Molecular Formal Concept Analysis for Compound Selectivity Profiling in Biologically Annotated Databases

Eugen Lounkine, Dagmar Stumpfe, and Jürgen Bajorath*

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

Received March 10, 2009

Molecular Formal Concept Analysis (MolFCA), an adaptation of formal concept analysis from information theory, is introduced for the systematic comparison of the selectivity of a compound against multiple targets and the extraction of compounds with complex selectivity profiles from biologically annotated databases. In MolFCA, multiple selectivity queries involving the comparison of an arbitrary number of targets and compound potency values or ratios can be defined and applied in a sequential manner to retrieve compounds with desired selectivity against targets of interest. In MolFCA applications, we extract compounds from a public domain database that share selectivity profiles with known drugs or drug candidates and study structure-selectivity relationships.

INTRODUCTION

High-throughput screening technology has provided the opportunity to screen large chemical libraries against many different biological targets and generate vast amounts of raw structure−activity data.[1,2] Chemogenomic approaches often utilize such data collections, in addition to optimized compound sets, in order to build biological activity profiles of individual compounds that report the compound's activity across series of targets, with the ultimate goal of exploring all ligand-target interactions.[3,4] Several key investigations have advanced this area of research. For example, affinity fingerprints that report ligand potencies across various targets have been utilized to compare compounds on the basis of their biological activity, rather than chemical structure.[5,6] By considering both activity profiles and chemical similarity of ligands, target relationships can also be established.[7] Furthermore, integration of activity and chemical space makes it possible to build ligand-target networks that can be applied to predict previously unobserved ligand-target interactions.[2,8−10] In the context of computational compound profiling, Bayesian modeling of ligand activity against multiple targets has become increasingly popular to identify targets for compounds with no reported biological activity, which is sometimes referred to as "target fishing".[11]

The analysis and prediction of ligand activity has also focused on preferred structural motifs including "privileged substructures" that are thought to be characteristic of sets of compounds with a specific biological activity,[12,13] and "activity class characteristic substructures" have been extracted from random fragment populations of bioactive molecules.[14] For the analysis of activity-relevant molecular fragments, we have previously adapted formal concept analysis (FCA),[15] a data mining technique from information theory.[16] The resulting approach, termed Fragment Formal Concept Analysis (FragFCA), has been applied, for example, to mine hierarchically generated fragment populations of G protein coupled receptor ligands for signature fragment combinations that occurred in compounds with defined activities.[15]

Going beyond ligand activity analysis and prediction, recent studies have begun to assess ligand selectivity by computational means.[17−23] In order to define ligand selectivity, potency values measured against multiple targets must be systematically compared, and selectivity threshold ratios must often be defined.[2,22] In particular, for closely related targets, ligand binding is often not an "all or nothing" event; rather, small molecules are found to bind with different potency against related targets, and, hence, differences in potency determine target selectivity. Computational approaches have been developed that address the issue of target selectivity on the basis of binary potency relationships.[20−23] For example, virtual selectivity searching has been introduced to identify compounds that are selective for one particular target over another.[21] However, computational methods for the exploration of more complex selectivity profiles involving more than two targets have thus far not been available.

Herein we introduce another variant of FCA termed Molecular Formal Concept Analysis (MolFCA) for the mining of selectivity profiles in biologically annotated compound databases. Different from FragFCA, which utilizes molecular fragments, MolFCA operates on the basis of entire molecules to identify compounds having a defined selectivity profile involving an arbitrary number of targets. MolFCA utilizes selectivity annotations derived from compound potency values measured against targets of interest. Importantly, MolFCA exclusively uses selectivity profiles for molecular representation but no structural descriptors or similarity criteria. As we show in this study, this makes it possible to identify structurally distinct compounds sharing a desired selectivity profile, without structural bias toward reference molecules. MolFCA enables the definition of complex queries for the identification of compound sets that display a particular selectivity profile and/or fall into a

* Corresponding author phone: +49-228-2699-306; fax: +49-228-2699-341; e-mail: bajorath@bit.uni-bonn.de.

10.1021/ci900095v CCC: $40.75
Published on Web 05/15/2009

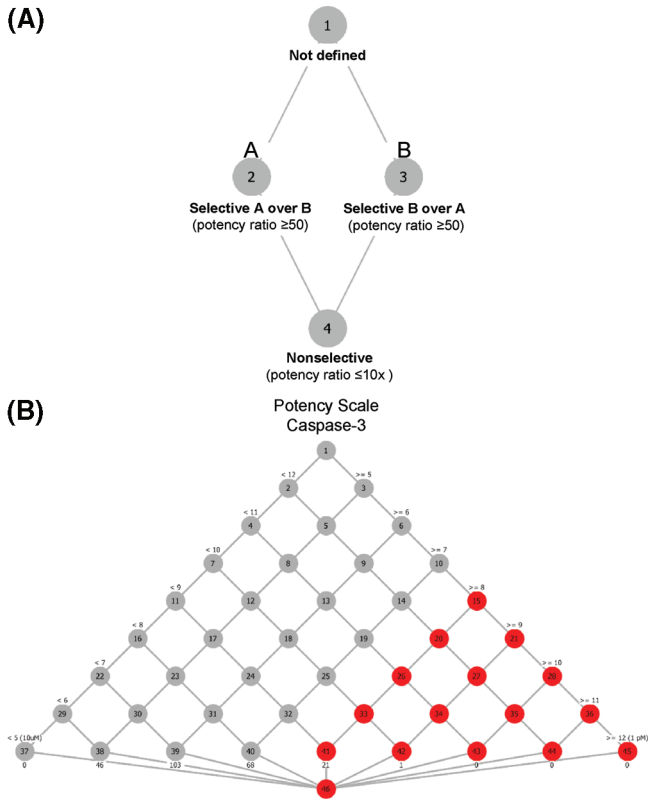

Figure 1. MolFCA scales. In (**A**), a prototypic selectivity scale is shown that discriminates between two targets A and B. Compounds selective for A over B are associated with node 2, whereas compounds with inverse selectivity are assigned to node 3. Nonselective compounds are found below node 4. The remaining compounds with no defined selectivity are associated with node 1. (**B**) A potency scale for caspase 3 inhibitors is shown. The selection (red nodes) yields a total of 22 (node 41: 21 + node 42: 1) highly potent inhibitors (≤100 nM).

defined potency range. So identified compounds can then be further analyzed to derive structure-selectivity relationships. Herein we describe the MolFCA methodology and its application to mine selectivity profiles in BindingDB.[24]

METHODS

Formal Concept Analysis. Formal concept analysis (FCA) is a data mining technique that derives relationships between defined objects and associated features (attributes). Typically, formal objects are database entries having multiple attributes.[16] FCA then operates on the basis of binary "have" and "does not have" or "occurs in" and "does not occur in" relationships between objects and attributes. Therefore, continuous properties of formal objects are binned into appropriate value ranges. For example, in the context of our study, exemplary relationships might be "molecule A has a potency value in the range of 10 nM−100 nM against target C", whereas "molecule B does not have a potency value in the range of 10 nM−100 nM against target C".

