JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

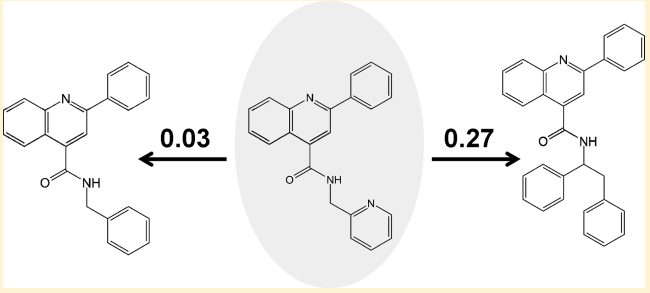
Development of a Method To Consistently Quantify the Structural Distance between Scaffolds and To Assess Scaffold Hopping Potential

Ruifang Li,[†] Dagmar Stumpfe,[†] Martin Vogt,[†] Hanna Geppert,[†] 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

Supporting Information

ABSTRACT: We introduce a method to determine a structural distance between any pair of molecular scaffolds. The development of this approach was motivated by the need to accurately evaluate scaffold hopping studies in virtual screening and medicinal chemistry and assess the degree of difficulty involved in facilitating a transition from one structure to another. In order to consistently derive structural distances, scaffolds of different composition and topology are subjected to molecular editing procedures that abstract from original scaffolds in a defined manner until compositional and topological equivalence can be established. Pairs of corresponding scaffold representations are transformed into one-dimensional atom sequences that are aligned using approaches adapted from biological sequence comparison. From best scoring atom sequence alignments, interscaffold distances are derived. The algorithm is evaluated at different levels including the analysis of a series of model scaffolds with defined chemical changes, a scaffold library, and scaffolds from reference compounds and hits of successful virtual screening applications. It is demonstrated that chemically intuitive scaffold distances are obtained for pairs of scaffolds with varying composition and topology. Distance threshold values for close and remote structural relationships between scaffolds are also determined. The methodology is made publicly available in order to provide a basis for a consistent assessment of scaffold hopping ability and to aid in the evaluation and comparison of virtual screening methods.



1. INTRODUCTION

Molecular scaffolds are extensively utilized in medicinal chemistry and chemoinformatics and scaffold analyses have been carried out from very different points of view.[1] The scaffold concept is based on the intention to define molecular core structures that represent different chemotypes. Thus, popular scaffold definitions such as the one originally put forward by Bemis and Murcko[2] define core structures by removing R-groups from ring systems and linker fragments between rings. Bemis and Murcko-type scaffolds also provide a basis for a systematic organization of core structures.[3] From such scaffolds, one can further abstract to so-called graph frameworks[2] or cyclic skeletons[4] by removing atom type and bond order information, which then focuses scaffold comparisons entirely on molecular topology. In addition, reduced cyclic skeletons[4] are obtained from cyclic skeletons by assigning unit size to all rings and unit lengths to linkers.

Scaffold hopping[5−12] is one of the most investigated and discussed topics in ligand-based virtual screening.[13,14] This term essentially refers to the ability to identify different structural classes of active compounds through computational screening. In virtual screening, one typically starts with known reference compounds and then applies similarity search and/or compound classification/machine learning methods to search virtually formatted compound databases and identify compounds having a bioactivity similar to the references. In typical benchmarking investigations (where known active compounds are added as potential hits to a background database), demonstrating scaffold hopping potential has become the gold standard for method comparison and "validation".[13,14] In addition, scaffold hopping is also the major criterion for the success of prospective virtual screening applications.[14]

Unfortunately, the assessment of scaffold hopping potential in virtual screening is often controversial and hampered by a number of complications.[8] For example, in many instances neither scaffolds nor what actually constitutes a successful scaffold hop are clearly defined. Moreover, in addition to such technical shortcomings, scaffolds meeting formal definitions can structurally be very similar, for example, they might only be distinguished by a single bond order or heteroatom position. In addition, topologically distinct scaffolds might often also be structurally similar to each other, e.g., they might only be distinguished by a meta- versus para-substitution at a ring. Thus, successful scaffold hops could range from the detection of very similar to truly distinct structures. However, the magnitude of structural differences between scaffolds is usually not considered when scaffold hopping studies are statistically analyzed.

Received: August 24, 2011
Published: September 28, 2011

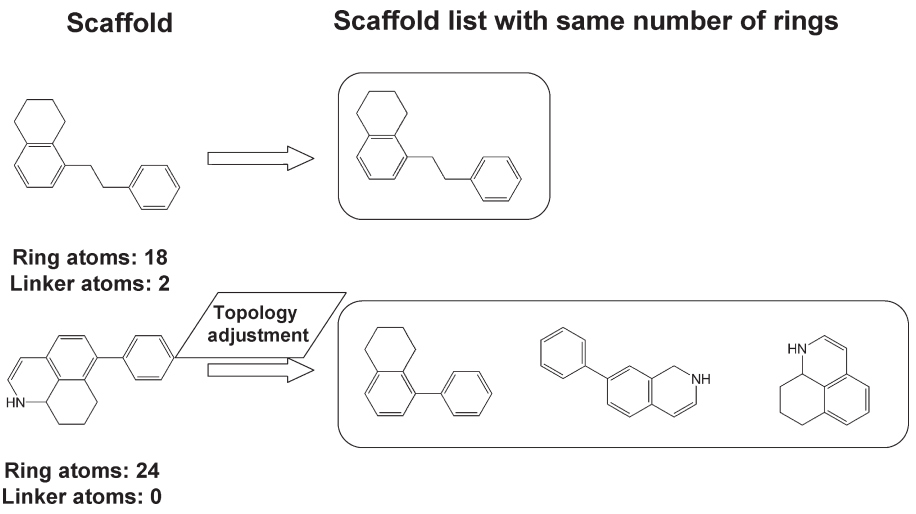


Figure 1. Topology adjustment. The structural decomposition of scaffolds to generate pairs of topologically equivalent scaffolds and/or substructures is illustrated. In this example, the first two of the three substructures generated from the scaffold at the bottom are topologically equivalent to the scaffold at the top.

