Contents lists available at SciVerse ScienceDirect

Bioorganic & Medicinal Chemistry

journal homepage: www.elsevier.com/locate/bmc

Database of bioactive ring systems with calculated properties and its use in bioisosteric design and scaffold hopping

Peter Ertl *

Novartis Institutes for BioMedical Research, Novartis Campus, CH-4056 Basel, Switzerland

ARTICLE INFO

Article history:
Available online 3 March 2012

Keywords:
Cheminformatics
Bioisosteric design
Scaffold
Scaffold hopping
Scaffold properties
Bioisosteric search

ABSTRACT

A system for identification of bioisosteric scaffolds is presented. The system uses a database of over 7000 scaffolds extracted from bioactive molecules. Scaffolds in the database are characterized by their size, shape, pharmacophore features and several ADME descriptors. Also properties characterizing electron-donating or -accepting power at connection vectors are considered. All these features are used as search criteria to find scaffolds with the most similar properties to the query. To guarantee fast processing the search is performed using topological descriptors only, but the system may be used to find optimal replacements of scaffolds also directly in the protein binding site. In this case a set of 3D conformations for the best 2D hits is generated and analogs optimally fitting the binding pocket are identified by overlap with the query ligand and by optimizing interactions with the protein. This tool is used at Novartis as an idea generator for identification of novel non-classical bioisosteric analogs in the drug discovery process.

1. Introduction

The concept of scaffold as the central part of a molecule is one of the basic concepts of medicinal chemistry.[1,2] The scaffold gives molecule its shape, determines whether the molecule is rigid or flexible and keeps substituents in their positions. Global molecular properties, such as hydrophobicity or polarity, which are important for bioavailability and the fate of molecules in organism, are also determined mostly by the composition of the scaffold. Electronic properties of the scaffold determine reactivity of a molecule which in turn is responsible for its metabolic stability and toxicity. Scaffolds also play an important role in several commonly used medicinal chemistry techniques. One of these is combinatorial chemistry, where various ring systems are used as central cores of combinatorial libraries. Another popular technique applied in the drug discovery process is 'scaffold hopping'[3,4] where the goal is to 'jump' in chemistry space, that is, to discover a new bioactive structure with improved properties starting from a known active compound via modification of the central core.

Despite its importance, the term 'scaffold' is used in medicinal chemistry literature rather freely, without a clearly defined meaning. The exact interpretation of this term varies from publication to publication and depends also on the particular application area. In some cases the term scaffold is used for a part of molecule remaining after removal of all non-ring substituents (such scaffolds are often called also Bemis–Murcko scaffolds[5] or molecule frameworks). In other cases the term scaffold denotes the largest, most central ring system in the molecule. When used in the combinatorial chemistry applications the term scaffold is used to describe a substructure common to all molecules in a combinatorial library, and may contain rings as well as non-ring functionalities, in some cases even lacking the rings at all. Throughout this article terms 'scaffold' and 'ring system' are used interchangeably with the same meaning, denoting a single ring, or a collection of fused or spiro rings, including also exocyclic multiple bonds. Non-ring substituents and chains connecting ring systems are not considered as their parts. One molecule can therefore contain several scaffolds separated by non-ring chains.

As already mentioned, one of the most important use of scaffolds (in whatever meaning this term is used) in medicinal chemistry is their use in bioisosteric replacements and scaffold hopping. All practicing medicinal chemists know that the successful scaffold hop requires a lot of experience and even then, a long trial and error optimization is often needed to identify novel scaffolds with optimal balance of necessary structural features and good physicochemical properties. Computational chemistry and cheminformatics can provide useful help to chemists in their effort to identify optimal scaffold replacements. In the following section several techniques used for this purpose are described.

Program CAVEAT,[6] a pioneering application in the field of automatic scaffold replacement developed by Lauri and Bartlett allowed identification of compatible linkers by searching a 3D database of scaffolds, where their connection bonds were encoded as vectors and selecting those matching position and orientation of

* Corresponding author.
E-mail address: peter.ertl@novartis.com
URL: http://peter-ertl.com

0968-0896/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2012.02.058

exit vectors of a query. Program Recore[7] is based on similar principle. In this case, however, also various filters to select only drug-like fragments are applied. Similar scaffolds are identified by considering compatibility in exit connection vectors and also pharmacophore features. Scientists from GlaxoSmithKline developed a web-based system allowing medicinal chemists to identify bioisosteric ring systems.[8] The program uses database of rings with up to three connections extracted from several in-house and external molecular collections. Identified rings were characterized by various calculated properties, including ADME characteristics, and by geometric features representing distances between connection points and angles between connection bonds. Broughton and Watson[9] described a database of ring systems extracted from drugs that have reached development Phase II or later. Rings were characterized by size, shape, hydrogen bonding features and quantum chemically calculated properties, enabling rational selection of entries from the database to mimic properties of the query system. Another system for ring template searching was described by scientists from Tripos.[10] Its database with 19000 scaffolds was built by extracting scaffolds from the Cambridge Structural Database. Scaffolds were characterized by so called centroid connecting path representation, a structure descriptor considering ring centroids, linker atoms and other important points on the paths between the ring centroids. Three different similarity metrics were used to identify similar scaffolds that could be used as starting points in scaffold-hopping applications. Program SHOP[11] identifies similar scaffolds not only by considering their geometrical features expressed as distances and dihedral angles between the anchor points, but also by their shape properties expressed as alignment-independent GRID molecular interaction fields. SARANEA[12] is an application allowing to identify molecules similar in structure, but differing in potency or selectivity (exhibiting so called activity or selectivity cliffs). One can use it to find scaffolds that are structurally similar to the query, but have better potency or no selectivity issues. SARANEA is available as open source software. Program MORPH[13] systematically modifies aromatic rings in molecules without altering molecule coordinates. Each carbon, nitrogen, oxygen or sulfur atom in each ring is systematically replaced by atoms of other types adjusting at the same time also bond orders and number of attached hydrogens. This is in principle similar to 'pyridine scan' often applied by medicinal chemists when they systematically replace carbons in phenyl ring by nitrogen. The advantage of this relatively simple approach is that the geometries of newly created systems do not need to be modified, while in more complex approaches new compounds derived de novo must be fit back or docked into the binding site what increases the complexity of the process and probability to miss the actual binding hypothesis.

