ARTICLES

Toward an Improved Clustering of Large Data Sets Using Maximum Common Substructures and Topological Fingerprints

Alexander Böcker*

Boehringer Ingelheim (Canada) Ltd. Research & Development, 2100 Cunard Street, Laval, Quebec, Canada H7S 2G5

Received March 13, 2008

A new clustering algorithm was developed that is able to group large data sets with more than 100,000 molecules according to their chemotypes. The algorithm preclusters a data set using a fingerprint version of the hierarchical *k*-means algorithm. Chemotypes are extracted from the terminal clusters *via* a maximum common substructure approach. Molecules forming a chemotype have to share a predefined number of rings, atoms, and non-carbon heavy atoms. In an iterative procedure, similar chemotypes and singletons are fused to larger chemotypes. Singletons that cannot be assigned to any chemotype are then grouped based on the proportion of overlap between the molecules. Representatives from each chemotype and the singletons are used in a second round of the hierarchical *k*-means algorithm to provide a final hierarchical grouping. Results are reported to an interactive graphical user interface which allows initial insights about the structure activity relationship (SAR) of the molecules. Example applications are shown for two chemotypes of reverse transcriptase inhibitors in the MDDR database and for the evaluation of descriptor-based similarity searching routines. A special focus was laid on the chemotype hopping potential of each individual routine. The algorithm will allow the analysis of high-throughput and virtual screening results with improved quality.

INTRODUCTION

Clustering methods belong to the family of unsupervised classification techniques, grouping data according to inherent properties only.[1] These methods are widely used in the pharmaceutical industry to cluster molecule data sets in order to decipher chemotypes present in the data.[2] These chemotype definitions may then be used to define lead series resulting from high-throughput screening (HTS),[2,3] to compare two compound repositories[4] or to define a representative subset by selecting a member from each chemotype.[5−7] Numerous descriptor-based and fingerprint-based clustering methods exist[8] and are usually separated into hierarchical methods like Ward's clustering,[9,10] and nonhierarchical methods like Jarvis-Patrick,[11] *k*-means,[1] self-organizing maps,[12] or Bayesian unsupervised clustering.[8] Depending on the descriptor or fingerprint set used, the selected similarity/distance threshold of a given metric and the clustering algorithm, two general limitations may occur: (i) chemotypes can be oversplit or (ii) clusters may contain different chemotypes. This is exemplified in Figure 1 showing five MDDR[13] human immunodeficiency virus (HIV) protease inhibitors (Identifiers (IDs): 263164, 263404, 162809, 162811, and 172410) and a gpIIb/IIIa receptor antagonist (ID: 228151). Clusters 1−3 were obtained using an in-house implementation of the hierarchical *k*-means algorithm[14,15] in combination with the Chemaxon chemical fingerprints[16] and a maximum Tanimoto dissimilarity coefficient of 0.85 (*vide infra*), while the lower section of Figure 1 shows the assumed optimum outcome of an ideal clustering (clusters 4 and 5). Intuitively the singleton in cluster 2 belongs to the same chemotype as the compounds in clusters 1 and 3; this chemotype is oversplit, creating an artificial singleton. Moreover, the molecules in clusters 1 and 3 also belong to the same chemotype (with the exception of molecule 228151 in cluster 1); again, oversplitting of the chemotype occurred. Cluster 1 is a typical example of a heterogeneous cluster containing molecules from two different chemotypes.

Several approaches have been proposed in the literature to circumvent these shortcomings and to attain optimum clustering (as in clusters 4 and 5).[17−20] One method is to add singletons into existing clusters in a "fuzzy" way.[17] Another is to collect singletons and recluster those using milder boundary conditions.[18] More recently, approaches based on maximum common substructures (MCS) have been introduced.[19−21] Stahl and Mauser calculated maximum common substructures for the clusters obtained by an exclusion sphere clustering algorithm. By mapping singletons onto the MCS, it was possible to substantially reduce the number of "artificial" singletons and improve the clustering homogeneity.[20] Gardiner et al. also combined the exclusion sphere algorithm with MCS calculations. By using an iterative MCS approach in combination with a graph-based similarity threshold (RASCAL score), they defined and separated the chemotypes in a cluster and reduced cluster heterogeneity.[21] The application of such MCS approaches can also be combined with R-group decomposition, providing

* Corresponding author phone: +(450) 682-4641 ext. 4505; e-mail: Alexander.bocker@boehringer-ingelheim.com.

