Journal of Medicinal Chemistry

ARTICLE

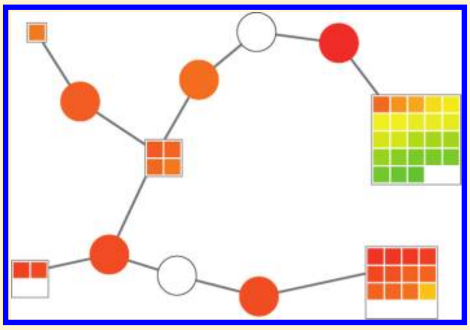
pubs.acs.org/jmc

Local Structural Changes, Global Data Views: Graphical Substructure–Activity Relationship Trailing

Mathias Wawer 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 systematic extraction of structure–activity relationship (SAR) information from large and diverse compound data sets depends on the application of computational analysis methods. Irrespective of the methodological details, the ultimate goal of large-scale SAR analysis is to identify most informative compounds and rationalize structural changes that determine SAR behavior. Such insights provide a basis for further chemical exploration. Herein we introduce the first graphical SAR analysis method that globally organizes large compound data sets on the basis of local structural relationships, hence providing an immediate access to important structural modifications and SAR determinants.



INTRODUCTION

Computer-aided analysis and visualization of SAR information contained in large compound data sets have been increasingly investigated topics in recent years.[1,2] For this purpose, different graphical analysis methods have been introduced, such as SAR maps,[3] structure–activity landscape index graphs,[4] or network-like similarity graphs.[5] Some of these methods are designed to globally view similarity and potency relationships in large compound data sets, identify activity cliffs,[4,5] or study the relationship between global and local SAR characteristics.[5] Informative local SAR environments can be further studied, for example, using a data structure termed a similarity-potency tree[6] that monitors structural nearest neighbor and potency relationships in a compound-centric, rather than global, manner.

Regardless of the methodological details, all SAR analysis methods must take into account similarity relationships between active compounds. To represent individual analogue series, standard R-group decomposition can be applied and numerical similarity measures are not essential. However, when compound data sets grow in size and become structurally diverse, the requirements change. All currently available numerical or graphical analysis methods that provide SAR views of large data sets have in common that they account for compound similarity on a whole-molecule basis, usually by calculating Tanimoto similarity (using different molecular representations) between active compounds in a pairwise manner. As a consequence, although compound subsets that are rich in SAR information are detected and visualized using these methods, structural changes that yield interpretable SAR patterns must generally be analyzed subsequently, following the preselection of compound subsets that introduce local or global SAR discontinuity.[7] Of course, uncovering structural modifications that yield defined SAR phenotypes and highly potent compounds is of cardinal importance for medicinal chemistry applications.

Therefore, we have designed a methodology for large-scale SAR analysis that does not rely on numerical compound similarity assessment but directly accounts for structural relationships between active compounds as an organizing principle. Therefore, we have initially generalized the matched molecular pair (MMP)[8] formalism as a compound similarity criterion. An MMP is defined as a pair of compounds that only differ at a single site such as a specific R-group or ring system. Hence, compounds forming an MMP are distinguished by a defined substructure, and the exchange of this substructure represents a converting chemical transformation. Applying this compound similarity criterion, we have then designed a potency-annotated bipartite graph representation that, for the first time, globally organizes compound data sets focusing on local compound substructure relationships. Herein, we describe the design of this data structure and illustrate its utility in an exemplary application on a large compound set.

MATERIALS AND METHODS

Matched Molecular Pairs. MMPs were calculated according to Hussain and Rea.[9] The algorithm generates molecular fragments by deleting acyclic single bonds and stores them as key–value pairs in an index table. If one single bond is deleted, a molecule is separated into two fragments. Each of these fragments is inserted once as a key in the index table and the other as the associated value. In the simplest case, two molecules forming an MMP differ only in one R-group attached to a common core via a single bond. During fragmentation and indexing, these R-groups are associated with the same key (common core). Thus, once the entire data set has been processed, all MMPs can be identified from the index table by searching for keys with more than one value. In addition to single bonds, bond pairs and triplets are also deleted, resulting in the formation of a core fragment and two ("double cut")

Received: January 11, 2011
Published: March 28, 2011

or three ("triple cut") substituents. These substituents are then collectively stored as a key and the core as the value. In our current implementation, keys are permitted to consist of maximally 10 heavy atoms. Additionally, the value fragment is not allowed to contain more heavy atoms than the corresponding key. Both thresholds can be easily modified to meet a specific analysis objective. MMP generation and molecule visualizations were implemented in Java using the OpenEye chemistry toolkit.[10]

