JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

Article

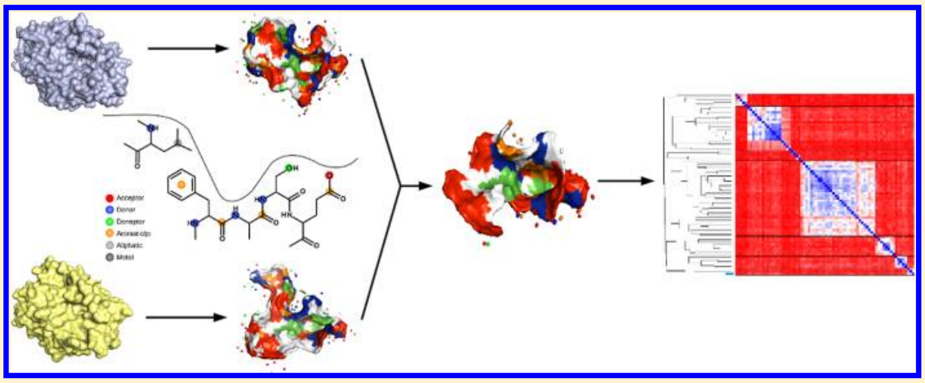
pubs.acs.org/jcim

Cavities Tell More than Sequences: Exploring Functional Relationships of Proteases via Binding Pockets

Serghei Glinca and Gerhard Klebe*

Institute of Pharmaceutical Chemistry, Philipps-Universität Marburg, Marbacher Weg 6, D-35032 Marburg and Associated with the Center of Synthetic Microbiology, Synmikro, University of Marburg, Germany

S *Supporting Information*



ABSTRACT: Computational approaches play an increasingly important role for the analysis and prediction of selectivity profiles. As most of the successfully administered small molecule drugs bind in depressions on the surface of proteins, physicochemical properties of the pocket-exposed amino acids play a central role in ligand recognition during the binding event. Cavbase is an approach to describe binding sites in terms of the exposed physicochemical properties and to compare them independent of their sequence and fold homology. Classification of proteins by means of their binding-site properties is a promising approach to obtain information necessary for selectivity modeling. For this purpose, the workflow *clusterScore* has been developed to explore the important parameters of a clustering procedure, which will allow an accurate classification of proteins. It has been successfully applied on two diverse and challenging data sets. The predicted number of clusters, as suggested by *clusterScore* and the subsequent clustering of proteins are in agreement with the EC and Merops classifications. Furthermore, putative cross-reactivity mapped between calpain-1 and cysteine cathepsins on structural level has so far only been described based on ligand data. In a benchmark study using ligand topology, binding site, and sequence information of eleven serine proteases, the emerging clusters indicate a pronounced correlation between the cavity and ligand data. These results emphasize the importance of binding-site information which should be considered for ligand design during lead optimization cycles. The program *clusterScore* is freely available and can be downloaded from our Web site www.agklebe.de.

■ INTRODUCTION

Greatly desired properties of a drug are either high affinity and selectivity toward one particular target or, dependent on the mode-of-action, also in special cases a promiscuous binding to a set of multiple targets might be important.[1] The latter situation can be given, e.g. for kinases where a set of proteins of the kinome has to be downregulated in a disease situation. During lead optimization, it is difficult to assign clear-cut criteria to the optimization strategy. In particular, the information about the target and the target family have to be analyzed and considered. General rules to follow in structure-based selectivity optimization such as shape and electrostatic complementary, flexibility, role of water, and allosteric or noncompetetive binding have been suggested.[2] Methods designed for selectivity analysis and prediction have gained an increasing importance over the last years. Ligand-based approaches have been shown to be valuable tools for pharmacological profiling mostly due to the availability of a vast amount of experimental data.[3,4] Structure-based approaches need to utilize more sophisticated data; therefore, more efficient algorithms are required to exploit the potential of these methods for routine applications.[5]

Various studies have used structural pocket similarity considerations leading to an accurate functional classification of kinases.[6−9] They found that on low sequence identity level binding sites can be highly conserved; on the contrary, kinases related by high sequence similarity can still expose significant differences in their pockets. Thereby experimentally observed cross-reactivities of known kinase inhibitors could be rationalized. Apart from cross-reactivity considerations regarding members of the same proteins family the prediction of potent binding to structurally and sequentially remote proteins

Received: November 15, 2012

is of utmost challenge. Therefore, the binding pocket of a given query protein can be screened against a database of pockets classified in the same way. Local binding site similarities of remote proteins contributing to ligand binding can be detected.[10−14] Bearing in mind the work of Mestres et al. that a drug interacts on average with six targets in a cell,[15] both approaches, either the functional classification of protein families and the broad database screening for similarities in sequentially remote proteins provide indispensable information to be considered for ligand selectivity profiling.

A wide range of different approaches for pocket detection and comparison have been developed up to now, which have been comprehensively reviewed elsewhere.[5] Since, our present study is based on Cavbase, a brief introduction of this methodology follows. Cavbase is able to detect, describe, and compare protein binding pockets independent of their sequence and fold geometry.[16] The pocket detection is performed using the Ligsite algorithm,[17] which exploits geometric data about the protein structure only, and during this step any information about a possibly bound ligand is neglected. After pocket detection, pseudocenters are assigned to the cavity-flanking residues according to predefined rules.[16,18] The pseudocenters encode physicochemical properties that are exposed on the surface of the detected pocket. Currently, seven different types of pseudocenters are implemented in Cavbase, covering the following properties: metal, H-bond donor, acceptor, mixed donor−acceptor, hydrophobic, π (ability to form pi−pi interactions), and aromatic. The pocket comparison is computed by a clique-detection algorithm which relies only on the information stored by means of the pseudocenters. After the maximum common subgraph is found, cavities are superimposed and a scoring function evaluates the overlap of the surfaces of the aligned pockets.

For a given data set of protein binding pockets, an all-against-all comparison can be performed, and based on the resulting similarity matrix, a clustering procedure can be applied. Functional classifications based on Cavbase similarity scores have been presented for α-carbonic anhydrases (α-CAs) and kinases.[6,18] In case of the α-CAs,[18] a separation on subfamily level was achieved and conformational or mutational differences were easily detectable. Kinases could be clustered on superfamily level and different activation states in the subfamilies could be distinguished. In both studies, the Cluto tool-kit[19] was applied for the clustering procedure. However, several limitations can occur when using Cluto. First, cavities that share only a marginal similarity are included and might end-up in the same cluster, which will bias the clustering. Second, Cluto provides a limited number of methods to evaluate the obtained clustering structure in order to choose the most suitable clustering strategy for the given problem. Third, Cluto requires as a prerequisite a predefined number of expected clusters, an assignment which usually appears quite arbitrary as the number of expected clusters is a priori not known.

In the present study, we introduce a new clustering workflow which was designed and validated for clustering data sets in terms of the Cavbase similarity metric, but the implemented routines can be applied to any similarity or distance matrix. The proposed procedure estimates the number of expected clusters, filters cavities using a user-defined threshold, and compares different clustering strategies. In case the user is unfamiliar with the clustering methods, application of cluster validation statistics can assist detecting the most appropriate clustering algorithm.

Structural data of binding sites can provide relevant information with respect to classification and prediction of ligand promiscuity and selectivity. On the basis of the developed clustering procedure we perform a comparative analysis of the cavity space of proteases. The selection of proteases for this case study is rather arbitrary; however, it was borne in mind that the authors have some experience with this protein class due to their extensive experimental structural work with these enzymes. Evaluating the architecture in terms of similarity of the cavity, sequence, and ligand spaces for a subset of human serine proteases provides some important insights into the question of whether a ligand-based classification correlates better with a cavity- or a sequence-based classification, as this issue is of utmost importance for the prediction of cross-reactivity among targets in computer-assisted drug design.

