752 J. Med. Chem. 2010, 53, 752–758
DOI: 10.1021/jm9014229

Journal of Medicinal Chemistry Article

Systematic Analysis of Public Domain Compound Potency Data Identifies Selective Molecular Scaffolds across Druggable Target Families

Ye Hu, Anne Mai Wassermann, Eugen Lounkine, 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

Received September 24, 2009

Molecular scaffolds that yield target family-selective compounds are of high interest in pharmaceutical research. There continues to be considerable debate in the field as to whether chemotypes with a priori selectivity for given target families and/or targets exist and how they might be identified. What do currently available data tell us? We present a systematic and comprehensive selectivity-centric analysis of public domain target−ligand interactions. More than 200 molecular scaffolds are identified in currently available active compounds that are selective for established target families. A subset of these scaffolds is found to produce compounds with high selectivity for individual targets among closely related ones. These scaffolds are currently underrepresented in approved drugs.

Introduction

Twenty years ago Evans et al.[1] first put forward the idea that chemotypes might exist that preferentially bind to a given target class, and the characterization of molecular scaffolds active against individual target classes has ever since been a topic of intense research in pharmaceutical settings.[2] The notion of "privileged substructures"[1] is highly attractive for drug discovery and chemical biology because they might ultimately be evolved into chemical entities that are selective for individual targets. However, it has been shown that substructures thought to be target class-characteristic typically also appeared in compounds active against other target families[3] and exclusive binding of known chemotypes to given target classes has not been confirmed to this date.

The concept of privileged substructures touches upon a much more general question in molecular probe and drug discovery, namely, how to generate small molecules that are selective for a target of interest within a target family.[4] Currently, only little is known about the relationship between molecular selectivity at the level of target families and individual targets[5] and it is not understood what the likelihood might be to discover selective compounds for different target classes.

Target selectivity (TS[a]) is typically explored on a case-by-case or family basis, and systematic analyses of compound selectivity data across different families are currently not available. With the growing availability of small molecule structure−activity data in the public domain, we are now in a position to explore molecular selectivity in a way that fundamentally differs from traditional case-by-case studies. This is accomplished by focusing, in an unbiased manner, on what data currently available for different target families might tell us about the selectivity of known molecular scaffolds and compounds. Such an analysis also provides a basis for the identification of new selective compounds.

To these ends, we have designed and carried out a systematic computational selectivity profile analysis of the BindingDB database,[6] a major public domain repository of activity information of small molecules, which we have found to represent by far the currently most comprehensive source of activity annotations that can be transformed into compound selectivity data. BindingDB contains ~31000 compound entries with ~57000 activity measurements taken from the scientific literature. Because of the ensuing high level of accuracy of the activity annotations, BindingDB is particularly suitable for a large-scale exploration of molecular selectivity. It represents an up-to-date view of the current scientific literature and knowledge in the field. The results of our analysis are reported herein and offer some surprising insights into the availability of target class-selective molecular scaffolds that might be evolved into target-selective compounds.

Results

Compounds, Targets, and Selectivity Sets. A total of 6343 compounds active against 259 human targets (Supporting Information Table S1) were extracted from BindingDB. Many of these compounds were active against multiple targets, yielding a total of 17 929 compound−target combinations, and we identified 520 target pairs that shared at least five active compounds (with an average of 34 molecules per pair). For each molecule active against a target pair, its target selectivity was calculated as $TS = pK_i^A - pK_i^B$ (where pK_i^A and pK_i^B refer to the logarithmic potency value of the compound against targets A and B, respectively). Absolute TS values of selected compounds ranged from 0 to 6.86, i.e., from equal potency (and thus no selectivity) to potency differences of nearly 7 orders of magnitude (i.e., highest selectivity for one of two targets). Each pair of targets and the compounds they shared represented 1 of 520 selectivity sets for further analysis.

*To whom correspondence should be addressed. Phone: +49-228-2699-306. Fax: +49-228-2699-341. E-mail: bajorath@bit.uni-bonn.de.

[a] Abbreviations: TS, target selectivity.

