Bioorganic & Medicinal Chemistry 23 (2015) 3183–3191

Contents lists available at ScienceDirect

Bioorganic & Medicinal Chemistry

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

Systematic assessment of scaffold hopping versus activity cliff formation across bioactive compound classes following a molecular hierarchy

Dagmar Stumpfe, Dilyana Dimova, Jürgen Bajorath *

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

ARTICLE INFO

Article history:
Received 25 February 2015
Revised 21 April 2015
Accepted 24 April 2015
Available online 1 May 2015

Keywords:
Molecular scaffolds
Topology
Compound-scaffold-skeleton hierarchy
Scaffold hopping
Activity cliffs
Compound data mining

ABSTRACT

Scaffold hopping and activity cliff formation define opposite ends of the activity landscape feature spectrum. To rationalize these events at the level of scaffolds, active compounds involved in scaffold hopping were required to contain topologically distinct scaffolds but have only limited differences in potency, whereas compounds involved in activity cliffs were required to share the same scaffold but have large differences in potency. A systematic search was carried out for compounds involved in scaffold hopping and/or activity cliff formation. Results obtained for compound data sets covering more than 300 human targets revealed clear trends. If scaffolds represented multiple but fewer than 10 active compounds, nearly 90% of all scaffolds were exclusively involved in hopping events. With increasing compound coverage, the fraction of scaffolds involved in both scaffold hopping and activity cliff formation significantly increased to more than 50%. However, ~40% of the scaffolds representing large numbers of active compounds continued to be exclusively involved in scaffold hopping. More than 200 scaffolds with broad target coverage were identified that consistently represented potent compounds and yielded an abundance of scaffold hops in the low-nanomolar range. These and other subsets of scaffolds we characterized are of prime interest for structure–activity relationship (SAR) exploration and compound design. Therefore, the complete scaffold classification generated in the course of our analysis is made freely available.

1. Introduction

The concept of molecular *scaffolds*[1,2] is of central relevance in medicinal chemistry. Scaffolds are extracted from molecules to approximate core structures and are explored as building blocks for compound design,[2] as privileged structural motifs to address specific target families,[3] or to organize compound classes in a systematic manner.[4,5] Following the definition that is most widely applied in medicinal chemistry, scaffolds are obtained from compounds by removing all substituents, while retaining ring systems and linker fragments between rings.[1] From scaffolds, one can further abstract through the generation of so-called cyclic skeletons (CSKs)[6] by converting all heteroatoms to carbon and setting all bond orders to 1. These abstract molecular representations[1,6] make it possible to organize scaffolds according to topological criteria.[2]

For the identification of structurally diverse active compounds, the *scaffold hopping* concept[7] is widely applied. Scaffold hopping refers to the search for pairs of compounds sharing the same activity but containing different core structures, thereby establishing chemical novelty within a confined bioactivity space. Therefore, through the application of computational methods, for example, pharmacophore or whole-molecule similarity search, it is attempted to extrapolate from known active compounds and identify others that share the same activity but are structurally distinct.[7–9] As such, scaffold hopping is typically considered the ultimate goal of virtual screening campaigns[7–10] and of high-throughput screening data analysis.[11]

Through large-scale compound data mining, it has been shown that many target-specific compound activity classes contain large numbers of different scaffolds[12] and that diverse scaffolds often represent specifically active compounds with comparably high potency.[13] These findings suggest that the computational identification of scaffold hops might often be less challenging than frequently assumed. Furthermore, these findings also indicate that many current targets are capable of binding a fairly wide spectrum of different core structures and hence represent promising small molecular targets.

* Corresponding author. Tel.: +49 228 2699 306; fax: +49 228 2699 341.
E-mail address: bajorath@bit.uni-bonn.de (J. Bajorath).

http://dx.doi.org/10.1016/j.bmc.2015.04.067
0968-0896/

Scaffold hops can also be considered in the context of *activity landscape* design and analysis.[14] The activity landscape concept is also popular in medicinal chemistry. An activity landscape is generally defined as a graphical representation that integrates structural similarity and potency relationships between compounds sharing the same bioactivtiy.[14] In the context of activity landscape analysis, scaffold hops have been rationalized as *similarity cliffs*, i.e., pairs of active compounds containing different scaffolds but having comparable potency.[15] So-defined similarity cliffs represent a prevalent activity landscape feature[16] and contrast *activity cliffs*,[17,18] another cardinal (yet rare) feature of activity landscapes that represents a focal point of structure–activity relationship (SAR) analysis.[18] Activity cliffs are generally defined as pairs of structurally similar compounds sharing the same activity but having a large difference in potency.[17] Accordingly, the comparison of compounds forming activity cliffs is likely to reveal SAR determinants. When rationalized on the basis of scaffolds, a necessary condition for activity cliff formation is that both compounds contain the same scaffold.[18]

Similarity cliffs/scaffolds hops and activity cliffs encode opposite scaffold/potency relationships: Pairs of compounds constituting a scaffold hop must contain distinct scaffolds but have comparable potency, whereas compounds forming activity cliffs must share the same scaffold but exhibit a large difference in potency. Hence, similarity cliffs/scaffolds hops and activity cliffs define opposite ends of the activity landscape feature spectrum, have fundamentally different SAR information content, and are usually not viewed in context.