10.1021/ci8000887 CCC: $40.75
Published on Web 10/29/2008

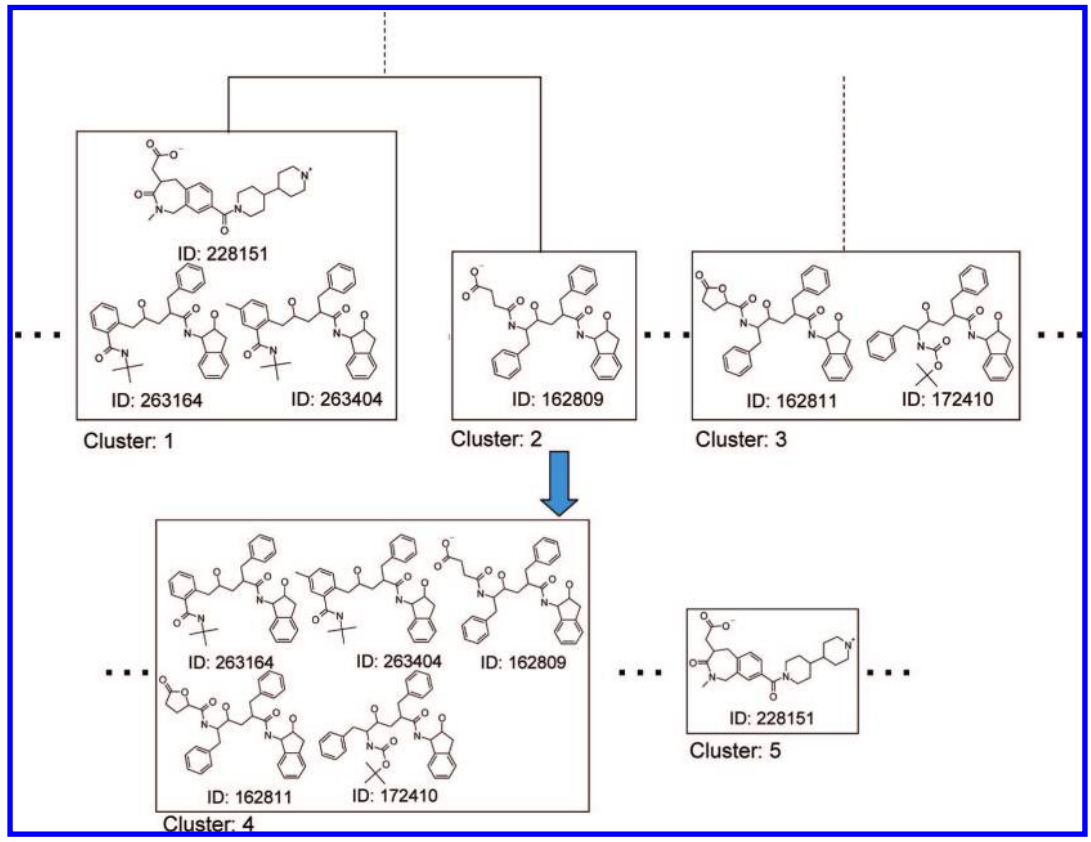

Figure 1. Clustering example of five HIV protease inhibitors (263164, 263404, 162809, 162811, and 172410) and a gpIIb/IIIa receptor antagonist (228151) from the MDDR database. Clusters 1−3 were obtained using a fingerprint version of the hierarchical *k*-means algorithm (*vide infra*). A Tanimoto coefficient of 0.85 was applied as a similarity threshold. Clusters 4 and 5 show the hypothetical optimum clustering of the molecules.

additional information on the structure activity relationship (SAR) of a chemotype for a given target.[21]

A limitation of MCS-based clustering approaches is that these calculations are computationally expensive and are thus not directly applicable to large data sets. Preclustering of large data sets using a fingerprint or descriptor-based algorithm may be helpful in such cases. In this work a new, high-performance clustering algorithm is proposed that performs well on large data sets of more than 100,000 molecules while addressing oversplitting of chemotypes and cluster heterogeneity. The algorithm is similar to the approach proposed by Stahl and Mauser in that the data set is preclustered using a fingerprint-based clustering algorithm (here a fingerprint based version of the hierarchical *k*-means algorithm[14,15] is applied), the obtained terminal clusters are used for calculating MCS,[20,21] and the oversplitting of a chemotype is minimized by using MCS calculations.[20] However different in this study is that (i) the cluster heterogeneity and the oversplitting of a chemotype is addressed by defining a chemotype based on parameters and rules which were established in close cooperation with the hit-to-lead group at Boehringer Ingelheim (Canada) Ltd., (ii) the minimization of the oversplitting of a chemotype is performed by equally considering singletons and clusters as a chemotype, and (iii) the representatives of each MCS-based cluster are finally clustered with the hierarchical *k*-means algorithm to provide a hierarchical grouping of the resulting chemotypes. This algorithm can rearrange clusters 1−3 in Figure 1 into clusters 4 and 5. The outcome of the algorithm is exemplified using HIV reverse transcriptase (RT) inhibitors in the MDDR. An application of immediate benefit is also shown in the context of evaluating descriptor-based similarity searching routines.

