A Hierarchical Clustering Approach for Large Compound Libraries

Alexander Böcker,[†,‡] Swetlana Derksen,[†] Elena Schmidt,[†] Andreas Teckentrup,[‡] and Gisbert Schneider*[,†]

Johann Wolfgang Goethe-Universität, Institut für Organische Chemie und Chemische Biologie,
Marie-Curie-Str. 11, D-60439 Frankfurt, Germany, and Boehringer Ingelheim Pharma GmbH & Co. KG,
Department of Lead Discovery, D-88397 Biberach a.d. Riss, Germany

Received January 3, 2005

A modified version of the *k*-means clustering algorithm was developed that is able to analyze large compound libraries. A distance threshold determined by plotting the sum of radii of leaf clusters was used as a termination criterion for the clustering process. Hierarchical trees were constructed that can be used to obtain an overview of the data distribution and inherent cluster structure. The approach is also applicable to ligand-based virtual screening with the aim to generate preferred screening collections or focused compound libraries. Retrospective analysis of two activity classes was performed: inhibitors of caspase 1 [interleukin 1 (IL1) cleaving enzyme, ICE] and glucocorticoid receptor ligands. The MDL Drug Data Report (MDDR) and Collection of Bioactive Reference Analogues (COBRA) databases served as the compound pool, for which binary trees were produced. Molecules were encoded by all Molecular Operating Environment 2D descriptors and topological pharmacophore atom types. Individual clusters were assessed for their purity and enrichment of actives belonging to the two ligand classes. Significant enrichment was observed in individual branches of the cluster tree. After clustering a combined database of MDDR, COBRA, and the SPECS catalog, it was possible to retrieve MDDR ICE inhibitors with new scaffolds using COBRA ICE inhibitors as seeds. A Java implementation of the clustering method is available via the Internet (http://www.modlab.de).

INTRODUCTION

Over the past two decades, the early drug discovery process has focused on two main concepts, rational design[1] and screening of large compound collections.[2,3] High-throughput screening (HTS) provides a rich primary source of compound activity data, particularly when fed by combinatorial and parallel synthesis.[4,5] Consequently, software requirements for data mining and model building have changed toward methods that are able to handle large data sets.[6,7] Exploiting the knowledge of both the "actives" and "nonactives" to extract structure−activity relationships (SARs) from HTS data is pivotal because it helps formulate crude SAR hypotheses at early stages of a drug discovery project. Unsupervised data clustering provides one possibility to get a quick first overview of potential groups of active and inactive compounds and to visualize and analyze decisive molecular features.

Numerous clustering methods have been proposed for such a purpose, for example, hierarchical distance-based methods such as Ward's clustering, as one prominent member,[8,9] and nonhierarchical methods such as Jarvis−Patrick,[10−13] *k*-means,[14−16] or Bayes' unsupervised clustering.[17] Hierarchical methods have the advantage of building up an interpretable relationship between the clusters. Many such algorithms have, at least, squared running time and memory requirements, which renders them unfeasible for very large data sets. On the other hand, nonhierarchical methods have linear running time and space requirements but can lead to either large and heterogeneous or small and exclusive clusters, depending on the number of predefined clusters and the distance boundary conditions. To circumvent these shortcomings, several extensions have been proposed in the literature such as the introduction of "fuzzy clusters"[10,15] or the re-clustering of singletons using milder distance boundary conditions.[12] Irrespective of these developments, new methods are constantly being introduced, for example, to build up a phylogeny-like tree employing maximum common substructures[18] or to cluster a data set according to the frequency of substructure elements.[19,20]

This article presents the implementation of the hierarchical *k*-means clustering method, which has been known in the bioinformatics and chemoinformatics fields for some time[21,22] and has been put into practice as the "Divisive K-means" method by researchers at Barnard Chemical Information Ltd. (http://www.bci.gb.com). The algorithm is able to hierarchically cluster very large data sets with presumably more than 1 million data points in a high-dimensional descriptor space. Our software was designed in such a way that the clusters can be graphically represented by a dendrogram and activity data (e.g., IC_{50} values, class labels) can be assigned to the individual clusters. The results are displayed in an interactive window, enabling the user to quickly navigate through the tree and interpret the results. The method was validated with two data sets, the MDL Drug Data Report (MDDR)[23] and

[†] Johann Wolfgang Goethe-Universität.

[‡] Boehringer Ingelheim Pharma GmbH & Co. KG.

* Corresponding author phone: +49 (0)69 798 29822; fax: +49 (0)69 798 29826; e-mail: gisbert.schneider@modlab.de.

10.1021/ci0500029 CCC: $30.25 © 2005 American Chemical Society
Published on Web 05/03/2005

Table 1. MDDR Compound Processing

filter	MDDR entries
none	141 692
remove entries lacking structure information	139 037
remove counterions[a]	138 584
neutralize and remove nonorganic compounds[b]	136 702
remove reactive and unsuited compounds[c] and compounds with nondrug activity record[d]	109 528
remove entries without "activity" record	59 173

[a] If it was not clear which was the counterion, entries were discarded. [b] Scitegig Pipeline Pilot was used to neutralize the structures and discard nonorganic compounds. [c] The Filter program from OpenEye[32] was adopted according to Hann and co-workers.[29] [d] A Python script was used to remove entries having a nondrug activity record.

the Collection Of Bioactive Reference Analogues (COBRA)[24] with the aim to reveal compound ontologies that may help interpret and explore HTS data. A retrospective study showed the ability of the method to find putatively new compounds having different scaffolds. The clustering method was implemented in Java and is freely available from http://www.modlab.de.

DATA AND METHODS

Data Preparation. Four data sets were used to test the method: one artificial data set, Fisher's Iris data set,[25] two data sets containing small organic molecules (COBRA[24] and MDDR[23]), and the SPECS catalog.[26] COBRA (version 3.1) is a data set containing 5 375 pharmacologically active molecules taken from the literature. It provides information about the target receptor class, receptor name, receptor subtype, and the indication field for each entry. Because our version of COBRA contains only desalted, neutralized, and "drug-like" molecules, no further filtering steps were applied.

The SPECS catalog (version of June 2003) is a vendor database consisting of 229 658 small organic molecules that can be purchased to build up diverse screening libraries for HTS and lead discovery programs.[26] The used version contains only desalted and neutralized compounds.

