairs on the score and the resulting series ranking. It should also be noted that the smaller the standard deviation of pairwise potency differences along the alignment, the higher is the SAR transfer potential of a parallel series, which is a desired feature of this scoring scheme. For applications carried out thus far, the score has been a rather stable indicator of SAR transfer potential.

Table 1. Data Sets, Parallel Series, and SAR Transfer Series[a]

data set	target	#compounds	#parallel series	#SAR transfer series
1	serotonin receptor 2C	506	7	0
2	urokinase	513	3	0
3	carbonic anhydrase IX	517	27	3
4	neutrophil elastase	534	24	1
5	nociceptin receptor	558	14	1
6	serotonin receptor 2A	638	5	0
7	sigma opioid receptor	641	36	4
8	adenosine receptor A2B	654	35	0
9	serotonin receptor 6	663	98	5
10	norepinephrine transporter	715	26	3
11	trypsin	719	4	0
12	melanin-concentrating hormone receptor 1	747	21	1
13	histamine H3 receptor	803	13	3
14	cannabinoid receptor 1	835	73	1
15	serotonin transporter	865	35	5
16	cannabinoid receptor 2	867	53	5
17	corticotropin releasing factor receptor 1	900	60	1
18	dopamine D3 receptor	970	60	3
19	dopamine D4 receptor	984	71	5
20	serotonin receptor 1A	1002	30	1
21	melanocortin receptor 4	1026	45	12
22	delta opioid receptor	1352	32	2
23	adenosine receptor A2A	1418	133	11
24	adenosine receptor A1	1453	145	22
25	mu opioid receptor	1503	35	3
26	kappa opioid receptor	1558	34	0
27	carbonic anhydrase I	1572	678	206
28	adenosine receptor A3	1597	291	23
29	carbonic anhydrase II	1700	521	178
30	dopamine receptor D2	1771	59	6
31	factor Xa	1972	42	8
32	thrombin	2037	60	13

[a] For the 32 sets of inhibitors or antagonists taken from BindingDB, the targets are reported. Compound data sets are sorted in increasing order of the number of compounds (#compounds) they contain. Furthermore, for each data set, the number of qualifying parallel series (#parallel series) we identified are reported and the number of SAR transfer series (#SAR transfer series) yielding a score >15.

According to this scoring function, parallel series with many compound pairs, little variation in pairwise potency differences along the series alignment, and large potency ranges are ranked highly. In addition, to avoid the comparison of structural series that represent a “flat” SAR, i.e., series in which all analogs have similar potencies, at least one of the two series must span a potency range of more than 1 order of magnitude to qualify for ranking.

Implementation. All calculations required to generate MMPs and BMMSGs, extract parallel series, and identify SAR transfer series were carried out using in-house written Java programs. Routines to generate MMPs were implemented using the OpenEye chemistry toolkit[10] and graph structures utilizing the Java package JUNG.[11]

Data Sets. We extracted all ring-containing compounds annotated with defined K_i values for human targets from BindingDB.[12] For all molecules with multiple potency records within an order of magnitude against the same target the arithmetic mean was calculated as the final potency. If available K_i values differed by