Herein, we generate a compound-scaffold-CSK hierarchy for compounds with activity against more than 300 human targets to systematically explore the capacity of molecular core structures to engage in scaffold hopping or the formation of activity cliffs. Because scaffolds can often be chemically very similar (e.g., only be distinguished by individual heteroatom replacement in rings or variation of aliphatic linker fragments), which might bias scaffold hopping assessment,[2] our analysis was focused on scaffolds with well-defined differences in molecular topology.[13,19] For comparison, activity cliff formation was determined within the same analysis context, by individually focusing on each scaffold and compounds containing this scaffold. Only high-confidence activity data were considered in our analysis. The results of our systematic assessment of scaffold hopping versus activity cliff formation are presented herein.

2. Materials and methods

2.1. Compound data sets

Compounds and activity data were assembled from ChEMBL (version 20).[20] Only compounds with precisely defined equilibrium constants (K_i values) for human targets at the highest confidence level (ChEMBL confidence score 9) were selected.

Compounds with multiple activity annotations for the same target were only selected if all values fell within one order of magnitude. In this case, the average potency value was calculated and used as the final activity annotation.

On the basis of these selection criteria, a total of 84,080 compounds with activity against 747 targets were obtained. Compound data mining and data analysis were performed with in-house generated KNIME workflows[21] and the aid of OpenEye toolkits.[22]

2.2. Scaffolds and cyclic skeletons

From all qualifying compounds, scaffolds were extracted according to the most widely used definition of scaffolds originally introduced by Bemis and Murcko (BM).[1] Following this approach, scaffolds are obtained from compounds by removing all substituents while retaining ring systems and linkers between them. From BM scaffolds, CSKs were obtained by converting all heteroatoms to carbon and setting all bond orders to 1. Thus, by definition, each CSK covered a set of topologically equivalent scaffolds.[2] Calculation of scaffolds and CSKs from compounds established a compound-scaffold-CSK hierarchy.[2]

Two CSKs were considered 'topologically distinct' if they differed in the number of rings and, in addition, if they were not involved in a substructure relationship (i.e., a CSK was not a substructure of another).[13] Differences in linker length alone were not sufficient to classify CSKs as topologically distinct. These restrictions were introduced to ensure that scaffolds involved in hopping events had substantial structural differences, thereby ruling out 'easy' scaffold hopping instances (involving, e.g., scaffolds only distinguished by minor heteroatom replacements in rings). We also note that relationships between structurally distinct scaffolds contained in compounds sharing the same activity might often not be straightforward to rationalize from a medicinal chemistry perspective without detailed knowledge of compound binding modes.

Only target-based compound data sets containing at least 10 different scaffolds were retained for further analysis. In addition, scaffolds consisting of a single six-membered ring were omitted from the analysis, due to their generic nature and large numbers of small compounds containing these scaffolds.

Taken together, these restrictions reduced the pool of qualifying compounds to 78,150 with activity against 347 different targets (instead of 747 for the original compound selection). These 347 target sets (activity classes) yielded a total number of 34,589 scaffolds and 21,817 CSKs (scaffolds and CSKs were counted multiple times if they occurred in multiple activity classes) including 18,374 and 8654 unique scaffolds and CSKs, respectively.

Scaffold-to-compound ratios were also determined. A total of 24,108 scaffolds (nearly 70%) represented only a single active compound while the remaining 10,481 scaffolds represented multiple compounds. Among these, 1265 scaffolds were detected that represented 10 or more compounds active against the same target.

Scaffolds and CSKs were isolated with in-house generated python scripts based upon the OpenEye OEChem Toolkit.[22]

2.3. Scaffold hopping and activity cliff criteria

Two compounds formed an activity cliff if they shared the same scaffold and had a potency difference of at least two orders of magnitude (100-fold) against the same target (i.e., within a given activity class).[18] By contrast, for the formation of a scaffold hop, two compounds were required to have topologically distinct scaffolds (i.e., scaffolds yielding topologically distinct CSKs, as defined above) and a potency difference within one order of magnitude against the same target.[13]

2.4. Scaffold hopping versus SAR transfer

It should be noted that this conventional definition of scaffold hopping does not take into consideration how different scaffolds might be substituted. Core structures and substitution patterns make contributions to compound activity. Hence, if one would like to quantify the difference in contribution to activity for two distinct scaffolds, substitutions patterns in active compounds containing these scaffolds must be identical. Furthermore, series of active compounds with distinct scaffolds but pairwise corresponding (identical) substitution patterns represent SAR transfer events and make it possible to monitor SAR progression on the basis of different scaffolds. However, quantitative scaffold contributions to activity or SAR transfer/progression are not part of the