The MDDR database (version of August 2003) contains 141 692 biologically relevant structures taken from patent literature, scientific journals, and meeting reports. Each entry contains a 2D molecular structure field (for 2655 entries, structural information was unavailable), an activity class field, and a corresponding activity class index (one molecule can be assigned to different activity classes). The MDDR was prepared using the following steps: (1) Entries lacking structural information were removed, leaving 139 037 compounds. (2) Counterions were removed using a statistical in-house approach implemented in the Kensington Discovery Edition[27] software package at Boehringer Ingelheim, and (3) structures were neutralized using Scitegic Pipeline Pilot.[28] For 453 entries, the counterion was indistinguishable. These entries were removed. Because we were only interested in small, organic, drug-like molecules, several drug-likeness filters were applied to the MDDR.[29,30] When all of the filtering steps were applied to the MDDR, the database was reduced to 109 528 entries (Table 1).

Applied "drug-likeness" filters:

- Remove the blood supplements, vaccines, monoclonal antibodies, molecules for cancer immune therapy, chemopreventives, chemoprotectives, molecules for gene therapy, radio sensitizer, diagnostic agents, antidotes, antibiotics, and antineoplastica.
- The molecular weight has to be between 150 and 1000 Da.
- Remove the molecules with reactive functional groups.
- Remove the molecules with more than six halogen atoms.
- Molecules have to contain a least one carbon atom and one nitrogen, oxygen, or sulfur atom.
- Molecules should only contain H, C, N, O, F, P, S, Cl, Br, and I (this step was done in combination with the neutralization).

Data preparation steps:

- Removal of the molecules without structures, adding explicit hydrogens, and setting the atom ionization to a formal charge were done using Molecular Operating Environment (MOE).[31]
- Removal of the counterions was performed using a statistical in-house approach implemented in the Kensington Discovery Edition software package (InforSense)[27] at Boehringer Ingelheim.
- Neutralization was done using pipeline pilot (Scitegic)[28] or MOE-SVL scripts for unrecognized cases.
- Removal of the entries that are not associated with a drug target in the MDDR database annotation (e.g., radio sensitizers) was done using standard functions of the Kensington Discovery Edition software.[27]
- Removal of the entries with undesired properties was done using the program FILTER (OpenEye)[32] and in-house Pearl scripts.

Descriptor Calculation and Pruning. All MOE 2D[31] and CATS2D[33] (CATS = pharmacophore atom types) descriptors were calculated for MDDR and COBRA. The MOE 2D descriptor set contains 147 descriptors describing physical properties, subdivided surface areas, atom and bond counts, Kier and Hall connectivity and kappa shape indices, adjacency and distance matrices, and pharmacophore features (for details, see http://www.chemcomp.com). The correlation vector descriptor CATS2D (150 dimensions) is based on potential pharmacophore points (PPPs).[34] Atoms are assigned to five different PPPs (hydrogen donor, hydrogen acceptor, ionizable or positively charged, ionizable or negatively charged, and lipophilic) and correlated with the respective distance counted in bond lengths (ranging from zero to nine bonds). All descriptors were mean-centered and scaled to unit variance.

To reduce the dimensionality, two consecutive approaches were followed: (1) the removal of descriptors having a low information content, measured following the Shannon entropy (SE) concept,[35] and (2) the removal of redundant descriptors using the unsupervised forward selection (UFS) algorithm.[36]

SE is defined by eq 1.[35]

$$\mathrm{SE} = -\sum_i p_i \log_2 p_i \tag{1}$$

with p_i giving the probability of the number of data points c_i within a data range i (eq 2):

$$p_i = \frac{c_i}{\sum c_i} \tag{2}$$

For SE scaling to a bin-independent data range, the obtained

Table 2. Results of Descriptor Pruning and Final Data Sets

	COBRA		MDDR				SPECS + COBRA + MDDR
descriptor set	MOE2D		CATS2D		MOE2D		MOE2D
original number of descriptors	147		150		147		147
entropy-based pruning[a]	111		45		110		111
UFS R^2 threshold	0.8	0.99	0.8	0.99	0.8	0.99	0.99
UFS-based pruning	22	53	31	45	24	56	60
final data set name	COBRA08	COBRA099	MDDRCATS08	MDDRCATS099	MDDRMOE08	MDDRMOE099	

[a] Descriptors having a standardized Shannon entropy below 0.3 were removed.

values were divided by the logarithm of the total number of bins (eq 3).

$$\mathrm{sSE} = \frac{\mathrm{SE}}{\log_2(\mathrm{bins})} \qquad (3)$$

Following the proposal of Bajorath and co-workers, we set the number of bins to 100 and discarded descriptors having a scaled SE (sSE) value equal to or less than 0.3 as "information-poor".[37−39]

UFS was performed to remove redundant dimensions from the remaining descriptor set, as published by Whitely and co-workers.[36] Starting with the two least-correlated descriptors, this method builds up a descriptor space by choosing the next descriptor having the lowest multiple correlation coefficient R^2 to the current descriptor set. This is repeated until a predefined threshold is reached. We used two thresholds (T), a conservative value of $T = 0.99$ and $T = 0.8$ as a more stringent value. This resulted in six data sets encoded by 22 (COBRA, MOE08), 24 (MDDR, MOE08), 31 (MDDR, CATS08), 45 (MDDR, CATS099), 53 (COBRA, MOE099), and 56 (MDDR, MOE099) dimensions (Table 2). In the remainder of this paper, the short names of the data sets will be used. UFS version 1.8 (http://www.cmd.port.ac.uk/cmd/software.shtml) was employed for descriptor pruning.

For evaluation purposes, an additional data set was built up, combining molecules from SPECS, COBRA, and MDDR. MOE 2D descriptors were calculated for the resulting 344 561 molecules, which were pruned employing the above-described SE and UFS approach ($T = 0.99$). This resulted in 60 remaining descriptors (Table 2).

Hierarchical k-Means Clustering. The k-means algorithm represents a nonhierarchical clustering technique.[17] It requires $O(kn)$ computation time and space, with n being the number of data points and k being the number of clusters. The k-means algorithm randomly selects k data points as initial cluster centroids (step 1). k clusters are formed by assigning each data point to its nearest centroid (step 2). New virtual centroids are then calculated for each cluster (step 3). The second and third steps are iterated until a predefined number of iterations is reached or the clusters do not change anymore. Although the algorithm was shown to produce reliable results, there are several features that have to be dealt with carefully:

(i) The number of cluster centroids k has to be predefined. For a large k, the resulting clusters tend to be small and exclusive, whereas for a low k, clusters tend to become large and heterogeneous. The optimal choice of k depends on the inherent structure of the data set and the aim of the particular study.