 Published on Web 12/09/2009

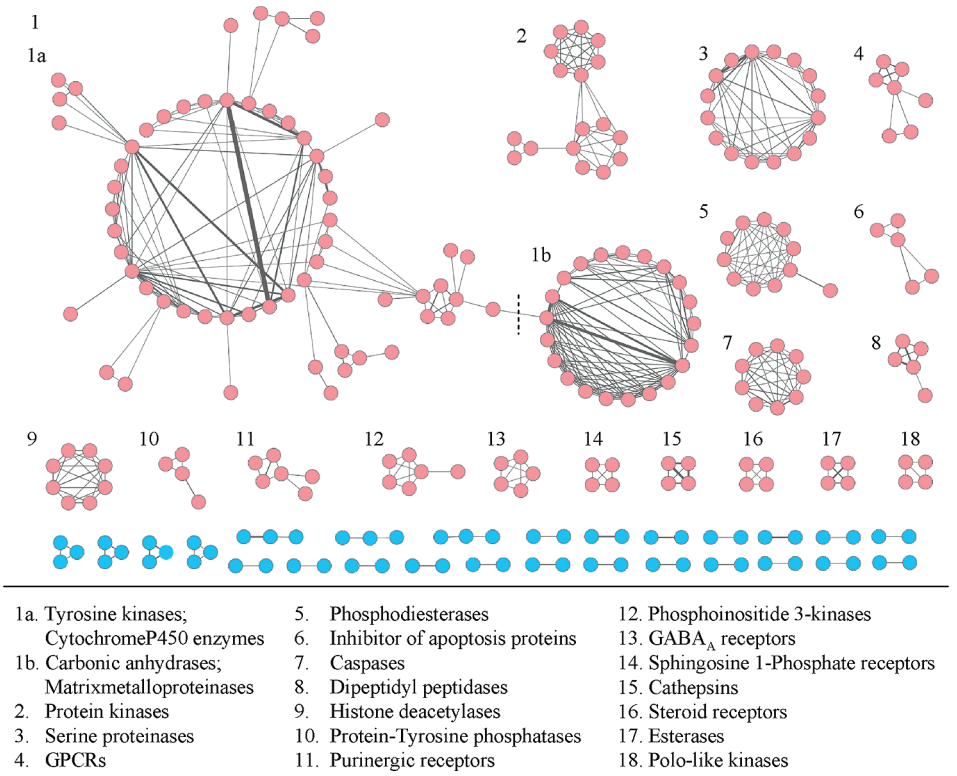

Figure 1. Target pair network. Nodes represent targets, and edges are drawn between nodes if they share at least five compounds. The network representation reveals a total of 18 communities containing at least four targets. Community 1 is subdivided (dashed vertical line) on the basis of target family membership. Nodes in communities are colored light-red and others light-blue.

Target Pair Network and Target Communities. The 259 human targets participated in multiple target pairs, and a network representation was generated to analyze target relationships (Figure 1). In the network, nodes represent targets and edges are drawn between nodes if they share at least five molecules. This number of molecules was chosen to control network noise and ensure the reliability of selectivity profiling. The width of edges is scaled according to the number of active compounds shared by a target pair. The network reveals the presence of 18 separate and in part densely connected communities containing at least four targets (smaller communities were not considered). These communities are found to represent different target families (Figure 1). Thus, known biological activities of small molecules organize targets into functional families, as has been observed in drug−target networks based on chemical drug similarity.[7,8] For the purpose of our selectivity studies, network analysis was only required to organize and preselect target communities. The largest community identified in our network (community 1) contains 82 targets that mainly belong to three target families, i.e., tyrosine kinases, carbonic anhydrases (CAs), and matrix metalloproteinases (MMPs). Tyrosine kinases form a large subset (1a) on the left in Figure 1, while CAs and MMPs form a densely connected subset (1b) on the right (i.e., they share many active compounds). These two subsets are linked by cytochrome P450 enzymes and steroid sulfatase. By removal of the edge connecting steroid sulfatase and CA2, community 1 was divided into subsets 1a and 1b, hence producing a total of 19 communities for further analysis. These communities consisted of 4−59 targets and 8−2252 active compounds. Details for each community are provided in Supporting Information Figure S1 and Supporting Information Table S2.