In principle, FCA is capable of operating on an arbitrary number of objects and attributes. Concepts are defined as complete sets of objects sharing a set of attributes so that no object or attribute can be added or excluded.[16] Concepts are not independent of each other but share objects and/or attributes. These relationships are visualized in concept lattices. Figure 1**A** shows a concept lattice that discriminates between four concepts defined by two independent selectivity attributes. In concept lattices, attributes are written above and objects below nodes. In general, for N independent attributes, the number of nodes in a concept lattice is 2^N. While the FCA methodology is capable of extracting all concepts for a large number of attributes, the respective lattices become difficult to read and navigate and hence are of limited practical utility. In order to design lattices that are easily interpretable, scales are defined that focus on a subset of attributes. The corresponding concept lattices are simple in their design and easy to navigate. However, scales can be combined in order to extract subsets of objects sharing specific attributes that are distributed over several scales, as further described below.

In our previous study,[15] we have adapted FCA for the systematic analysis of molecular fragment combinations in order to extract sets of signature fragments that were characteristic of individual activity classes. Furthermore, fragment combinations specific to highly potent compounds within individual activity classes were identified. In FragFCA, molecular fragment combinations were formal objects and attributes included qualitative activity and potency information. As a molecular representation, fragment fingerprints of test compounds were utilized. Therefore, in FragFCA scales were defined that focused on up to three activities at a time. Applying these scales, fragment combinations were selected that occurred in molecules with defined activity profiles. We had limited the number of different activities covered by one scale to three in order to avoid the use of lattices with 16 or more nodes that would have been required for four or more independent attributes and would have been difficult to navigate.

Here we introduce MolFCA for selectivity profiling of compounds in biologically annotated databases. Different from FragFCA, MolFCA formal objects are entire molecules, rather than fragments. Compound attributes utilized by MolFCA include binned potency ranges and compound selectivity relationships calculated from potency information. Hence, MolFCA permits the analysis of chemical databases without making reference to compound structure. Rather, it compares selectivity relationships against multiple targets in a systematic manner using target pair information as input and organizes compounds according to shared selectivity attributes or potency ranges.

This type of molecular organization can be visualized using concept lattices. Figure 1**A** shows a prototypic concept lattice that distinguishes compounds selective for target A over target B from nonselective compounds and compounds having inverse selectivity, i.e. target B over target A. Each node in a concept lattice represents a set of compounds sharing a defined set of selectivity attributes. Nodes are consecutively numbered, and compound numbers (rather than a list of compound names) are written below the nodes and attributes above the nodes. In order to extract a subset of compounds sharing a defined set of attributes, the corresponding node is selected. For example, in Figure 1**A**, compounds that are selective for target B over target A are identified by node three.

Concept lattices containing too many attributes are difficult to read and navigate. Therefore, scales are defined that focus on subsets of attributes. In MolFCA, two types of scales are utilized. Selectivity scales distinguish between selective, inverse selective, and nonselective compounds for a target

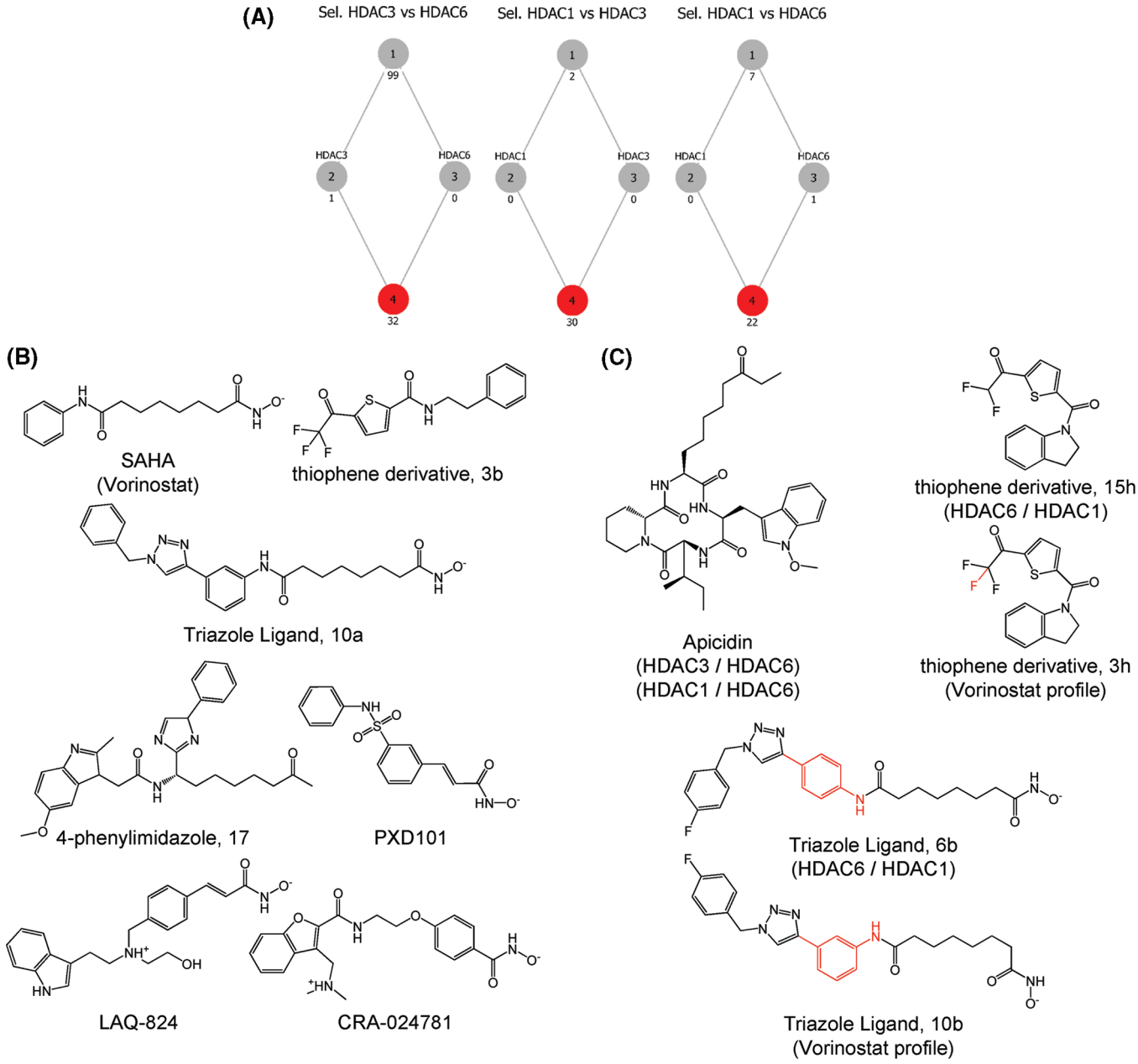

Figure 2. Vorinostat selectivity profile mining. (**A**) Selectivity profile query. Three selectivity scales are combined in order to select compounds (red nodes) active with comparable potency against HDAC1, HDAC3, and HDAC6. Seven compounds with undefined selectivity are discarded by the last scale and also one compound selective for HDAC6 over HDAC1. (**B**) The reference compound SAHA (Vorinostat) is shown together with two representative thiophene derivatives and two triazole ligands. Four additional compounds not belonging to these structural classes are also shown. The compound names correspond to the BindingDB identifiers. (**C**) Three compounds are shown that deviate from the Vorinostat profile. For the thiophene derivative and the triazole ligand, structurally highly similar compounds with different selectivity profiles are displayed. Structural differences between each pair of similar compounds with distinct selectivity profiles are highlighted in red.