At Novartis several new methodologies to navigate the scaffold universe and to identify novel bioactive scaffolds were developed. In the 'Quest for the rings' study[14] all possible conjugated ring systems with one, two, and three, 5- and 6-membered rings were constructed creating a database of 600,000 virtual scaffolds. Bioactive regions in this large scaffold space were identified by self-organizing neural network trained on a set of scaffolds extracted from known bioactive molecules. The study has shown that the bioactivity is sparsely distributed in the scaffold universe, forming only several relatively small 'bioactivity islands' that are the most promising places to search for novel bioactive scaffolds. The Scaffold Tree methodology[15] is another way to classify scaffolds. By this approach scaffolds are organized into a hierarchical tree where the scaffolds with one ring are located at the root level of the tree, scaffolds with two rings are in the next level and so on. This classification allows easy and intuitive visualization of large collections of molecules, as well as supports identification of novel bioactive scaffolds by 'jumping' from a branch containing known bioactive molecules to neighboring branches.[16] At Novartis also several interactive web-based tools were developed, with goal to assist medicinal chemists in identifying non-classical bioisosteric analogs.[17,18] These interactive web tools are quite popular among Novartis researchers and are used by more than 1600 registered users from six Novartis research sites, helping them to identify bioisosteric substituents and linkers, that is, fragments connected by one and two connections, respectively, to the rest of the molecule. When looking for bioisosteric replacements of complex molecules and particularly in scaffold hopping applications, it is necessary to be able to identify also analogs of ring systems containing multiple connection points. Development of a system to perform this task and creation of necessary database of multi-connected scaffolds is the topic of this publication.

2. Results

2.1. Creation of the scaffold database

Goal of the present study was to analyze ring systems in bioactive molecules and create a database of scaffolds that could be used as a basis for identification of compatible bioisosteric rings in drug discovery projects. A similar system to identify substituents and linkers (i.e., groups connected to the rest of the parent molecule by one or two connection bonds, respectively) is already available at Novartis.[17,18] Many bioactive molecules, however, have a central core with more connections points. To document this we analyzed scaffolds in approved drugs from the DrugBank,[19] bioactive molecules from the ChEMBL database[20] and compared them with scaffolds extracted from the ZINC database.[21] The ChEMBL database is publicly available database of molecules extracted from medicinal chemistry literature and other sources, including also bioactivity data and information about respective targets. The ZINC database contains commercially available drug-like molecules from various vendor catalogs. Results of the scaffold analysis are summarized in Table 1, showing percentage of scaffolds with different number of connection points in various databases.

One can see that scaffolds having three and four connection points form indeed a large portion of drugs (27.4%) as well as general bioactive molecules (24.8%). When developing an automatic system for bioisosteric design it is therefore necessary to cover also this type of scaffolds. The most common scaffolds with one to four connection points extracted from the ChEMBL database are shown in Figs. 1–4. Even in this quite limited set of the most common rings one can see differences between various groups. Only six ring systems are present in all four groups, namely benzene, pyridine, thiophene, indole, pyrrolidine and piperidine. Ring systems with one connection are mostly property modifying substituents, including solubilizing rings present only in this set like morpholine, benzodioxole and benzodioxan or hydrophobic cycloalkanes. Scaffolds having two connections act mostly like linkers joining

Table 1
Percentage of scaffolds with different number of connection points in drugs, bioactive molecules and drug-like molecules

Number of connections	% of the all scaffolds		
	DrugBank	ChEMBL	ZINC
1	27.13	28.70	27.33
2	37.80	45.35	50.14
3	16.68	19.30	18.42
4	10.74	5.45	3.73
5	5.80	0.85	0.30
6	1.07	0.26	0.06
7	0.04	0.04	0.00
8	0.00	0.01	0.00