Furthermore, by definition, the addition of any ring moiety to a structural core results in a new Bemis and Murcko-type scaffold (and graph framework), although such structures might often better be rationalized as analogs of a given scaffold, rather than distinct scaffolds, which represents another general complication.[15] In order to address some of the possible complications involved in the assessment of scaffold hopping, well-defined scaffold sets have been generated for benchmark investigations.[16,17] However, currently there is no approach available to consistently evaluate "how far one has jumped"[6] to facilitate a scaffold hop, i.e., how different the structures of known and detected scaffolds really are and how challenging a scaffold hopping exercise might be. In light of this situation we have developed a generally applicable method to determine the structural distance between any two scaffolds, regardless of their chemical composition and topology, which is presented herein.

2. METHODOLOGY

In the following, we outline the principal challenge of consistent distance calculations on scaffolds with different composition and/or topology, describe the components of the algorithm designed to establish topological correspondence between different scaffolds, introduce the distance calculation for pairs of aligned scaffolds, and report the implementation of the method.

2.1. Principles of Scaffold Comparison. The major challenge in quantifying structural distances between any two scaffolds is to account for differences in chemical composition (e.g., different numbers of differently sized or -composed rings) and topology (e.g., linkers connecting rings at different ring atom positions). Two scaffolds are considered to have the same topology if their relative arrangement of rings and linkers is equivalent, but ring sizes, composition, and linker length are permitted to vary. Taking these requirements into account, procedures are initially required to establish topological equivalence between scaffolds having similar or distinct structures. In the latter case, it is necessary to further abstract from original scaffolds, and we follow the concept of reduced cyclic skeletons,[4] which compare molecular composition and topology by stepwise abstraction from heteroatom, bond order, ring size, and linker length information. Once correspondence between two scaffolds is established, atom alignment can be carried out from which a similarity score is obtained that can then be converted into a distance. The final distance measure must also take into account the different abstractions made to arrive at topological equivalence. Taken together, these requirements rationalize the different steps of the algorithm, as discussed in the following.

We explain the approach on the basis of classical Bemis and Murcko scaffolds,[2] although other scaffold definitions could also be utilized. Different from the original scaffold definition by Bemis and Murcko, we consider here double-bonded atoms such as carbonyl oxygens in rings and linkers as a part of the scaffolds in order to avoid the introduction of ambiguous atomic hybridization state upon scaffold generation.

2.2. Topology Adjustment. The initial comparison of different scaffolds might reveal three basic scenarios, i.e. the scaffolds might

(i) differ in the number of ring systems they contain.
(ii) have otherwise different topology, i.e., differ in the relative arrangement of ring systems and linkers.
(iii) have the same topology but corresponding rings and linkers differ in size and/or composition.

In the third case, one could directly proceed to generate a molecular alignment. By contrast, in the first two cases, molecular transformations must be applied in order to establish topological equivalence, which we term "topology adjustment", as outlined in Figure 1.

Specifically, if two scaffolds differ in the number of ring systems, individual rings (including fused rings) are iteratively removed from the larger scaffold until it contains the same number of rings as the smaller one, which produces a list of possible substructures. A stringent condition for topology adjustment is that removal of a ring must produce a coherent substructure. In the example given in Figure 1, the larger scaffold contains four and the smaller three rings. Hence, removal of one ring from the larger scaffold yields the desired set of substructures. However, from this scaffold, only three rings can be removed to obtain one of three coherent substructures. Alternatively, if the number of rings is the same in two compared scaffolds but their topology differs, both scaffolds are subjected to

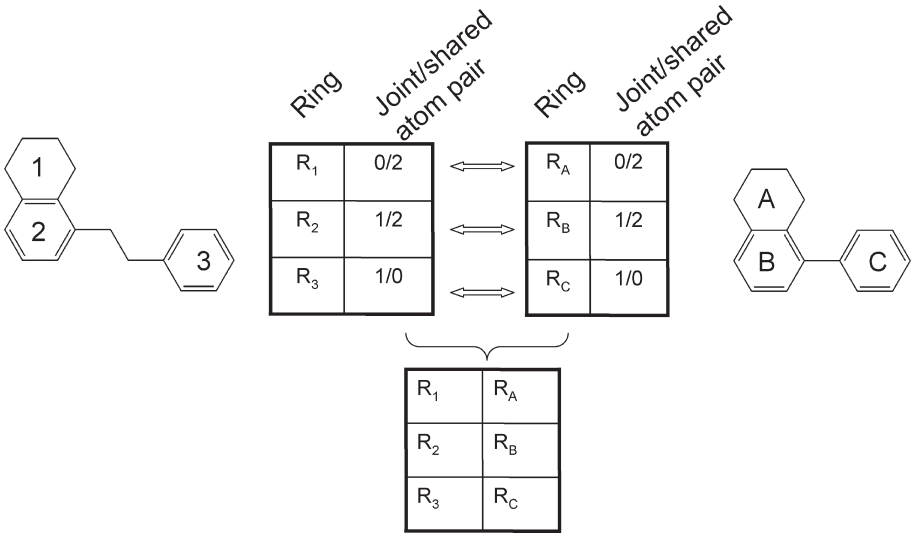


Figure 2. Ring correspondence. The ring mapping procedure to establish topological equivalence between two scaffolds is illustrated. In this example, the corresponding ring pairs are 1-A, 2-B, and 3-C.

topology adjustment by iterative removal of one or more rings until a pair of substructures with equivalent topology is obtained, as further discussed in the following. Regardless of the number of rings present in compared scaffolds, this systematic approach ensures that all possible substructures with equivalent topology are generated and taken into account.