■ MATERIALS AND METHODS

Data Set Selection and Benchmark. The first data set is used for the validation of the clustering workflow and contains 502 cavities from 16 different proteins covering all 6 principal classes of enzymes according to the Enzyme Commission.[20] This data set will be referred to in the following as *EC data set* (Table 1). It is worth mentioning that the data are challenging for classification issues due to two aspects. First, the number of individual entries accounted in the classes deviates strongly, ranging from 5 up to 70. Second, four classes are represented

Table 1. Sixteen EC Classes Are Represented in the EC Data Set Composed by 502 Binding Sites[a]

EC number	name	remarks (e.g., organism)	number
1.1.1.21	aldose/xylose reductase	human, pig, C. tenius[b]	62
1.1.1.42	isocitrate dehydrogenase	E. coli	21
1.1.1.62	estradiol 17β-dehydrogenase	human	16
1.14.13.2	hydroxybenzoate-monooxygenase	P. fluorescens	30
2.7.1.37	cyclin-dependent kinase 2	human	46
2.7.1.112	C-Src tyrosine kinase	human, mouse[b]	20
2.7.4.9	thymidilate kinase	human, M. tuberculosis, S. cerevisiase[b]	35
3.4.21.5	thrombin	human	41
3.4.23.16	HIV-1 protease	HIV	48
3.4.24.86	TNF-α converting enzyme	human	16
4.1.1.23	COMP-decarboxylase	human, S. cerevisiae[b]	36
4.2.1.1	α-carbonic anhydrase I, II, III, IV[c]	human	70
5.3.1.5	xylose isomerase	A. missouriensis	13
5.4.2.1	phosphoglycerate mutase	S. cerevisiae	5
6.3.2.1	pantoate-β-alanine ligase	M. tuberculosis	27
6.3.4.4	adenylosuccinate synthase	E. coli	16

[a] For PDB codes of the considered examples, see the Supporting Information. [b] Four proteins are regarded that originate from more than one organism. [c] The α-carbonic anhydrase group comprises four different human isoforms.

by proteins originating from multiple organisms and one group consists of four different enzyme isoforms; therefore, common binding site motifs must be detected independent from any given sequence identity.

The second data set comprises human proteases with one structure from bovine trypsin as an exception. These data are termed the *proteases data set* (see Supporting Information Table 1). In order to retrieve a reliable and methodologically orthogonal reference classification the Merops database,[21] release 8.4, has been consulted. Merops is a manually curated database that classifies proteases in a hierarchical manner and assigns proteins to families and clans. A Merops family contains proteins for which relationships to a representative protease or another family members can be shown in terms of sequence comparison using a subset of residues only that are responsible for the catalyzed reaction. If possible, families are grouped into clans. A clan contains proteins for which relationships can be established and considers the three-dimensional arrangement of catalytic and noncatalytic residues. Hence, Merops clans include proteins for which relationships cannot be established merely based on sequence comparison, Cavbase should be able to detect sequence-independent relationships, which would then be reflected by the emerged clustering structure. The proteases data set is also used for the workflow validation, but in addition we investigate the differences between the computed clustering and the original Merops classification. The proteases data set considers 90 individual proteases from 12 Merops clans.

The last part of the present study is focused on the serine proteases, a subset of the proteases. We generated and compared ligand-, cavity-, and sequence-based clustering. For this purpose, we selected 11 proteins for which sufficient public data on ligand inhibition are available (Table 2). Ligand data for the regarded proteins have been retrieved from the ChEMBL database.[22] Only ligands have been included in the data set that fulfilled the following criteria: molecular weight should be below 600 Da, inhibition constant K_i better than 1 μM, achiral, and a maximum of 100 compounds per protein were considered.

Table 2. Ligands of 11 Serine Proteases Retrieved from the ChEMBL Database[a]

PDB ID	serine protease	number of retrieved ligands
1klt	chymase	52
1kli	factor VIIa	100
2jkh	factor Xa	100
1spj	kallikrein 1	23
1eax	matriptase	19
2any	plasma kallikrein	27
1vzq	thrombin	100
1a5h	tissue-type plasminogen activator	36
2zft	trypsin (bovine)	62
2bm2	β2-tryptase	72
1gj7	urokinase-type plasminogen activator	100

[a] For PDB codes of the considered examples, see the Supporting Information.

Similarity Matrices. The pocket similarity matrix has been generated by Cavbase. The comparison of sequences was carried out by the fasta35 program.[23] The sequence identity values were examined the same way as the Cavbase similarity scores. A sequence identity matrix is constructed and used as input for the clustering procedure.

The computation of the ligand-based similarity matrix for serine proteases has been performed as follows. 691 ligands were mutually compared using the RDKit topology fingerprint[24] and a Tanimoto similarity measure.[24] Since it is known which ligand is associated with which target, the computed similarity matrix can be divided in 121 groups (11 × 11 proteins). For each group, the average similarity is calculated disregarding the trivial matches of identical ligands. This step leads to a more compact matrix that is ready to compare with the other matrices (Figure 1b).

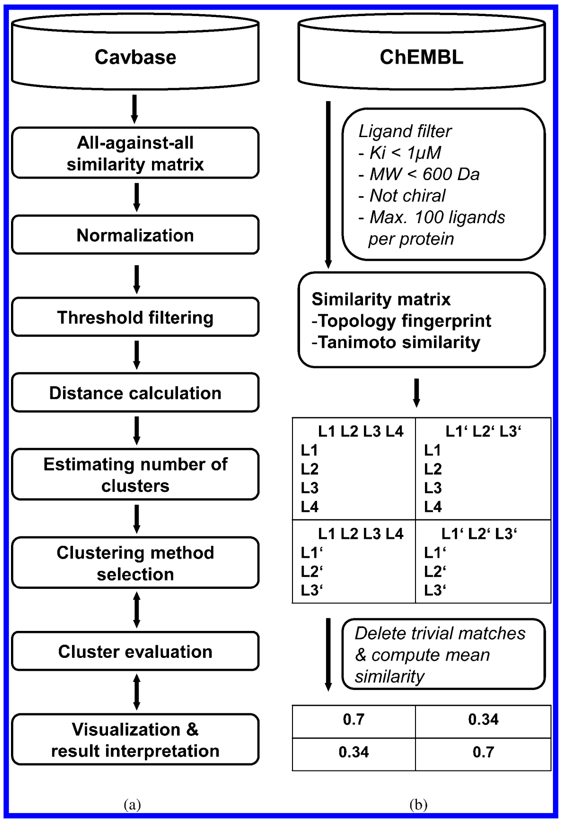


Figure 1. (a) Clustering workflow exemplified for Cavbase. (b) Computation of the similarity matrix for serine proteases using the inhibition data accessible on the ChEMBL database.

Clustering Workflow. The basic process of a cluster analysis implies steps like feature selection or extraction, clustering algorithm design or selection, cluster validation, and result interpretation.[25] These essential steps have been adopted to create a clustering workflow, which has been validated for the Cavbase similarity metric (Figure 1a). Each step is described in detail in the following sections.

Distance Calculation. The generated raw Cavbase similarity scores exhibit positive values. Hence, the resulting similarity matrix has to be symmetrized and normalized. Applying a strategy of normalization implicates that all variables are given an equal weight and they can be converted to distances. A vector of normalization factors is determined by $F_N = 1/(S_D)^{1/2}$, where S_D are the similarity values from the main diagonal. Rows and columns are subsequently multiplied by this vector. The resulting normalized similarity matrix contains 1 on the main diagonal and other values are between 0 and 1. As clustering algorithms depend on the input given in the distance matrix, different clusterings will emerge applying different distance measures. In order to cover the search space as efficient as possible, four different distance measures (D_1–D_4) were computed from the normalized input scores S_N.

$$D_1 = 1 - S_N$$

$$D_2 = \frac{1}{S_N} - 1$$

$$D_3 = (S_N - 1)^2$$

$$D_4 = \sqrt{1 - S_N}$$

Threshold Filtering. In general, variables with no information content in a data set will make the clustering less clear-cut; therefore, they should be assigned a zero weight, which virtually discards them from the analysis.[26] In order to avoid any unreasonable bias in our clustering we check for data points that fall below a predefined threshold, e.g. such a data point could correspond to a cavity that shares marginal similarity to any other members of the data set except itself. In the following, a threshold of 20% was set as minimal mutual similarity for all data sets. This idea of setting a threshold limit of 20% to all data sets was adapted from the so-called twinlight zone definition used in sequence comparisons[27] beyond which most approaches fail to find a conclusive alignment of proteins, at least in sequence space.