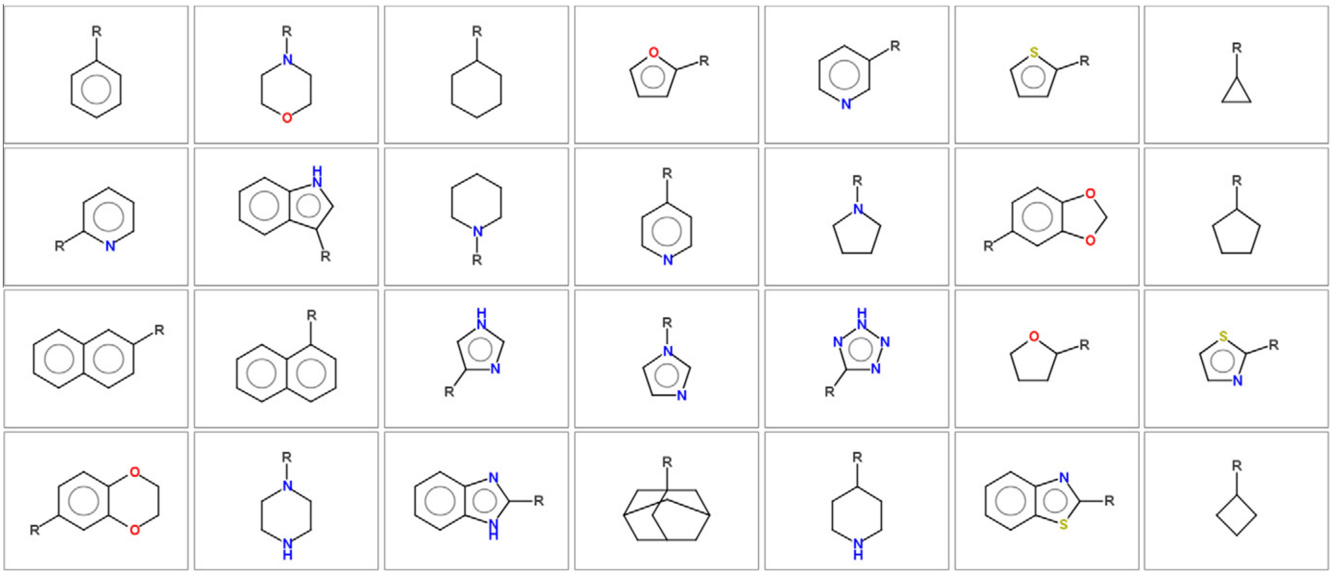

Figure 1. The most frequent ring systems with one connection.

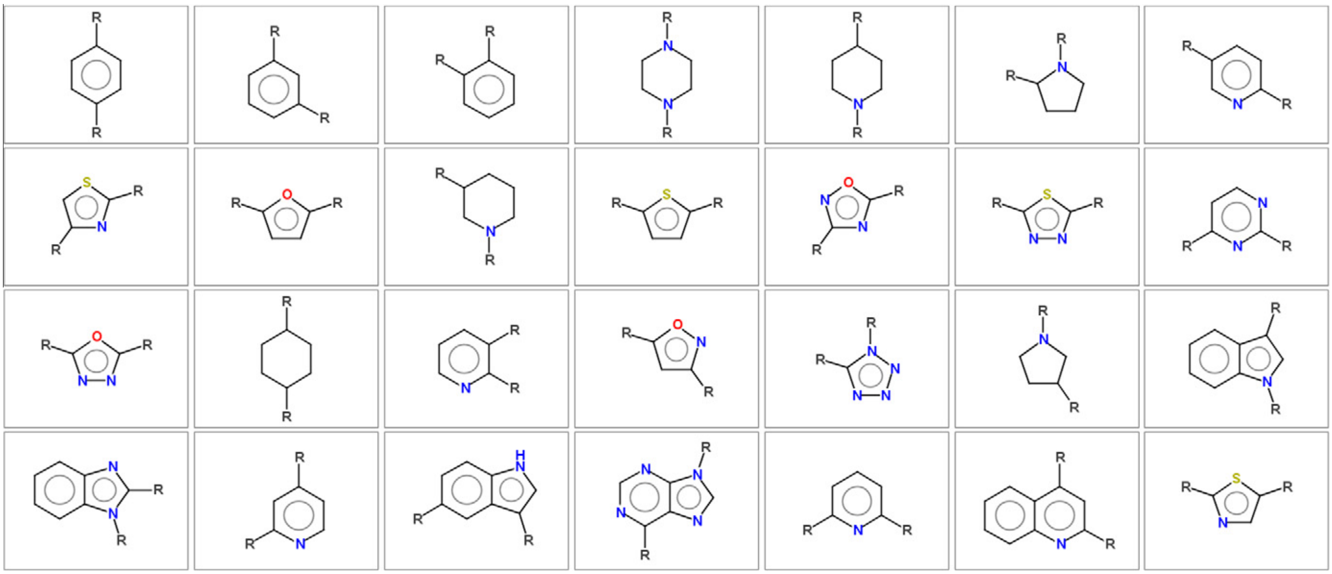

Figure 2. The most frequent ring systems with two connections.

two parts of the parent molecule together, and scaffolds with three and particularly four connection bonds act as central hubs, giving molecules their shape and keeping substituent at their proper positions. This situation sets also different requirements on the scaffold properties to be used as search criteria in scaffold similarity searches. While for scaffolds with one connection point descriptors characterizing their ADME properties are the most important and for linker scaffolds one has to consider particularly the distance between their two connection points, for more complex scaffolds with three and four connections points one has to find good balance between ADME properties and structural descriptors characterizing mutual positions of bond exit vectors.