2.3. Ring Mapping. In order to establish topological equivalence between scaffolds and decomposition products, corresponding rings are mapped through the comparison of structural "ring environments". For this purpose, ring atoms are divided into three different types, termed here "joint", "shared", and "free" atoms. The first two types refer to ring atoms that connect to a linker (or form a linker) or a fused ring, respectively. The remaining atoms are free atoms not involved in topological constraints. As illustrated in Figure 2, each ring is then encoded as a joint/shared atom value pair. Since the size and composition between corresponding rings in two scaffolds is permitted to differ, rings can be mapped on the basis of shared value pairs. If the initial mapping of corresponding rings remains ambiguous, different from the example in Figure 2, an exhaustive search is carried out by iterative removal of one ring at a time, followed by comparison of reduced cyclic skeletons of the remaining substructures to identify an identical pair. In this case, a tree structure is used to keep track of all the possible mappings.

2.4. Atom Sequence Transformation and Alignment. Once topological equivalence between two scaffolds has been established, one can proceed to the scaffold alignment stage, which is ultimately required for distance calculations. Therefore, an atom sequence transformation is carried out. For computational efficiency, corresponding rings and linkers are separately transformed and aligned. If correspondence between two rings is established, the identification of corresponding linker fragments is straightforward. In order to consistently transform ring structures into a one-dimensional atom sequence, "atom environments" are systematically assigned to define possible "anchor atoms" as a starting point for sequence generation. Therefore, in addition to the joint, shared, and free atoms discussed above, a fourth atom type is defined as a "double-bonded atom" that forms a double bond to another atom attached to a ring (e.g., a carbonyl oxygen). All ring atoms are assigned to these types and their counts are determined, i.e., the number of atoms per ring belonging to each type. Possible anchor atoms are atoms in the ring that belong to the atom type with lowest count. The procedure is illustrated in Figure 3. In this example, the joint atom type has the lowest count and there is only a single anchor atom candidate. Starting from this anchor atom, two alternative atom sequences are derived based on clockwise or counterclockwise ring traversal. If there is more than one possible anchor atom, each candidate atom is used once as an anchor, and all possible sequences are derived for both reading directions.

For the alignment of corresponding ring and linker sequences, we have adopted the Needleman-Wunsch algorithm[18] for biological sequence comparison and the atom substitution matrix introduced by Berglund et al.[19] This substitution matrix is based on Tripos MOL2 atom types and defines substitution costs according to similarity of atom types, bond orders, and hydrogen bond donor and acceptors (with scores between 0 and 1 for highest and lowest costs).[19] We have further extended the matrix through addition of the four topological atom types defined herein with constant substitution costs according to standard MOL2 atom type replacements. Furthermore, atom insertions and deletions have also been permitted applying the highest possible cost. A subset of most frequently observed atom type replacements and associated costs is reported in Table 1, and the complete substitution matrix is provided in Table S1 of the Supporting Information. The granularity of atom types in the substitution matrix and the substitution costs determine the alignment results and are thus an important component of the distance calculations. A reduction of atom types would be expected to lower the accuracy of the atom alignments.

For each pair of rings or linkers, alignments between all possible atom sequences are calculated, and the best scoring individual alignments are selected. Their combination then yields the scaffold alignment.

2.5. Distance Calculation. Following atom sequence alignment, scaffold distances are determined, as shown in Figure 4. The similarity of aligned atom sequences is calculated as the

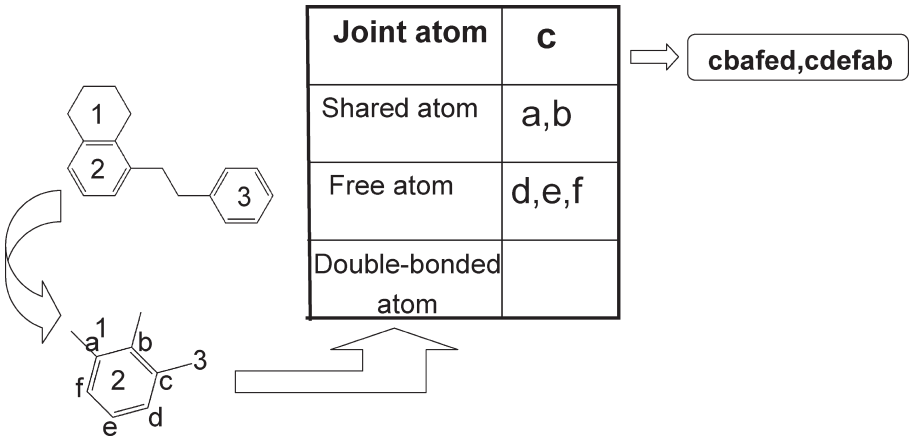


Figure 3. Atom sequence transformation. Possible atom sequences of ring 2 of the exemplary scaffold are derived. Ring atoms are labeled a-f and atom c serves as the anchor atom.