Estimating Number of Clusters. A crucial parameter in a clustering procedure is the number of expected clusters; therefore, most algorithms require a predefined value given by the user during data set compilation. In case the number of clusters is not predefined or the data set is not appropriately evaluated, silhouettes can be applied. We have selected two rather complementary approaches: average silhouettes (AS)[28] and median split silhouettes (MSS)[29] (see the Supporting Information). AS is a reliable global measure of the relevance of clustering results, whereas MSS analyzes the local structure within a cluster by calculating the average homogeneity of the clusters in the clustering result. Both methods were implemented for the partitioning around medoids[26] clustering. We will stress the predictive power of these two methods when applied to Cavbase similarity matrices.

Clustering Algorithms and Cluster Validity Assessment. There is a wide range of clustering algorithms, which makes selection of the most appropriate algorithm difficult. We considered the most commonly used hierarchical agglomerative methods (Ward's method, single, complete, group average, median, and centroid linkage), the hierarchical divisive analysis, and the partitioning around medoids method.[26] For a given k, different clustering methods can lead to different results. Therefore, internal and external criteria can be applied for cluster validity assessment. Detailed description of these approaches can be found elsewhere.[30] In general, external criteria evaluate the results of a clustering algorithm using a predefined structure and internal criteria validate the results in terms of quantities using the proximity matrix itself. We made use of the adjusted Rand index[31] (ARI) as an external measure to compare the clustering results to an independent reference classification or for the comparison between each other (see Supporting Information section 1.2). Furthermore, we tested the ability of nine internal cluster validity criteria (see Supporting Information section 1.1) to discriminate between meaningful and less significant clustering structures with respect to the reference classifications.

RESULTS AND DISCUSSION

Clustering Workflow Validation. *Threshold Filter.* The application of a threshold of 20% for the normalized similarity score to discard data points, leads only in case of the proteases data set to an elimination of 16 cavities, which will be discussed in detail below. Subsequently the resulting proteases set contains 74 entries.

Number of Clusters. In our approach, the estimated number of clusters depends on one user-specified parameter, namely the maximum number of clusters. This means, the program subsequently computes AS and MSS for the possible number of clusters, from 2 to k_{max}. In the case of the EC data set, AS is able to find the number of 16 EC classes taking D_2 as distance. A representative cluster of the α-carbonic anhydrase subset (70 entries) is depicted in Figure 2. Interestingly, MSS performs better on the proteases data and suggests for three distance measures seven clusters, which is the number of clans for the data points remaining after the filtering step (Table 3). For further analysis, the number of clusters was set to 16 for the EC data set and for the proteases to 7, which is in accordance with the reference classifications.

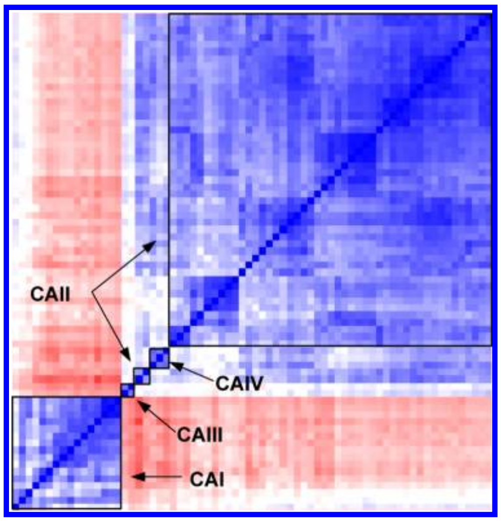
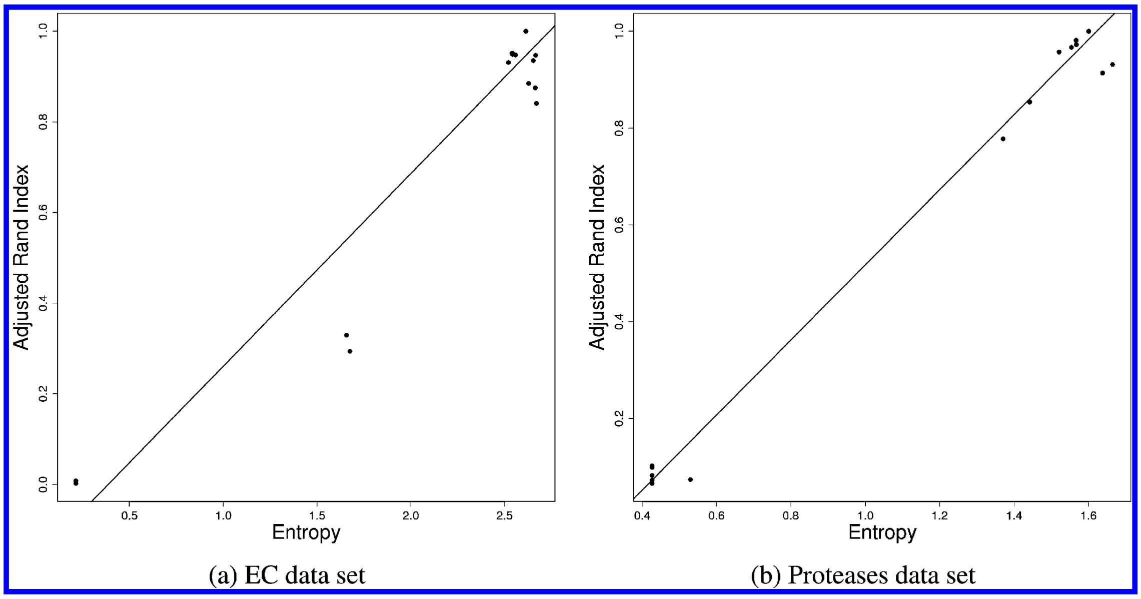


Figure 2. Representative cluster of 70 crystal structures determined for four α-carbonic anhydrases (CA) I, II, III, IV. Individual structures corresponding to one of the four isoforms are found in the same cluster among the EC data set, although the differences between them are apparent within the cluster itself as highlighted by black borders.

Table 3. Estimated Number of Clusters for the EC and Proteases Data Set with k_{max} = 25

data set	distance	average silhouette	median split silhouette
EC	D_1	17	20
	D_2	**16**	21
	D_3	22	18
	D_4	14	23
proteases	D_1	8	7
	D_2	8	24
	D_3	10	7
	D_4	9	7



(a) EC data set

(b) Proteases data set

Figure 3. Pearson correlation of the clustering entropy to the ARI.

Cluster Validity. As mentioned above, we were interested whether for a given k any cluster validity criterion is able to discriminate between significant and less significant clustering structures. For this purpose, the internal cluster validity criteria were checked for correlation with the ARI. An ARI of 1 means that the generated clustering matches a predefined splitting. The lower the ARI, the less is the agreement with the external classification. A correlation could be found for the entropy of the clustering,[32] which was derived from the information theory (see Supporting Information Table 2). The correlation plot in Figure 3 shows that clustering structures with high ARI values have also a high entropy values. This finding can guide a user to consider only clustering methods best-ranked according to the entropy measure instead of considering all possible clusterings and investigate them in more detail.