Graph Generation. The graphs are constructed on the basis of an MMP index table. They contain two different types of objects as nodes: (1) *keys* that correspond to the key fragments of the MMP index table and (2) *molecules*. Only keys associated with more than one value are considered. Keys are connected by an edge to all compounds that contain the respective key fragment. The size threshold of 10 heavy atoms is not applied in this step to also include structures with larger substituents attached to the key fragment. Therefore, in our implementation, we ultimately include all molecules by adding relevant value fragments above the size threshold to the index in a subsequent step. However, the second constraint that limits the size of the value relative to the size of the key must generally be met. Edges are associated with the value fragment of the respective key−molecule pair. Because connections are only formed between two different types of objects, keys and molecules, this data structure represents a bipartite graph. If two keys are connected to the same set of molecules, the less specific key (i.e., the one associated with the larger value fragment for each of the compounds) is removed, which reduces the complexity of the graph by omitting redundant information. In addition, key nodes that connect to compound subsets of another node and nonconnected nodes (singletons) are removed. Subset relationships are stored in a separate hierarchical treelike graph that contains only key nodes. Here, a key is the successor of another if it connects to a subset of its neighbors. The graph structures were implemented using the Java package JUNG.[11]

Graph Visualization. For graph visualization, the molecule nodes are colored by potency according to a continuous gradient from green to red, reflecting the lowest and highest potencies in the data set, respectively. Key nodes are colored according to their "cut level": white, light blue, and dark blue nodes indicate keys resulting from single, double, and triple cuts, respectively. For clarity, molecules connected to only one key are not shown as separate nodes but are combined to a "supernode" that represents this key as a rectangle containing a square for each molecule that is colored by potency. Additionally, edges are colored according to the cut level of the corresponding key node. The graph layout is generated using the JUNG implementation of a self-organizing map (SOM) algorithm, and every connected component of the graph is laid out separately. The graph layout can be interactively edited.

RESULTS AND DISCUSSION

Methodological Concept. The method introduced herein is designed to represent the global composition of a compound data set and its potency distribution by focusing on local substructure matches. On the basis of the MMP index table, molecules are organized into structural sets. Each set contains all compounds that differ only by a single modification at a specific site. In the following, we refer to these sets as matching molecular series (MMS). These sets often overlap because a compound that differs at one site from a number of molecules might differ at another site from others. Such a compound would then belong to two MMS. By systematically generating all MMS for a compound data set, structural relationships contained in this set are comprehensively accounted for. A bipartite graph structure has been designed to represent the composition of MMS and the relationships formed between them. Furthermore, the bipartite graph is annotated with compound potency information. The complete graph representation is termed a bipartite MMS (BMMS) graph.

In addition to the MMP concept that provides the basis for MMS and bipartite graph generation, other structural organization schemes have also been introduced. These include classical R-group decomposition of analogue series (as utilized, for example, for the generation of SAR maps[3] or combinatorial analogue graphs[12]), hierarchical scaffold generation,[13] and the scaffold tree data structure.[14] In the scaffold tree, rings are iteratively removed from initially generated hierarchical scaffolds according to predefined chemical rules until only an individual ring remains. Hence, the scaffold tree captures hierarchical substructure relationships between scaffolds along rule-based decomposition pathways. For our major purpose, i.e., the replacement of calculated molecular similarities in SAR-relevant compound network representations with directly accessible structural relationships, the MMP concept has been the preferred choice due to its generality.

Bipartite Graph Representation. In the BMMS graph, "key nodes" represent MMS. A key is the substructure common to all molecules in a series. Individual compounds are represented by "molecule nodes". Each molecule of a series is connected to the corresponding key node by an edge. Key nodes are graphically annotated with their substructure, and edges are annotated with the substitution that distinguishes a molecule from its key. All MMS a molecule belongs to are identified by the keys it is connected to in the graph. Molecule nodes are color-coded according to compound potency. The BMMS graph provides a global view of the structural and potency relationships contained in a data set. Figure 1 shows how this data structure is generated and how the graph representation looks. In Figure 1a, the graph of a model compound set is shown that includes eight possible keys. To simplify the graph structure, redundant key nodes are removed from the graph. In this example, keys 2 and 3 as well as 5 and 6 describe the same sets of molecules. In both cases, the key associated with the more general substructure is removed (3 and 6, respectively) because it does not provide additional structural information. Hence, such keys are considered redundant. Furthermore, key 5 describes a subset of the compounds connected to key 7. Therefore, key 5 is also removed from the graph. The final reduced BMMS graph is shown in Figure 1b. To further simplify the graphical representation, molecules only connected to a single key are not drawn as individual nodes but combined into a supernode that represents this key as a rectangle containing squares (molecules) colored by potency (this symbol is also used for a single compound that is only connected to one node). Although key 5 is removed from the graph, as discussed above, the subset relationship between key 5 and key 7 is recorded in the subset hierarchy, as shown in Figure 1c. The hierarchy exclusively consists of key nodes and is part of the data structure. Its graphical representation complements the information contained in the BMMS graph, as further illustrated below.

SAR Patterns. A characteristic feature of the BMMS graph and its associated hierarchy is that these graph representations contain *signature patterns* (subgraphs) that reveal detailed SAR information. This feature is of central relevance for SAR analysis. The signature patterns are schematically illustrated in Figure 2.