Table 1. Subset of the Atom Substitution Matrix[a]

	C.2	C.3	C.ar	N.3	N.ar	O.2	S.3	C.ar.3	-
C.2	1	0.6	0.7	0.2	0.1	0	0.1	0.6	0
C.3	0.6	1	0.6	0.1	0	0	0.2	0.5	0
C.ar	0.7	0.6	1	0.2	0.1	0	0.1	0.9	0
N.3	0.2	0.1	0.2	1	0.7	0	0.1	0.05	0
N.ar	0.1	0	0.1	0.7	1	0.1	0	0.05	0
O.2	0	0	0	0	0.1	1	0	0	0
S.3	0.1	0.2	0.1	0.1	0	0	1	0.05	0
C.ar.3	0.6	0.5	0.9	0.15	0.05	0	0.05	1	0
-	0	0	0	0	0	0	0	0	1

[a] Atom types are defined as follows: C.2, sp2 carbon; C.3, sp3 carbon; C.ar, aromatic carbon of degree 2; C.ar.3, aromatic carbon of degree 3; N.3, sp3 nitrogen; N.ar, aromatic nitrogen; =.2, sp2 oxygen; S.3, sp3 sulfur. In addition, "-" denotes an insertion or deletion.

score resulting from atom type matches.[19] The dissimilarity or distance value is then obtained as *Dist = 1-similarity*. Because the objective is to derive the distance between the original scaffolds, the calculated similarity is normalized with respect to the sum of the largest number of ring and linker atoms in the original scaffolds, as illustrated in Figure 4. In addition, alignments between a scaffold and all alternative topologically adjusted decomposition products are compared, and the smallest distance value is selected as the final scaffold distance.

2.6. Implementation. In Figure 5, a summary of the complete algorithm is presented, which embeds the major steps of initial scaffold comparison, topology adjustment, atom sequence transformation, pairwise sequence alignment, and distance calculations. The method was implemented in Java utilizing the OpenEye chemistry toolkit.[20] As input format for scaffolds, SMILES strings[21] are used.

3. EVALUATION

In the following, the methodology is evaluated at different levels including comparison of model scaffolds, analysis of a scaffold library, and comparison of reference compounds and hits from successful virtual screening applications. Initially, it has been most important to assess whether our methodology produces chemically intuitive distance relationships between different scaffolds.

3.1. Chemically Intuitive Scaffold Distances. In Figure 6a, series of benzene-derived scaffolds of increasing size are shown. One series includes oligomers of the biphenyl scaffold and the other corresponding forms of diphenyl-methane. Scaffold distances between benzene and these scaffolds are reported as well as distances between scaffolds of incrementally increasing size in each series. As one would expect, the distances between benzene and oligomeric scaffolds that increase in size by one ring at a time also increase in a stepwise manner. In addition, the distance between benzene and diphenyl-methane is slightly larger than the distance between benzene and biphenyl. Furthermore, within each series, pairwise distances between scaffolds become smaller with increasing size, which is a consequence of similarity score normalization by scaffold atom numbers, as discussed above. The observed score reduction is desirable because with increasing scaffold size, the addition of a single benzene ring corresponds to a smaller proportion of added structure. In Figure 6b, the influence of linker length and heteroatom content on calculated scaffold distances is evaluated. In horizontal direction, the linker length increases by one carbon atom per step starting from a comparison of biphenyl and phenyl-pyridine, which yields gradually slightly decreasing scores, consistent with the observations made above. In vertical direction, the size of the diphenyl-methane scaffold and the phenyl-pyridyl-methane derivatives remains constant as one nitrogen (ring or linker) atom is added per step. Thus, in each pairwise comparison, the difference is exactly one nitrogen atom and the calculated scaffold distances remain constant, as to be expected. In Figure 6c, different heteroaliphatic and heteroaromatic rings are compared. Among the five-membered rings on the left of this figure, pairwise distances between corresponding rings with defined structural differences are fully consistent, with the exception of furane compared to cyclopentadienone, which produces a distance of 0.42 that is considerably larger than the distance of 0.18 between the fully aliphatic counterparts of these scaffolds. However, this is also a meaningful result because furane is an aromatic system, whereas cyclopentadienone is not. Thus, larger distance should be obtained in this case than for the comparison of the fully aliphatic structural counterparts. Among the six-membered rings on the right of this figure, different atomic substitution costs result in modified structural distances (for example, for nitrogen versus sulfur), and there are combinations of aromatic and heteroatom contributions. For example, the distance between