pair. For capturing selectivity relationships, the type of scale depicted in Figure 1**A** is applied throughout our analysis. The selectivity scales focus on two independent attributes and therefore distinguish between four different concepts on the basis of which compound subsets are extracted. For our analysis, lattices of higher complexity are not required. In MolFCA, compound selectivity is calculated on the basis of potency ratios and used to annotate compounds, while in FragFCA qualitative activity scales are used. This has the advantage that no activity threshold (such as, for example, a 10 μM threshold used in FragFCA) is required. Therefore, selectivity can be assessed independently of a specific compound potency range: a compound with 20 μM potency against two targets is considered nonselective and also one with 20 nM potency against these two targets. Scales capturing selectivity relationships between more than two targets become rather complex because four different relationships can exist between every target pair, i.e. selective, inverse selective, nonselective, or not defined, as shown in Figure 1**A**. Furthermore, potency scales are designed to report the distribution of compounds among nine potency bins ranging from 100 μM to 10 pM. Each range contains compounds within one log unit (pKi), e.g. compounds having a potency value falling into the subrange 10 nM−100 nM. An example for the type of potency scale utilized in our analysis is shown in Figure 1**B**.

Table 1. Target Families and MolFCA Queries[a]

target family	reference	no. of scales	BindingDB targets
histone deacetylase	Vorinostat (BindingDB)	3	histone deacetylase 1 (HDAC1) histone deacetylase 3 (HDAC3) histone deacetylase 6 (HDAC6)
phosphodiesterase	Cilomilast (BindingDB)	6	phosphodiesterase type 4 (PDE4B) phosphodiesterase type 4 (PDE4D) phosphodiesterase type 10 (PDE10A) phosphodiesterase type 11 (PDE11A) phosphodiesterase type 1 (PDE1B) phosphodiesterase type 2 (PDE2A) phosphodiesterase type 3 (PDE3B)
inosine monophosphate dehydrogenase	mycophenolic acid (BindingDB)	3	inosine monophosphate dehydrogenase type 1 (IMPDH1) inosine monophosphate dehydrogenase type 2 (IMPDH2)
caspase	IDN 6556 (ref.[29])	3	caspase-3 caspase-7 caspase-8

[a] Reported are the four target families and the compounds that were used as references to define selectivity profile queries. "no. of scales" reports the number of individual scales used to design each query. Names of individual targets reported in BindingDB are given under "BindingDB targets".

The central feature of MolFCA is the combination of different selectivity and potency scales for the definition of increasingly complex queries that yield compounds with specific selectivity and potency profiles. This enables the assessment of compound distributions over different selectivity classes without the need to utilize complex lattices that are difficult to read and navigate. Rather, our selectivity scale design renders queries and their information content highly variable. For example, selectivity scales can also be combined to identify nonselective compounds, as illustrated in Figure 2**A**. Here three scales are combined into a query that yields 22 histone deacetylase (HDAC) inhibitors that have comparable potency against three histone deacetylases (HDAC1, HDAC3, and HDAC6). The total number of compounds reported in the second scale corresponds to the number of compounds selected by the first one and the number of compounds reported in the third scale to the number of compounds selected by the second one. Thus, preselected compounds are sequentially transferred from scale to scale. Queries can consist of an arbitrary number of different scales. This permits the definition of increasingly complex selectivity profiles involving multiple targets and target selectivity comparisons. MolFCA has been implemented in the Molecular Operating Environment (MOE).[25] The implementation enables the definition of scales, interactive browsing of the compound databases using these scales, generation of concept lattice representations, and storage of queries. This permits requerying of updated compound databases without the need to reassemble individual queries.

Compound Selectivity Annotation. For attribute generation, compound selectivity was calculated on a target pair basis. Given a compound with reported potencies against targets A and B, three selectivity categories were defined: (1) if compound potency for A was ≥50-fold higher than for B, then the compound was considered selective for A over B (e.g., A: 1 nM, B: 70 nM); (2) if compound potency for B was ≥50-fold higher than for A, then the compound was selective for B over A (e.g., A: 70 nM, B: 1 nM); (3) if the compound potency ratio for target A and B was ≤10, then the compound was nonselective (e.g., A: 1 nM, B: 7 nM). Different thresholds for selectivity (≥50x) and nonselectivity (≤10x) were applied in order to avoid boundary effects associated with using only a single threshold. Thus, compounds could be reliably classified as selective or nonselective despite possible fluctuations in potency measurements. Only compounds with potencies reported against both targets and falling into one of the three categories were considered for selectivity annotation. Remaining compounds were assigned to the topmost node in MolFCA selectivity scales (node 1 in Figure 1**A**) that was not used in query definitions.

Selectivity Profile Mining. We have applied MolFCA to the BindingDB[24] database in order to identify inhibitors with defined selectivity profiles. BindingDB is very suitable for this analysis because it is publicly available and contains compound potency annotations (in the form of K_i or IC50 values) for targets that can be grouped into different target families. In order to obtain single potency values for each compound, multiple potency annotations for human targets, if available, were combined using the geometric mean of provided potency values.

MolFCA queries were generated in order to find inhibitors with defined selectivity profiles directed against one of the following four target families (for which sufficient compound information was available): histone deacetylases, phosphodiesterases, inosine monophosphate dehydrogenases, and caspases. Table 1 reports individual targets for each family.

Utilizing reference compounds taken from BindingDB or literature sources, MolFCA has been applied to search for BindingDB compounds with corresponding selectivity profiles. Furthermore, specific MolFCA queries were relaxed by omitting or softening individual selectivity constrains in order to identify additional compounds with defined deviations from the reference profile.

Drug Selectivity Profile Mining. MolFCA has been applied to search for inhibitors of the four different target families described above that match selectivity profiles of four known drugs or drug candidates that act on these targets: suberoylanilide hydroxamic acid (SAHA, marketed as Vorinostat, brand name Zolinza), a histone deacetylase inhibitor approved for the treatment of cutaneous T-cell lymphoma;[26] SB 207499 (Cilomilast, brand name Ariflo), a selective phosphodiesterase type 4 (PDE4) inhibitor used for the treatment of asthma and chronic obstructive pulmonary disease;[27] mycophenolic acid (MPA, brand name Myfortic), a reversible, noncompetitive inosine monophosphate dehy-

Table 2. Selected BindingDB Compounds[a]