Detailed Analysis of the Proteases Clustering. Cavbase is able to cluster successfully the proteases data on the Merops clan level, whereas sequence approaches fail (Figure 4, within the sequence similarity matrix of the 90 entry-protease data set a mean of 31.7, a median of 28.9, and a standard deviation of 12.1 is found). This result demonstrates the advantage of a binding site-based classification particularly of remote or unrelated proteins. An all-against-all sequence comparison of proteins in such a data set reveals only a low signal-to-noise ratio, and the obtained clustering disagrees with available knowledge. Although the binding site-based clustering of proteases can be matched to the Merops clan classification, there are remarkable differences. Due to the initial threshold filtering, 16 cavities are removed from the data set (Table 4). In order to illustrate the impact of this step, besides prevention of clustering bias, the relationships of the discarded proteins to one another within their group and the relationships to other clustered proteins is described in the following.

Four proteins that represent an entire Merops clan on themselves were discarded from the data set. These proteins are the following: membrane-bound dipeptidase, AMSH-like peptidase, tripeptidyl-peptidase-1, and taspase-1. Three other proteins which are cysteine proteases from the same clan belong to the group of deubiquitinating enzymes (DUBs). DUBs have different topologies and mechanisms of substrate recognition, but the spatial arrangement of the catalytic triad and the oxyanion hole are highly conserved.[34] The resulting selectivity and uniqueness of DUBs' binding sites is reflected by the filtering step of the Cavbase similarity matrix. UCH-L1, TNFα-IP3, and otubain-2 are from the same clan, but apparently, the differences in their binding sites are significant. In Figure 5, the mutually matched binding site surface patches of both enzymes are superimposed and shown in a side-by-side view. The enzymes share only the conserved catalytic core in common. Although the arrangement of the proposed catalytically active histidine (His161) of UCH-L1 is too remote to create a catalytically active His-Cys diad,[35] Cavbase is able to match the conserved environment around the catalytically active cysteines.

UCH-L1 enzyme is associated with Parkinson's disease and lung cancer,[35] and active-site inhibitors of this enzyme show antiproliferative effects in the H1299 lung cancer cell line.[36] Even though the three DUBs are only sparsely populating the cysteine protease family, the application of the filtering step before clustering the matrix can provide valuable information about the uniqueness and singularity of particular binding sites which can help to resolve the selectivity issues of a specific target protein. A cavity screening of the UCH-L1 binding site against the entire Cavbase containing about 275 000 cavity entries revealed that the most similar binding sites, apart from UCH-L1 itself, show a similarity of only 20% or lower.

Interestingly, also proteins from the same Merops family and clan are discarded applying the 20% threshold filter. This indicates high specificity of these enzymes toward their substrates, owing to differently exposed active-site properties, which is not reflected in the sequence space. An example is valacylovir hydrolase and protein phosphatase methylesterase-1 which both fall into the same Merops family. Valacyclovir hydrolase displays high specificity for cleavage of amino acid esters,[37] whereas the protein phosphatase methylesterase-1 binds selectively the carboxy-terminal residues of the catalytic subunit of protein phosphatase-2A.[38]

The cysteine protease human bleomycin hydrolase (hBH) is also removed in cavity space although it shares high sequence identity of about 50% to the cathepsins B, K, and S. hBH is a

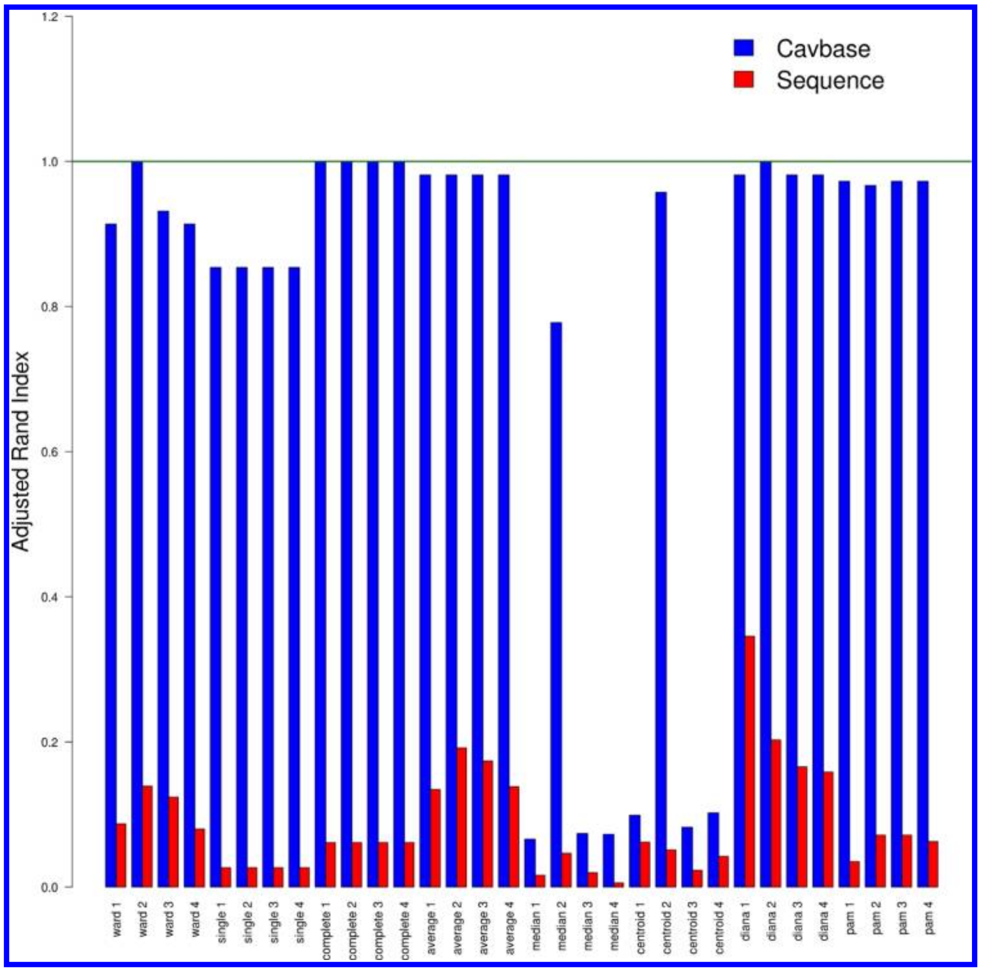


Figure 4. Cavbase and sequence-based clustering of proteases. On the *x*-axis all implemented clustering methods and used distances are shown. The *y*-axis represents the adjusted Rand index (ARI). As the Merops clan classification was defined as reference, a clustering that matches the reference classification, has an ARI of 1, which is depicted as a green line intersecting the *y*-axis. ARIs of the Cavbase approach are shown in blue, and the results from the sequence analysis are shown in red. Cavbase is able to reproduce the Merops clan classification in the case of complete-linkage for all calculated distances, in the case of Ward's method and divisive analysis for the D_2 distance measure. In contrast, the sequence-based analysis of the proteases data set delivers rather poor performance. Used clustering methods: ward = Ward's method, single = single linkage, complete = complete linkage, average = group average, median = median linkage, centroid = centroid linkage, diana = hierarchical divisive analysis, pam = partitioning around medoids.

representative of a self-compartmentalizing protease. The flexibility of its C-terminus contributes to the active site and controls the activity of the enzyme.[39] The C-terminus near the catalytic cysteine is involved in substrate binding and forms a specific cavity that barely shares any similarity with the regarded cysteine cathepsins.