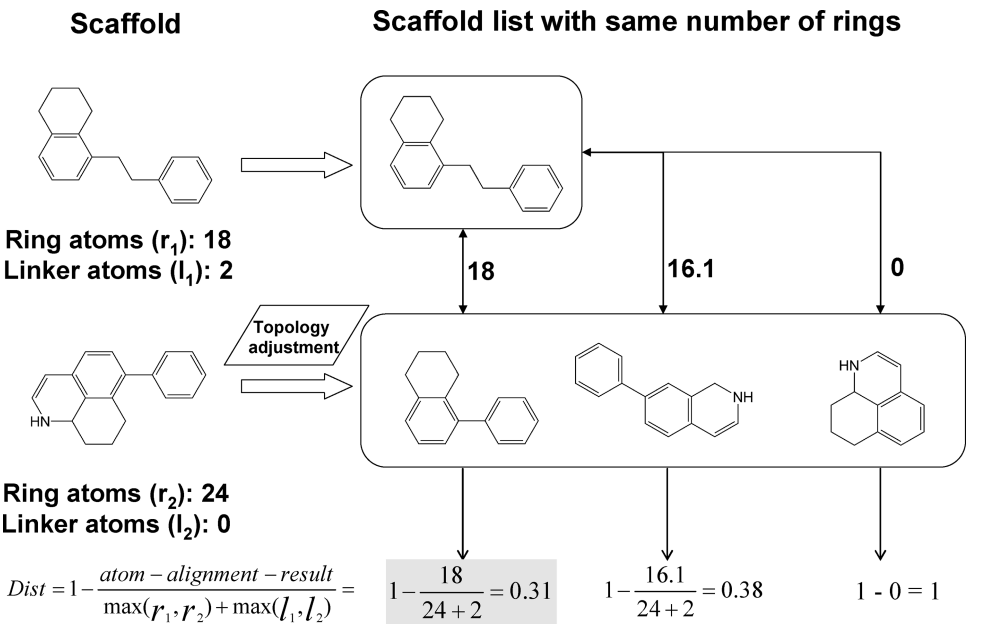


Figure 4. Distance calculation. The figure illustrates how normalized distance values are calculated from alternative atom sequence alignments. Numbers next to arrows between scaffolds report atom alignment scores. In this example, the largest number of ring atoms is 24 (for lower scaffold) and the largest number of linker atoms is 2 (upper scaffold), yielding a sum of 26 for normalization of atom sequence similarity scores.

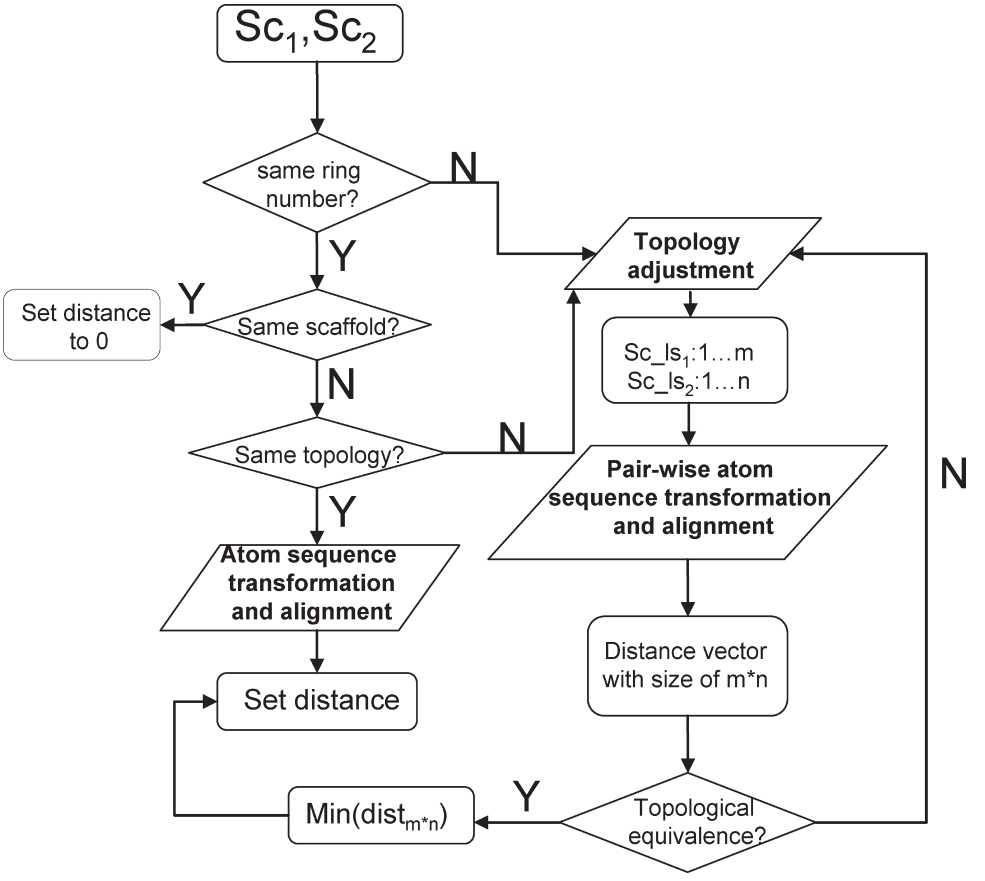


Figure 5. Scaffold comparison, modification, and distance calculation. An outline of the complete algorithm for generation of interscaffold distances is shown ("Sc" means scaffold).

benzene and cyclohexane (0.40) is smaller than the distance between benzene and piperidine (0.47), which is reasonable because compared to benzene, piperidine is not only aliphatic (like cyclohexane) but also contains a heteroatom.