query	compound name	note
HDAC	4-phenylimidazole, 17	
	apicidin	selective HDAC1/HDAC6, HDAC3/HDAC6
	CRA-024781	
	LAQ-824	
	PXD101	
	SAHA	reference compound
	thiophene derivative, 15h	selective HDAC6/HDAC1
	thiophene derivative, 19c	
	thiophene derivative, 3b	
	thiophene derivative, 3c	
	thiophene derivative, 3d	
	thiophene derivative, 3f	
	thiophene derivative, 3g	
	thiophene derivative, 3h	
	thiophene derivative, 3i	
	triazole ligand, 10a	
	triazole ligand, 10b	
	triazole ligand, 10c	
	triazole ligand, 10d	
	triazole ligand, 12b	
	triazole ligand, 14a	
	triazole ligand, 14b	
	triazole ligand, 14c	
	triazole ligand, 14d	
	triazole ligand, 6b	selective HDAC6/HDAC1
PDE	(R,S)-mesopram	selective PDE2B/PDE10A
	cilomilast	reference compound
	filaminast	selective PDE2B/PDE11A
	piclamilast	
	roflumilast	selective PDE2B/PDE11A
	rolipram	selective PDE2B/PDE10A
	zardaverine	selective PDE2B/PDE11A
IMPDH	C2-mycophenolic adenine dinucleotide (C2-MDA)	
	mycophenolic acid (MPA)	reference compound, ≤ 100 nM
	mycophenolic adenine dinucleotide (MAD) analogue, 37	
	mycophenolic adenine dinucleotide (MAD) analogue, 38	≤100 nM
	tiazofurin adenine dinucleotide (TAD)	
	tiazofurin adenine dinucleotide (TAD) analogue, 25	≤100 nM
	tiazofurin adenine dinucleotide (TAD) analogue, 26	
	tiazofurin adenine dinucleotide (TAD) analogue, 27	≤100 nM
	tiazofurin adenine dinucleotide (TAD) analogue, 28	
	tiazofurin adenine dinucleotide (TAD) analogue, 29	
	tiazofurin adenine dinucleotide (TAD) analogue, 30	
	tiazofurin adenine dinucleotide (TAD) analogue, 31	≤100 nM
	tiazofurin adenine dinucleotide (TAD) analogue, 32	≤100 nM
caspases	Ac-DEVD-CHO	
	Burnham Institute compound 1	low potency
	Burnham Institute compound 2	low potency
	inhibitor 3	
	inhibitor 66a	
	isoquinoline-1,3,4-trione 13f	
	isoquinoline-1,3,4-trione 9k	
	valine aspartyl ketone 14	
	valine aspartyl ketone 35	

[a] For all four queries, the identified BindingDB compounds are reported. "Compound name" corresponds to the BindingDB field "BindingDB Monomer Display Name" and can be used to retrieve compounds via the BindingDB web interface. In addition to compounds matching the respective queries, reference molecules and compounds identified with relaxed selectivity profiles are reported (see text).

drogenase (IMPDH) inhibitor used as an immunosuppressant to prevent transplant rejection;[28] and IDN 6556, a pan-caspase inhibitor that is used as an apoptosis (i.e., programmed cell death) inhibitor to prevent liver tissue damage during liver transplantation.[29]

RESULTS AND DISCUSSION

Molecular Formal Concept Analysis. We introduce MolFCA for compound selectivity profile mining in biologically annotated databases. Additionally, compounds were selected based on defined potency ranges against individual targets. Accordingly, two types of scales were generated: selectivity and potency scales. Selectivity scales were designed to distinguish between three selectivity relationships, i.e. selective for one target over another, inverse selective, and nonselective. Potency scales were designed for the selection of compounds falling into a defined potency range for an individual target. Potency scales complement

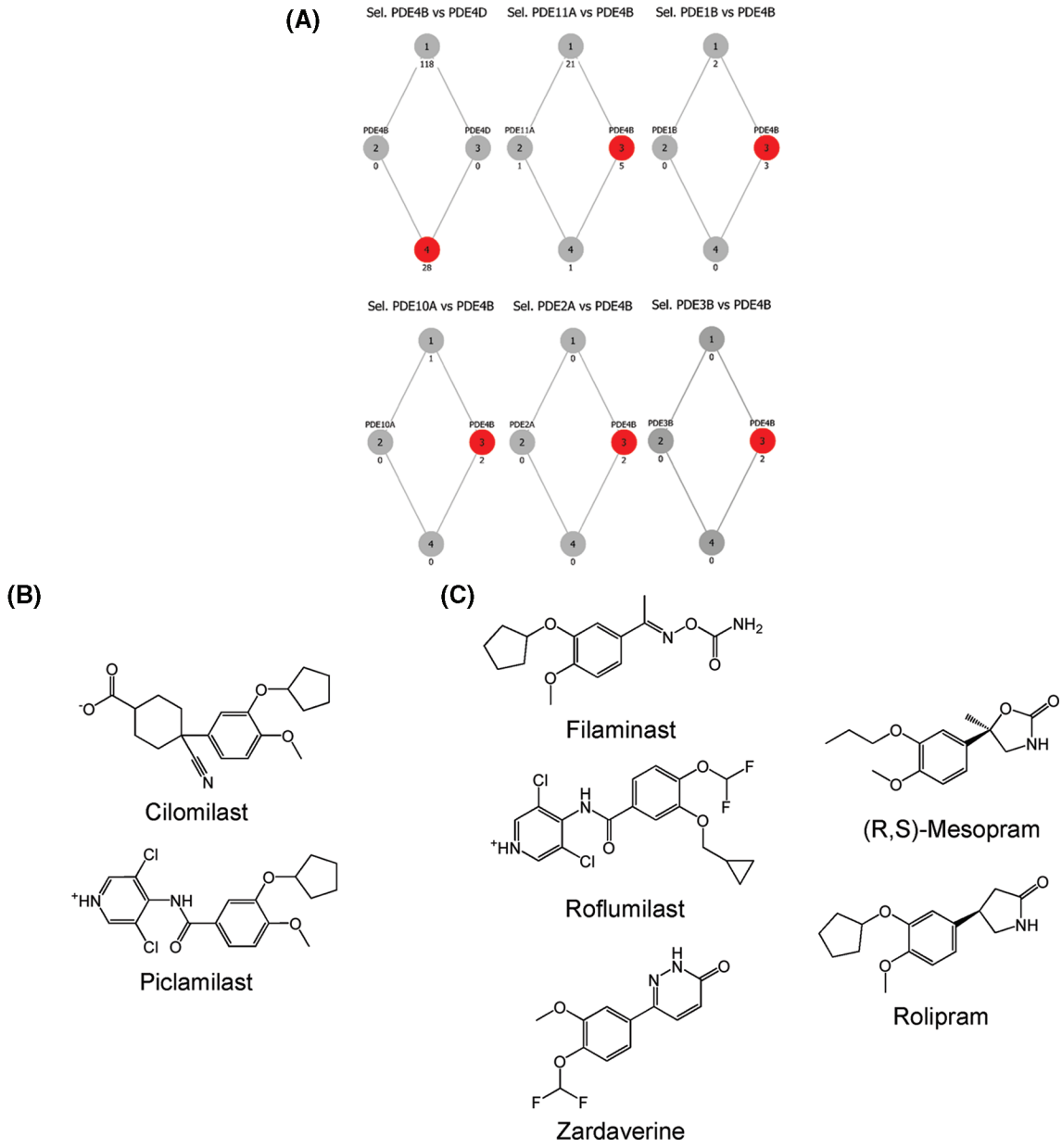

Figure 3. Cilomilast selectivity profile mining. (**A**) Six scales used to define the Cilomilast selectivity query are shown. (**B**) Cilomilast and Piclaminast are identified by MolFCA to share a selectivity profile against seven phosphodiesterase targets. (**C**) Three compounds that partly match the Cilomilast query are shown on the left that are also used to treat respiratory disorders. On the right, two additional compounds that partly match the query are shown that are currently evaluated for the treatment of neurodegenerative disorders.

selectivity scales because selectivity is assessed based on potency ratios, rather than specific potency ranges. Hence, two compounds might share the same selectivity profile but differ in potency. The design of the two prototypes of scales, selectivity scales and potency scales, is knowledge-based and optimized for selectivity queries. Only these two scale types had to be designed manually, whereas individual scales for each target or pair of targets were then calculated automatically for each relevant target family. In principle, the design of alternative selectivity scales simultaneously accounting for more than two targets is possible, similar to the activity scales utilized in FragFCA comparing three targets. An advantage of such scales would be that many selectivity relationships within a target family could be explored through one scale. However, we found that the associated increase in concept lattice complexity outweighs their practical utility, for two reasons. First, selectivity is defined for a pair of targets, and thus a selectivity scale would have to be defined for each individual target. These scales would have to contain all other targets as attributes. Second, a similarly complex selectivity scale would also be required in order to discriminate between compounds for which selectivity was defined on the basis of compounds that were active but nonselective. Consequently, this would require generating and processing many more nodes than for the scales reported herein. For example, five targets can be compared with our simple selectivity scales using 10 concept lattices with a total number of 40 nodes. The alternative more complex scales would also use 10 concept lattices, but each lattice would contain 16 nodes, amounting to a total number of 160 nodes. This level of lattice complexity makes the design of selectivity queries very difficult.