Selectivity of Calpain-1 Inhibitors over Cysteine Cathepsins. Calpains are also cysteine proteases that participate in many calcium regulated functions, e.g. cell proliferation, differentiation, and apoptosis. Their activity depends on the presence of calcium ions. All considered calpains are assigned to the same Merops family C2 and clan CA. In our cavity cluster analysis, calpain-2 and calpain-9 are removed from a shared cluster whereas calpain-1 is assigned to the cluster comprising the cysteine cathepsins from the Merops C1 family. From this finding, several conclusions can be drawn. First, separation of individual calpains reflects the structural flexibility and diversity of active sites in calpains.[40] Second, detecting calpain-1 in the same cluster with cysteine cathepsins leads to the assumption that their binding site properties are similar, despite the low sequence identity of calpain-1 with the other cathepsins, which varies from 25 to 44%. It is interesting to see whether this assumption is also reflected by independent ligand data published in literature. Indeed, the majority of calpain-1 inhibitors lack selectivity over corresponding cathepsins.[41] The overlapping binding site surfaces are visualized for calpain-1 and three cysteine cathepsins V, B, and S in Figure 6. The Ligsite algorithm detects as geometrically most distinct subsites of calpains S1, S2, S1′, and S2′ which were used for the analysis.

The S1 and S1′ pockets of calpain-1 are highly similar to the corresponding subsites of cathepsins, but the S2 and S2′ pockets differ in terms of their exposed properties. This observation suggests for the design of putatively selective ligands a stronger focus on specific interactions with the residues in the S2/S2′ subpockets. Commonly used P2 residues to be considered in ligands are valine and leucine, which provide affinity toward calpains however do not facilitate selectivity.[42] For instance, Cuerrier et al. reported that placement of groups capable of hydrogen-bond formation at the P2 position improves ligand selectivity of calpain-1 over the cysteine cathepsins.[43] Similar effects to achieve selectivity by

Table 4. Following Protease Entries That Comprise a Similarity Value below 0.2 with Any Other Members of the Data Set Were Discarded

catalytic mechanism[a]	Merops clan	protein name
CP	CA	human bleomycin hydrolase
CP	CA	calpain-2
CP	CA	calpain-9
CP	CA	ubiquitin carboxy-terminal hydrolase L1
CP	CA	TNFα-induced protein-3
CP	CA	otubain-2
MP	MC	carboxypeptidase U
MP	MS	membrane dipeptidase
MP	MG	methionyl aminopeptidase-2
MP	MG	Xaa-Pro dipeptidase
MP	MP	AMSH-like protease
SP	SC	serine carboxypeptidase A
SP	SC	valacyclovir hydrolase
SP	SC	protein phosphatase methylesterase-1
SP	SB	tripeptidyl-peptidase-1
TP	PB	taspase-1

[a] CP = cysteine protease, MP = metallo protease, SP = serine protease, TP = threonine protease.

exchanging hydrophobic for hydrogen bonding groups have been also described for aspartyl proteases.[1]

Ligand, Cavity, and Sequence Data: Cluster Analysis of Serine Proteases. Weskamp et al. have shown that the cavity space correlates well with ligand binding data and fold space.[44] A more detailed analysis of Stegemann et al. concentrated on a data set of proteins with mutual sequence identity below 25% that bind cofactors as ligands.[45] In the latter study, we performed a comprehensive analysis of the large body of proteins accommodating cofactors. The analysis revealed high similarity in the cluster pattern of the target protein space if the classification was based either on Cavbase or fold information. Obviously, both spaces regard quite similar or highly correlated information. In the ligand space addressing conformational differences, remarkable differences could be observed. Comprehensive studies on kinases showed significant correlation between similarity of binding sites and the respective ligands they bind.[7,9] In the present work, we were interested whether the previously observed trends can be extended to serine proteases considering a broad range of bioactive ligands. Therefore, we faced the ligand similarity matrix with those obtained from cavity and sequence space and performed a similar clustering.

In order to extract information by comparable means from the different matrices, k and the applied clustering method must be identical. The determination of an optimal k was performed for the three spaces. Our routine suggested for the ligand data a marked value of five, and for the sequence data, the number of clusters was estimated to four and six, respectively (Table 5). The values suggested for the cavity data are less clear-cut, as k = 2 is too small to come up with a meaningful clustering and k = 8 leads to a clustering comprising a large number of singletons. Hence, selection of the most appropriate clustering was performed for each k on sequence and ligand data using the above introduced entropy measure as evaluation criterion (Figure 7). The results have been mutually compared by the ARI. Sequence and cavity data reveal clusterings that deviate significantly, as indicated by the ARI. However, the emerging clustering based on cavity data correlates well with the clustering based on ligand topology, and the obtained similarity is significantly more pronounced than on the corresponding sequence data level (see Supporting Information Table 3).

A more detailed analysis of the investigated spaces has been carried out using the following clustering settings: k = 6 with the complete-linkage clustering method and the previously introduced distance measure D_3. The results are depicted in terms of heatmaps in Figure 8. The more bluish the color, the more similar are the data points, whereas red color indicates increasing dissimilarity.

The overall structuring of three heatmaps suggests much higher discriminative power for the ligand and cavity data compared to the sequence data. The latter shows hardly any discriminative power apart from urokinase-type (uPA) and tissue-type plasminogen activator (tPA) or trypsin and kallikrein-1 which end up in joint clusters. On the ligand heatmap trypsin is of special evidence as strikingly an extended blue bar demonstrates the unspecific character of this particular enzyme. With respect to substrate cleavage, trypsin is one of the most promiscuous enzymes in this family.[46] Two clusters are identically indicated in the three input spaces based on ligand,

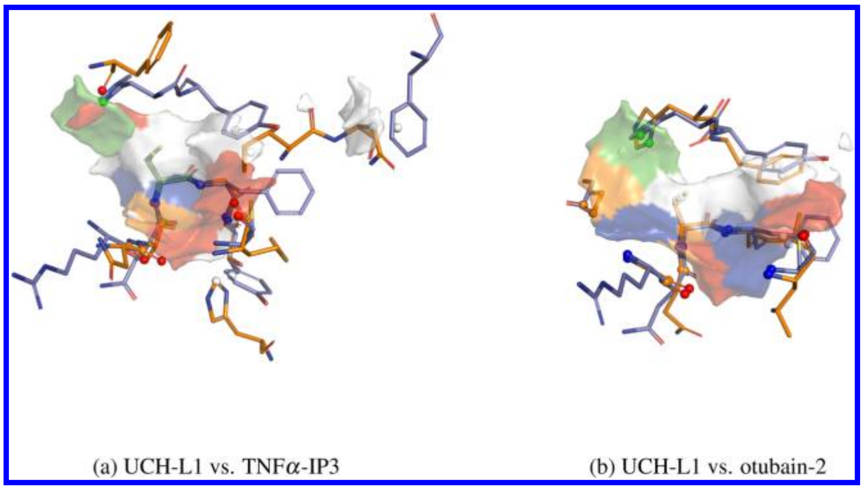


Figure 5. Overlapping active site surface regions of ubiquitin carboxy-terminal hydrolase L1 (UCH-L1) are superimposed onto the cavities of (a) TNFα-induced protein-3 (TNFα-IP3) and (b) otubain-2. Cavities were rendered using PyMOL.[33]