dx.doi.org/10.1021/ci2003945 |J. Chem. Inf. Model. 2011, 51, 2507–2514

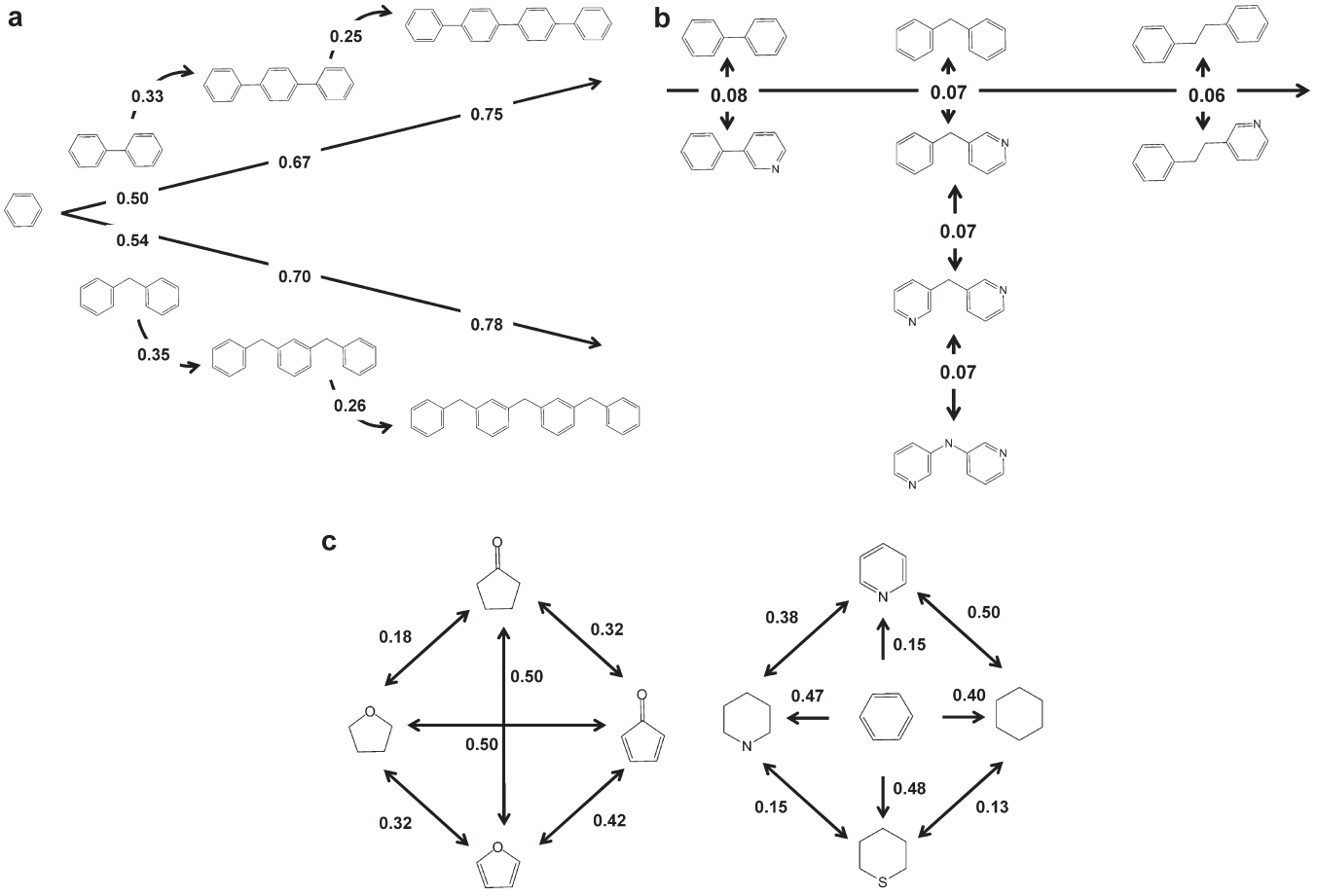


Figure 6. Comparison of model scaffolds. Distances are calculated between series of scaffolds that are systematically distinguished by different types of chemical changes including (**a**) varying size degrees, (**b**) varying linker length and heteroatom content, and (**c**) aromatic versus aliphatic character and/or varying heteroatom content.

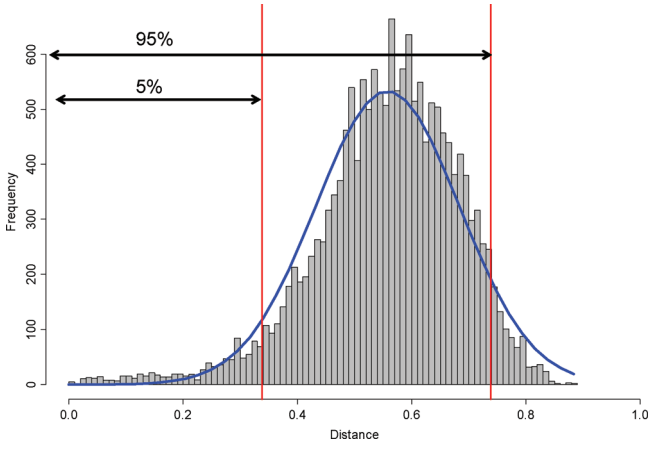


Figure 7. Scaffold distance distribution. The distribution of all pairwise distance values within a library of bioactive scaffolds is reported. The fifth and 95th percentile of the distribution are marked.

Thus, taken together, these examples illustrate that scaffold distances calculated are chemically intuitive and in accord with observed structural differences.

3.2. Scaffold Library Analysis. We next analyze a scaffold library reported by Vogt et al.[17] This library was designed to consist of 10 scaffold sets derived from compounds with different activities. Each scaffold set contained between 13 and 33 unique Bemis and Murcko-type scaffolds, with a total of 184 Bemis and Murcko scaffolds representing 10 different bioactivities. We utilize this library to estimate the scaffold distance value distribution between bioactive scaffolds. In Figure 7, the distribution of all pairwise scaffold distances is reported. The theoretically possible value range for the scaffold distances is the interval [0,1]. In practice, no scaffold distances larger than 0.90 are found. This can be rationalized by the fact that all scaffolds contain at least one ring. This means that topological adjustment of a scaffold will always yield topologically equivalent scaffolds consisting of at least one corresponding ring. In other words, the observed limit of 0.9 indicates that all library scaffolds contain rings that share at least a certain degree of similarity. The histogram in Figure 7 shows a typical bell shaped distribution, given the constraints that the range is limited from 0 to 0.9 and that the mean is shifted toward higher values. Such distributions can be modeled either by normal distributions or, alternatively, by distributions enforcing the constraints on the range of possible distances such as the beta distribution (that limits the range to [0,1]). In Figure 7 a fit to a normal distribution is shown with a mean of 0.56 and a standard deviation of 0.13. From the histogram the 5% and 95% levels are determined as 0.34 and 0.74, respectively. Thus, from the point of statistical significance,