Selectivity and potency scales were combined in order to build complex queries covering selectivity relationships for multiple targets and focusing on defined potency ranges. In MolFCA, subsets of compounds selected by a specific scale

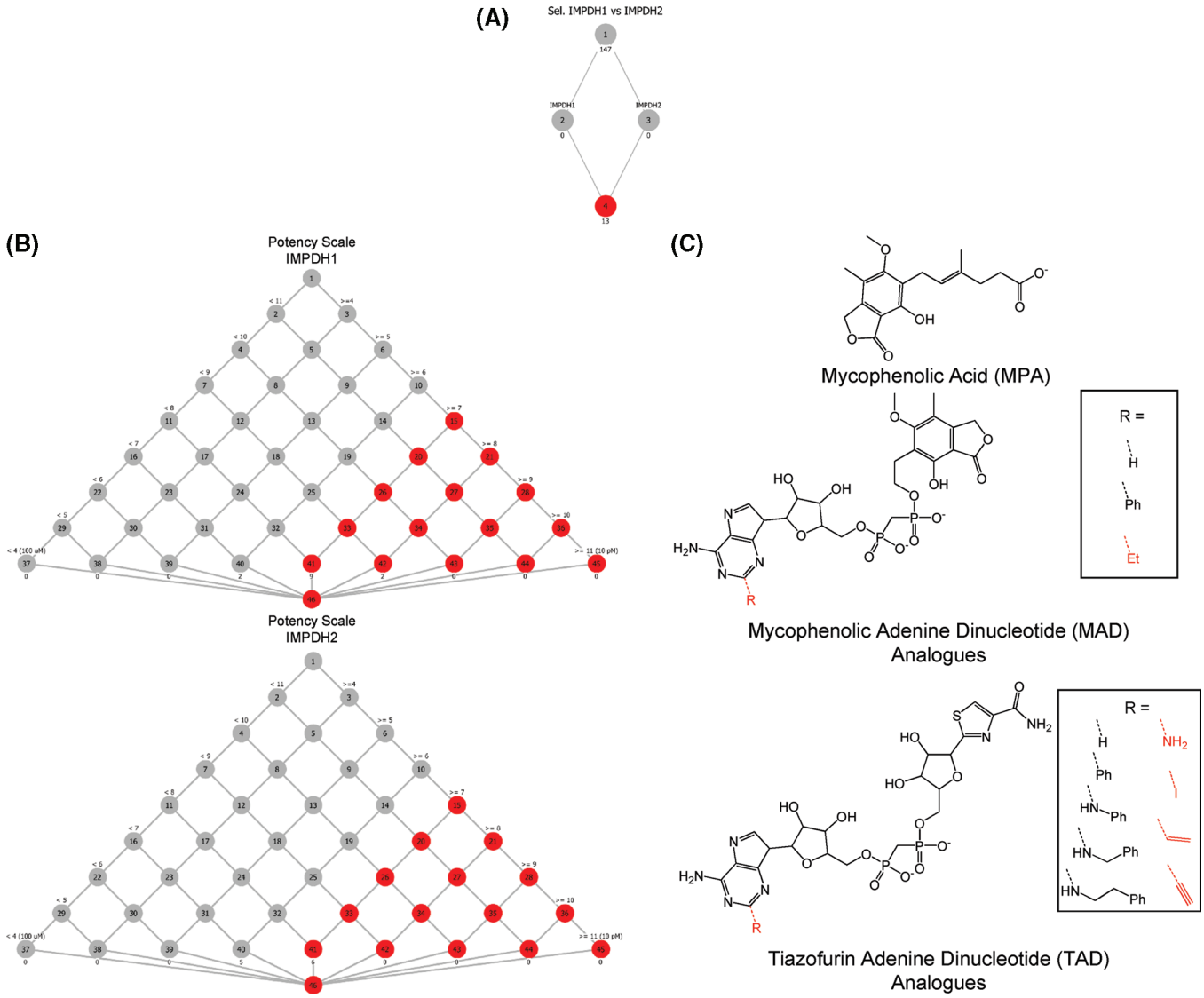

Figure 4. Identification of potent pan-IMPDH inhibitors. (**A**) The standard selectivity scale used to identify 13 nonselective IMPDH inhibitors is shown. (**B**) Two potency scales are combined to identify potent IMPDH1 and IMPDH2 inhibitors. (**C**) The reference compound MPA is shown together with the core structures of the 12 additional nonselective IMPDH inhibitors. The R group of each inhibitor is shown in the boxes on the right, and R groups of the five highly potent compounds selected using potency scales are colored red.

are transferred to another scale with a different selectivity focus, thereby facilitating the selection of sequentially reduced compound subsets. Queries are built in an interactive manner, which permits changing the order of scales that form a query. Although the scale order has no effect on the final set of filtered compounds, it allows assessing the distribution of compounds over all selectivity relationships without the necessity of utilizing very complex concept lattices, as described above.

Our scale design strategy and the possibility to assemble queries in an interactive manner (and thus "browse" different scales in an arbitrary order) provides a flexible but well-defined way of mining of biologically annotated databases. The molecular representation used as the only input for MolFCA is similar to an affinity fingerprint or vector representation, where each component corresponds to a target-associated potency. However, MolFCA provides two advantages over a direct comparison of such fingerprints. First, MolFCA explicitly accounts for well-defined pairwise target selectivity relationships based on two selectivity thresholds, as described in the Methods section. These relationships also include nondefined selectivity relationships resulting from, for example, the lack of relevant data, which is often the case when combining different biological screening sets, like in BindingDB. By contrast, fingerprint comparison could only indirectly account for selectivity (for example, through the use of specialized similarity metrics) and does not take the potential lack of data into account. Second, MolFCA makes it possible to focus on a subset of targets during query design because each selectivity scale captures two targets. By contrast, for this purpose, fingerprints would need to be redesigned and recalculated prior to comparison, whereas scales—once defined—do not need to be modified.