Scaffolds and Selectivity Profiles. From the initial pool of 6343 active compounds, hierarchical molecular scaffolds[9] were isolated that represented at least five active compounds, yielding a total of 210 distinct scaffolds, listed in Supporting Information Table S3. For each target within a community with at least five ligands having the same scaffold, the active compounds were collected. The TS values for target pairs containing this target and the active compounds were calculated. The median of these TS values is an indicator of scaffold selectivity for the particular target. A high median TS value means that a scaffold shows high selectivity toward the target over other targets within the community. A negative median TS value indicates that the scaffold produces compounds that are selective for other members of the community. On the basis of median TS values, a scaffold−target heat map was generated to represent the *target selectivity profile* of each scaffold within a community. Furthermore, for each scaffold found in a community, all relevant compounds used in the generation of the target−scaffold heat map were pooled, and the median of their absolute TS values was calculated. In this case, high median values indicate that a scaffold produces many compounds with different potency against individual targets and hence a differentiated selectivity profile within a community. A scaffold−community heat map was also generated to represent the *community selectivity profile* of each scaffold. Supporting Information Figure S1 reports the number of scaffolds in each community. For two communities (6 and 13), no relevant scaffolds were found. For the other communities, the number of scaffolds ranged from 1 to 102. For individual targets, between 1 and 32 scaffolds were found.

Target and Community Selectivity of Scaffolds. The scaffold−target heat map for community 3 representing serine