First, substitution sites having a large effect on compound potency ("SAR hot spots") can be identified by searching for *key nodes connected to compounds that cover a broad potency range*

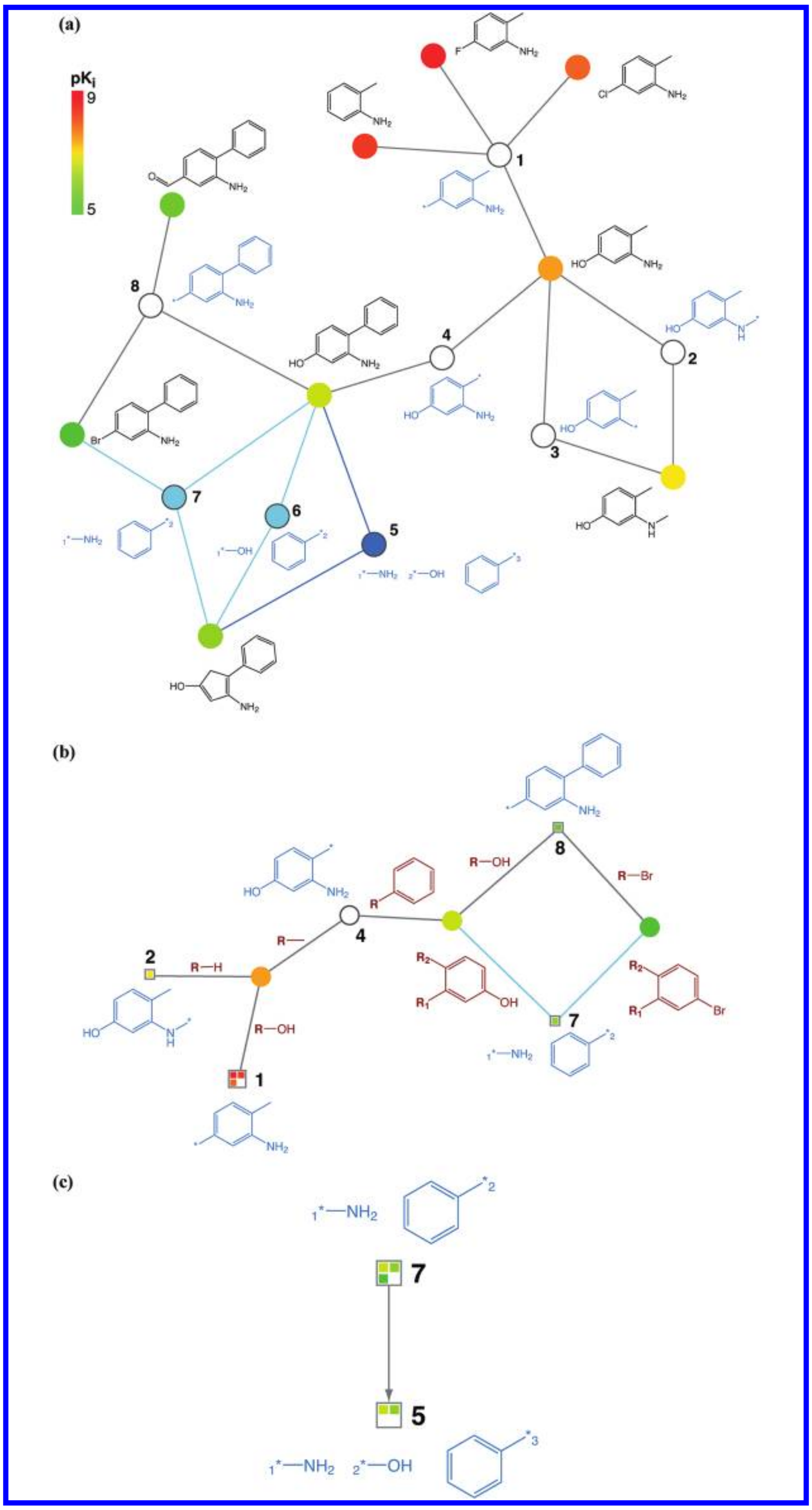


Figure 1. BMMS graph structure. Schematic illustrations of the BMMS graph structure are shown. (a) An unprocessed graph containing all possible key nodes was calculated for a model data set. Key nodes (numbered from 1 to 8) are colored according to their cut level in white, light blue, or dark blue for single, double, and triple cuts, respectively (see the Materials and Methods). Furthermore, molecule nodes are colored by potency according to a color spectrum from green (lowest potency) to red (highest potency) as indicated by the color bar on the left. All compound structures (black) and shared substructures (blue) that are associated with key nodes are shown next to the corresponding nodes. Asterisks in key node substructures mark attachment points for variable substituents that occur in the compound series. (b) The processed graph is shown after removal of (1) key nodes that describe the same compound set or (2) a subset of another node. In addition, (3) molecule nodes only attached to one key are combined into a multicompound key symbol (supernodes). The substructures of all keys are shown in blue, and the variable chemical groups that distinguish a molecule from its key are drawn in red next to their connecting edge. (c) The key subset hierarchy for the model data set is displayed and annotated with substructures.