A characteristic feature of MolFCA is that it is a data mining technique that utilizes standardized scale types to assemble flexible queries for subsets of targets. The MolFCA query design and information content would be difficult to mimic using other filtering techniques such as, for example, binary filter components that are combined in a consecutive manner. As mentioned above, scales must be carefully designed. They should be easy to navigate, and scale types

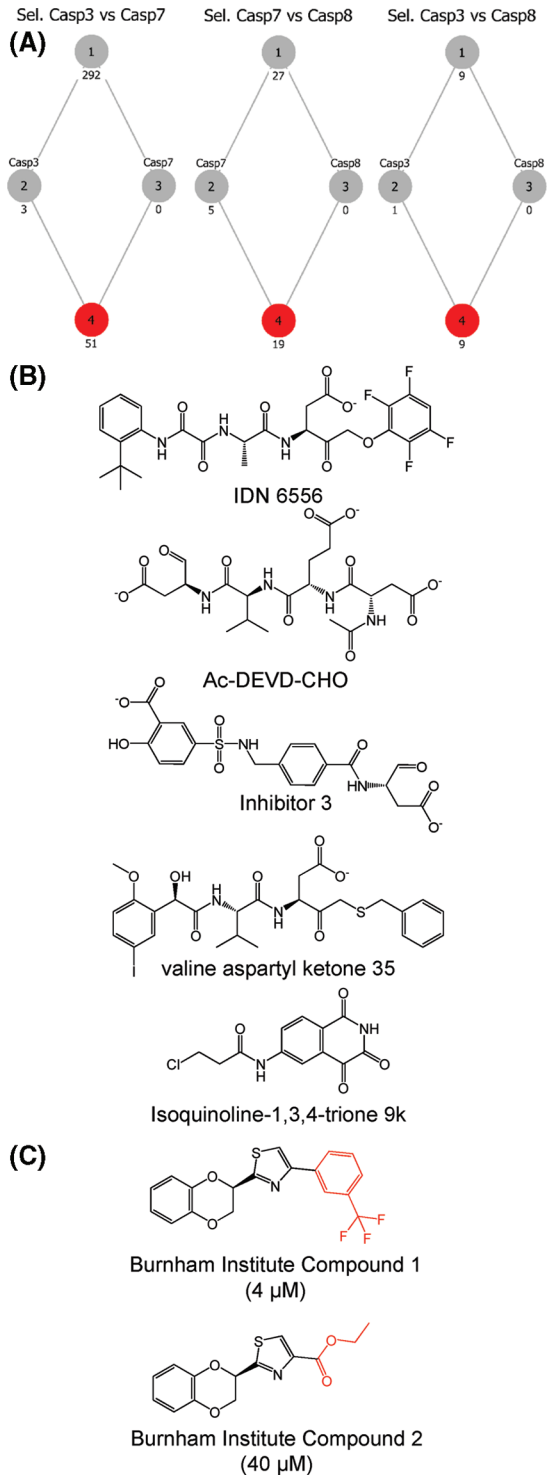

Figure 5. Identification of pan-caspase inhibitors. (**A**) The query utilizing three combined selectivity scales is shown. (**B**) IDN 6556, a literature compound with indicated pan-caspase inhibitory activity, is shown at the top. Shown are four representative compounds with high to medium potency that match the query. (**C**) Two weakly potent compounds with a 10-fold difference in potency are shown that also match the query. Structural differences between the two molecules are highlighted in red.

must be applicable to different databases and target families. This has been achieved by designing selectivity scales focusing on pairwise target comparisons and potency scales for each target. The simplistic scale design might, at first glance, represent simple binary filters. However, each selectivity scale discriminates four concepts. Although the specific applications of MolFCA presented herein could in principle be reproduced using such stepwise filtering techniques, such filtering protocols would be rigid compared to the flexible combination of MolFCA scales that allow the analysis of compound distributions during query design.

In this study, MolFCA has been applied to search for inhibitors of four different target families in the BindingDB using reference selectivity profiles of known drugs or drug candidates, as described in the following.

Selectivity Profile Mining - Vorinostat. The selectivity profile of Vorinostat was derived from BindingDB potency data against histone deacetylases HDAC1 (93 nM), HDAC3 (52 nM), and HDAC6 (43 nM). Thus, according to our selectivity classification, this compound was nonselective for each target pair, i.e. HDAC1 vs HDAC3, HDAC1 vs HDAC6, and HDAC3 vs HDAC6. In order to mine the database for compounds matching this selectivity profile, we have combined three selectivity scales (i.e., one scale for each pair). Figure 2**A** illustrates the query. Only one possible arrangement of scales is shown for clarity. However, in query design, scales can be combined in an arbitrary order. On the first scale, 32 compounds nonselective for targets HDAC3 and HDAC6 were extracted. These compounds were then projected onto the second scale. For two compounds, the selectivity profile for these two targets was not defined. The remaining 30 compounds were transferred to the third scale, and the final set of 22 compounds (including Vorinostat) was selected from node 4. Thus, 21 additional compounds matching the selectivity profile of Vorinostat were identified. Nine of these compounds belonged to the triazole structural class, and eight other compounds were thiophene derivatives. The remaining four compounds were distinct and included LAQ-824, CRA-024781, PXD101 (BindingDB identifiers), and a 4-phenylimidazole derivative.

Figure 2**B** shows Vorinostat, a representative triazole ligand, a thiophene derivative, and the four additional inhibitors. As can be seen, these chemotypes are structurally distinct. Strikingly, the thiophene derivatives and the 4-phenylimidazole derivative lack the hydroxamic acid group that is a hallmark of Vorinostat and other HDAC inhibitors. Nevertheless, the most potent thiophene derivative is comparable to Vorinostat (i.e., HDAC1: 210 nM, HDAC3: 120 nM, and HDAC6: 240 nM). This example illustrates that MolFCA is capable of identifying structurally distinct compounds with corresponding selectivity profiles.

In order to find other compounds that only partly matched the Vorinostat selectivity profile, we relaxed the HDAC query by systematically modifying the selectivity category on each scale or by removing an individual scale, leading to the identification of three additional compounds. First, Apicidin, a cyclic peptide antibiotic acting through protozoal HDAC inhibition and also inhibiting tumor proliferation,[30] was found to be nonselective for HDAC1 over HDAC3 but selective for these two targets over HDAC6. Second, an additional triazole ligand was identified as selective for HDAC6 over HDAC1 but nonselective for the target pairs HDAC1/HDAC3 and HDAC6/HDAC3 (HDAC1: 97.8 nM, HDAC3: 13.7 nM, HDAC6: 1.9 nM). Third, another thiophene derivative was also found to be selective for HDAC6 over HDAC1 (HDAC1: 9.7 μM, HDAC6: 15 nM) and identified by omitting HDAC3 from the query (potency data for HDAC3 was not provided for this compound). Figure 2**C**

shows these three compounds together with structurally similar molecules that were identified to match the Vorinostat profile. It is evident that only subtle structural deviations between these compounds altered their selectivity profiles. Thus, compounds identified with original and relaxed MolFCA queries were structurally similar but had different profiles and could be used to analyze structure-selectivity relationships. This again illustrates an advantage of molecular function-oriented data mining in the absence of structural representations: there are no structural constraints involved in the detection of selectivity profiles and hence any structural similar or diverse subset of molecules can be identified.

Selectivity Profile Mining - Cilomilast. As a second profile mining application, we focused on Cilomilast, a potent and selective PDE4 inhibitor used for the treatment of respiratory disorders.[27] Based on BindingDB potency data provided for Cilomilast (PDE4B: 25 nM, PDE4D: 11 nM, PDE1B: 87 μM, PDE2A: 160 μM, PDE3B: 87 μM, PDE10A: 73 μM, and PDE11A: 21 μM) a selectivity query combining six selectivity scales was assembled (Figure 3**A**). Compounds nonselective for PDE4B and PDE4D were selected by the first scale. The subsequent scales assessed the selectivity of these compounds for PDE4B over the other targets. This query yielded one additional compound, Piclamilast, which shared the Cilomilast selectivity profile but was overall more potent (i.e., PDE4B: 41 pM, PDE4D: 21 pM, PDE1B: 68 μM, PDE2A: 54 μM, PDE3B: 11 μM, PDE10A: 21 μM, PDE11A: 1.6 μM). Both compounds are shown in Figure 3**B**.