METHODS

Data Set. The MDDR database (December 2006 version) was used in this work.[13] It contains 166,698 biologically relevant molecules from patent literature, scientific journals, and meeting reports. Each entry provides a 2D molecular structure field, an activity class field, and a corresponding activity class index (a molecule can be assigned to multiple activity classes). The MDDR database was prepared using an in-house Web interface implemented in Pipeline Pilot (version 6)[22] at Boehringer Ingelheim (Canada) Ltd. This Web interface removes (1) entries lacking structural information, (2) entries where the molecule contains non-organic-like atoms, and (3) entries with molecular weight below 150 dalton. Counter ions are removed, and the molecules are ionized using the MOE wash routine.[23] Descriptors for topological polar surface area (TPSA), rotational bonds, molecular weight, number of hydrogen bond donors and acceptors, and SlogP are calculated in MOE.[23] Violations occur if the molecular weight is ≥500, TPSA ≥ 140, SlogP ≥ 5 or ≤ 1, number of hydrogen bond donors ≥ 5, number of hydrogen bond acceptors ≥10, or number of rotational bonds ≥ 10. If more than two out of six properties are violated, the molecule is flagged. Finally, an in-house list of unwanted substructures is mapped onto each molecule. If a substructure is detected, the molecule is flagged. At the end, all such flagged molecules were removed leaving a total of 118,691 entries for the analysis.

209 2D descriptors were calculated for these molecules using MOE.[23] These descriptors comprise the MOE 2D[23] and a few proprietary in-house descriptors. The in-house descriptors range from different ring counts over substructure, atom, or connectivity counts to molecule properties like volume or density. They are mostly complementary to the MOE 2D descriptors. As described in a recent publication by Whitley and co-workers "relevant" descriptors were selected having a standard deviation >0.0005, and nonredundant descriptors were identified using unsupervised forward selection (UFS).[24] The maximum squared multiple correlation coefficient was left at the default value of 0.99. The final set comprised 94 2D descriptors. From this set, two different descriptor representations were generated. In the first representation (MDDR_MOE2D_NORM) all descriptors were normalized to the range of 0 and 1. In the second representation (MDDR_MOE2D_SCALED) all descriptors were centered to the mean and scaled to unit variance.

The Chemaxon chemical fingerprints were calculated for all molecules. These fingerprints are of topological nature and consider all linear and cyclic paths in a molecule up to a given bond length. Both atom and bond types are considered.[16] For this study the bit length was defined as 1024, and the number of bonds under consideration was set to 7. All other parameters were left at default. Throughout this publication these fingerprints are referred to as topological fingerprints.

The Algorithm. The MCS-based clustering algorithm has been designed to cluster large data sets with more than 100,000 data points and to provide the analyst with the different chemotypes present in the data set. Finding chemotypes in a data set first requires the definition of a chemotype. Different parameters have been investigated that may define a chemotype. Those can be set by the user and have to be matched by a MCS: The minimum "distance" Φ between the topological fingerprints[16] of the molecules of a chemotype and the fingerprints of the MCS (measured as Tanimoto coefficient, see eq 2), the number of rings Γ (smallest set of smallest rings), the number of non-carbon heavy atoms Ψ, the number of heavy atoms Π, and the percentage Ω of heavy atoms which are present in a MCS compared to the heavy atoms of the individual molecules belonging to the chemotype. Suitable values were established for these parameters with the hit-to-lead group at Boehringer Ingelheim (Canada) Ltd. using several in-house hit to lead projects. According to this, Φ was set to ≥ 0.7, Γ was defined as ≥ 3, and Ψ was set to ≥ 2. For Π two alternative definitions were used ($\Pi_1 \geq 13$ and $\Pi_2 \geq 25$) and Ω was defined as 80%. These parameters are applied in combination or alone at different stages of the algorithm. The algorithm can be subdivided into five stages: (1) fingerprint based preclustering of the data set, (2) formation of MCS-based clusters from the preclustering, (3) fusion of the MCS-based clusters and singletons, (4) processing of the remaining singletons, and (5) fingerprint based hierarchical postclustering of the identified chemotypes. These stages are demonstrated in the center and on the left side of Figure 2 as a cartoon and text representation, respectively. Additionally, on the right side, the intended effect of each stage on the homogeneity of the clustering results is outlined in orange. The cartoon representation is illustrated according to five different classes (light blue circle, green triangle, yellow square, red x, and blue plus). The hierarchical relationship between the entries is—if present—schematically shown. In contrast to singletons, terminal clusters ($N > 1$ entry) are indicated by a box. The parameters and rules which are used to define a chemotype are additionally mentioned within the arrows.