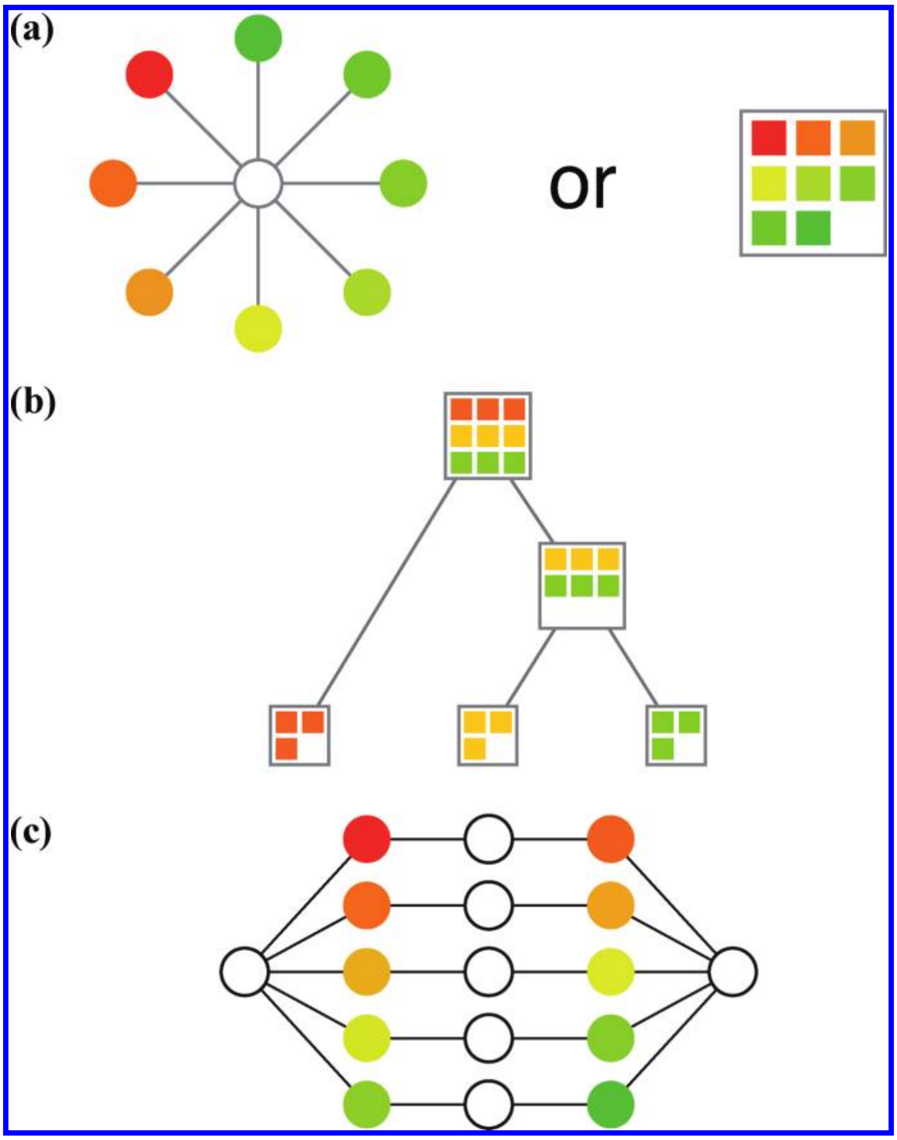


Figure 2. BMMS graph SAR signature patterns. (a) SAR hot spots appear as key nodes connected to molecules that cover a broad potency range (left). These key nodes might be represented as supernodes (right; see also Figure 1b). (b) An exemplary subset hierarchy pattern is shown where an MMS is separated into four subseries that distinguish highly (orange/red), moderately (yellow), and weakly (green) potent compounds from each other. Because each key in the hierarchy is associated with a substructure, increasingly subset-specific substructures along the hierarchy reveal potency-determining structural changes. (c) An exemplary pattern describing two "parallel series" is shown, i.e., sets of molecules that differ in one site and have additionally been modified at another site with the same set of substituents. In the graph, such series are easily identified by repeating molecule—key—molecule paths of length three. The key node in the path always connects the corresponding molecule pairs and marks the site where the two series differ.

(Figure 2a). The position of the substitution site is provided by the substructure associated with the corresponding key node.

Second, structural changes responsible for observed potency effects are revealed by *hierarchical supernode patterns* (Figure 2b). In this pattern, series containing highly and weakly potent compounds yield successively smaller subsets that ultimately separate molecules with different potencies from each other. The substructures associated with the key nodes then reveal favorable substitution sites and R-groups.

Third, the occurrence of multiple series of compounds modified at the same site with overlapping sets of substituents can be detected. These series occur as *key nodes connected by several molecule—key—molecule paths* of length three (Figure 2c). Thus, subsets of compounds modified at distinct substitution sites can be immediately identified, and how substitutions at different sites alter compound potency can be examined.

It is important to note that these characteristic SAR patterns are an intrinsic feature of the BMMS data structure. Their detection in the graphs is sufficient to extract interpretable SAR information from compound sets, if it is available. As further discussed below, these patterns immediately identify structural modifications that are responsible for potency alterations.