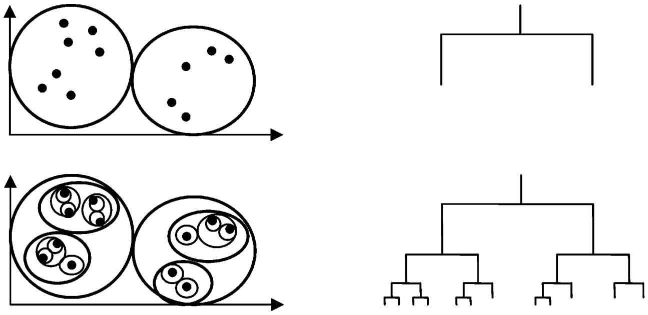


Figure 1. Example of hierarchical clustering ($k = 2$). Data objects (left) are represented by a hierarchical tree structure (right).

(ii) Depending on the initially chosen cluster centroids, small clusters that are present in the data distribution could be missed.

Because of the nature of the algorithm, a hierarchical relationship between the clusters is not assigned. Still, visualizing hierarchical relationships can be helpful for the analysis of large data sets. To address this in the context of the above-mentioned issues, we implemented a modified form of the k-means algorithm, forming k clusters at each level of a hierarchical tree.[22] Figure 1 gives an example of a tree with $k = 2$ (binary tree). The idea is that, for hierarchical clustering, no predefinition of the number of clusters is required. In contrast, it must be defined until which distance threshold molecules are treated as similar and are fused to form one cluster.

The basic steps of the modified algorithm are

Step 1: Define k. (For a binary tree, $k = 2$.)

Step 2: Select a distance threshold Θ. (In the present study, the Euclidian metric was employed to define "distance".)

Step 3: Perform data clustering for the actual tree level. (Starting with the root node, k child nodes are created and the data set is partitioned according to the k-means algorithm.)

Step 4: Check for each cluster: if the maximum distance of a data point to the created virtual mean exceeds the threshold, repeat Step 3 for this cluster. Otherwise, terminate the procedure.

It should be noticed that the hierarchical k-means approach is a technique that uses a randomization step during the initialization of the centroid vectors. Thus, identical trees will not necessarily result from multiple runs on the same data set.

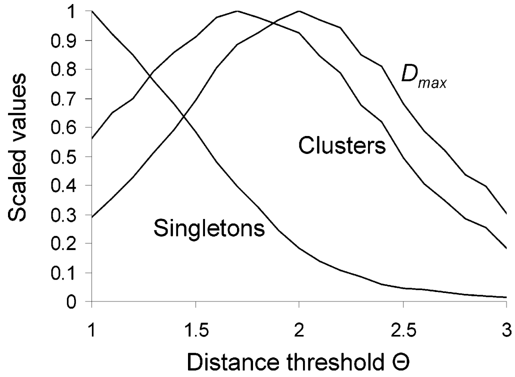


Figure 2. Identification of an optimal value for the distance threshold Θ. Clustering was done for the COBRA data set using a subset of all MOE 2D descriptors (22 descriptors).

Θ is an arbitrary value. It was used to terminate the clustering process at a stage where both the total number of clusters (a cluster contains at least two data points) and the total number of singletons (a cluster containing exactly one data point) do not assume extreme values. To help find a useful threshold value, a preliminary experiment was performed. For a specified distance range, the number of singletons, the number of clusters, and the sum of the maximum distances in each terminal cluster ("leaf") ($D_{\max}$; see eq 4) were calculated, scaled to [0,1] and plotted in one graph (Figure 2). $D_{\max}$ can be interpreted as the sum of the cluster "radii" of the i leaf clusters.

$$D_{\max} = \sum_i \max[d(x_j, c_i)] \quad \text{with } 1 \le j \le N_i \tag{4}$$

where d is the Euclidian distance metric, x_j represents data points that are a member of the same terminal cluster i, c_i represents the centroid of cluster i, and N_i is the number of members of cluster i. The Θ value that led to the maximal $D_{\max}$ value was used for our clustering experiments (Figure 2). It represents a compromise between fine- and coarse-grained clustering of the data space, which is free of the subjective assignment of Θ values. Table 3 summarizes the results obtained for the different data sets used in the present work.

Enrichment Factor. The main points of interest were the purity of individual clusters and whether molecules that interact with the same target (or receptor) or a family of related targets fall in the same sub tree. For each cluster node, we calculated an enrichment factor (EF).[16] This gives an estimate of how well compounds that bind to the same target (or target class) are clustered in a tree node i (eq 5).

$$\mathrm{EF}_{i,c} = \frac{\dfrac{\#\mathrm{entry}_{i,c}}{\#\mathrm{entry}_{\mathrm{root},c}}}{\dfrac{\#\mathrm{entry}_{i}}{\#\mathrm{entry}_{\mathrm{root}}}} \tag{5}$$

with $\#\mathrm{entry}_{i,c}$ being the number of entries in the node i belonging to class c, $\#\mathrm{entry}_i$ being the total number of entries in node i, $\#\mathrm{entry}_{\mathrm{root},c}$ being the total number of entries of class c in the data set, and $\#\mathrm{entry}_{\mathrm{root}}$ being the total number of entries. EF > 1 indicates that more compounds belonging to an activity class c were clustered in a tree node than would be expected from an equal distribution. Note that the largeness of this value depends on the size of the tree section under consideration. On upper tree levels, where clusters are big, EF values can be low, whereas EF values on the lower tree levels can get bigger at random.

RESULTS AND DISCUSSION

In a first experiment, Fisher's Iris data set was used to test the performance of the clustering algorithm. This data set was used by Fisher (1936) in his initiation of the linear-discriminant-function technique.[25] It consists of 150 random samples of flowers from the Iris species *setosa*, *versicolor*, and *virginica*. For each species, there are 50 observations for sepal length, sepal width, petal length, and petal width in centimeters, yielding a four-dimensional data space. Three class labels were used, representing the three different species. Figure 3 shows the projection of the data on the first two principal components (PCs) and the corresponding binary k-means tree that was constructed using the original data. It is evident that the tree representation is in agreement with the PC projection in that three distinct classes are shown that occupy different branches of the tree. Three entries of Class 2 were assigned to the right side of the tree, which is dominated by Class 1 examples. These points are indicated by little arrows in Figure 3. This observation reveals a disadvantage of the algorithm: the number of clusters on a tree level, k, forces the data space to be split into k subregions on each tree level. Local data densities lying at an interface region between two such subregions bear the danger of being torn apart. As can be seen in Figure 3b, after the third split of the Iris data set, the three "outliers" form a pure cluster again. This outcome of our preliminary experiment is promising because although data points were assigned to the "wrong" side of the tree, in the end, pure but smaller clusters were obtained. It should be kept in mind that although an optimum solution exists for such a problem, when dealing with large data sets, we can only use algorithms that try to reach the optimum.