(1) Preclustering of the Data Set. The calculation of a MCS between two molecules belongs to the class of NP-complete problems (*i.e.* no nondeterministic solution can be found in polynomial time). Consequently, for the clustering of large molecule data sets based on MCS, the number of MCS calculations should be as low as possible. To achieve this, the data set is preclustered using a fingerprint-based version of our recently published hierarchical *k*-means algorithm.[14,15] Only the molecules of the obtained clusters are then used for the MCS calculation. The preclustering algorithm is based on the *k*-means algorithm:[1]

Step 1: Select k data points randomly as initial cluster centroids.

Step 2: Form k clusters by assigning each data point A to its nearest centroid B.

Step 3: Calculate new virtual centroids for each cluster.

Step 4: Iterate the second and third steps until a predefined number of iterations is reached or the clusters do not change anymore.

For the application of the *k*-means algorithm in combination with fingerprints, the bits of entry A and the centroid B are converted into double variables. The nearest data point is then determined using the numerical version of the Tanimoto coefficient C (eq 1)

$$C(A,B) = \frac{\sum_{i=1}^{n} x_{i,A} \cdot x_{i,B}}{\sum_{i=1}^{n} (x_{i,A})^2 + \sum_{i=1}^{n} (x_{i,B})^2 - \sum_{i=1}^{n} x_{i,A} \cdot x_{i,B}} \quad (1)$$

with $x_{i,A}$ representing the value of position i of molecule A and $x_{i,B}$ representing the value of bit position i of molecule B.[25] It should be noted that in B the positions i are no longer bits but averages.

To convert the algorithm into a hierarchical clustering the following steps are applied:

Step 0: Define $k = 2$ to obtain a binary dendrogram. Specify a similarity coefficient Θ. In the present study, a conservative Tanimoto coefficient of 0.85 was employed to define similarity.[26]

Step 1: Perform data clustering employing the *k*-means algorithm. k child clusters are created, and the data set is partitioned according to the *k*-means algorithm.

Step 2: Check for each cluster: If the maximum Tanimoto coefficient between the data points is below the threshold Θ, repeat Step 1 for this cluster. Otherwise terminate.

The hierarchical *k*-means algorithm has its limitations in its classification. Singletons (terminal cluster of size 1) might exist that can be assigned to a terminal cluster with more than one entry. To cope with these cases for each singleton, the Tanimoto coefficient (eq 1) is calculated to the centroids of the terminal clusters with more than one entry. If Θ is exceeded, the singleton is added to the cluster. It should be noted that preclustering (and singleton assignment) can also be performed with any other clustering algorithm and any