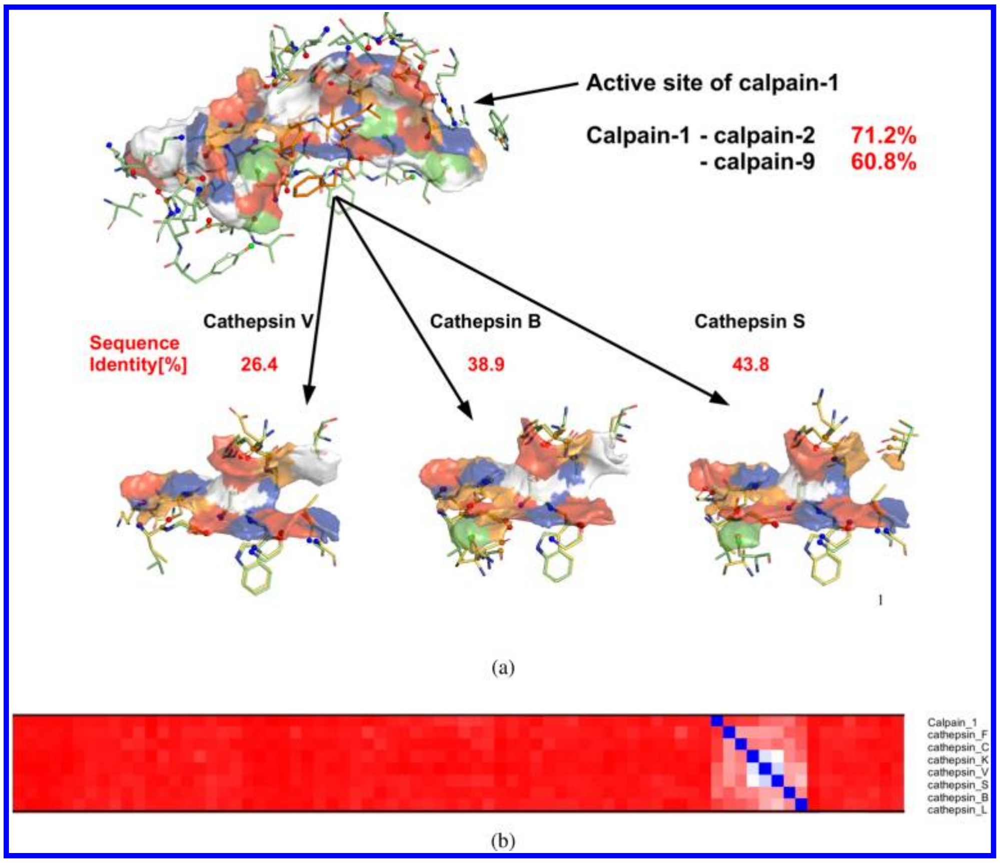


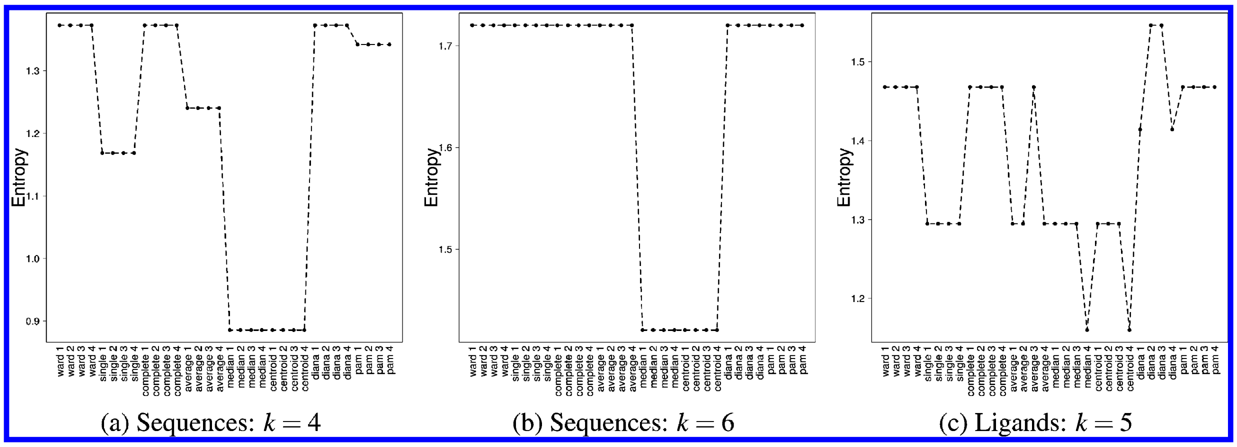
Figure 6. Cross-reactivity between calpain-1 and cysteine cathepsins. (a) Sequence independent similarity of binding site surfaces is detected. Although calpain-2 and calpain-9 exhibit a high sequence identity to calpain-1, the binding site of calpain-1 shares a higher similarity to the cathepsins. Graphical representations were prepared using PyMOL.[33] (b) Section from the entire cavity similarity matrix. High similarity is indicated in blue, and low similarity in red. Remarkably, calpain-1 is found in the same cluster with the cathepsins.

Table 5. Serine Proteases[a]

data set	distance	average silhouette	median split silhouette
ligands	D_{1-4}	5	5
cavities	D_{1-4}	2	8
sequences	D_{1-4}	4	6

[a] Estimated number of clusters based on the ligand similarity, Cavbase score, and sequence identity matrices of serine proteases with $k_{\max}$ = 11.

cavity, and sequence data. FactorXa (fXa) and thrombin share a common cluster whereas chymase ends up in all cases as a singleton. Thrombin and fXa are closely related members of the blood coagulation cascade,[47] and even the successful development of dual inhibitors acting equally potent against both enzymes has been accomplished.[48] The human chymase, which is the only representative chymotrypsin-like protease in the data set, is found as a singleton, reflecting its distinct properties in



(a) Sequences: $k = 4$ (b) Sequences: $k = 6$ (c) Ligands: $k = 5$

Figure 7. Entropy measure for the (a and b) sequence and (c) ligand clusterings of serine proteases. The highest entropy values for the sequence data using k = 4 are found for three and using k = 6 for six of the applied clustering methods. Ligand data suggest for k = 5, the divisive analysis clustering for D_2 and D_3.

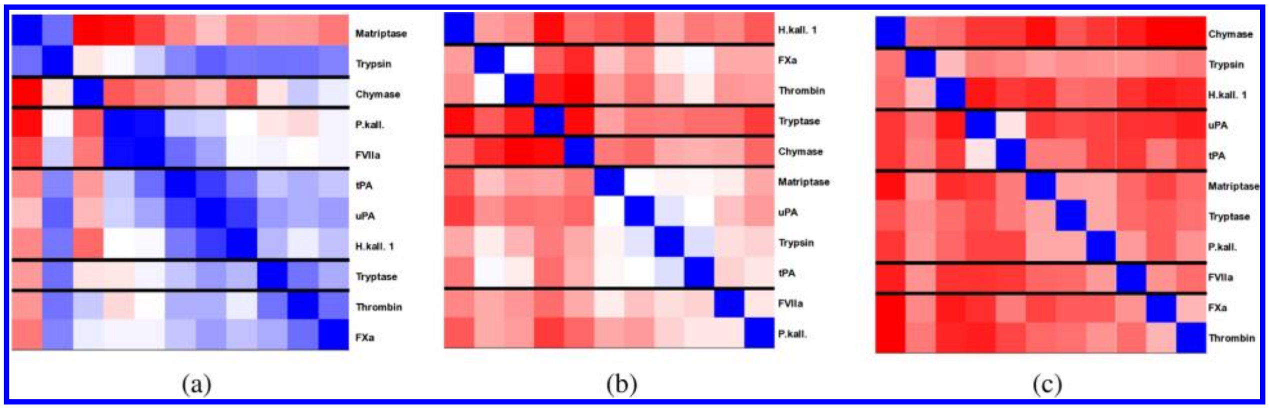


Figure 8. Heatmaps obtained for clustering (a) ligand, (b) cavity, and (c) sequence data applying the complete-linkage method and $k = 6$. Deep blue color represents maximum possible similarity, and deep red color, maximum dissimilarity in the corresponding data set. The range between the maxima is mixed with white color. For reasons of clarity, the six individual clusters are separated only by horizontal black lines and the entries are labeled to the right of the heatmap.

the data set and its high selectivity toward its biological substrates.[49]

Focusing on the more discriminating spaces based on ligand and cavity information, related cluster patterns are found for the 2-fold clusters formed by factor VIIa (fVIIa) and plasma kallikrein (PK) and the singleton created by β-tryptase. The latter β-tryptase is a mast cell serine protease that has been directly linked to the pathology of asthma.[50] Its clustering as a singleton is supported by the fact that successful design of selective β-tryptase inhibitors could be achieved in several studies.[51−53] On the contrary, only a few attempts have been described to address the selectivity problem depicted in the heatmaps of fVIIa and PK, which is particularly indicated for the ligand clustering.[54,55] Interestingly, the latter selectivity issue reflected by the properties of known ligands is already suggested by our comparative analysis in cavity space as fVIIa and PK show the highest mutual similarity in this space. As a consequence and reflecting the current state of inhibitor development, a similar clustering is therefore proposed for ligand topology information and exposed physicochemical properties of the binding pockets.