These results encouraged us to apply hierarchical k-means clustering to MDDR and COBRA. We focused on two examples, (i) caspase 1 [interleukin 1 (IL1) cleaving enzyme, ICE; EC number 3.4.22.36] inhibitors[40] from COBRA and (ii) glucocorticoid receptor ligands[41] from MDDR.

ICE inhibitors[40] prevent IL1 cleavage, which plays a major role in a wide range of inflammatory and autoimmune diseases, like rheumatoid arthritis, osteoarthritis, chronic obstructive pulmonary disease, and asthma.[40,42] ICE belongs to the family of cysteine proteases and specifically cleaves Asp116−Ala117 and Asp27−Gly28. Inhibitors typically mimic this residue motif.[42]

Glucocorticoid receptors[43,41] bind glucocorticoids and induce gene transcription. This leads to catabolic reactions in extrahepatic tissues, anabolic reactions in the liver, immune-suppressive reactions in the lymphatic system, and, under stress, to elevated cortisol levels having inflammation blocking effects.[43] Because of the various functions of glucocorticoids, drugs that bind to glucocorticoid receptors have implications in a lot of therapeutic areas, for example, rheumatic disease or allergic reactions.[43]

COBRA Clustering (ICE Inhibitors). We clustered two versions of COBRA (COBRA08 and COBRA099, Table 3) using the hierarchical k-means approach algorithm ($k = 2$).

Table 3. Calculated Distance Thresholds for the Different Data Sets

data set	descriptor set	UFS threshold, *T*	distance threshold, Θ
MDDRCATS08	CATS2D	0.8	2.3
MDDRCATS099	CATS2D	0.99	2.6
MDDRMOE08	MOE2D	0.8	1.8
MDDRMOE099	MOE2D	0.99	2.8
COBRA08	MOE2D	0.8	2.0
COBRA099	MOE2D	0.99	2.8
SPECS + COBRA + MDDR	MOE2D	0.99	2.6

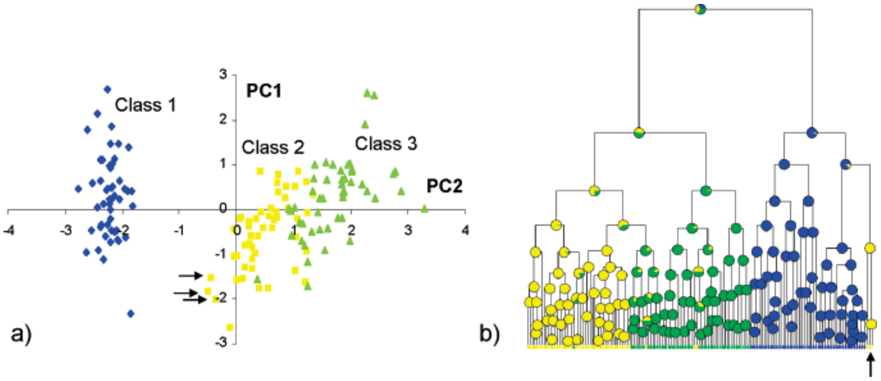


Figure 3. Clustering of Fisher's Iris data set. (a) Score plot of the data according to the first two PCs (explained variance > 95%). The different data classes are colored in blue (Class 1), yellow (Class 2), and green (Class 3). (b) Binary tree ($k = 2$, $\Theta = 0$). Each tree node (cluster) is represented by a pie chart showing its relative class composition. Arrows indicate the location of three Class 2 data points in the PC plot (part a) and the *k*-means tree (part b).

For both data sets, all ICE inhibitors were assigned to one side of the tree on the first tree level. On subsequent tree levels, the inhibitors separate into different branches. When the emerging clusters obtained with the different descriptor sets were compared, different sets of ICE inhibitors were grouped together in the clusters. This is expected because the larger descriptor set (COBRA099) should emphasize properties in a different way than the smaller one. For COBRA08, three large clusters emerge containing mainly ICE inhibitors. Cluster one consists of 15 entries, with nine ICE inhibitors (EF = 82.7), cluster two consists of 12 entries, with six ICE inhibitors (68.9), and cluster three consists of 14 entries, again with nine ICE inhibitors (EF = 88.6; in total, 39 ICE inhibitors are present in COBRA). The molecules in the three clusters, which are not defined as ICE inhibitors, fall into two classes. Class one consists of other protease inhibitors, like matrix metalloproteinase inhibitors,[44] human rhinovirus 3C protease inhibitors,[45] or hepatitis C virus NS3 protease inhibitors.[46] This is not surprising because small organic protease inhibitors try to mimic peptide sequences using similar design strategies,[47] and these peptide sequences can be, in turn, very similar. Class two consists of $\alpha_4\beta_1$ intregrin (also known as very late antigen 4, VLA-4) antagonists, which have potential for the treatment of allergic diseases like asthma and other chronic inflammatory diseases.[48] Although both ICE inhibitors and VLA-4 antagonists play a role in the treatment of allergic diseases, no interconnection is known to the authors between both. Regarding the ICE inhibitors in the three clusters, two out of these three clusters are composed of structurally similar molecules; the third cluster contains less-related compounds. Representative structures are shown in Figure 4. We also observed small individual clusters with only one or two ICE inhibitors. These might have resulted from unsuitable cluster boundaries (cf. Figure 2) or might be a consequence of shortcomings of the chosen descriptor set.

The representatives contain mutual substructure elements that are found in identical or only slightly different forms in all other structures of the cluster. A peptidic moiety (Figure 4, yellow) represents a substructure motif that is present in all ICE inhibitors: a modified aspartic acid, alanine, valine, and a peptide bond to a carbonyl residue. Clusters A and B are more closely related to each other, which can be explained by the shared ethyl phenol group (Figure 4, magenta). They differ in the occurrence of an acetamide group in cluster A (Figure 4, green) and a propyl benzene group in cluster B (Figure 4, green). The more distant cluster C accounts for two unique substructures, a toluene group and a ring closure connecting the alanine and valine residues (Figure 4, green). The observed distribution of the ICE inhibitors among three main clusters is a consequence of the chosen distance threshold Θ.