Table 2. Scaffold Distances for Virtual Screening Applications[a]

no.	mean pairwise scaffold distance	no. of hits	no. of references
1	0.82	1	1
2	0.78	2	6
3	0.72	1	6
4	0.72	2	21
5	0.69	2	5
6	0.68	8	23
7	0.68	1	1
8	0.67	1	4
9	0.62	20	80
10	0.60	1	8
11	0.60	3	6
12	0.60	2	43
13	0.60	17	8
14	0.53	9	2
15	0.53	9	6
16	0.53	8	26
17	0.52	3	8
18	0.52	1	6
19	0.52	1	16
20	0.50	1	16
21	0.50	6	1
22	0.46	2	5
23	0.45	2	88
24	0.39	12	2
25	0.07	1	33

[a] For 25 ligand-based virtual screening (scaffold hopping) applications extracted from the literature, the number of reference compounds and confirmed hits and the mean pair-wise scaffold distance between them are reported (and ranked in the order of decreasing mean distance).

scaffolds with a distance less than 0.34 can be considered similar, while scaffolds with a distance larger than 0.74 are dissimilar.

3.3. Scaffold Hopping. From the above analysis, one can infer that an average scaffold hopping distance would be around 0.5. How does this relate to the scaffold hopping potential of virtual screening calculations? In order to address the question we analyze a subset of 25 successful ligand-based virtual screening applications that we previously extracted from original literature sources.[22] For these virtual screens, we confirmed that they identified (and experimentally evaluated in appropriate assays) previously unknown compounds that met the formal scaffold hopping criterion on the basis of Bemis and Murcko-type scaffolds isolated from reference compounds and hits. These virtual screening exercises were carried out targeting different protein families and applying different ligand-based methods and can be considered to represent the current state-of-the-art in this field. For the purpose of our analysis, individual target and method information is not required, and we thus consider the union of these studies. In Table 2, we report the number of reference compounds and hits and the mean scaffold distance between them for each investigation. Interestingly, in 14 of 25 cases, reference compounds and hits yield distances falling into the interval [0.4−0.6], i.e. close to the average scaffold hopping distance deduced from library analysis. There are only two

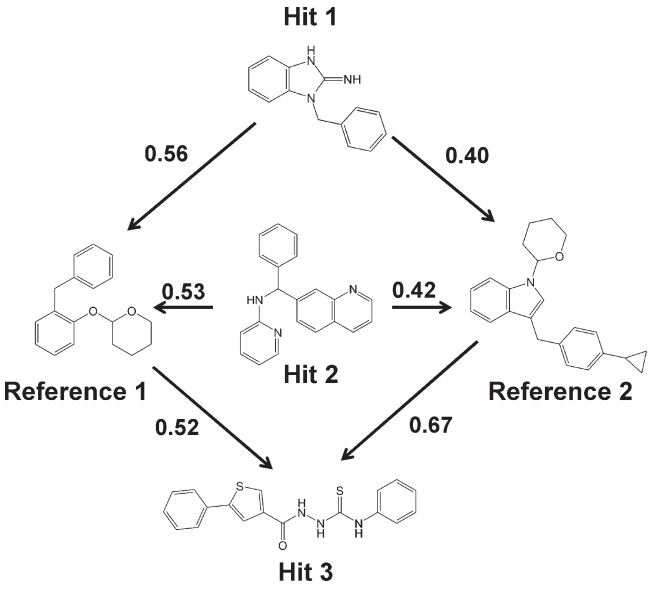


Figure 8. Exemplary scaffold relationships. Distances between scaffolds extracted from reference compounds and hits of a prospective virtual screening application[23] are reported.

instances of an average scaffold hopping distance below 0.4 and one below 0.1. However, Table 2 also reveals three instances of average scaffold hopping distances between reference compounds and hits of greater than 0.7 and one case of a single reference compound and hit producing a scaffold distance of greater than 0.8. For all compound sets, the minimum, maximum, and median scaffold distance was 0.03, 0.89, and 0.62, respectively. Hence, a scaffold distance of ~0.6 would represent a meaningful expectation value for a prospective virtual screen producing novel hits. However, this value is lower than the threshold value for structurally dissimilar scaffolds derived from the distance distribution in Figure 7. Furthermore, any scaffold hopping exercise resulting in a scaffold distance >0.8 would be considered a top result in terms of structural novelty.

In Figure 8, an example of scaffolds extracted from reference compounds and hits is shown that illustrates the structural meaning of close to average scaffold hopping distances. On the basis of our analysis, the majority of prospective virtual screens summarized in Table 2 is characterized by scaffold hopping distances between reference compounds and hits that are close to average or larger than distances obtained for random comparisons of bioactive scaffolds.

4. DISCUSSION AND CONCLUSIONS

We have introduced a methodology to define and consistently quantify a structural distance between any two scaffolds, regardless of their chemical composition and topology, which represents a nontrivial task. Standard similarity measures such as the Tanimoto coefficient calculated on the basis of fingerprint representations have insufficient resolution to accurately account for structural differences between scaffolds. Currently, there is no other methodology available to quantify structural distances between scaffolds. In order to compare distinct scaffolds, we represent them at varying degrees of abstraction through iterative ring removal in order to establish topological correspondence. On this basis, rings and linkers comprising original or reduced

scaffolds are then transformed into atom sequences. Approaches adapted from biological sequence comparison are then used to align these atom sequences and score their similarity. From the highest similarity value of possible alignments, the scaffold distance is then derived using a function that normalizes similarity values relative to original scaffold atom counts. Following our approach, calculated scaffold distances have been shown to be chemically intuitive for series of model scaffolds with defined structural modifications. The development of our scaffold distance method was originally motivated by the need to consistently and quantitatively assess the structural distance between scaffolds in the context of virtual screening studies. The current absence of well-defined and generally applicable scaffold hopping criteria and the difficulties involved in comparing different scaffold hopping studies present a major conundrum for this field. Scaffold distances can be used to complement the conventional assessment of recall performance in retrospective scaffold hopping investigations. In addition, they can be utilized to compare individual reference compounds and newly identified hits and quantify the degree of difficulty involved in facilitating a scaffold hop. By comparing reference compounds and hits from a sample of state-of-the-art prospective virtual screening applications, we have shown that the majority of reported scaffold hops yield distances close to the average distance between structurally diverse bioactive scaffolds. For the assessment of virtual screening applications, it would be meaningful to determine the smallest distance between scaffolds of newly identified hit(s) and known active scaffolds as a quality criterion. In order to provide a basis for community-wide scaffold comparisons, the distance method is made freely available for download upon publication via the following URL: http://www.lifescienceinformatics.uni-bonn.de.