The scaffolds with three and four connections extracted from the ChEMBL served as a basis for construction of our scaffold replacement database. After cleaning, removal of scaffolds containing non-organic atoms and too large scaffolds the dataset contained 4834 SMILES strings for scaffolds with three connection points and 2516 SMILES representations for scaffolds with four connection points.

2.2. Scaffold similarity searches

To be able to perform similarity searches, scaffolds in the database need to be characterized by appropriate numerical descriptors. The descriptors should represent properties that are medicinal chemistry relevant and play role in ligand-receptor interactions and are also easy to calculate. Since the scaffolds we are dealing with are actually fragments with several connections to the rest of the molecule, it is necessary to consider also descriptors characterizing properties of these connection points. Based on our experience with system for substituent and linker similarity searches we selected the descriptors that are listed in Table 2.

The composition of scaffolds is described by simple topological count descriptors. Slightly more complex descriptors are used to

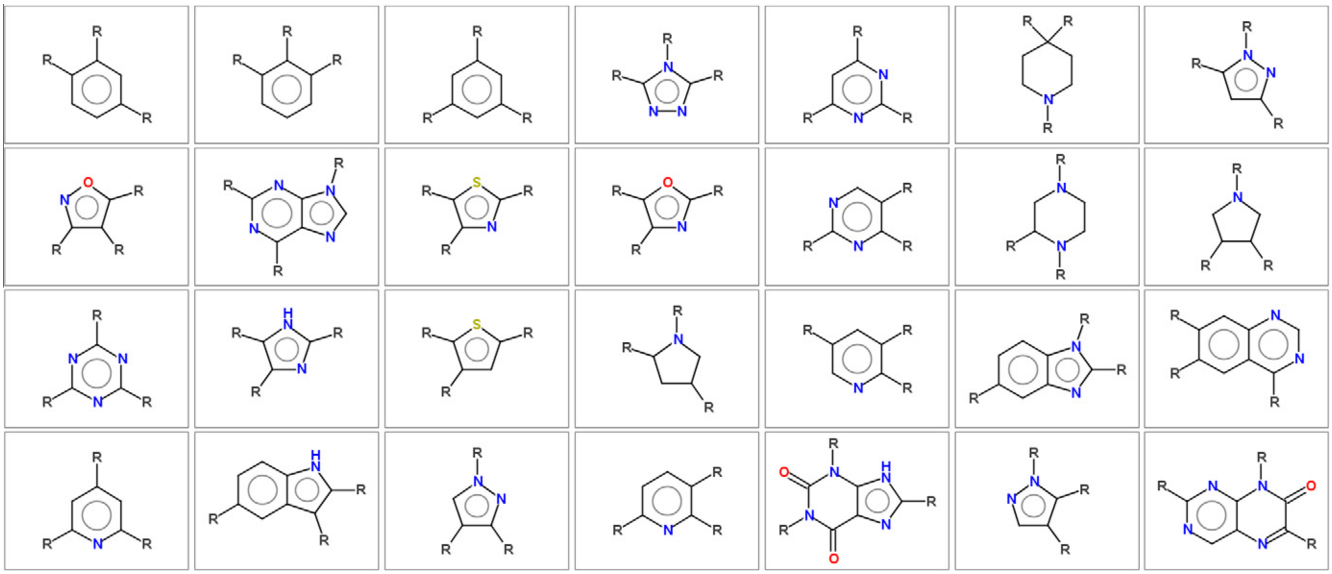

Figure 3. The most frequent ring systems with three connections.

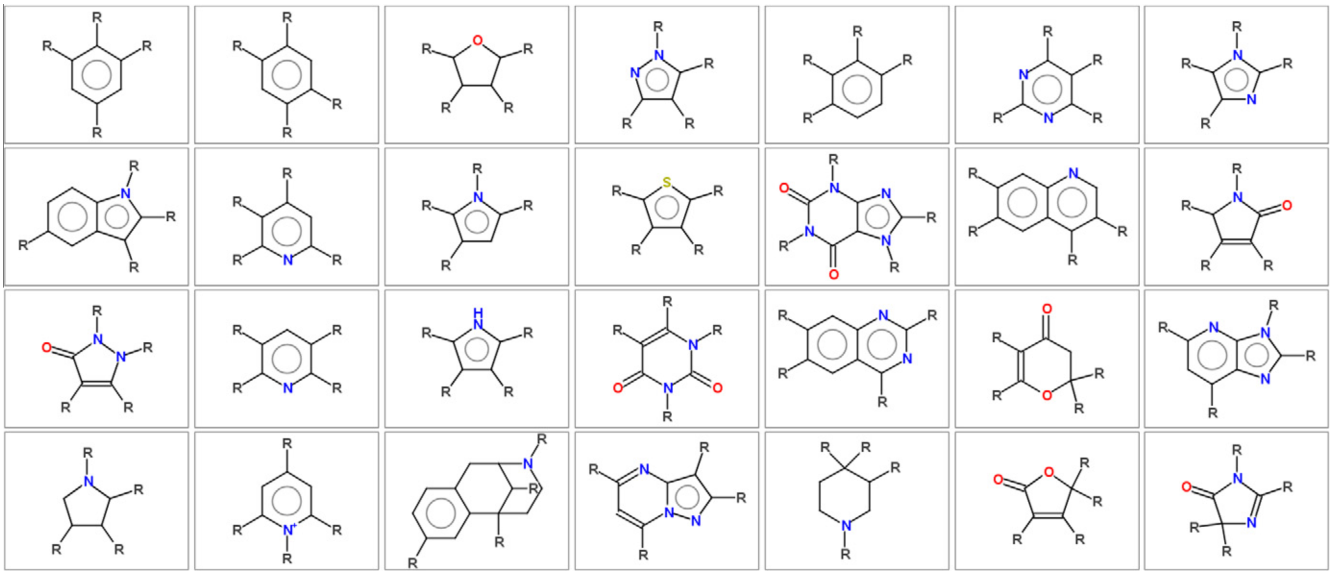

Figure 4. The most frequent ring systems with four connections.