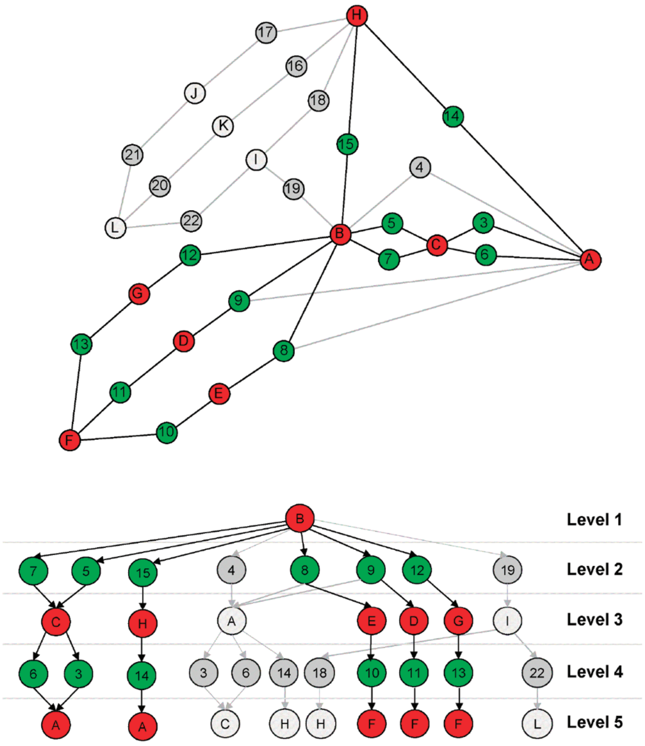


Figure 3. Identification of parallel series. At the top, a model BMMSG is shown. For two parallel series, key and molecule nodes are depicted in red and green, respectively. The remaining key and molecule nodes of the graph are shown in white and gray, respectively. The two red/green parallel series share key node B. The spindle in the lower left section of the graph has a regular topology, as described in Figure 2b. The other in the upper right section has an "intertwined" topology that is more difficult to recognize. At the bottom, a directed layered graph representation is derived from the BMMSG using key node B as the root. Graph levels are given on the right. Node C at level 3 is an example of a "merged" node that is connected to nodes 5 and 7 on level 2. Similarly, node A on level 3 and nodes A and C on level 5 are merged nodes. For both key nodes A and F in level 5, three paths are identified that lead to the root node. Hence, nodes A and B and nodes B and F are detected as "head" nodes of the two parallel series on the lower left and upper right of the BMMSG, respectively.

more than 1 order of magnitude, the compound-target pair was not further considered. These criteria led to the selection of 53,760 compounds that were organized according to their target activities. After removal of target sets with less than 500 ligands, a total of 32 different data sets remained for further analysis that contained between 506 and 2037 compounds, as summarized in Table 1. These compound sets were subjected to BMMSG calculation, extraction of parallel series, and SAR transfer ranking.

■ RESULTS AND DISCUSSION

BMMSG Design. A characteristic feature of the BMMSG is its bipartite nature because it contains molecule and key nodes that capture the underlying MMP-based structural organization. Another characteristic is that all compounds associated with the same key form an MMS whose members only differ by a single structural change at a specific site. Hence, each key node represents a unique MMS. However, a molecule can belong to more than one MMS because it might differ at one site from one subset of molecules and at another site from a second subset. Such MMS with different degrees of overlap determine the connectivity of the network representation.

(a) Construction of a directed layered subgraph (DLG)

1) Iteratively remove nodes with degree 1 from the BMMSG until no node with degree 1 can be found.

2) Randomly select a key node as the root (level 1) and define level 1 as the current level.

3) Add all neighbors of each node in the current level to the tree (unless the neighbor is a parent of the node) and assign them to a new level. Make the new level the current level.

4) If the current level is 3 or 5, find all instances of the same node that share a common predecessor in (current level - 2); merge these instances into one node while retaining all connections to the previous level.

5) Repeat steps 3 and 4 until the DLG contains five levels.

6) Identify parallel series (step b).

7) Repeat steps 2 – 6 until all key nodes have been selected as root once.

(b) Identification of parallel series

1) Select a node in level 5 of the DLG that has not been explored before.

2) Find all replicates of this node in level 5 and mark the node and its replicates as explored.

3) Build a set of paths between the root and the selected terminal node(s) in level 5.

4) Remove paths from the set that share molecule nodes.

5) If the set contains more than two paths, a parallel series has been found.

6) Repeat steps 1–5 until all nodes have been explored.

Figure 4. Search algorithm. An outline of the algorithm to search for parallel series is provided. The search procedure involves two nested stages including (a) construction of the directed layered graph and (b) identification of parallel series.

BMMSG Subgraph of Parallel Series. From the BMMSG illustrated in Figure 1, characteristic patterns (subgraphs) might emerge that reflect SAR information contained in a data set. Among these is a subgraph we term a spindle. A representative spindle subgraph is highlighted in the BMMSG representation in Figure 1. The spindle is particularly interesting because it is indicative of parallel analog series that might be associated with SAR transfer events, as detailed in the Methods section. Figure 2a shows an example of an SAR transfer series. In Figure 2b, the corresponding spindle is shown, and the structural changes within the parallel/SAR transfer series are specified. Spindle subgraphs do not always represent a perfect parallel series. There can be exceptions depending on the structural transformations that define the corresponding MMPs (see, for example, molecules A and G in Figure 2b). However, any parallel series available in a compound data set will appear as a spindle subgraph in a BMMSG. Thus, thorough analysis of spindles might occasionally identify false-positives but will inevitably reveal all parallel series that might be contained in a set. Importantly, because the appearance of spindle subgraphs can greatly vary in