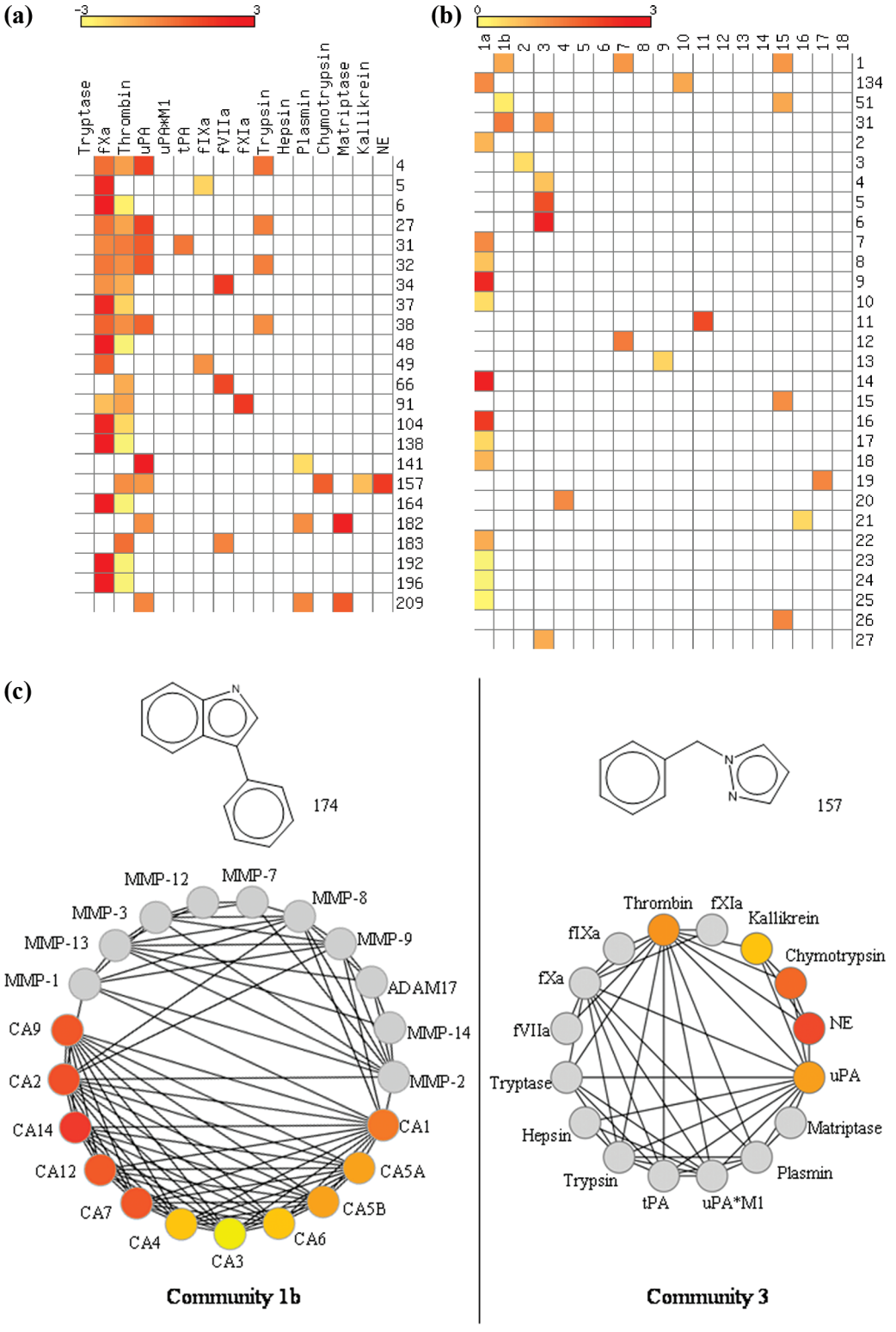

Figure 2. Target and community selectivity profiles. (a) The heat map representing the target selectivity profile of community 3 is shown. Targets form columns and scaffolds rows. A cell corresponding to a scaffold−target combination is filled if the scaffold is present in at least five compounds active against the target and color-coded according to median TS values. (b) A section of the community selectivity profiles is shown. Here, columns represent communities and rows scaffolds. Cells are color-coded according to absolute median TS values. (c) Shown are community-centric target selectivity profiles for two representative scaffolds (174 and 157) that are selective for communities 1b and 3, respectively. Nodes are color-coded by median TS values of active compounds according to part a. Thus, for targets with red nodes, the scaffold has highest potential to produce selective compounds. Targets for which fewer than five active compounds containing the scaffold exist are depicted as gray nodes. Edges between nodes are drawn according to Figure 1.

proteases is shown in Figure 2a as an example (Supporting Information Figure S2 shows the corresponding heat maps for all communities). Median TS values are represented via a continuous color spectrum ranging from −3 (yellow) to 3

(red). A key observation in Figure 2a is that individual scaffolds mostly display different selectivity against related targets, and this trend is observed for all communities (Supporting Information Figure S2). For example, scaffold 6 represents compounds that are active against factor Xa and thrombin but these inhibitors are much more potent against factor Xa and thus highly selective for this target. Similar observations are made for scaffolds 48, 104, 138, 164, 192, and 196, all of which differentiate between these two proteases. Other scaffolds represent compounds that inhibit proteases more broadly. For example, scaffold 157 represents inhibitors of five proteases. However, the compounds are more potent against neutrophil elastase than against the other targets. Supporting Information Figure S2 shows that selectivity-conferring scaffolds were found for many targets across all communities, and Supporting Information Table S4 lists the scaffolds that are most selective for individual targets. The number of scaffolds per target varies in part significantly, but for many targets only a single scaffold is found that yields selective compounds relative to the other targets of the communities.

Figure 2b shows a heat map representing the community selectivity profile of a subset of scaffolds (and Supporting Information Figure S3 shows the corresponding profiles for all 210 scaffolds). Here, median of absolute TS values are represented via a continuous color spectrum ranging from 0 (yellow) to 3 (red). A value of 0 means that the scaffold does not generate selective compounds across the community, and a value of 3 means that compounds containing the scaffold display at least a 1000-fold difference in potency against targets within the community. Figure 2c shows two representative examples of scaffolds that act on multiple targets within a community yielding substantial differences in compound selectivity. A key observation in Figure 2b is that only four scaffolds (1, 31, 51, and 134) are active against multiple communities. These scaffolds mainly correspond to compounds that are nonselective. By contrast, all other scaffolds are found to specifically act on only one community. However, these community-selective scaffolds display a distinctly different potential to yield target-selective compounds. Supporting Information Table S5 reports the potential of community-selective scaffolds to produce target-selective compounds. A total of 111 scaffolds display a target-selective tendency (median $|TS| \geq 1$), and 37 of these scaffolds represent compounds with at least 100-fold potency differences against other community targets.