Table 2
Descriptors used in the scaffold similarity searches

Descriptor	Details	Ref.
natoms	Size of the scaffold (number of non-hydrogen atoms)	
Bond types	Number of single, double, aromatic and exocyclic bonds in the scaffold	
rdistances	Topological distances between connection points	
log *P*	Calculated octanol-water partition coefficient	22
TPSA	Topological polar surface area	23
Shape features	Vector characterizing number of atoms at certain topological distance from the connection points	
Pharmacophore features	Vectors characterizing number of hydrogen bond donors and acceptors at certain topological distance from the connection points	
Calculated Hammett sigma parameters	Descriptor characterizing electron-donating -accepting power at connection bonds	24
Symmetry	Symmetry of connection points	
Frequency	Frequency of the scaffold in the ChEMBL database	

characterize scaffold shape and basic pharmacophore features. Positions of R groups are characterized by topological distances between them (three distances for system with three connection points and six distances for systems with four connection points).

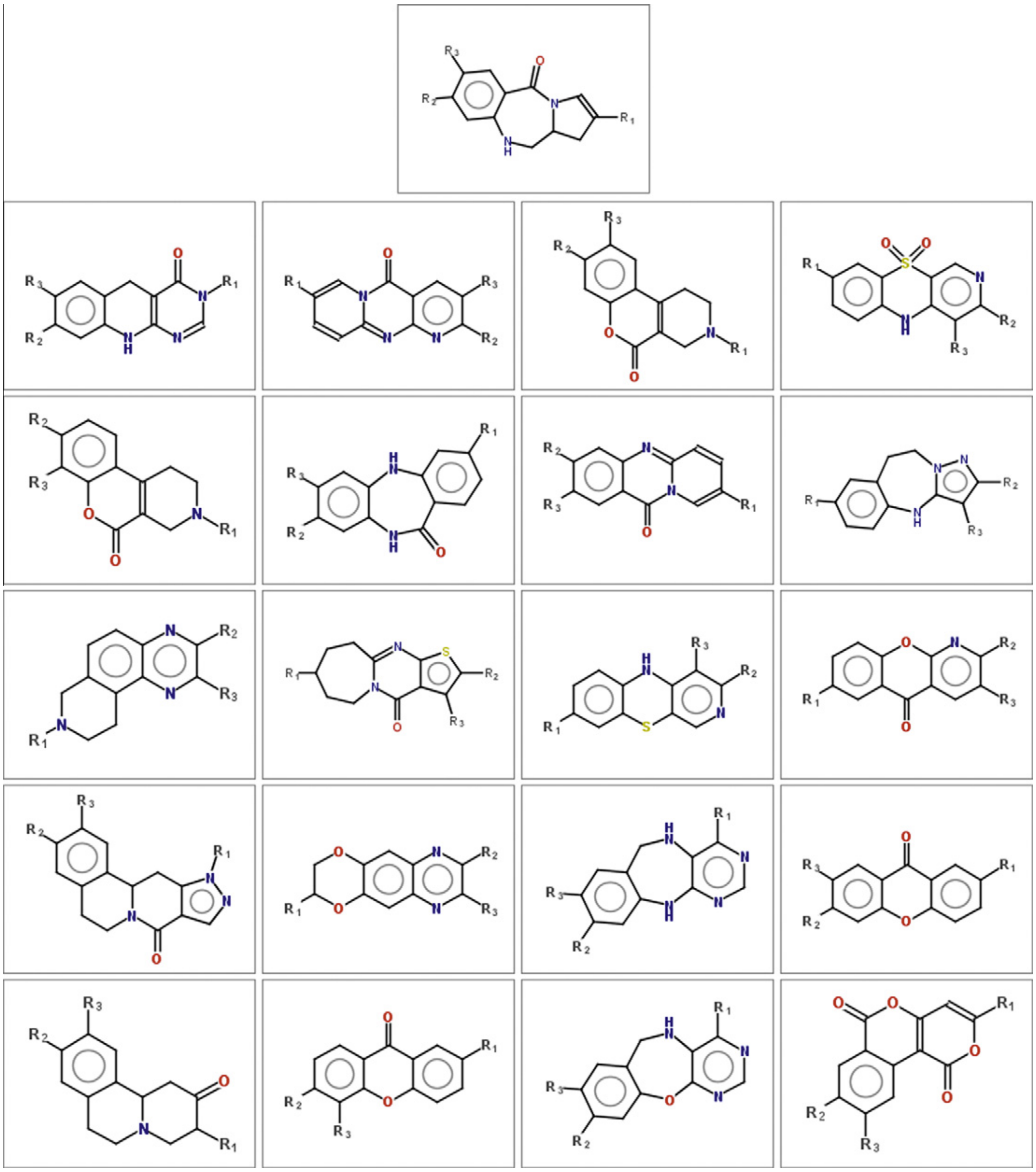

Figure 5. Results of the scaffold similarity search, query scaffold is on the top.