MDDR Clustering (Glucocorticoid Receptor Ligands). We clustered the MDDR with our algorithm ($k = 2$) using the four descriptor sets listed in Table 2 and the corresponding calculated stop thresholds listed in Table 3. The clustering process was fast, for example, 69.5 s for the MDDRCATS08 data set (45 dimensions; 59 173 entries) on a 2 GHz Mobile Intel Pentium CPU (1 GB RAM). When the different descriptor sets were compared according to their capability to separate glucocorticoid receptor ligands from the rest of MDDR, the MDDRCATS08 descriptor set yielded the best results at the root level. In total, 90% of the glucocorticoid receptor ligands were assigned to the right side. Glucocorticoid receptor ligands in the MDDR can be divided into three main lead classes: 23% class I, 66% class II, and 11% class III (see Figure 5A). When the different descriptor sets were judged according to their capability to separate the different lead classes, both descriptor sets using CATS2D separated class I from classes II and III on the root level. In contrast, the MDDRMOE099 set separated class III from classes I and II, and the MDDRMOE08 descriptor set showed

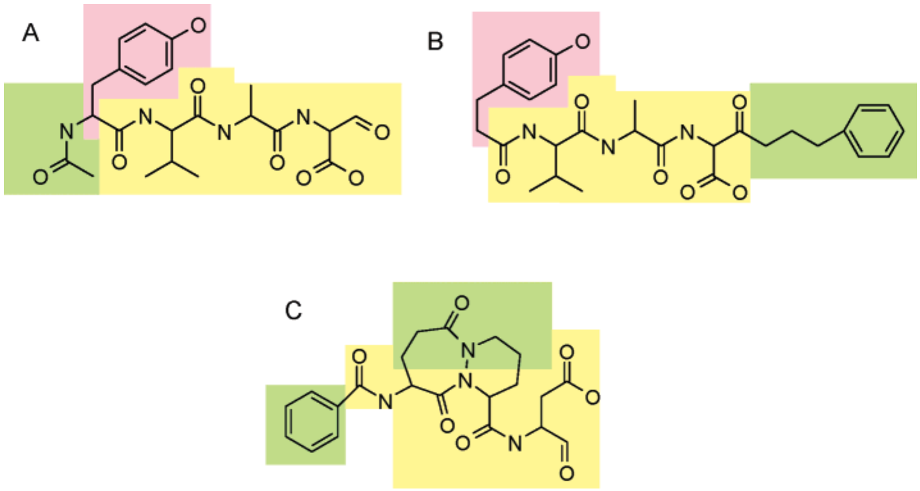


Figure 4. Representative ICE inhibitors from the three main emerging clusters (A, B, and C; COBRA and MOE 2D descriptors; UFS threshold = 0.8). Clusters A and B contain closely related molecular structures. Yellow: common substructure motif in all three clusters. Magenta: common motif in clusters A and B. Green: unique motifs.

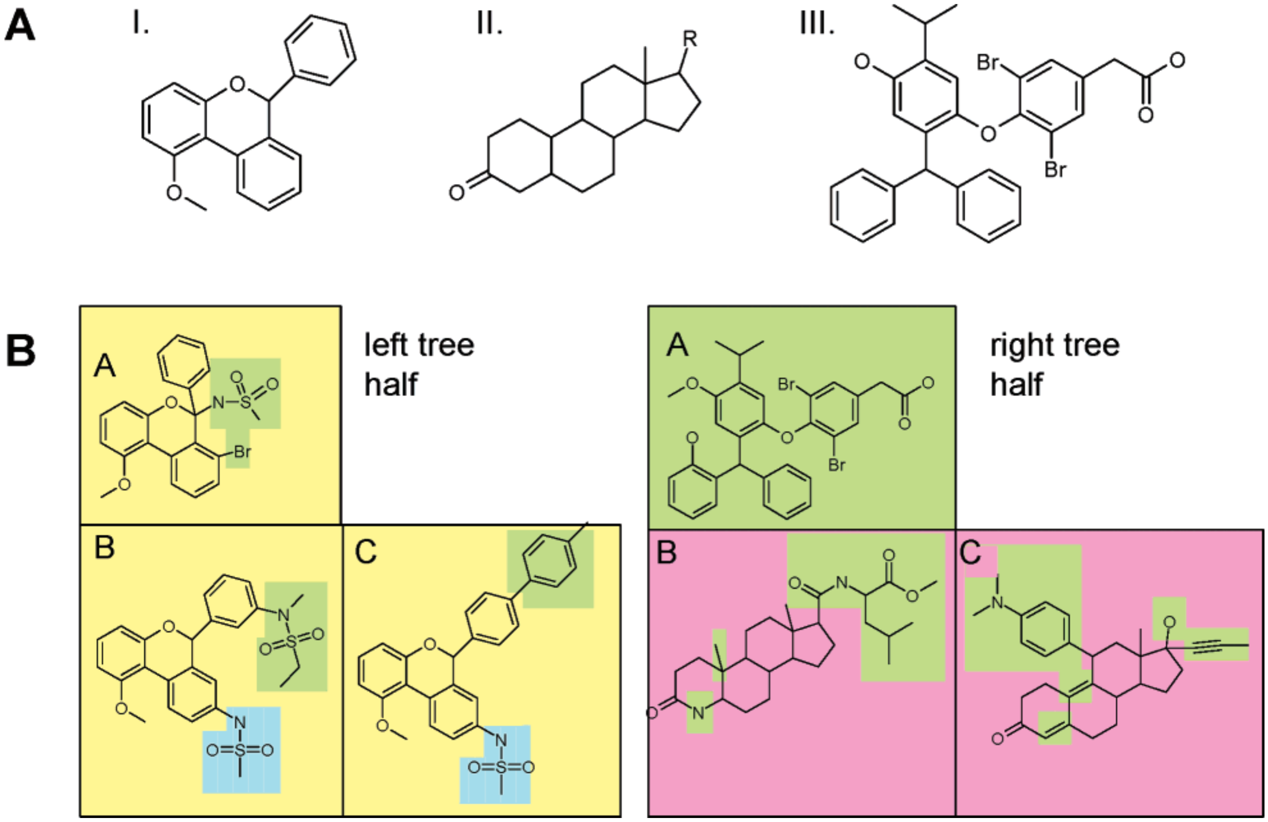


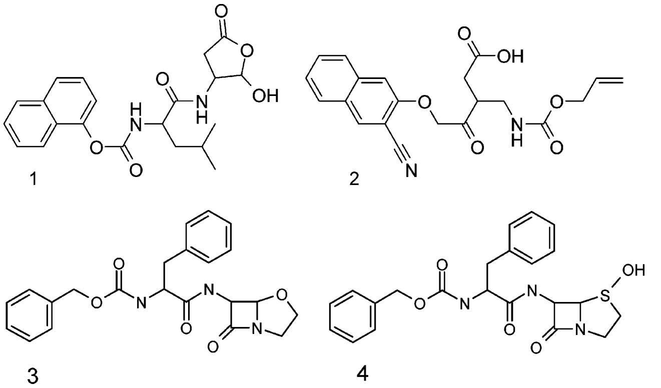
Figure 5. (A) The three core structures (I–III) of glucocorticoid receptor ligands present in the MDDR database. (B) Representative glucocorticoid receptor ligands in the three largest emerging clusters (A, B, and C) of the left and right tree halves on tree level 1 of the binary cluster tree. Results are shown for MDDR and CATS2D descriptors ($T = 0.8$; $\Theta = 2.3$). On both sides, clusters B and C lie in closer proximity to each other than they do to cluster A. Yellow: common motif in all three clusters on the left side. Blue: common motif in clusters B and C on the left side. Magenta: common motif in clusters B and C on the right side. Green: unique motifs.

no complete separation at all. When clusters were chosen from the descriptor set obtained with MDDRCATS08, which contained more than two glucocorticoid receptor ligands, 270 additional molecules were selected in total. Of these molecules, 86.3% are described as ligands of other nuclear hormone receptors or as ligands being involved in the synthesis of steroid hormones. The remaining 37 molecules are assigned to a wide variety of inhibitor classes, but they all have one structural feature in common, at least two six-membered rings.