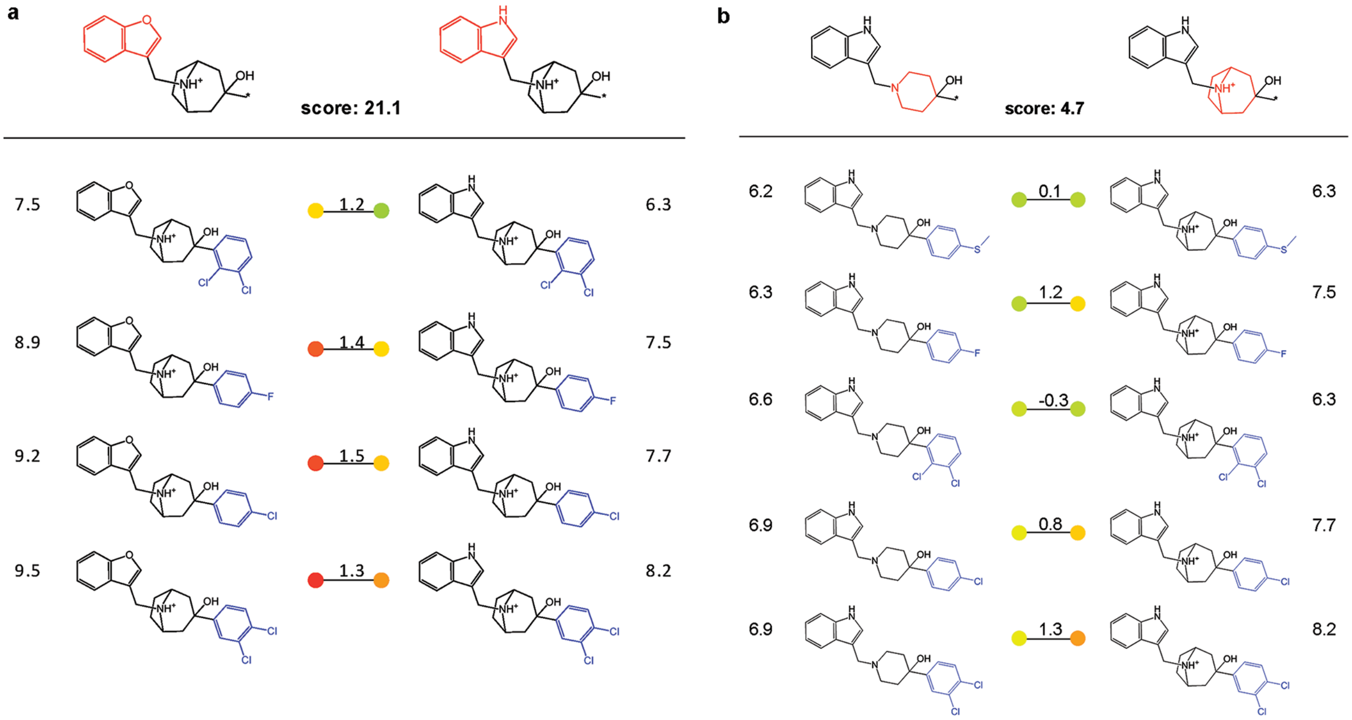


Figure 5. Scoring of series. Parallel series identified for a set of dopamine D3 receptor antagonists were scored and ranked according to their SAR transfer potential. Head nodes are displayed at the top (distinguishing substructures are colored red), and pairs of corresponding analogs are vertically aligned in order of increasing potency. Each compound is annotated with its pK_i value, and nodes are colored accordingly. Edges between nodes of compound pairs are annotated with the logarithmic potency difference of the analogs. For each parallel series, the score of the ranking function is reported. In (a) and (b), series with high and low SAR transfer scores are shown, respectively.

BMMSGs and they can occur in complex graph environments, as illustrated in Figure 3, spindles might often be difficult to identify, especially in graph representations of large and/or structurally homogeneous data sets. Therefore, a search methodology is highly desirable to systematically identify subgraphs with spindle topology from BMMSGs and extract all possible parallel series from data sets.

Search Algorithm. In order to identify spindle subgraphs with generalized topology from BMMSGs, a breadth-first search algorithm was designed and implemented. The algorithm proceeds as follows:

First, a random key node is chosen as the starting point that forms the root of a search tree (level 1 at the bottom of Figure 3). All of its neighbors are then added to the tree, forming level 2. To identify all paths of length five that originate from the root (and might form a spindle), the tree is iteratively expanded to a depth of five levels. Due to the bipartite nature of the graph, levels 1, 3, and 5 contain only key nodes, whereas levels 2 and 4 consist of molecule nodes, as illustrated in Figure 3.