In order to identify additional compounds that only partly matched the Cilomilast profile, we then focused on individual scales for PDE4B selectivity over the other related targets. In this case, three additional compounds were found that were nonselective for PDE4B and PDE4D but selective for PDE4B over PDE11A including Zardaverine (PDE4B: 930 nM, PDE4D: 390 nM, PDE11A: 140 μM), Filaminast (PDE4B: 960 nM, PDE4D: 1 μM, PDE11A: 57 μM), and Roflumilast (PDE4B: 840 pM, PDE4D: 680 pM, PDE11A: 25 μM). All of these compounds are currently also evaluated or used as bronchodilatory agents to treat respiratory disorders.[31−33] Furthermore, two other compounds were found that were nonselective for PDE4B and PDE4D but selective for PDE4B over PDE10A including Mesopram (PDE4B: 420 nM, PDE4D: 1.1 μM, PDE10A: 63 μM) and Rolipram (PDE4B: 915 nM, PDE4D: 1.1 μM, PDE 10A: 140 μM). Different from the inhibitors discussed above, Mesopram is evaluated for the treatment of multiple sclerosis,[34] and Rolipram is evaluated for the treatment of depression[35] but also has immunosuppressive properties.[36] Figure 3**C** shows these five compounds. Thus, through relaxation of a complex PDE4 selectivity query, compounds with related yet distinct selectivity profiles were identified that have in part different therapeutic indications.

Mining for Highly Potent Inhibitors - Mycophenolic Acid. Mycophenolic acid (MPA) inhibits IMPDH1 and IMPDH2 in a nonselective manner with high potency values of 33 nM and 11 nM, respectively. In order to identify other potent (i.e., $\leq$100 nM) IMPDH inhibitors, we have designed a MolFCA query consisting of one selectivity and two potency scales. The standard selectivity scale, shown in Figure 4**A**, was applied to select a total of 13 inhibitors that were nonselective against IMPDH1 and IMPDH2. These compounds were then transferred to the IMPDH1 potency scale shown at the top of Figure 4**B**. Eleven of these 13 compounds (including MPA) were found to fall into the desired IMPDH1 potency range (nodes 41 and 42). These eleven inhibitors were then projected onto the IMPDH2 potency scale (Figure 4**B**, bottom, nodes 40 and 41). Six of these compounds fell into the desired IMPDH2 potency range (node 41). The 12 additional inhibitors belong to two structural classes, nine tiazofurin adenine dinucleotide (TAD) analogues and three mycophenolic adenine dinucleotide (MAD) analogues, shown in Figure 4**C**. In addition to MPA, the six highly potent inhibitors included four TAD analogues and one MAD analogue.

De Novo Selectivity Profiles - IDN 6556. The caspase inhibitor IDN 6556 was not available in BindingDB and was taken from the literature.[31] This compound prevents apoptosis in liver transplants and has been indicated to act as a pan-caspase inhibitor by inhibiting both initiator and effector caspases.[29] In order to find potential pan-caspase inhibitors in BindingDB, we have assembled a selectivity query for pan-caspase inhibitors in the absence of a reference compound with reported potencies. This query was designed to identify caspase inhibitors with comparable potency against caspase 3, 7, and 8. Caspase 8 is an initiator caspase, i.e. it activates downstream caspases, whereas the other two caspases are effector caspases, i.e. they induce apoptosis by chromatin fragmentation. We used the three scales shown in Figure 5**A** to define the query and identified nine compounds representing five chemotypes, all of which were distinct from IDN 6556 (Figure 5**B**). Seven of these nine compounds had potency values in the range of 10 nM to 500 nM. The two remaining compounds represented a weakly active chemotype (i.e., with potency values of 40 μM and 4 μM) and are shown in Figure 5**C**. This example illustrates that effective MolFCA queries can also be defined in the absence of a reference compound on the basis of our selectivity criteria.

CONCLUSIONS

We have introduced Molecular Formal Concept Analysis (MolFCA) for mining of compound selectivity profiles in biologically annotated databases. MolFCA queries were designed to identify BindingDB compounds that shared defined selectivity profiles with reference molecules. The selectivity queries consisted of up to six MolFCA scales and involved up to seven targets. MolFCA identified structurally diverse compounds matching each selectivity profile including a de novo designed query. These compounds represented in part very different structure-selectivity relationships. Our findings indicate that MolFCA is capable of successfully detecting sets of compounds that are active against different target families and share complex selectivity profiles.

REFERENCES AND NOTES

(1) Bajorath, J. Integration of Virtual and High-Throughput Screening. *Nat. Rev. Drug Discovery* **2002**, *1*, 882–894.
(2) Bajorath, J. Computational Analysis of Ligand Relationships within Target Families. *Curr. Opin. Chem. Biol.* **2008**, *12*, 352–358.
(3) Bredel, M.; Jacoby, E. Chemogenomics: An Emerging Strategy for Rapid Target and Drug Discovery. *Nat. Rev. Genet.* **2004**, *5*, 262–275.
(4) Rognan, D. Chemogenomic Approaches to Rational Drug Design. *Br. J. Pharmacol.* **2007**, *152*, 38–52.