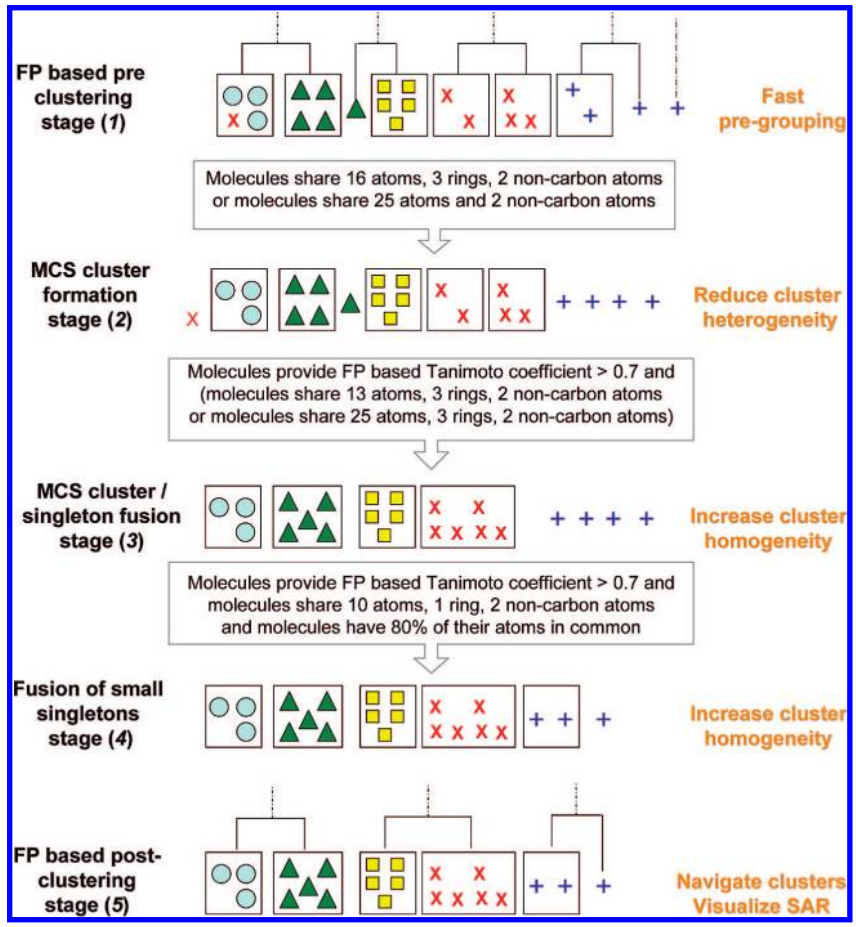

Figure 2. Schematic representation of the different stages of the combined fingerprint (FP) based and MCS based clustering algorithm. On the right side, the intended effect of the different stages on the cluster homogeneity is demonstrated. In the center this effect is illustrated as a cartoon display with five different classes (light blue circle, green triangle, yellow square, red x, and blue plus). The hierarchical relationship between the entries is schematically outlined. In contrast to singletons, terminal clusters ($N > 1$ entry) are indicated by a box. The parameters and rules which are used to define a chemotype are additionally mentioned within the arrows.

other similarity metric or coefficient which is able to manage large data sets.[19,21] For this study we decided to use the Tanimoto coefficient since the used topological fingerprints[16] are equivalent to the Daylight fingerprints,[27] and they were shown to work well in combination with the Tanimoto coefficient.[20,26]

(2) Creation of MCS-Based Clusters. The main goal in creating a MCS-based cluster (termed a MCS cluster) out of a fingerprint-based cluster (termed a FP cluster) consists of adding only those molecules to the MCS cluster that belong to the same chemotype. This is schematically outlined in Figure 2 for the cluster with the three light-blue circles and the red x. A second goal is to minimize the number of time-consuming MCS calculations. Two rules were established to define a chemotype with the hit-to-lead group at Boehringer Ingelheim (Canada) Ltd.: (i) A chemotype contains Γ rings, Ψ non-carbon heavy atoms, and $\Pi_1 + 3$ heavy atoms. (ii) A chemotype contains Π_2 heavy atoms and Ψ non-carbon heavy atoms. An example of the effect of rule (i) can be seen in Figure 3 for the dihydroquinazolinones 205091 and 197508. Both molecules are HIV RT inhibitors. The calculated MCS in Figure 3 represents a chemotype since it contains 3 rings (the dihydroquinazoline 2-membered ring and the cyclopropane), 16 heavy atoms, and 4 non-carbon heavy atoms. For the calculation only FP clusters with more than 1 entry are considered. FP clusters with only 1 entry are directly transformed into MCS clusters of size 1. The following steps are executed for the molecules of a cluster.

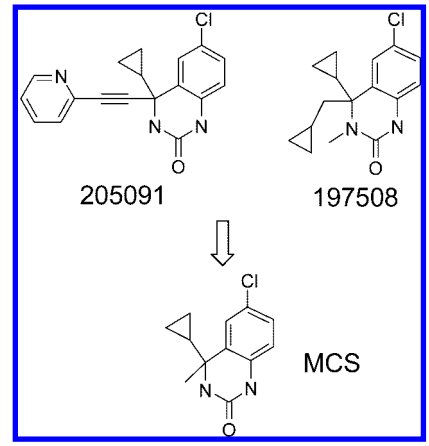

Figure 3. Two HIV RT inhibitors, 205191 and 197508, and their calculated MCS. The MCS represents a chemotype according to rule (i) in stage 2 and 3 of the algorithm since it contains 3 rings, 4 non-carbon heavy atoms, and 16 heavy atoms.