To avoid redundant exploration steps, key nodes on level 3 that occur multiple times are merged into a single node prior to generating levels 4 and 5. This step effectively preserves cycles of length four that are present in the original graph. Different from a formal tree structure, it is thus possible that multiple nodes at level 2 point to the same node at level 3. We therefore refer to this data structure as a "directed layered graph" (DLG). Similarly, replicate nodes in level 5 are merged if they are (indirectly) connected to the same node in layer 3, which again preserves cycles of length four.

For each key node in level 5 that consists of the same number of fragments as the root node, a set of unique five-node paths to the root node is determined. Only those paths are retained where edges connecting nodes at levels 1 and 2 and levels 4 and 5, respectively, are associated with the same value fragment. For a key node occurring multiple times at level 5, all sets of identified paths are merged. During this step, only paths exclusively consisting of molecule nodes not found in any other path are retained. If the final set contains more than two paths, a parallel series consisting of at least three compound pairs has been identified. The terminal key nodes of these paths form the head node pair of this parallel series. Then, the entire procedure is repeated by selecting the next unexplored key node as the root for the search that is iteratively continued until all key nodes have been explored. An outline of the search algorithm is provided in Figure 4. Applying this search routine, compounds forming all spindles with canonical topology are extracted from the BMMSG representation of a data set. On the basis of the MMP generation procedure, as discussed above, transfer series can also be identified in which the SAR is influenced by stereochemical criteria (data not shown).

Search Trials. We have applied the methodology to search for parallel series in 32 compound data sets from BindingDB that met our selection criteria. In all data sets, varying numbers of spindle subgraphs were identified, ranging from seven to 1462. A total of 5857 parallel series were identified, and 2770 of these series (~47%) qualified for transfer ranking (the remaining series did not cover a potency range of at least 1 order of magnitude). However, all data sets contained qualifying parallel series, and their numbers ranged from three to 678 per set. Hence, on the basis of our analysis, parallel series were frequently detected across different data sets. This suggests that