Exemplary Application. The method is applicable to large compound data sets. For example, it was applied herein to analyze a set of 881 factor Xa inhibitors from BindingDB.[15] The BMMS graph representing the entire data set is shown in Figure 3. It consists of 23 connected components containing a total of 858 compounds. Twenty-three compounds did not form

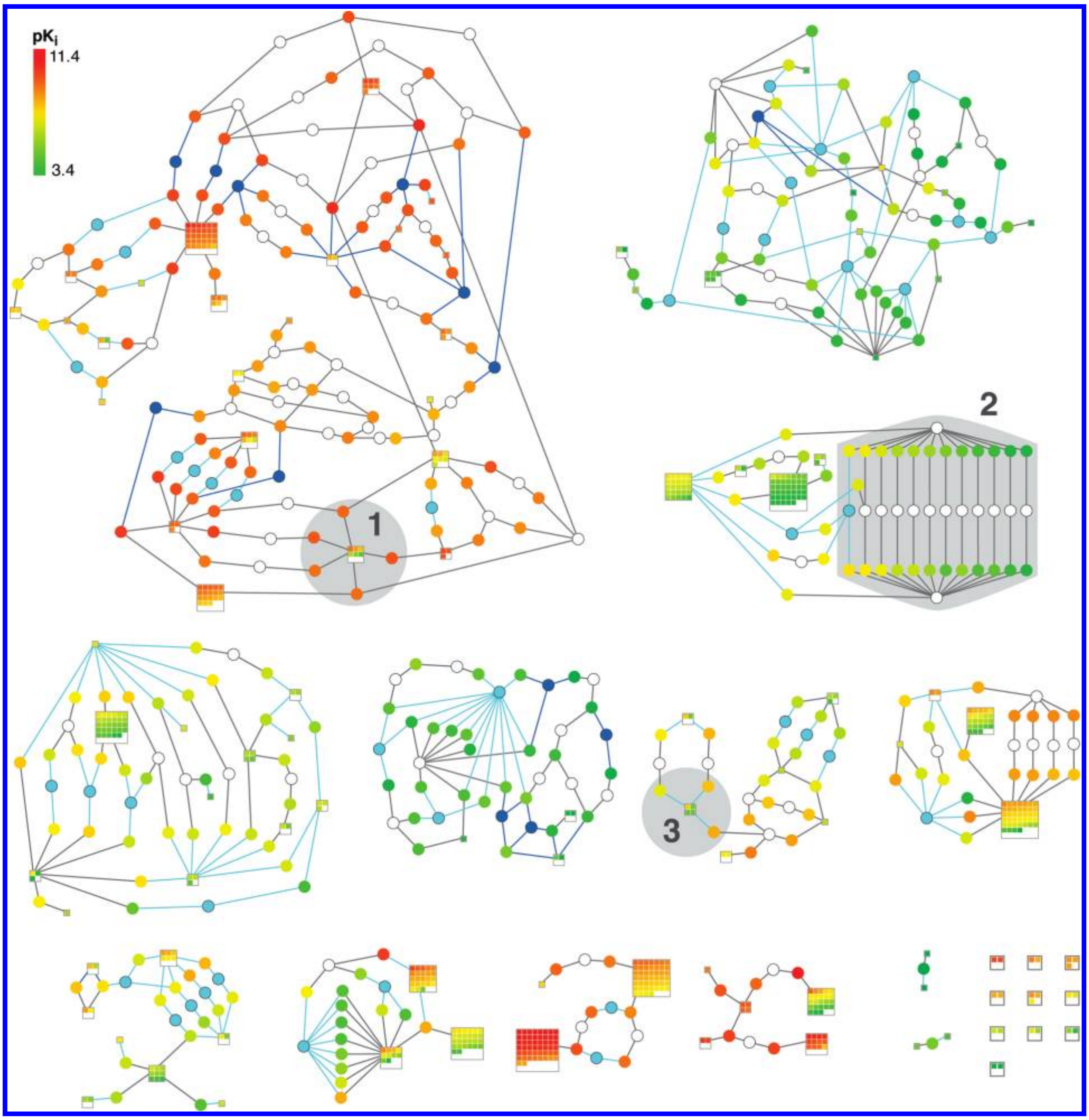


Figure 3. BMMS graph of a factor Xa inhibitor set. The graph contains 858 compounds distributed over 23 connected components. Selected regions are highlighted and shown in detail in Figures 4 and 5.

an MMP, and these singletons were omitted because they do not convey SAR information. Disjoint subgraphs are formed because molecules in distinct graph components differ by more than one structural modification and are hence not connected.

Many components of the factor Xa graph are found to predominantly contain similarly colored molecule nodes. In individual components, many molecules belonging to the same series show little potency variation. Only in a few cases, green and red nodes are connected to the same key. Such combinations of green and red nodes form "activity cliffs".[16] The regular potency distribution in the factor Xa graph indicates that changes in compound structure are in this case mostly (but not exclusively) accompanied by gradual changes in potency, consistent with the presence of substantial SAR continuity.[16]

Although large potency differences between structurally related compounds are rare in the factor Xa data set, several regions in the BMMS graph resemble the characteristic pattern outlined in Figure 2a and represent SAR hot spots. A representative example is the highlighted region 1 in Figure 3 shown in detail in Figure 4a. Here, the para substituent of a benzyl group emerges as an SAR hot spot. For a detailed analysis of individual substituents, the subset hierarchy is used to search for a pattern resembling the one in Figure 2b. The section of the hierarchy containing this SAR hot spot is shown in Figure 4b. The key node at the top represents a series of 11 compounds. Eight of these compounds have medium to high potency and three compounds only low potency. Following the branches down the hierarchy, a progressive separation of highly and weakly active compounds is observed.