Electron-donating or -accepting power at connection points is characterized by quantum chemical parameters developed in house[24] that are compatible with experimental Hammett σ constants. Calculated log$P$ and TPSA values characterize the ADME properties of fragments, such as hydrophobicity, permeability and solubility. And finally it is necessary to keep track also on the symmetry of connection points as discussed below.

Scaffolds are stored in the database as SMILES strings with connection points marked as [R] atoms in SMILES, together with the calculated descriptors described above. SMILES representations are in canonical form, including also canonicalization of connection points. During the similarity search scaffolds compatible with query are identified simply as scaffolds with the most similar properties. In the search one needs to consider, of course, also possible permutations of connection points. For system with two connection points two such orientations are possible (like amide and inverted amide), for systems with three connection points six orientations and for systems with four connection points 24 possible permutations of R groups are possible. For symmetrical scaffolds the number of unique combinations is respectively lower (therefore also symmetry information need to be stored in the database).

The most similar scaffolds are identified as those where the sum of differences in their descriptors is minimal. Difference in numerical properties is calculated simply as difference in their absolute values, similarity of shape and pharmacophore features as Tanimoto-like similarity of the respective vectors. Since the magnitudes of all descriptors are roughly the same, all descriptors are considered with the same weight. The exception is TPSA that is scaled down by the factor of 50 (because the magnitude of TPSA is much larger than of other descriptors) and similarity in topological distances between connection points that is up-scaled by five. Adding importance to distances between connection points is necessary to guarantee that the 3D structures that may be created from hits would have reasonable overlap in exit vectors with the query.

Since all scaffolds were extracted from existing bioactive molecules reported in medicinal chemistry literature, one should not expect any obvious problems with presence of undesired substruc-