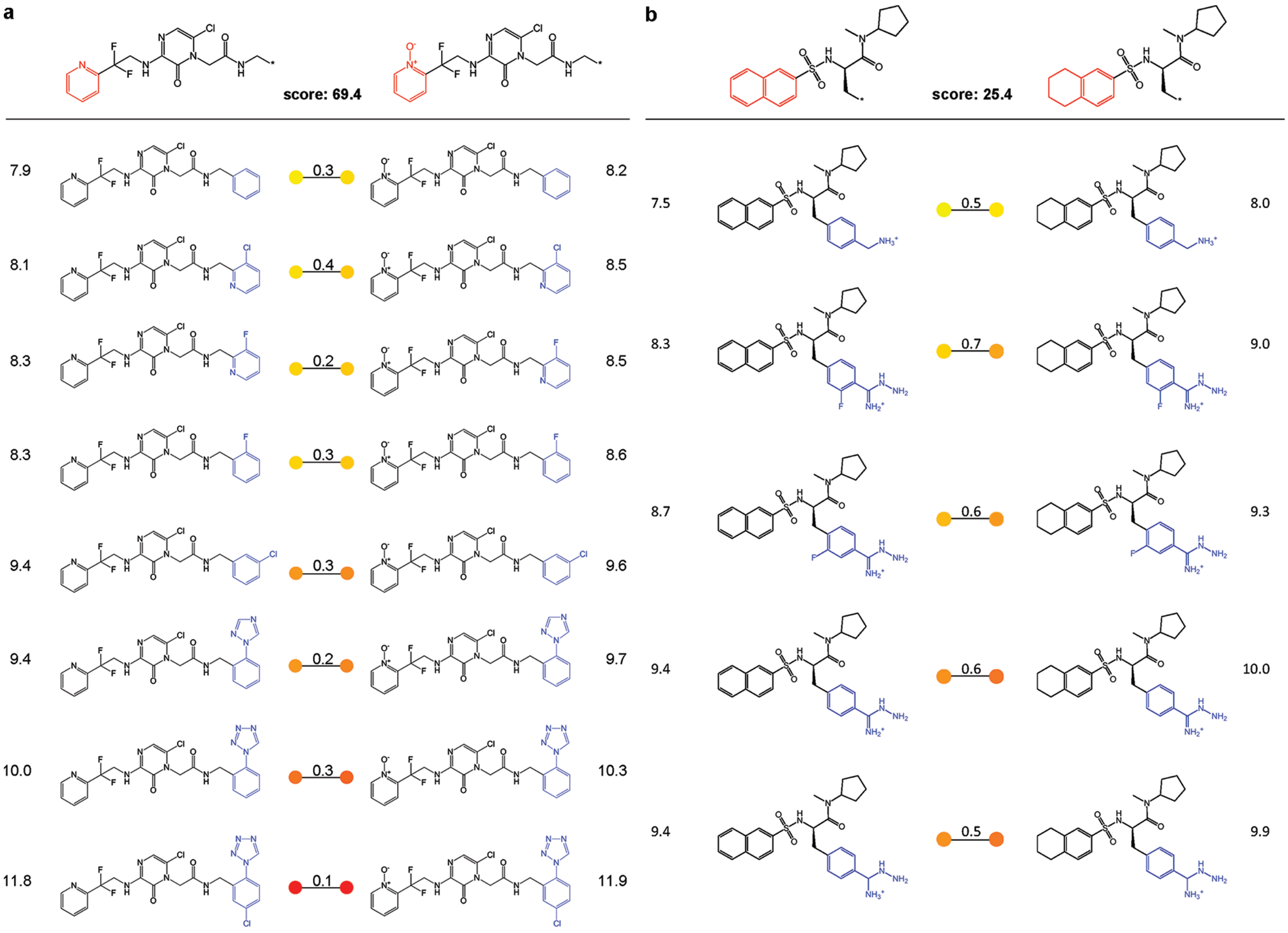


Figure 6. Thrombin inhibitor series. For a set of thrombin inhibitors, two parallel series with high SAR transfer potential (high scores) are shown. The representation is according to Figure 5.

parallel series should be considered as a source of potentially valuable SAR information.

SAR Transfer Potential. For each data set, qualifying parallel series were aligned and scored to assess their SAR transfer potential. Our scoring function effectively discriminated between series with high and low transfer potential, as illustrated in Figure 5 for parallel series extracted from a dopamine D3 receptor antagonist data set. The series in Figure 5a consists of four pairs of analogs, yields a high score (21.1), and shows very regular potency progression over 2 orders of magnitude. By contrast, the series in Figure 5b, which consists of five analog pairs (with one series spanning a similar potency range) obtains a low score (4.7). In this series, potency differences between corresponding analogs vary significantly, and no notable SAR transfer potential is observed. These examples are representative for all data sets. We generally found that the applied scoring scheme was sensitive to differences in SAR transfer potential and produced meaningful and chemically intuitive rankings of identified parallel series. Examining the score distribution over all data sets, a transfer score greater than 15 was generally found to be an indicator of clear SAR transfer events. This threshold was conservatively set; many parallel series yielding scores below this threshold also displayed at least partial SAR transfer potential. However, in 26 of our 32 data sets, SAR transfer series with scores greater than 15 were identified, and many of these series should be of immediate interest for medicinal chemistry.

SAR Transfer Series. In Table 1, qualifying parallel and SAR transfer series are listed for all activity classes. Many activity classes contained between approximately 5 and 60 qualifying parallel series. In a few cases, especially antagonists of adenosine receptor isoforms, several hundred parallel series were detected. In general, the numbers of high-priority SAR transfer series were much smaller, as one should expect. Here, sulfonamide-containing inhibitors of carbonic anhydrase I and II were an exception, with approximately 200 transfer series in each case. If these rather unusual data sets were excluded from the comparison, between one and 23 SAR transfer series were identified in 24 of the remaining 30 compound sets, with an average of close to six series per set. In six cases, only a single SAR transfer series was identified. In Figure 6, representative examples of highly ranked series from a thrombin inhibitor set are shown. The top-ranked series in Figure 6a (score 69.4) consists of eight analog pairs and reveals text book SAR transfer characteristics with relatively small and very similar incremental increases in potency along the alignment over a range of 4 orders of magnitude. The sixth-ranked parallel series in Figure 6b consists of five analog pairs with larger but also very similar potency increases over a range of approximately 2 orders of magnitude. Hence, this series also displays clear SAR transfer characteristics, as reflected by a high SAR transfer score of 25.4.