Overall, the different descriptor sets resulted in comparable MDDR clustering, in that, for all descriptor sets on the left and right sides, mainly three large clusters appeared containing the glucocorticoid receptor ligands. Figure 5B shows, as an example, the results obtained with the MDDRCATS08 descriptor set.

The structures shown in Figure 5B are representatives of the main emerging clusters (three on each side of the cluster tree viewed from the root). Cluster "importance" was rated and selected according to the EF of either lead class I, II, or

Chart 1



III (Figure 5A) and not according to the EF of glucocorticoid receptor ligands in total. The EFs in the three Class I clusters (Figure 5B, left tree half) were 2818, 2817, and 2192, whereas the EFs in the two class II clusters (Figure 5B, right tree half, Clusters B and C) were 493 and 401, and the EF of the class III cluster was 5917 (Figure 5B, right tree half, Cluster A). The shown structures are grouped according to their relationship in the tree, in that, in both tree halves, clusters B and C lie in closer proximity to each other than they do to cluster A. The three different lead classes are separated. Class I is present only on the left side in the three clusters, indicated by yellow-colored substructures. The green color shows the unique substructures across all of the clusters, and blue fragments represent a common substructure of the related clusters B and C on the left tree half of the tree. In clusters B and C, from the right tree half of the tree, lead class II shows the classical steroid backbone, whereas lead class III dominates cluster A (Figure 5). Despite the separation of the three lead classes across the clusters, differently substituted core structures were recognized and grouped.

Application Example of the Hierarchical k-Means Approach. To further evaluate the clustering algorithm, we chose one application that might be of immediate practical benefit. One can use a tree that was generated with reference data (e.g., MDDR or COBRA) for predicting the potential activity of molecules by projecting them onto the tree and analyzing the distribution of target classes in the activated clusters. This can also be performed simultaneously by reading-in compounds for which the pharmacological activity is unknown together with reference compounds. It provides the user with a quick and easy way to find new compounds that are putatively active. We performed one such study to illustrate the idea. The combined data set of MDDR, COBRA, and the SPECS catalog was used to build up a binary tree employing the hierarchical *k*-means approach ($k = 2$ and $\Theta = 2.6$). Then, all known caspase-1 inhibitors from COBRA (39) were marked, and the clusters containing these molecules were screened for co-located MDDR caspase-1 inhibitors (188). A challenging exercise is to look for “scaffold hops” within a cluster. One such pair is given by structures **1** (COBRA)[49] and **2** (MDDR),[50] which were grouped together (see Chart 1). Both are known caspase-1 inhibitors with different scaffolds. In this particular cluster, only four molecules were co-located (one from COBRA and three from MDDR), but all of them are known cysteine protease inhibitors. Compounds **3**[51] and **4**[52] represent the other two molecules from MDDR (Chart 1). They are both cathepsin L inhibitors[53] and share the peptide-like backbone part with the caspase-1 inhibitors. This example demonstrates a possible use of the hierarchical *k*-means approach for constructing focused screening libraries.

In summary, all three clustering examples demonstrated that a meaningful automatic grouping of chemical structures by the hierarchical *k*-means approach is feasible for large data sets. The hierarchical nature of the cluster relationships might provide a possibility to find SARs from the different clusters if activity data are present.

Technical Aspects. An advantage of the algorithm is that it combines the low computation time of *k*-means[17] with a hierarchical relationship between the data entries. Because such a tree does not necessarily have to be balanced, the worst case run time and space requirement would be $O(n^2)$. However, creating a very unbalanced tree is only possible if no structure in the data would exist, which is very unlikely. The best case run time and space requirement is $O(n \log n)$, which is also the expected average case. Reading, clustering, and visualizing a combinatorial library of 500 000 compounds described in 68 dimensions with the MOE 2D descriptor set was done in less than 13 min on a 2 GHz Xeon Processor with 3 GB RAM ($k = 2$; $\Theta = 0$).

One disadvantage of nonhierarchical clustering algorithms, either to form large and heterogeneous or small and exclusive clusters, was removed in creating the hierarchy. Here, the user has a choice to decide at which level to cut the tree. However, other specific problems of the *k*-means approach remain: because *k*-means randomly chooses the initial cluster centers, different clustering schemes may occur using the same data sets. Another problem is that *k*-means can only deal with hyperspherical clusters,[17] which means that data points located at interface regions between two hyperspheres may be torn apart. If $k = 2$ is chosen for a large data set, this problem occurs on each split point in the tree and can result in a multitude of wrongly clustered compounds. Applying fuzzy class assignment rules might be one possibility to address this issue.[15]

We proposed Θ as a threshold for splitting the tree. It bears the disadvantage that it has to be recalculated for each data set. The potential benefit is that each individual data set might be more appropriately clustered compared to those of the conventional *k*-means approach.

One feature of the clustering approach is visualization of and navigating in the resulting tree. By the usage of specific colors for the different class labels, it can be quickly determined where sub trees are present that contain these labeled entries. With the option to zoom into sub trees, this visualization is possible for deeper tree levels, where usually a multitude of nodes exists.

CONCLUSION

We presented a hierarchical clustering algorithm, hierarchical *k*-means, which can handle very large data sets in high-dimensional space. At each point in the tree, a user-defined number of descendent nodes (i.e., clusters) are created according to the *k*-means algorithm. Clustering stops if a maximum Euclidian distance to the virtual mean does not exceed a user-defined threshold. To guide the user in choosing this threshold, we proposed a value, where a maximum in the sum of the relative distances to the virtual means is achieved.