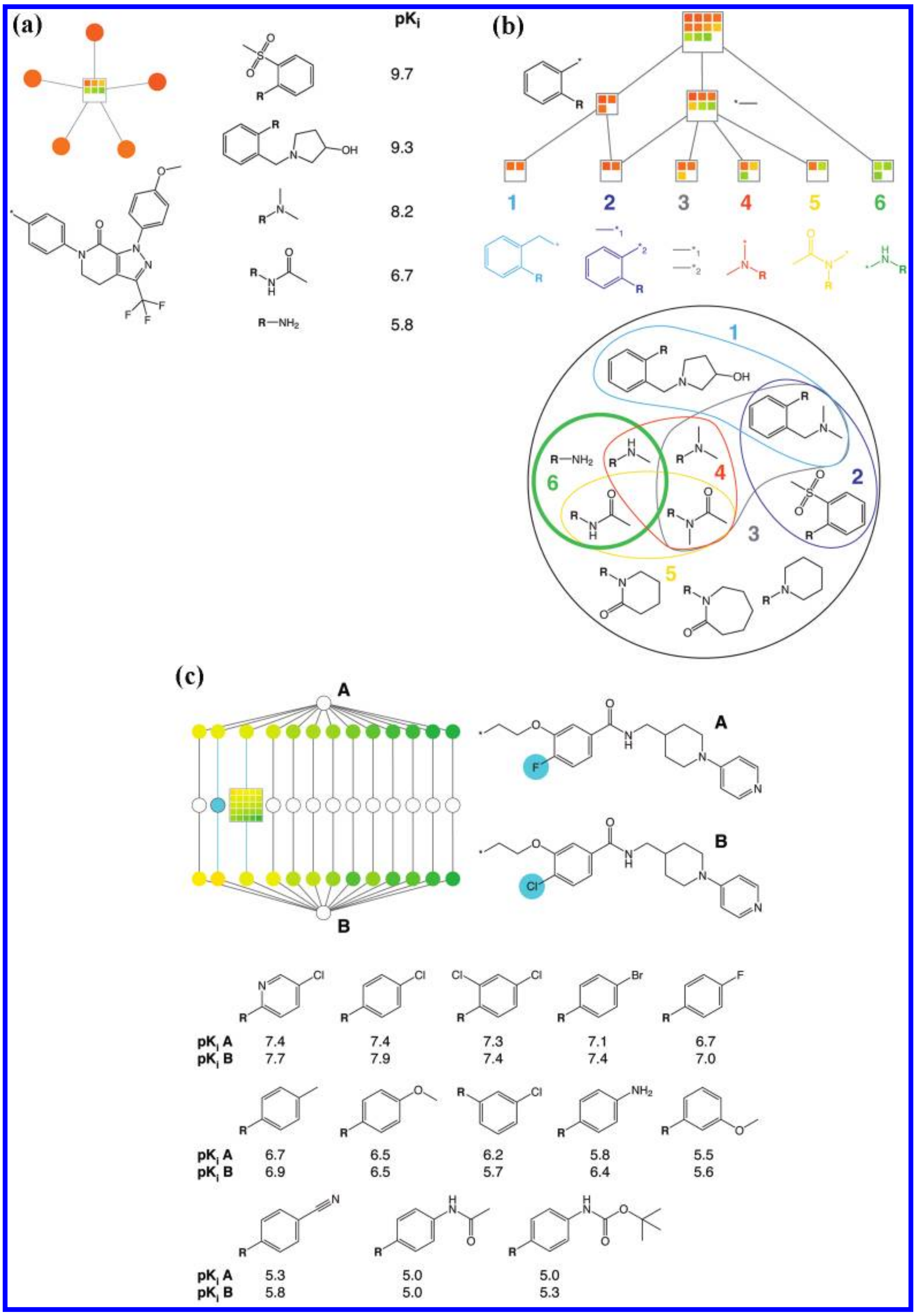

(a)
pKi
9.7
9.3
8.2
6.7
5.8
(b)
1
2
3
4
5
6
(c)
A
B
pKi A 7.4 7.4 7.3 7.1 6.7
pKi B 7.7 7.9 7.4 7.4 7.0
pKi A 6.7 6.5 6.2 5.8 5.5
pKi B 6.9 6.5 5.7 6.4 5.6
pKi A 5.3 5.0 5.0
pKi B 5.8 5.0 5.3

Figure 4. Informative SAR patterns for the factor Xa inhibitor set. The series shown here were identified by searching the graph for signature patterns presented in Figure 2. (a) An SAR hot spot is shown with its associated substructure and exemplary substituents. The potency of the corresponding compounds is reported as the pK_i value. (b) The subset hierarchy for the series in (a) is displayed. The top node of the hierarchy represents the entire series that is described by the substructure shown in (a). Each of the keys below represents a more specific substructure. For simplicity, only the groups added in each key node to the general substructure are reported. Sites where these groups are attached to the general substructure are marked with an "R". In addition, the substituents of the compounds in this series and their subset relationships are shown in a Venn diagram. For clarity, only the sets defined by the terminal key nodes are considered (and numbered). (c) Two parallel series are shown that correspond to the pattern in Figure 2c. The nodes have been ordered according to decreasing potency for series A from left to right. The common substructures of these series and their R-groups are displayed in the same order as the molecule nodes. The potency of the corresponding molecules is reported as the pK_i value.