In conclusion, on the basis of our analysis, the approach introduced herein is indicated to quantify different degrees of structural relationships between scaffolds in a meaningful way. It is hoped that the method will also be helpful to others to analyze scaffold hopping potential more consistently and at higher resolution than has been possible thus far.

■ ASSOCIATED CONTENT

Supporting Information. Complete atom type substitution matrix utilized for our study. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

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

■ REFERENCES

(1) Hu, Y.; Stumpfe, D.; Bajorath, J. Lessons Learned from Molecular Scaffold Analysis. *J. Chem. Inf. Model.* **2011**, *51*, 1742–1753.

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

(3) 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.

(4) Xu, Y.-J.; Johnson, M. Algorithm for Naming Molecular Equivalence Classes Represented by Labeled Pseudographs. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 181–185.

(5) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chem., Int. Ed.* **1999**, *19*, 2894–2896.

(6) Brown, N.; Jacoby, E. On Scaffolds and Hopping in Medicinal Chemistry. *Mini-Rev. Med. Chem.* **2006**, *6*, 1217–1229.

(7) Schneider, G.; Schneider, P.; Renner, S. Scaffold-Hopping: How Far Can You Jump? *QSAR Comb. Sci.* **2006**, *25*, 1162–1171.

(8) Tsunoyama, K.; Amini, A.; Sternberg, M. J. E.; Muggleton, S. H. Scaffold Hopping in Drug Discovery Using Inductive Logic Programming. *J. Chem. Inf. Model.* **2008**, *48*, 949–957.

(9) Wale, N.; Watson, I. A.; Karypis., G. Indirect Similarity Based Methods for Effective Scaffold-Hopping in Chemical Compounds. *J. Chem. Inf. Model.* **2008**, *48*, 730–741.

(10) Senger, S. Using Tversky Similarity Searches for Core Hopping: Finding the Needles in the Haystack. *J. Chem. Inf. Model.* **2009**, *49*, 1514–1524.

(11) Mackey, M. D.; Melville, J. L. Better than Random? The Chemotype Enrichment Problem. *J. Chem. Inf. Model.* **2009**, *49*, 1154–1162.

(12) Hu, Y.; Bajorath, J. Global Assessment of Scaffold Hopping Potential for Current Pharmaceutical Target. *Med. Chem. Commun.* **2010**, *1*, 339–344.

(13) Geppert, H.; Vogt, M.; Bajorath, J. Current Trends in Ligand-Based Virtual Screening: Molecular Representations, Data Mining Methods, New Application Areas, and Performance Evaluation. *J. Chem. Inf. Model.* **2010**, *50*, 205–216.

(14) Stumpfe, D.; Bajorath, J. Applied Virtual Screening: Strategies, Recommendations, and Caveats. In *Methods and Principles in Medicinal Chemistry. Virtual Screening. Principles, Challenges, and Practical Guidelines*; Sotriffer, C., Ed.; Wiley-VCH: Weinheim, 2011; pp 73−103.

(15) Katritzky, A. R.; Kiely, J. S.; Hebert, N.; Chassaing, C. Definition of Templates within Combinatorial Libraries. *J. Comb. Chem.* **2000**, *2*, 2–5.

(16) Rohrer, S. G.; Baumann, K. Impact of Benchmark Data Set Topology on the Validation of Virtual Screening Methods: Exploration and Quantification by Spatial Statistics. *J. Chem. Inf. Model.* **2008**, *48*, 704–718.

(17) Vogt, M.; Stumpfe, D.; Geppert, H.; Bajorath, J. Scaffold Hopping Using Two- Dimensional Fingerprints: True Potential, Black Magic, or a Hopeless Endeavor? Guidelines for Virtual Screening. *J. Med. Chem.* **2010**, *53*, 5707–5715.

(18) Needleman, S. B.; Wunsch, C. D. A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *J. Mol. Biol.* **1970**, *48*, 443–453.

(19) Berglund, A. E.; Head, R. D. PZIM: A Method for Similarity Searching Using Atom Environments and 2D Alignment. *J. Chem. Inf. Model.* **2010**, *50*, 1790–1795.

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

(21) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.

(22) Ripphausen, P.; Nisius, B.; Bajorath, J. State-of-the-Art in Ligand-based Virtual Screening. *Drug Discovery Today* **2011**, *16* 372–376.

(23) Wu, J.-S.; Peng, Y.-H.; Wu, J.-M.; Hsieh, C.-J.; Wu, S.-H.; Coumar, M. S.; Song, J.-S.; Lee, J.-C-; Tsai, C.-H.; Chen, C.-T.; Liu, Y.-W.; Chao, Y.-S.; Wu, S.-Y. Discovery of Non-glycoside Sodium-dependent Glucose Co-transporter 2 (SGLT2) Inhibitors by Ligand-based Virtual Screening. *J. Med. Chem.* **2010**, *53*, 8770–8774.