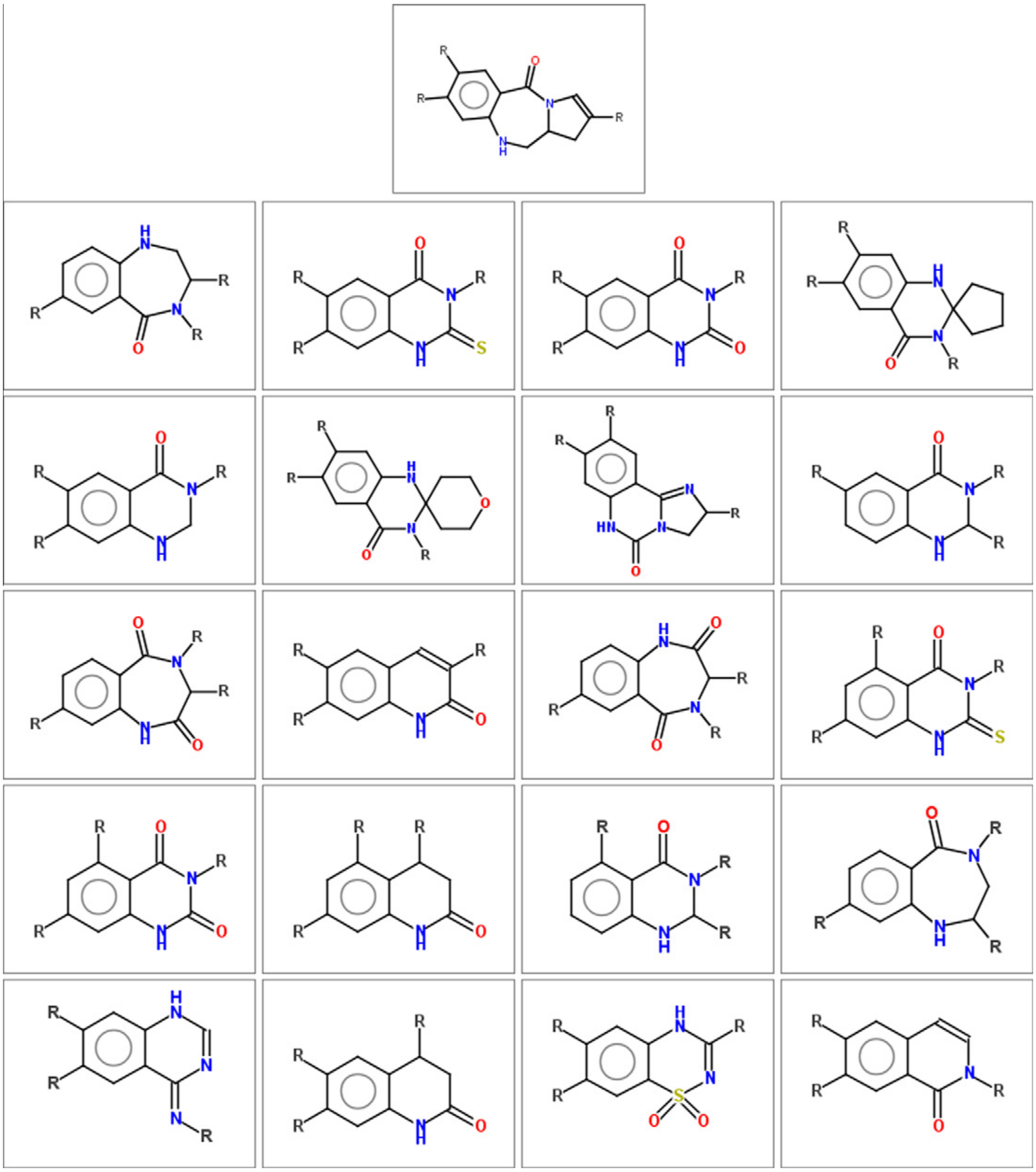

Figure 6. Results of the fingerprint based scaffold similarity search by PipelinePilot.

ture features or generally with their drug-likeness. Indeed, recent property analysis of the ChEMBL database (from which the scaffolds were extracted) documented[25] good ADME properties of molecules contained herein. Of course, some of the scaffolds, particularly those originating from natural products, are quite complex and may be not so easy to synthesize. If one is interested in getting only the most common or easy to make scaffolds, it is possible to filter the hits based on their frequency in the original ChEMBL database (this information is also stored as a parameter in the scaffold database) or additionally rank them using easy to calculate synthetic accessibility score[26] described recently.

To illustrate performance of our scaffold bioisosteric replacement procedure a similarity search to identify analogs of an interesting pyrrolo-benzodiazepine scaffold[27] with three connection points from the ligand of PDB complex 2K4L was performed. The results are shown in Figure 5. One can see that identified hits show good similarity to the query, both in pharmacophore features and shape, as well as in relative positions of the exit vectors. To see the differences between our approach and results of classical scaffold similarity search based on fingerprints we performed also a search with the same query and the same ring database using the FCFP_4 fingerprints as implemented in PipelinePilot.[28] Results of this search are shown in Figure 6. It is interesting to note that not a single scaffold is common between the two sets of hits. This, however, is not surprising, because two completely different sets of molecular descriptors were used in the similarity searches. The PipelinePilot search is based on generalized structural features, assuring that the hits will contain the same fragments as the query and will be structurally similar to it (and they indeed are). One can see, however, that the majority of the hits are smaller then the query, consisting of two instead of three fused rings, therefore overlap in the exit vectors cannot be so good. Our approach does not use any substructure features, and hits are identified according to their property and shape similarity to the query, as well as overlap in exit vectors. This comparison clearly documents that the method presented here allows identification of non-classical bioisosteric analogs not accessible through classical fingerprint-based similarity searches.

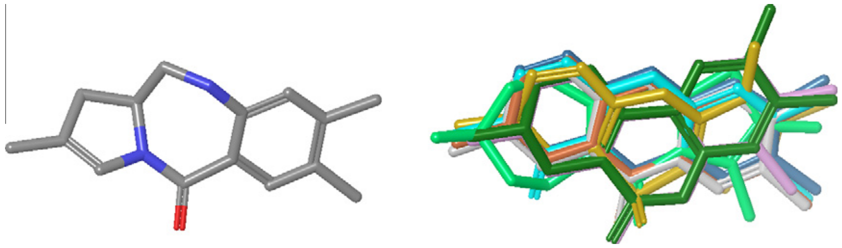

Figure 7. 3D structure of the query scaffold (left) and the best 10 bioisosteric analogs identified by topological similarity and subsequent 3D superimposition.

As already discussed, scaffold similarity searches by our approach are performed using 2D topological descriptors calculated from molecule connectivity information only (although 3D information is also included somehow by shape descriptors and particularly by considering distances between attachment points). Searching is therefore fast, because it does not require superimposition and matching of 3D exit vectors, what is the most time consuming part of other scaffold replacement approaches. Another advantage of using only topological descriptors as search criteria is the fact that one does not need to consider conformational flexibility, complicating otherwise similarity searches based on 3D superimposition. On the other hand, when it is necessary to identify bioisosteric analogs of ligands where the 3D structure of ligand-receptor complex is known, it is of advantage to use also this information. This may be done by a follow-up procedure where for the best 2D hits obtained by pure topological search all reasonable conformations are generated and conformations having the best 3D overlap with the query are selected. This is illustrated by Figure 7, where the 3D structure of query scaffold extracted from the PDB complex 2K4L is shown together with 10 top hits. One can see that suggested bioisosteric scaffolds have similar shape with the query and also exit vectors point to the same directions, providing reasonable candidates for construction of bioisosteric analogs of the original ligand.

2.3. Methodology details

The database of scaffolds used in this project was created by extracting scaffolds with three and four connections from bioactive molecules in the ChEMBL database,[20] version 11, where molecules with activity (IC_{50}, EC_{50} or K_i value) below 10 μm were considered to be ‘bioactive’. Scaffolds with up to 20 non-hydrogen atoms, containing only organic elements were considered. Actual extraction of scaffolds, their cleaning and canonicalization was done by the mib molecule processing engine from Molinspiration.[29] Calculation of scaffold properties was done by SmilesWorker, an in-house software written in Java. The web service to perform bioisosteric searches was implemented in Python, calling the SmilesWorker to do actual scaffold property comparison. Optional 3D similarity searches, including generation of conformations and alignment of scaffolds were done by FieldAlign software[30] from Cresset running in batch mode.