■ CONCLUSIONS

In the present study, we describe the development, validation, and application of a novel clustering workflow with particular focus on the Cavbase similarity metric. Owing to the implemented routines, any proximity matrix can be provided as input. The program is able to predict the correct number of clusters for two data sets of binding sites and clusters them automatically in accordance with expert classifications, based on orthogonal information such as EC numbers and Merops clans.

As a case study, human proteases were analyzed in more detail. Clustering based on cavity information indicates a cross-reactivity between the cysteine protease calpain-1 and cysteine cathepsins, which has been reported upon calpain-1 inhibitor development in the literature.[41] Unlike binding site information, the usage of a sequence identity matrix as input for clustering fails to produce any meaningful results, thereby making the detection of the described cross-reactivity virtually impossible.

Finally, we utilize our workflow in an attempt to investigate the relationships between ligand, cavity, and sequence spaces of serine proteases. Clustering of ligands, using solely similarities based on their topologies, leads to a pattern that shows higher correlation to the clustering of binding sites than to that of sequences. On the one hand, this result has to be treated with caution, as only eleven serine proteases were considered in the analysis. This fact results mainly from the limited access to the sparse ligand data stored in public databases. On the other hand, the evaluation of binding site information along with protein classification from orthogonal sources can deliver in a data mining approach valuable data to discriminate proteins with respect to selectivity criteria for the development of putative ligands that standard sequence comparison methods can hardly achieve.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

PDB codes of the data sets and mathematical definitions for selected cluster validity criteria.

This material is available free of charge via the Internet at http://pubs.acs.org/.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: Klebe@staff.uni-marburg.de.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the LOEWE program “Synmikro” of the State of Hesse, Germany. The authors thank Florian Meyer (Schober Information Group) for helpful discussion regarding different clustering and statistical approaches. The authors also thank Dr. Peter Kolb and Dr. Juan Carlos Mobarec (Philipps-Universität Marburg) for carefully reading and critically reviewing the manuscript.

■ REFERENCES

(1) Kawasaki, Y.; Freire, E. Finding a better path to drug selectivity. *Drug Discov. Today* **2011**, *16*, 985−990.

(2) Huggins, D. J.; Sherman, W.; Tidor, B. Rational approaches to improving selectivity in drug design. *J. Med. Chem.* **2012**, *55*, 1424−1444.

(3) 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.

(4) Gregori-Puigjané, E.; Mestres, J. SHED: Shannon entropy descriptors from topological feature distributions. *J. Chem. Inf. Model.* **2006**, *46*, 1615−1622.

(5) Pérot, S.; Sperandio, O.; Miteva, M. A.; Camproux, A.-C.; Villoutreix, B. O. Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery. *Drug Discov. Today* **2010**, *15*, 656−667.

(6) Kuhn, D.; Weskamp, N.; Hüllermeier, E.; Klebe, G. Functional classification of protein kinase binding sites using Cavbase. *ChemMedChem* **2007**, *2*, 1432−1447.

(7) Kinnings, S. L.; Jackson, R. M. Binding site similarity analysis for the functional classification of the protein kinase family. *J. Chem. Inf. Model.* **2009**, *49*, 318−329.

(8) Feldman, H. J.; Labute, P. Pocket similarity: are alpha carbons enough? *J. Chem. Inf. Model.* **2010**, *50*, 1466−1475.

(9) Spitzer, R.; Cleves, A. E.; Jain, A. N. Surface-based protein binding pocket similarity. *Proteins* **2011**, *79*, 2746−2763.

(10) Weber, A.; Casini, A.; Heine, A.; Kuhn, D.; Supuran, C. T.; Scozzafava, A.; Klebe, G. Unexpected nanomolar inhibition of carbonic anhydrase by COX-2-selective celecoxib: new pharmacological opportunities due to related binding site recognition. *J. Med. Chem.* **2004**, *47*, 550−557.

(11) An, J.; Totrov, M.; Abagyan, R. Pocketome via comprehensive identification and classification of ligand binding envelopes. *Mol. Cell. Proteomics* **2005**, *4*, 752−761.

(12) Milletti, F.; Vulpetti, A. Predicting polypharmacology by binding site similarity: from kinases to the protein universe. *J. Chem. Inf. Model.* **2010**, *50*, 1418−1431.

(13) Defranchi, E.; Schalon, C.; Messa, M.; Onofri, F.; Benfenati, F.; Rognan, D. Binding of protein kinase inhibitors to synapsin I inferred from pair-wise binding site similarity measurements. *PLoS ONE* **2010**, *5*, e12214.

(14) Ren, J.; Xie, L.; Li, W. W.; Bourne, P. E. SMAP-WS: a parallel web service for structural proteome-wide ligand-binding site comparison. *Nucleic Acids Res.* **2010**, *38*, W441−W444.

(15) Mestres, J.; Gregori-Puigjané, E.; Valverde, S.; Solé, R. V. The topology of drug-target interaction networks: implicit dependence on drug properties and target families. *Mol. Biosyst.* **2009**, *5*, 1051−1057.

(16) Schmitt, S.; Kuhn, D.; Klebe, G. A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.* **2002**, *323*, 387−406.

(17) Hendlich, M.; Rippmann, F.; Barnickel, G. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graphics Modell.* **1997**, *15*, 359−363.

(18) Kuhn, D.; Weskamp, N.; Schmitt, S.; Hüllermeier, E.; Klebe, G. From the similarity analysis of protein cavities to the functional classification of protein families using cavbase. *J. Mol. Biol.* **2006**, *359*, 1023−1044.

(19) Zhao, Y.; Karypis, G. Data Clustering in Life Sciences. *Molec. Biotechnol.* **2005**, *31*, 55−80.

(20) Bairoch, A. The ENZYME database in 2000. *Nucleic Acids Res.* **2000**, *28*, 304−305.

(21) Rawlings, N. D.; Barrett, A. J.; Bateman, A. MEROPS: the peptidase database. *Nucleic Acids Res.* **2010**, *38*, D227−D233.

(22) ChEMBL. http://www.ebi.ac.uk/chembl/ (accessed July 2011).

(23) Pearson, W. R.; Lipman, D. J. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci.* **1988**, *85*, 2444−2448.

(24) Landrum, G. RDKit: Open-source cheminformatics. http://www.rdkit.org (accessed June 2011).

(25) Xu, R.; Wunsch, D. C. *Clustering*; Wiley-IEEE Press: New Jersey, 2009.

(26) Kaufman, L.; Rousseeuw, P. J. *Finding Groups in Data An Introduction to Cluster Analysis*; John Wiley and Sons: New York, 1990.

(27) Rost, B. Twilight zone of protein sequence alignments. *Prot. Eng. Des. Sel.* **1999**, *12*, 85−94.

(28) Rousseeuw, P. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53−65.

(29) Pollard, K. S.; van der Laan, M. J. New methods for identifying significant clusters in gene expression data. *Proceedings of the American Statistical Association, Biometrics Section [CD-ROM]*; American Stastistical Association: Alexandria, VA, 2002.

(30) Halkidi, M.; Batistakis, Y.; Vazirgiannis, M. On Clustering Validation Techniques. *J. Intell. Inf. Syst.* **2001**, *17*, 107−145.