Taken together, the results of the target and community selectivity profile analysis reveal that community-selective scaffolds are consistently found and that a subset of these scaffolds has in part significant potential to yield target-selective compounds within their communities. Figure 3 shows examples of scaffolds having high potential to produce target-selective compounds for major drug targets including, among others, receptor tyrosine kinases, G-protein-coupled receptors, or caspases.

Community-selective scaffolds can also be utilized to identify new target-selective compounds, as illustrated in Figure 4. For example, the community and target selectivity profiles suggest that compounds containing scaffold 37 should have high potential to produce inhibitors that are selective for factor Xa over thrombin. When a nonpublic domain database was searched,[10] two compounds containing this scaffold were identified that are currently not available in BindingDB and both of these compounds are indeed

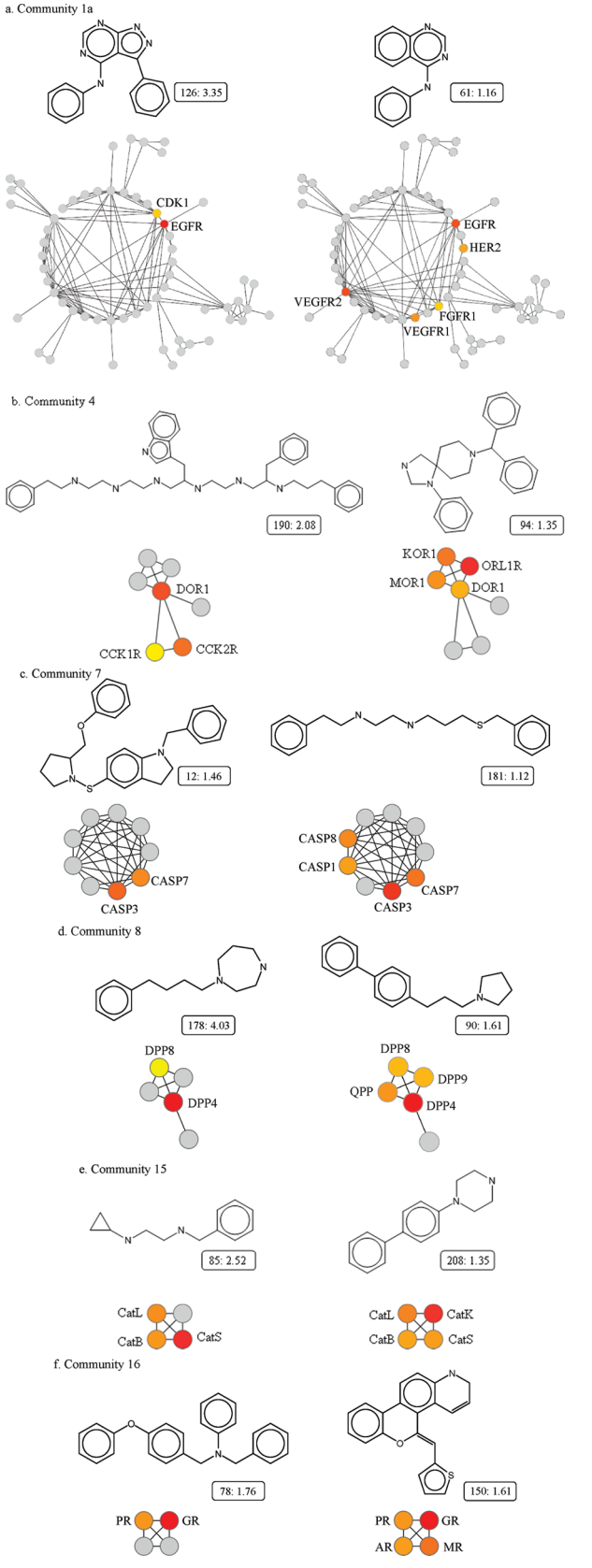

Figure 3. Community-selective scaffolds. For different target communities, selective scaffolds are shown that have high potential to yield target-selective compounds. Scaffolds have "scaffold number: median TS value" annotations. On the left of each figure, the scaffold with the highest median TS value in the community is shown. On the right, another scaffold with a broader selectivity profile is shown.