3. Conclusions

A methodology for identification of scaffolds with three and four connection points compatible in shape, pharmacophore features and ADME molecular properties with a query scaffold was described here. This method is implemented as a web service. It may be therefore accessed through an http web request with SMILES of a query (scaffold with 1–4 attachment points) as a parameter. A web interface allowing chemists to draw the query scaffold and mark its connection points directly within web page[31] provides convenient interactive access to this service. After submitting the request a list of compatible scaffolds is displayed that may be used as an idea generator for chemists to help them design new non-classical bioisosteric analogs. This is an enhancement of our earlier tool for identification of bioisosteric substituents and linkers.[17,18] The web service to identify compatible scaffolds is used also by a new in-house tool for automatic design and optimization of bioisosteric analogs. This tool, that we plan to describe in more details in a follow-up article, combines molecule fragmentation and identification of bioisosteric substituents and linkers reported previously with the system for scaffold replacements described here, followed by generation of conformations, 3D alignment and scoring of suggested analogs directly in the protein binding pocket. This tool may be used for identification of bioisosteric analogs having optimal interactions with the receptor, as well as for scaffold hopping applications, fragment growing and de novo molecule design.

References and notes

1. Langdon, S. R.; Brown, N.; Blagg, J. *J. Chem. Inf. Model* **2011**, *51*, 2174.
2. Hu, Y.; Stumpfe, D.; Bajorath, J. *J. Chem. Inf. Model.* **2011**, *51*, 1742.
3. Langdon, S. R.; Ertl, P.; Brown, N. *Mol. Inf.* **2010**, *29*, 366.
4. Hessler, G.; Baringhaus, K.-H. *Drug Discovery Today* **2010**, *7*, e263.
5. Bemis, G. W.; Murcko, M. A. *J. Med. Chem.* **1996**, *39*, 2887.
6. Lauri, G.; Bartlett, P. A. *J. Comput. Aided Mol. Des.* **1994**, *8*, 51.
7. Maass, P.; Schulz-Gasch, T.; Stahl, M.; Rarey, M. *J. Chem. Inf. Model* **2007**, *47*, 390.
8. Lewell, X.; Jones, A.; Bruce, C.; Harper, G.; Jones, M.; McLay, I.; Bradshaw *J. Med. Chem.* **2003**, *6*, 3257.
9. Broughton, H.; Watson, I. *J. Mol. Graphics Modell* **2004**, *23*, 51–58.
10. Bohl, M.; Loeprecht, B.; Wendt, B.; Heritage, T.; Richmond, N. J.; Willett, P. *J. Chem. Inf. Model* **1882**, *2006*, 46.
11. Bergmann, R.; Linusson, A.; Zamora, I. *J. Med. Chem.* **2007**, *50*, 2708.
12. Lounkine, E.; Wawer, M.; Wassermann, A. M.; Bajorath, J. *J. Chem. Inf. Model* **2010**, *50*, 6.
13. Beno, B. R.; Langley, D. R. *J. Chem. Inf. Model* **2010**, *50*, 1159.
14. Ertl, P.; Jelfs, S.; Muehlbacher, J.; Schuffenhauer, A.; Selzer, P. *J. Med. Chem.* **2006**, *49*, 4568.
15. Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. A.; Waldmann, H. *J. Chem. Inf. Model* **2007**, *47*, 47.
16. Ertl, P.; Schuffenhauer, A.; Renner, S. In *Chemoinformatics and Computational Chemical Biology*; Bajorath, J., Ed.; Humana Press, 2010; p 245.
17. Ertl, P. *J. Mol. Graphics Modell* **1998**, *16*, 11.
18. Ertl, P. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 374.
19. Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V.; Djoumbou, Y.; Eisner, R.; Guo, A. C.; Wishart, D. S. *Nucleic Acids Res.* **2011**, *39*, D1035.
20. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. *Nucl. Acids Res.* **2012**, *40*, D1100.
21. Irwin, J. J.; Shoichet, B. K. *J. Chem. Inf. Model* **2005**, *45*, 177.
22. Wildman, S. A.; Crippen, G. M. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868.
23. Ertl, P.; Rohde, B.; Selzer, P. *J. Med. Chem.* **2000**, *43*, 3714.
24. Ertl, P. *Quant. Struct.-Act. Relat.* **1997**, *5*, 377.
25. Gleeson, M. P.; Hersey, A.; Montanari, D.; Overington, J. *Nat. Rev. Drug Disc.* **2011**, *10*, 197.
26. Ertl, P.; Schuffenhauer, A. *J. Cheminf.* **2009**, *1*, 8.
27. Antonow, D.; Barata, T.; Jenkins, T. C.; Parkinson, G. N.; Howard, P. W.; Thurston, D. E.; Zloh, M. *Biochemistry* **2008**, *47*, 11818.
28. PipelinePilot 8.0, Accelrys, Inc., http://www.accelrys.com.
29. mib 2010.10, Molinspiration Cheminformatics, http://www.molinspiration.com.
30. FieldAlign 3.0, Cresset Group, http://www.cresset-group.com.
31. Ertl, P. *J. Cheminf.* **2010**, *2*, 1.