The three weakly potent compounds are found to contain a common substructure that distinguishes them from the other more potent compounds. It is evident that primary and secondary amines and amides are unfavorable substituents (set 6 in Figure 4b, highlighted

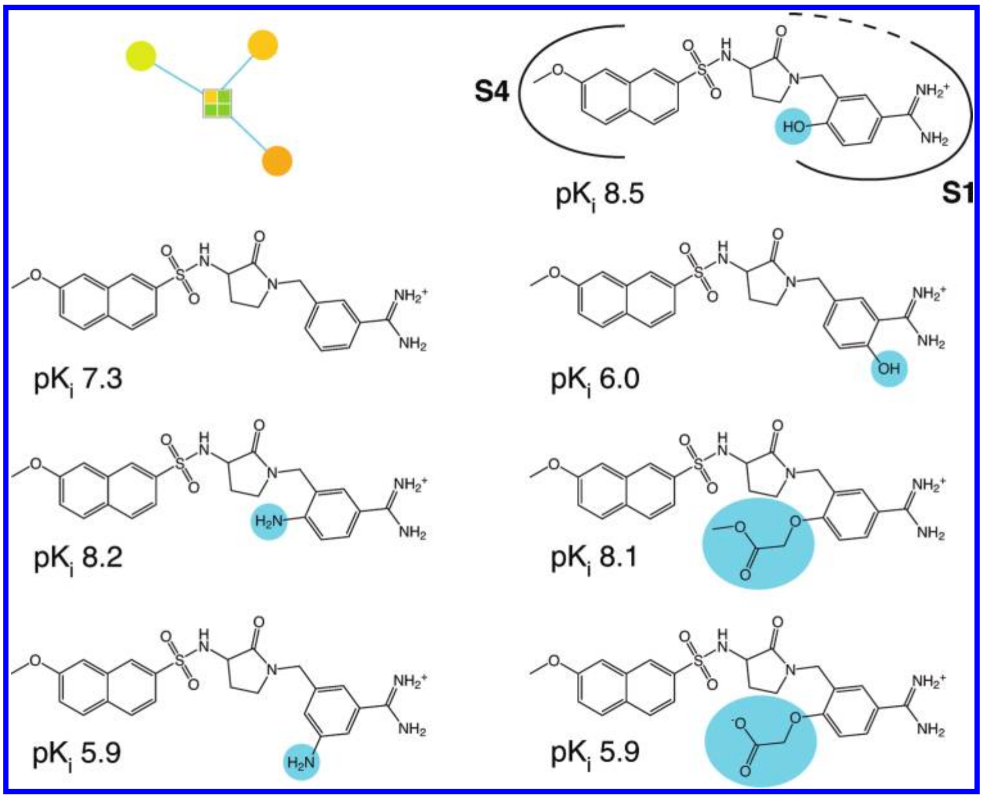


Figure 5. Structural changes leading to potency alterations. Shown is the BMMS subgraph (region 3 in Figure 3) for a series of factor Xa inhibitor analogues whose structural differences are highlighted. X-ray data reveal that major interaction sites of these compounds in the active site of factor Xa include the S1 and S4 pockets, as indicated for a crystallographically characterized analogue.

by thick green circle). Other R-groups are less critical in this case because the remaining compounds carry structurally diverse substituents but are all moderately to highly potent.

In the hierarchy, a node with more than one predecessor combines the individual features of its parental key nodes and thus describes the overlap between the corresponding compound series, as illustrated for node 2 in Figure 4b. Thus, for SAR hot spots identified in BMMS graphs, the analysis of their hierarchical ordering reveals detailed structural relationships between compounds having different potencies.

In Figure 4c, two compound series are shown with parallel modifications at the same site, yielding the pattern in Figure 2c. They correspond to the highlighted region 2 in Figure 3. The difference between the two series is a halogen substitution of fluorine (series A) by chlorine (series B), which is represented by each key node that connects corresponding compounds. The color distribution reveals that the potencies of molecules carrying the same substituent are generally similar. In both series the potency of the compounds changes in the same direction and gradually increases, as revealed by the pattern. It becomes clear that para-substituted benzyl (or pyridinyl) groups are preferred substituents and that these groups in most potent compounds carry a halogen substituent, preferentially chlorine. The potency difference between a meta- and para-substituted chlorobenzyl is approximately 1 (series A) or 2 (series B) orders of magnitude.

Two of the keys that link the two series encode a substructure obtained by deleting two single bonds (double cut), i.e., the blue node in Figure 4c and the adjacent supernode. In these cases, not only the halogen substitution occurs, but the entire substituted ring structure might be replaced. Thus, more extensive modifications at the second site can be explored by analyzing these series. Thus, this parallel series pattern is rich in SAR information and provides direct access to structural changes at defined sites that gradually increase compound potency.