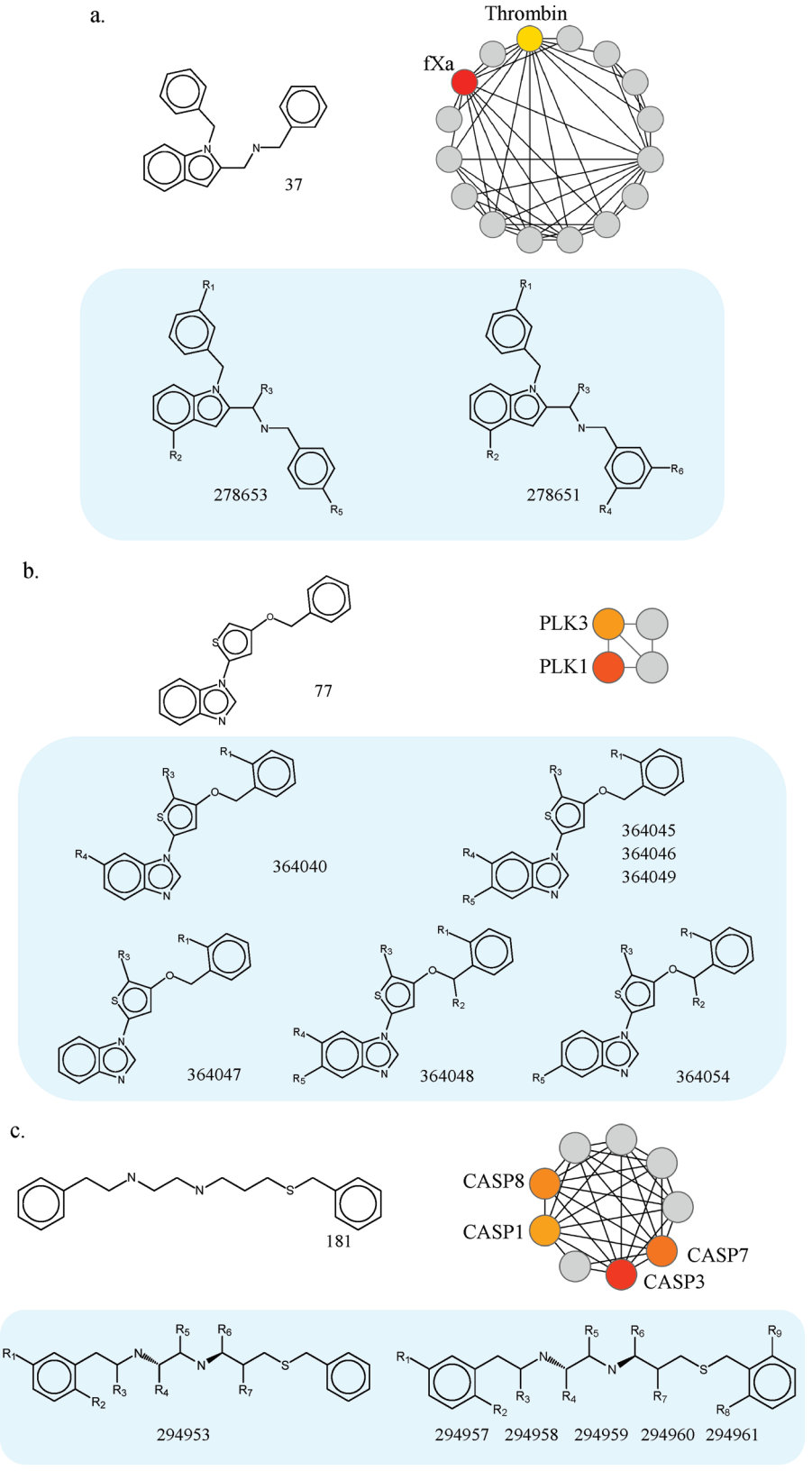

Figure 4. Searching for selective compounds. Examples of scaffolds (and their community selectivity profiles) are shown that were utilized to search the MDDR database. Compounds found to have the predicted selectivity are shown on a blue background. MDDR compounds are license-protected and therefore represented as Markush structures. Each Markush structure is annotated with MDDR identifiers of the compounds it represents.

reported to be highly selective for factor Xa (Figure 4a). Similarly, compounds were found containing scaffold 77 (Figure 4b) and 181 (Figure 4c) that were inhibitors of polo-like kinase 1 and caspase 3, respectively, with no reported activity against other community targets. The target selectivity profile for the caspase community also

suggested that compounds containing scaffold 12 should have comparable potency for caspase 3 and 7 but not for other members of the caspase family. This prediction is confirmed by a recent study aiming at the development of isatin sulfonamides as caspase inhibitors.[11] Four compounds containing scaffold 12 were reported that inhibited both caspase 3 and 7 with nanomolar potency and were ~200-fold selective over caspases 1, 6, and 8.

Distribution of Community-Selective Scaffolds in Drugs. We have also determined the distribution of community-selective scaffolds in known drugs. Therefore, 1247 approved drugs were retrieved from DrugBank[12] and a total of 726 unique scaffolds were isolated from them. Only 11 of these drug scaffolds were also found within the set of 206 target community-selective scaffolds, illustrating that these scaffolds are currently underrepresented in approved drugs. Because a subset of community-selective scaffolds is target-selective, as discussed above, chemical exploration of these scaffolds might be expected to provide further opportunities for drug discovery.

Discussion

The focal point of our study has been the exploration of small molecule selectivity on a large scale. Ligand preferences of target families have thus far been explored by calculating the frequency of occurrence of selected substructures in compounds active against individual target families. Such statistical approaches are based on a binary formulation of biological activity (i.e., active vs inactive) and do not take selectivity into account. The approach reported herein is specifically focused on exploring the selectivity of active compounds at the level of molecular scaffolds. It is data-driven and does not employ any preconceived notions of structure-selectivity relationships or target family assignments. Rather, the target pair network provides a data structure to organize known targets into communities based on shared ligands. Moreover, community and target selectivity profiles make it