(5) Kauvar, L. M.; Higgins, D. L.; Villar, H. O.; Sportsman, J. R.; Engqvist-Goldstein, A.; Bukar, R.; Bauer, K. E.; Dilley, H.; Rocke, D. M. Predicting Ligand Binding to Proteins by Affinity Fingerprinting. *Chem. Biol.* **1995**, *2*, 107–118.
(6) Fabian, M. A.; Biggs, W. H.; Treiber, D. K.; Atteridge, C. E.; Azimioara, M. D.; Benedetti, M. G.; Carter, T. A.; Ciceri, P.; Edeen, P. T.; Floyd, M.; Ford, J. M.; Galvin, M.; Gerlach, J. L.; Grotzfeld, R. M.; Herrgard, S.; Insko, D. E.; Insko, M. A.; Lai, A. G.; Lélias, J.; Mehta, S. A.; Milanov, Z. V.; Velasco, A. M.; Wodicka, L. M.; Patel, H. K.; Zarrinkar, P. P.; Lockhart, D. J. A Small Molecule-Kinase Interaction Map for Clinical Kinase Inhibitors. *Nat. Biotechnol.* **2005**, *23*, 329–336.
(7) Mestres, J.; Martin-Couce, L.; Gregori-Puigjane, E.; Cases, M.; Boyer, S. Ligand-Based Approach to In Silico Pharmacology: Nuclear Receptor Profiling. *J. Chem. Inf. Model.* **2006**, *46*, 2725–2736.
(8) Paolini, G. V.; Shapland, R. H. B.; van Hoorn, W. P.; Mason, J. S.; Hopkins, A. L. Global Mapping of Pharmacological Space. *Nat. Biotechnol.* **2006**, *24*, 805–815.
(9) 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.
(10) Yildirim, M. A.; Goh, K.; Cusick, M. E.; Barabási, A.; Vidal, M. Drug-Target Network. *Nat. Biotechnol.* **2007**, *25*, 1119–1126.
(11) Nidhi; Glick, M.; Davies, J. W.; Jenkins, J. L. Prediction of Biological Targets for Compounds Using Multiple-Category Bayesian Models Trained on Chemogenomics Databases. *J. Chem. Inf. Model.* **2006**, *46*, 1124–1233.
(12) Schnur, D. M.; Hermsmeier, M. A.; Tebben, A. J. Are Target-Family-Privileged Substructures Truly Privileged. *J. Med. Chem.* **2006**, *49*, 2000–2009.
(13) Aronov, A. M.; McClain, B.; Moody, C. S.; Murcko, M. A. Kinase-Likeness and Kinase-Privileged Fragments: Toward Virtual Polypharmacology. *J. Med. Chem.* **2008**, *51*, 1214–1222.
(14) Batista, J.; Bajorath, J. Mining of Randomly Generated Molecular Fragment Populations Uncovers Activity-Specific Fragment Hierarchies. *J. Chem. Inf. Model.* **2007**, *47*, 1405–1413.
(15) Lounkine, E.; Auer, J.; Bajorath, J. Formal Concept Analysis for the Identification of Molecular Fragment Combinations Specific for Active and Highly Potent Compounds. *J. Med. Chem.* **2008**, *51*, 5342–5348.
(16) Priss, U. Formal Concept Analysis in Information Science. *Annu. Rev. Inform. Sci. Technol.* **2006**, *40*, 521–543.
(17) Frye, S. V. Structure-Activity Relationship Homology (SARAH): a Conceptual Framework for Drug Discovery in the Genomic Era. *Chem. Biol.* **1999**, *6*, R3–R7.
(18) Schuffenhauer, A.; Floersheim, P.; Acklin, P.; Jacoby, E. Similarity Metrics for Ligands Reflecting the Similarity of the Target Proteins. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 391–405.
(19) Xia, X.; Maliski, E. G.; Gallant, P.; Rogers, D. Classification of Kinase Inhibitors Using a Bayesian Model. *J. Med. Chem.* **2004**, *47*, 4463–4470.
(20) Stumpfe, D.; Ahmed, H. E. A.; Vogt, I.; Bajorath, J. Methods for Computer-Aided Chemical Biology. Part 1: Design of a Benchmark System for the Evaluation of Compound Selectivity. *Chem. Biol. Drug. Des.* **2007**, *70*, 182–194.
(21) Vogt, I.; Stumpfe, D.; Ahmed, H. E. A.; Bajorath, J. Methods for Computer-Aided Chemical Biology. Part 2: Evaluation of Compound Selectivity Using 2D Molecular Fingerprints. *Chem. Biol. Drug. Des.* **2007**, *70*, 195–205.
(22) Stumpfe, D.; Geppert, H.; Bajorath, J. Methods for Computer-Aided Chemical Biology. Part 3: Analysis of Structure-Selectivity Relationships Through Single- or Dual-Step Selectivity Searching and Bayesian Classification. *Chem. Biol. Drug. Des.* **2008**, *71*, 518–528.
(23) Ahmed, H. E. A.; Geppert, H.; Stumpfe, D.; Lounkine, E.; Bajorath, J. Methods for Computer-Aided Chemical Biology. Part 4: Selectivity Searching for Ion Channel Ligands and Mapping of Molecular Fragments as Selectivity Markers. *Chem. Biol. Drug. Des.* **2009**, *73*, 273–282.
(24) 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.
(25) *Molecular Operating Environment (MOE) 2007.09*; Chemical Computing Group, Inc.: Montreal, Canada, 2007.
(26) Haberland, M.; Montgomery, R. L.; Olson, E. N. The Many Roles of Histone Deacetylases in Development and Physiology: Implications for Disease and Therapy. *Nat. Rev. Genet.* **2009**, *10*, 32–42.
(27) Giembycz, M. A. Cilomilast: A Second Generation Phosphodiesterase 4 Inhibitor for Asthma and Chronic Obstructive Pulmonary Disease. *Expert Opin. Invest. Drugs* **2001**, *10*, 1361–1379.
(28) Sanford, M.; Keating, G. M. Enteric-Coated Mycophenolate Sodium: A Review of its Use in the Prevention of Renal Transplant Rejection. *Drugs* **2008**, *68*, 2505–2533.
(29) Natori, S.; Higuchi, H.; Contreras, P.; Gores, G. J. The Caspase Inhibitor IDN-6556 Prevents Caspase Activation and Apoptosis in Sinusoidal Endothelial Cells During Liver Preservation Injury. *Liver Transplant* **2003**, *9*, 278–284.
(30) Han, J. W.; Ahn, S. H.; Park, S. H.; Wang, S. Y.; Bae, G. U.; Seo, D. W.; Kwon, H. K.; Hong, S.; Lee, H. Y.; Lee, Y. W.; Lee, H. W. Apicidin, a Histone Deacetylase Inhibitor, Inhibits Proliferation of Tumor Cells via Induction of p21WAF1/Cip1 and Gelsolin. *Cancer Res.* **2000**, *60*, 6068–6064.
(31) Kips, J. C.; Joos, G. F.; Peleman, R. A.; Pauwels, R. A. The Effect of Zardaverine, an Inhibitor of Phosphodiesterase Isoenzymes III and IV, on Endotoxin-Induced Airway Changes in Rats. *Clin. Exp. Allergy* **1993**, *23*, 518–523.
(32) Spina, D. PDE4 Inhibitors: Current Status. *Br. J. Pharmacol.* **2008**, *155*, 308–315.
(33) Heaslip, R. J.; Lombardo, L. J.; Golankiewicz, J. M.; Ilsemann, B. A.; Evans, D. Y.; Sickels, B. D.; Mudrick, J. K.; Bagli, J.; Weichman, B. M. Phosphodiesterase-IV Inhibition, Respiratory Muscle Relaxation and Bronchodilation by WAY-PDA-641. *J. Pharmacol. Exp. Ther.* **1994**, *268*, 888–896.
(34) Dinter, H.; Tse, J.; Halks-Miller, M.; Asarnow, D.; Onuffer, J.; Faulds, D.; Mitrovic, B.; Kirsch, G.; Laurent, H.; Esperling, P.; Seidelmann, D.; Ottow, E.; Schneider, H.; Tuohy, V. K.; Wachtel, H.; Perez, H. D. The Type IV Phosphodiesterase Specific Inhibitor Mesopram Inhibits Experimental Autoimmune Encephalomyelitis in Rodents. *J. Neuroimmunol.* **2000**, *108*, 136–146.
(35) Wachtel, H. Potential Antidepressant Activity of Rolipram and other Selective Cyclic Adenosine 3′,5′-Monophosphate Phosphodiesterase Inhibitors. *Neuropharmacology* **1983**, *22*, 267–272.
(36) Sommer, N.; Löschmann, P. A.; Northoff, G. H.; Weller, M.; Steinbrecher, A.; Steinbach, J. P.; Lichtenfels, R.; Meyermann, R.; Riethmüller, A.; Fontana, A. The Antidepressant Rolipram Suppresses Cytokine Production and Prevents Autoimmune Encephalomyelitis. *Nat. Med.* **1995**, *1*, 244–248.

CI900095V