Figure 5 summarizes how SAR information is practically extracted from BMMS graphs. A subgraph representing an interesting SAR pattern, as discussed above, is shown for the factor Xa data set (region 3 highlighted in Figure 3). This subgraph contains a series of analogous inhibitors where R-group modifications at a single site lead to potency variations spanning nearly 3 orders of magnitude. The X-ray structure of one of these analogues bound to factor Xa[17] reveals that these compounds intensively interact with the S1 and S4 pockets in the active site of the enzyme, as indicated in Figure 5. As can be seen, the modifications within this series of analogues that cause a significant degree of SAR discontinuity[16] predominantly affect interactions in the S1 site of factor Xa. Hence, the SAR trend revealed by the BMMS subgraph can also be rationalized in light of the experimentally observed binding mode of one of these inhibitors.

CONCLUSIONS

Herein we have introduced a graphical SAR analysis method that systematically organizes compound data sets on the basis of local substructure relationships. The underlying data structure does not depend on whole-molecule similarity calculations. The BMMS graph representation contains characteristic subgraph patterns that capture detailed SAR information and reveal important structural modifications. Associated graphs of key node hierarchies complement the SAR information obtained from

BMMS graphs. Subgraphs representing well-defined SAR patterns are an intrinsic feature of this data structure. Hence, in practical applications, BMMS graphs and key hierarchies of compound data sets are searched for such patterns. If they are present, a data set contains interpretable SAR information and the underlying structural modifications can be readily accessed. The exemplary analysis of the large factor Xa data set presented herein illustrates all components and analysis steps required to extract SAR information from BMMS graphs of compound data sets.

AUTHOR INFORMATION

Corresponding Author

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

ACKNOWLEDGMENT

The authors would like to thank Anne Mai Wassermann and Peter Haebel for helpful discussions. M.W. is supported by Boehringer Ingelheim.

ABBREVIATIONS USED

MMP, matched molecular pair; MMS, matching molecular series; BMMS graph, bipartite matching molecular series graph; SAR, structure−activity relationship; SOM, self-organizing map

REFERENCES

(1) Peltason, L.; Bajorath, J. Systematic Computational Analysis of Structure-Activity Relationships: Concepts, Challenges and Recent Advances. *Future Med. Chem.* **2009**, *1*, 451–466.

(2) Bajorath, J.; Peltason, L.; Wawer, M.; Guha, R.; Lajiness, M. S.; Van Drie, J. H. Navigating Structure-Activity Landscapes. *Drug Discovery Today* **2009**, *14*, 698–705.

(3) Agrafiotis, D.; Shemanarev, M.; Connolly, P.; Farnum, M.; Lobanov, V. SAR; Maps, A; New, S. A. R. Visualization Technique for Medicinal Chemists. *J. Med. Chem.* **2007**, *50*, 5926–5937.

(4) Guha, R.; Van Drie, J. H. Structure-Activity Landscape Index: Identifying and Quantifying Activity Cliffs. *J. Chem. Inf. Model.* **2008**, *48*, 646–658.

(5) Wawer, M.; Peltason, L.; Weskamp, N.; Teckentrup, A.; Bajorath, J. Structure-Activity Relationship Anatomy by Network-like Similarity Graphs and Local Structure-Activity Relationship Indices. *J. Med. Chem.* **2008**, *51*, 6075–6084.

(6) Wawer, M.; Bajorath, J. Similarity-Potency Trees: A Method To Search for SAR Information in Compound Data Sets and Derive SAR Rules. *J. Chem. Inf. Model.* **2010**, *50*, 1395–1409.

(7) 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.

(8) Kenny, P. W.; Sadowski, J. Structure Modification in Chemical Databases. In *Chemoinformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH: Weinheim, Germany, 2004; 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) Peltason, L.; Weskamp, N.; Teckentrup, A.; Bajorath, J. Exploration of Structure-Activity Relationship Determinants in Analogue Series. *J. Med. Chem.* **2009**, *52*, 3212–3224.

(13) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.

(14) Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. A.; Waldmann, H. The Scaffold Tree—Visualization of the Scaffold Universe by Hierarchical Scaffold Classification. *J. Chem. Inf. Model.* **2007**, *47*, 47–58.

(15) 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.

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

(17) Choi-Sledeski, Y. M.; McGarry, D. G.; Green, D. M.; Mason, H. J.; Becker, M. R.; Davis, R. S.; Ewing, W. R.; Dankulich, W. P.; Manetta, V. E.; Morris, R. L.; Spada, A. P.; Cheney, D. L.; Brown, K. D.; Colussi, D. J.; Chu, V.; Heran, C. L.; Morgan, S. R.; Bentley, R. G.; Leadley, R. J.; Maignan, S.; Guilloteau, J.-P.; Dunwiddies, C. T.; Pauls, H. W. Sulfonamidopyrrolidinone Factor Xa Inhibitors: Potency and Selectivity Enhancements via P-1 and P-4 Optimization. *J. Med. Chem.* **1999**, *42*, 3572–3587.