According to two shown retrospective examples (ICE inhibitors in COBRA and glucocorticoid receptor ligands in MDDR), suitable clustering was achieved, projecting the structural relationship of the compounds onto the hierarchical relationship of the emerging clusters. Clustering a combined data set, consisting of COBRA, MDDR, and the SPECS catalog, it was possible to retrieve MDDR ICE inhibitors with new scaffolds using COBRA ICE inhibitors as seeds. This demonstrates the ability of hierarchical *k*-means to construct focused screening collections. The approach might also be feasible for the analysis of large compilations of peptide binding data obtained in chemical genomics approaches.[54,55]

In the context of HTS data analysis, we judge the method to be useful in deriving fuzzy SARs already on the level of primary screening results. Identifying clusters, significantly enriched with primary hits, provides rules to detect putative false negative hits, compounds that are active but are missed in the screening experiment. Not losing these compounds guarantees a broader and more reliable knowledge base for therapeutic projects.

Abbreviations. 2D, two dimensional; 3D, three dimensional; COBRA, Collection of Bioactive Reference Analogues; CPU, central processing unit; GPCR, G-protein coupled receptor; HTS, high-throughput screening; IL1, interleukin 1; ICE, interleukin 1 cleaving enzyme; MDDR, MDL Drug Data Report; MOE, Molecular Operating Environment; PC, principle component; PPP, potential pharmacophore points; RAM, random access memory; SAR, structure−activity relationship; SE, Shannon entropy; sSE, standardized Shannon entropy; SVL, Support Vector Language; UFS, unsupervised forward selection; VLA-4, Very Late Antigen 4.

ACKNOWLEDGMENT

Dr. G. M. Downs is thanked for helpful discussions and pointing out alternative implementations of the *k*-means approach. This research was supported by the Beilstein Institut zur Förderung der Chemischen Wissenschaften and the Fonds der Chemischen Industrie (FCI).

REFERENCES AND NOTES

(1) Schneider, G.; Böhm, H. J. Virtual Screening and Fast Automated Docking Methods. *Drug Discovery Today* **2002**, *7*, 64−702.

(2) Bajorath, J. Integration of Virtual and High-Throughput Screening. *Nat. Rev. Drug Discovery* **2002**, *1*, 882−894.

(3) Valler, M. J.; Green, D. Diversity Screening Versus Focused Screening in Drug Discovery. *Drug Discovery Today* **2000**, *5*, 286−293.

(4) Croston, G. E. Functional Cell-Based uHTS in Chemical Genomic Drug Discovery. *Trends Biotechnol.* **2002**, *20*, 110−115.

(5) Walters, W. P.; Namchuk, M. Designing Screens: How to Make Your Hits a Hit. *Nat. Rev. Drug Discovery* **2003**, *2*, 259−266.

(6) Bleicher, K. H.; Böhm, H. J.; Müller, K.; Alanine, A. I. Hit and Lead Generation: Beyond High-Throughput Screening. *Nat. Rev. Drug Discovery* **2003**, *2*, 369−378.

(7) Böcker, A.; Schneider, G.; Teckentrup, A. Status of HTS Data Mining Approaches. *QSAR Comb. Sci.* **2004**, *23*, 207−213.

(8) Brown, R. D.; Martin, Y. C. Use of Structure−Activity Data to Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.

(9) Ward, J. H. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.* **1963**, *58*, 236−244.

(10) Doman, T. N.; Cibulskis, J. M.; Cibulskis, M. J.; McCray, P. D.; Spangler, D. P. Algorithm5: A Technique for Fuzzy Similarity Clustering of Chemical Inventories. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1195−1204.

(11) Jarvis, R. A.; Patrick, E. A. Clustering Using a Similarity Measure Based on Shared Nearest Neighbors. *IEEE Trans. Comput.* **1973**, *22*, 1025−1034.

(12) Menard, P. R.; Lewis, R. A.; Mason, J. S. Rational Screening Set Design and Compound Selection: Cascaded Clustering. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 497−505.

(13) Willett, P.; Winterman, V.; Bawden, D. Implementation of Nonhierarchic Cluster Analysis Methods in Chemical Information Systems: Selection of Compounds for Biological Testing and Clustering of Substructure Search Output. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 109−118.

(14) Duda, R. O.; Hart, P. E.; Stork, D. G. *Pattern Classification*; John Wiley & Sons: New York, 2000.

(15) Holliday, J. D.; Rodgers, S. L.; Willett, P.; Chen, M.; Mahfouf, M.; Lawson, K.; Mullier, G. Clustering Files of Chemical Structures Using the Fuzzy *k*-Means Clustering Method. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 894−902.

(16) Otto, M. *Chemometrics. Statistics and Computer Application in Analytical Chemistry*; Wiley-VCH: Weinheim, Germany, 1998.

(17) Jain, A. K.; Murty, M. N.; Flynn, P. J. Data Clustering: A Review. *ACM Comput. Surveys* **1999**, *31*, 265−323.

(18) Nicolaou, C. A.; Tamura, S. Y.; Kelley, B. P.; Bassett, S. I.; Nutt, R. F. Analysis of Large Screening Data Sets Via Adaptively Grown Phylogenetic-Like Trees. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1069−1079.

(19) Richon, A. LeadScope: Data Visualization for Large Volumes of Chemical and Biological Screening Data. *J. Mol. Graphics Modell.* **2000**, *18*, 76−79.

(20) Roberts, G.; Myatt, G. J.; Johnson, W. P.; Cross, K. P.; Blower, P. E., Jr. LeadScope: Software for Exploring Large Sets of Screening Data. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1302−1314.

(21) Steinbach, M.; Karypis, G.; Kumar, V. *A Comparison of Document Clustering Techniques*. Technical Report 00-034; Department Computer Science & Engineering: University of Minnesota, 2000.

(22) (a) Barnard, J. M.; Downs, G. M.; Wild, D. J.; Wright, P. M. Better Clusters Faster. *Third Joint Sheffield Conference on Chemoinformatics*, 2004. (b) Downs, G. M.; Barnard, J. M. *Clustering Methods and Their Uses in Computational Chemistry*, vol. 18; Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley-VCH: Weinheim, Germany, 2002; pp 1−40. (c) Sultan, M.; Wigle, D. A.; Cumbaa, C. A.; Maziarz, M.; Glasgow, J.; Tsao, M. S.; Jurisica, I. Binary Tree-Structured Vector Quantization Approach to Clustering and Visualizing Microarray Data. *Bioinformatics* **2002**, *18*, 111−119.

(23) MDL Drug Data Report; Elsevier MDL: San Leandro, CA. http://www.mdl.com.

(24) Schneider, P.; Schneider, G. Collection of Bioactive Reference Compounds for Focused Library Design. *QSAR Comb. Sci.* **2003**, *22*, 713−718.

(25) Fisher, R. A. The Use of Multiple Measurements in Axonomic Problems. *Ann. Eugenics* **1936**, *7*, 179−188.