possible to assign molecular scaffolds to communities and explore their potential to produce target-selective compounds. Key findings of our analysis include the following: more than 200 scaffolds exist in currently available public domain compounds that are selective for communities of closely related targets, and a majority of these scaffolds yield compounds that are either selective for individual targets or display a target-selective tendency. These scaffolds can also be utilized to search for other active compounds having a desired selectivity profile. Hence, currently available data suggest that a substantial molecular knowledge base exists to generate target class- or target-selective small molecular probes or leads. Because we focus on currently available activity data of small molecules, the scaffold and selectivity information we report should provide many alternative starting points for a further experimental evaluation of scaffold selectivity profiles and the chemical exploration of molecular selectivity.

Methods

In order to comprehensively cover public domain compound data that could be utilized to extract target selectivity information relevant for our analysis, we also analyzed bioassays available in Pubchem[13] as a potential source. Compound screens were analyzed for appropriate selectivity information. However, given the target pair criteria applied in our study, only three relevant target pairs could be identified in Pubchem. The results of our analysis are reported in Supporting Information Figure S4. From BindingDB, compounds with reported activity (IC_{50} and/or K_i values) against human targets were extracted. If multiple potency measurements were reported in a BindingDB entry, their geometric mean was calculated to yield a single potency value. For each molecule active against a target pair, its target selectivity (TS) was calculated as $TS = pK_i^A - pK_i^B$. TS and median TS values are simple, intuitive, and continuous measures of target selectivity that do not require the definition of selectivity thresholds. Conventional hierarchical scaffolds were derived according to Bemis and Murcko.[9] These scaffolds represent ring systems connected by linkers after removal of substituents. Compounds and scaffolds were recorded and processed in SMILES format.[14] The target pair network was generated using Cytoscape.[15] Target communities connected only by intra- but no intercommunity edges and comprising a minimum of four targets were isolated. Community- and target-selective scaffolds were searched in the MDDR database[10] and compared to scaffolds extracted from DrugBank.[12] Nonselective scaffolds were not further analyzed. The community- and target-based selectivity profile analysis was carried out with in-house generated Pipeline Pilot[16] and Perl programs.