(31) Hubert, L.; Arabie, P. Comparing partitions. *J. Classif.* **1985**, *2*, 193−218.

(32) Meilă, M. Comparing clusterings-an information based distance. *J. Multivar. Anal.* **2007**, *98*, 873−895.

(33) Schrödinger, LLC. *The PyMOL Molecular Graphics System*, version 1.3r1; 2010.

(34) Nanao, M. H.; Tcherniuk, S. O.; Chroboczek, J.; Dideberg, O.; Dessen, A.; Balakirev, M. Y. Crystal structure of human otubain 2. *EMBO Rep.* **2004**, *5*, 783−788.

(35) Das, C.; Hoang, Q. Q.; Kreinbring, C. A.; Luchansky, S. J.; Meray, R. K.; Ray, S. S.; Lansbury, P. T.; Ringe, D.; Petsko, G. A. Structural basis for conformational plasticity of the Parkinson's disease-associated ubiquitin hydrolase UCH-L1. *Proc. Natl. Acad. Sci.* **2006**, *103*, 4675−4680.

(36) Liu, Y.; Lashuel, H.; Choi, S.; Xing, X.; Case, A.; Ni, J.; Yeh, L. A.; Cuny, G. D.; Stein, R. L.; Lansbury, P. J. Discovery of Inhibitors that Elucidate the Role of UCH-L1 Activity in the H1299 Lung Cancer Cell Line. *Chem. Biol.* **2003**, *10*, 837−846.

(37) Lai, L.; Xu, Z.; Zhou, J.; Lee, K.-D.; Amidon, G. L. Molecular basis of prodrug activation by human valacyclovirase, an alpha-amino acid ester hydrolase. *J. Biol. Chem.* **2008**, *283*, 9318−9327.

(38) Xing, Y.; Li, Z.; Chen, Y.; Stock, J. B.; Jeffrey, P. D.; Shi, Y. Structural mechanism of demethylation and inactivation of protein phosphatase 2A. *Cell* **2008**, *133*, 154−163.

(39) O'Farrell, P. A.; Gonzalez, F.; Zheng, W.; Johnston, S. A.; Joshua-Tor, L. Crystal structure of human bleomycin hydrolase, a self-compartmentalizing cysteine protease. *Structure* **1999**, *7*, 619−627.

(40) Davis, T. L.; Walker, J. R.; Finerty, P. J.; Mackenzie, F.; Newman, E. M.; Dhe-Paganon, S. The crystal structures of human calpains 1 and 9 imply diverse mechanisms of action and auto-inhibition. *J. Mol. Biol.* **2007**, *366*, 216−229.

(41) Donkor, I. O. Calpain inhibitors: a survey of compounds reported in the patent and scientific literature. *Expert. Opin. Ther. Pat.* **2011**, *21*, 601−636.

(42) Choe, Y.; Leonetti, F.; Greenbaum, D. C.; Lecaille, F.; Bogyo, M.; Brömme, D.; Ellman, J. A.; Craik, C. S. Substrate profiling of cysteine proteases using a combinatorial peptide library identifies functionally unique specificities. *J. Biol. Chem.* **2006**, *281*, 12824−12832.

(43) Cuerrier, D.; Moldoveanu, T.; Campbell, R. L.; Kelly, J.; Yoruk, B.; Verhelst, S. H. L.; Greenbaum, D.; Bogyo, M.; Davies, P. L. Development of calpain-specific inactivators by screening of positional scanning epoxide libraries. *J. Biol. Chem.* **2007**, *282*, 9600−9611.

(44) Weskamp, N.; Hüllermeier, E.; Klebe, G. Merging chemical and biological space: Structural mapping of enzyme binding pocket space. *Proteins* **2009**, *76*, 317−330.

(45) Stegemann, B.; Klebe, G. Cofactor-binding sites in proteins of deviating sequence: Comparative analysis and clustering in torsion angle, cavity, and fold space. *Proteins* **2011**, 626−648.

(46) Hilpert, K.; Ackermann, J.; Banner, D. W.; Gast, A.; Gubernator, K.; Hadvary, P.; Labler, L.; Mueller, K.; Schmid, G. Design and synthesis of potent and highly selective thrombin inhibitors. *J. Med. Chem.* **1994**, *37*, 3889−3901.

(47) Sanderson, P. E. J. Small, noncovalent serine protease inhibitors. *Med. Res. Rev.* **1999**, *19*, 179−197.

(48) Nar, H.; Bauer, M.; Schmid, A.; Stassen, J.-M.; Wienen, W.; Priepke, H. W.; Kauffmann, I. K.; Ries, U. J.; Hauel, N. H. Structural basis for inhibition promiscuity of dual specific thrombin and factor Xa blood coagulation inhibitors. *Structure* **2001**, *9*, 29−37.

(49) McGrath, M. E.; Mirzadegan, T.; Schmidt, B. F. Crystal structure of phenylmethanesulfonyl fluoride-treated human chymase at 1.9 Å. *Biochemistry* **1997**, *36*, 14318−14324.

(50) Molinari, J. F.; Scuri, M.; Moore, W. R.; Clark, J.; Tanaka, R.; Abraham, W. M. Inhaled tryptase causes bronchoconstriction in sheep via histamine release. *Am. J. Respir. Crit. Care Med.* **1996**, *154*, 649−653.

(51) Combrink, K. D.; Gülgeze, H. B.; Meanwell, N. A.; Pearce, B. C.; Zulan, P.; Bisacchi, G. S.; Roberts, D. G.; Stanley, P.; Seiler, S. M. 1,2-Benzisothiazol-3-one 1,1-dioxide inhibitors of human mast cell tryptase. *J. Med. Chem.* **1998**, *41*, 4854−4860.

(52) Hopkins, C. R.; Czekaj, M.; Kaye, S. S.; Gao, Z.; Pribish, J.; Pauls, H.; Liang, G.; Sides, K.; Cramer, D.; Cairns, J.; Luo, Y.; Lim, H.-K.; Vaz, R.; Rebello, S.; Maignan, S.; Dupuy, A.; Mathieu, M.; Levell, J. Design, synthesis, and biological activity of potent and selective inhibitors of mast cell tryptase. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 2734−2737.

(53) Lee, C.-S.; Liu, W.; Sprengeler, P. A.; Somoza, J. R.; Janc, J. W.; Sperandio, D.; Spencer, J. R.; Green, M. J.; McGrath, M. E. Design of novel, potent, and selective human beta-tryptase inhibitors based on alpha-keto-[1,2,4]-oxadiazoles. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 4036−4040.

(54) Olivero, A. G.; Eigenbrot, C.; Goldsmith, R.; Robarge, K.; Artis, D. R.; Flygare, J.; Rawson, T.; Sutherlin, D. P.; Kadkhodayan, S.; Beresini, M.; Elliott, L. O.; DeGuzman, G. G.; Banner, D. W.; Ultsch, M.; Marzec, U.; Hanson, S. R.; Refino, C.; Bunting, S.; Kirchhofer, D. A selective, slow binding inhibitor of factor VIIa binds to a nonstandard active site conformation and attenuates thrombus formation in vivo. *J. Biol. Chem.* **2005**, *280*, 9160−9169.

(55) Young, W. B.; Mordenti, J.; Torkelson, S.; Shrader, W. D.; Kolesnikov, A.; Rai, R.; Liu, L.; Hu, H.; Leahy, E. M.; Green, M. J.; Sprengeler, P. A.; Katz, B. A.; Yu, C.; Janc, J. W.; Elrod, K. C.; Marzec, U. M.; Hanson, S. R. Factor VIIa inhibitors: chemical optimization, preclinical pharmacokinetics, pharmacodynamics, and efficacy in an arterial baboon thrombosis model. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 2037−2041.