(26) SPECS, Delft, The Netherlands. http://www.specs.net/.

(27) InforSense Ltd., London, U.K. http://www.inforsense.com.

(28) SciTegic, San Diego, CA. http://www.scitegic.com.

(29) Hann, M.; Hudson, B.; Lewell, X.; Lifely, R.; Miller, L.; Ramsden, N. Strategic Pooling of Compounds for High-Throughput Screening. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 897−902.

(30) Muegge, I. Selection Criteria for Drug-Like Compounds. *Med. Res. Rev.* **2003**, *23*, 302−321.

(31) Chemical Computing Group (CCG). http://www.chemcomp.com.

(32) OpenEye Scientific Software, Santa Fe, NM. http://www.eyesopen.com.

(33) Fechner, U.; Franke, L.; Renner, S.; Schneider, P.; Schneider, G. Comparison of Correlation Vector Methods for Ligand-based Similarity Searching. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 687−698.

(34) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. Scaffold-Hopping by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chem., Int. Ed.* **1999**, *38*, 2894−2896.

(35) Shannon, C. E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379−423.

(36) Whitley, D. C.; Ford, M. G.; Livingstone, D. J. Unsupervised Forward Selection: A Method for Eliminating Redundant Variables. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1160−1168.

(37) Godden, J. W.; Bajorath, J. Shannon Entropy−A Novel Concept in Molecular Descriptor and Diversity Analysis. *J. Mol. Graphics Modell.* **2000**, *18*, 73−76.

(38) Godden, J. W.; Bajorath, J. Differential Shannon Entropy As a Sensitive Measure of Differences in Database Variability of Molecular Descriptors. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1060−1066.

(39) Godden, J. W.; Stahura, F. L.; Bajorath, J. Variability of Molecular Descriptors in Compound Databases Revealed by Shannon Entropy Calculations. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 796−800.

(40) Braddock, M.; Quinn, A. Targeting IL-1 in Inflammatory Disease: New Opportunities for Therapeutic Intervention. *Nat. Rev. Drug Discovery* **2004**, *3*, 330−340.

(41) Norman, A. W.; Mizwicki, M. T.; Norman, D. P. Steroid-Hormone Rapid Actions, Membrane Receptors and a Conformational Ensemble Model. *Nat. Rev. Drug Discovery* **2004**, *3*, 27−41.

(42) Talanian, R. V.; Brady, K. D.; Cryns, V. L. Caspases as Targets for Anti-inflammatory and Anti-Apoptotic Drug Discovery. *J. Med. Chem.* **2000**, *43*, 3351−3371.

(43) Brody, T. M.; Larner, J.; Minneman, K. P. *Human Pharmacology. Molecular to Clinical*; Mosby: St. Louis, MO, 1998.

(44) Baker, A. H.; Edwards, D. R.; Murphy, G. Metalloproteinase Inhibitors: Biological Actions and Therapeutic Opportunities. *J. Cell Sci.* **2002**, *115*, 3719−3727.

(45) Johnson, T. O.; Hua, Y.; Luu, H. T.; Brown, E. L.; Chan, F.; Chu, S. S.; Dragovich, P. S.; Eastman, B. W.; Ferre, R. A.; Fuhrman, S. A.; Hendrickson, T. F.; Maldonado, F. C.; Matthews, D. A.; Meador, J. W., III.; Patrick, A. K.; Reich, S. H.; Skalitzky, D. J.; Worland, S. T.; Yang, M.; Zalman, L. S. Structure-Based Design of a Parallel Synthetic Array Directed Toward the Discovery of Irreversible Inhibitors of Human Rhinovirus 3C Protease. *J. Med. Chem.* **2002**, *45*, 2016−2023.

(46) Goudreau, N.; Cameron, D. R.; Bonneau, P.; Gorys, V.; Plouffe, C.; Poirier, M.; Lamarre, D.; Llinas-Brunet, M. NMR Structural Characterization of Peptide Inhibitors Bound to the Hepatitis C Virus NS3 Protease: Design of a New P2 Substituent. *J. Med. Chem.* **2004**, *47*, 123−132.

(47) Böhm, H. J.; Klebe, G.; Kubinyi, H. *Wirkstoffdesign*; Spektrum Akademischer Verlag: Heidelberg, Germany, 2002.

(48) Lin, L. S.; Lanza, T. J.; Castonguay, L. A.; Kamenecka, T.; McCauley, E.; Van Riper, G.; Egger, L. A.; Mumford, R. A.; Tong, X.; MacCoss, M.; Schmidt, J. A.; Hagmann, W. K. Bioisosteric Replacement of Anilide with Benzoxazole: Potent and Orally Bioavailable Antagonists of VLA-4. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 2331−2334.

(49) Edwards, P. Combinatorial Chemistry. *Drug Discovery Today* **2003**, *8*, 326−327.

(50) Hagmann, W. K.; MacCoss, M.; Mjalli, A. M.; Zhao, J. J. Substd. ketone derivs. as inhibitors of interleukin 1-beta-converting enzyme for treatment of inflammation in, e.g., lung, central nervous system, kidney, joints, endocardium, pericardium, eyes, ears, skin, gastrointestinal tract and urogenital system. [WO 9505192], 1994.

(51) Cameron, A.; Guo, D.; Kaleta, J.; Menard, R.; Micetich, R. G.; Purisima, E.; Zhou, N. E. 6-substituted amino 4-oxa 1-aza-bicyclo (3.2.0) heptan-7-one derivatives which inhibit cysteine protease are used to treat muscular dystrophy, arthritis, myocardial infarction, Alzheimer’s disease, bacterial infection, common cold, osteoporosis and cancer metastasis. [WO 9738008], 1997.

(52) Guo, D.; Micetich, R. G.; Singh, R.; Zhou, N. E.; Zhou, N. New substituted amino bicyclic-beta-lactam penam and cepham derivatives are inhibitors of cysteine protease and are useful in treatment of cancer, rheumatoid arthritis, osteoporosis and muscular dystrophy. [US 6232305] 2001.

(53) Turk, B.; Turk, D.; Turk, V. Lysosomal Cysteine Proteases: More Than Scavengers. *Biochim. Biophys. Acta* **2000**, *1477*, 98−111.

(54) Mitopoulos, G.; Walsh, D. P.; Chang, Y. T. Tagged Library Approach to Chemical Genomics and Proteomics. *Curr. Opin. Chem. Biol.* **2004**, *8*, 26−32.

(55) Frank, R. High-Density Synthetic Peptide Microarrays: Emerging Tools for Functional Genomics and Proteomics. *Comb. Chem. High Throughput Screening* **2002**, *5*, 429−440.

CI0500029