Supporting Information Available: Tables S1−S5 listing all investigated targets, target communities, scaffolds, target-selective scaffolds, and community-selective scaffolds, respectively; Table S6 listing the results of scaffold overlap analysis between selectivity-conferring and current drug scaffolds; Figures S1−S3 showing target and scaffold distributions, target selectivity profiles, and community selectivity profiles, respectively; Figure S4 showing the results of target pair analysis in PubChem. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Evans, B. E.; Rittle, K. E.; Bock, M. G.; Dipardo, R. M.; Freidinger, R. M.; Whitter, W. L.; Lundell, G. F.; Veber, D. F.; Anderson, P. S. Methods for Drug Discovery: Development of Potent, Selective, Orally Effective Cholecystokinin Antagonists. *J. Med. Chem.* **1988**, *31*, 2235–2246.

(2) Horton, D. A.; Bourne, G. T.; Smythe, M. L. The Combinatorial Synthesis of Bicyclic Privileged Structure or Privileged Substructures. *Chem. Rev.* **2003**, *103*, 893–930.

(3) Schnur, D. M.; Hermsmeier, M. A.; Tebben, A. J. Are Target-Family-Privileged Substructures Truly Privileged? *J. Med. Chem.* **2006**, *39*, 2000–2009.

(4) Bajorath, J. Computational Analysis of Ligand Relationships within Target Families. *Curr. Opin. Chem. Biol.* **2008**, *12*, 352–358.

(5) Karaman, M. W.; Herrgard, S.; Treiber, D. K.; Gallant, P.; Atteridge, C. E.; Campbell, B. T.; Chan, K. W.; Ciceri, P.; Davis, M. I.; Edeen, P. T.; Faraoni, R.; Floyd, M.; Hunt, J. P.; Lockhart, D. J.; Milanov, Z. V.; Morrison, M. J.; Pallares, G.; Paterl, H. K.; Pritchard, S.; Wodicka, L. M.; Zarrinkar, P. P. A Quantitative Analysis of Kinase Inhibitor Selectivity. *Nat. Biotechnol.* **2008**, *26*, 127–132.

(6) 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.

(7) Paolini, G. V.; Shapland, R. B. H.; van Hoorn, W. P.; Mason, J. S.; Hopkins, A. L. Global Mapping of Pharmacological Space. *Nat. Biotechnol.* **2006**, *24*, 805–815.

(8) Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. Relating Protein Pharmacology by Ligand Chemistry. *Nat. Biotechnol.* **2007**, *25*, 197–206.

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

(10) *MDL Drug Data Report* (*MDDR*), version 2005.2; Symyx Software: San Ramon, CA, 2005.

(11) Smith, G.; Glaser, M.; Perumal, M.; Nguyen, Q. D.; Shan, B.; Arstad, E.; Aboagye, E. O. Design, Synthesis, and Biological Characterization of a Caspase 3/7 Selective Isatin Labeled with 2-[^{18}F]fluoroethylazide. *J. Med. Chem.* **2008**, *51*, 8057–8067.

(12) Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: A

Knowledgebase for Drugs, Drug Actions and Drug Targets. *Nucleic Acids Res*. **2008**, *36*, D901–D906.
(13) PubChem. http://pubchem.ncbi.nlm.nih.gov/ (accessed September 1, **2009**).
(14) 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.
(15) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res*. **2003**, *13*, 2498–2504.
(16) *Scitegic Pipeline Pilot, Student Edition*, version 6.1; Accelrys, Inc.: San Diego, CA, 2007.