■ CONCLUDING REMARKS

Herein we have introduced an approach for the identification of parallel and SAR transfer series in compound data sets. The assessment of SAR transfer potential of analog series is a topic of high interest in medicinal chemistry, especially in the context of lead optimization. Although SAR transfer considerations typically focus on individual compound series, the assessment of SAR transfer events is also interesting from a data mining perspective. In this case, the aim is to systematically search available compound data for potential SAR transfer series and thereby provide knowledge for medicinal chemistry applications. For this purpose, we have developed a graph mining algorithm that makes it possible to extract all possible parallel series from compound network representations. The graph mining algorithm currently is the only generally applicable computational approach to systematically investigate SAR transfer. Potential limitations of the approach include that subgraphs might be detected that do not represent SAR transfer events, which is compensated for by the introduction of a scoring scheme. In addition, in the present version of the algorithm, the structural environment of head nodes is not explored. The design of a graph mining method for the assessment of SAR transfer has in part been motivated by our interest in substructure relationship-based SAR networks that are readily interpretable in chemical terms. The automated extraction of all parallel series complements network analysis. Candidate series were assessed with our newly designed scoring function and ranked on the basis of their SAR transfer potential. In total, we identified 142 high-priority SAR transfer series in 24 different compound data sets. In two other (unusual) cases, ~200 transfer series were detected, and six compound data sets did not contain SAR transfer series. However, in the majority of data sets studied here, at least a few transfer series were identified (on average, close to six per set). Thus, SAR transfer series are frequently found in compound data sets and present a valuable source of SAR information. Attractive transfer series can also be mapped back on networks from which they originate to study their structural environments and associated SAR information. In our analysis, we have shown that the combination of graph mining and scoring effectively identifies chemically intuitive SAR transfer series in large compound data sets. Thus, the methodology introduced herein is expected to complement the analysis of SAR networks and further improve our ability to identify SAR transfer events through large-scale data mining.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +49-228-2699-306. Fax: +49-228-2699-341. E-mail: bajorath@bit.uni-bonn.de.

Present Address

#Chemical Biology Program and Platform, Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

A.M.W. is supported by Boehringer Ingelheim.

■ REFERENCES

(1) *The Practice of Medicinal Chemistry*, 3rd ed.; Wermuth, C. G., Ed.; Academic Press-Elsevier: Burlington, San Diego, USA; London, UK, 2008.

(2) Wess, G.; Urmann, M.; Sickenberger, B. Medicinal Chemistry: Challenges and Opportunities. *Angew. Chem., Int. Ed.* **2001**, *40*, 3341−3350.

(3) Wawer, M; Lounkine, E.; Wassermann, A. M.; Bajorath, J. Data Structures and Computational Tools for the Extraction of SAR Information from Large Compound Sets. *Drug Discovery Today* **2010**, *15*, 630−639.

(4) Wassermann, A. M.; Wawer, M.; Bajorath, J. Activity Landscape Representations for Structure-Activity Relationship Analysis. *J. Med. Chem.* **2010**, *53*, 8209−8223.

(5) Wawer, M.; Bajorath, J. Local Structural Changes, Global Data Views: Graphical Substructure-Activity Relationship Trailing. *J. Med. Chem.* **2011**, *54*, 2944−2951.

(6) Wassermann, A. M.; Bajorath, J. A Data Mining Method to Facilitate SAR Transfer. *J. Chem. Inf. Model.* **2011**, *51*, 1857−1866.

(7) Sheridan, R. P. The Most Common Replacements in Drug-like Compounds. *J. Chem. Inf. Model.* **2002**, *42*, 103−108.

(8) Kenny, P. W.; Sadowski, J. Structure Modification in Chemical Databases. In *Chemoinformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH: Weinheim, Germany, 2005; pp 271−285.

(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) *OEChem TK* version 1.7.4.3; OpenEye Scientific Software Inc.: Santa Fe, NM, 2010.

(11) Java Universal Network/Graph Framework, version 2.0.1. http://jung.sourceforge.net/ (accessed Jan 11, 2010).

(12) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: a Web-Accessible Database of Experimentally Determined Protein−Ligand Binding Affinities. *Nucleic Acids Res.* **2007**, *35*, D198−D201.

■ NOTE ADDED AFTER ASAP PUBLICATION

This paper was published ASAP on April 3, 2012 with incorrect versions of Figures 2, 4, 5, and 6. The corrected paper was published ASAP on April 4, 2012.