Step 0: Test rule (i) or (ii) for each molecule n. All violating molecules are transformed to MCS clusters of size 1 and are removed from consideration at this stage of the algorithm. This step removes all molecules from the MCS calculation that cannot form a chemotype.

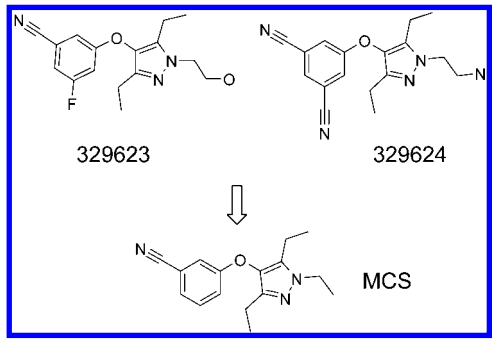

Figure 4. Two HIV RT inhibitors, 329623 and 329624, and their calculated MCS. The MCS represents a chemotype according to the definitions in stage 4 of the algorithm since it contains 2 rings, 4 non-carbon heavy atoms, and 20 heavy atoms and the MCS and both 329623 and 329624 have more than 80% of their atoms in common.

Step 1: Calculate Tanimoto coefficients between all pairs of molecules in the cluster (eq 2, *vide infra*) and determine nearest neighbor pair *i* and *j*.

Step 2: Calculate MCS between *i* and *j*. For the MCS calculation the maximum common edge substructure algorithm was used which is implemented in Chemaxon's JChem java package.[16] The MCS forms a new representative molecule *c*.

Step 3: Test *c*. If *c* violates both rules (i) and (ii) all *n* molecules are transformed to MCS clusters of size 1, and the next FP cluster is analyzed.

Step 4: Randomly select next molecule *i*.

Step 5: Map *c* onto *i* using the substructure search algorithm implemented in Chemaxon's JChem Java package. This algorithm performs a topological fingerprint based prescreening, followed by an atom-by-atom comparison, whereas both atom types and bond types are considered.[16] If *c* maps onto *i,* add *i* to the MCS cluster and continue with step 4. This (fast) premapping routine reduces the cost-intensive MCS calculations.

Step 6: Calculate MCS between *i* and *c*. The MCS forms a new molecule *t*. If *t* violates both rules (i) and (ii), *i* is transformed to MCS cluster of size 1. Otherwise *i* is added to the MCS cluster. *t* forms the new representative *c* of the chemotype. Continue with step 4.

(3) Fusion of MCS-Based Clusters. The result of stage (2) of the MCS algorithm is a list *m* of MCS clusters (including singletons). In *m*, several clusters or singletons might exist where the entries belong to the same chemotype (i.e. oversplitting of a chemotype). Examples are shown in Figure 2 for the singleton with red x and the two clusters with red x or the singleton with green triangle and the cluster with green triangle. To address this oversplitting the same rules are applied to define a chemotype as in stage (2). The only exception is that for rule (i) Π_1 heavy atoms are now required. It should be noted that a large proportion of the molecules in the data set are presumably represented by a few MCS. Still unnecessary and time-consuming MCS calculations have to be avoided to be applicable to large data sets. For this, topological fingerprints are generated for all MCS in *m*. For MCS clusters with one entry, the molecule is duplicated, and one copy is considered as the MCS. In the following only MCS pairs are considered whose Tanimoto coefficient exceeds Φ. Φ is calculated according to eq 2

$$\Phi(A,B) = \frac{N_{A\&B}}{N_A + N_B - N_{A\&B}} \quad (2)$$

where N_A and N_B are the number of bits set in the bit strings of MCS *A* and *B*, respectively, and $N_{A\&B}$ is the number of bits that are common to both.[25] The reason for selecting the bit version of the Tanimoto coefficient instead of the numerical version (eq 1) lies in the fact that it is faster to execute since only bit shift operations are used. For the hierarchical *k*-means this version of the Tanimoto coefficient is not applicable since the calculated centroids are no longer fingerprints with bits. Instead each position represents an average over the bits set or not set at the corresponding position in the fingerprint of the entries of a cluster. For the fusion step, the most logical approach would be to analyze the closest MCS pairs first. To save calculation time, the algorithm does not directly determine the closest neighbor pair. Instead, Φ is set to 0.85 and systematically lowered to its original (0.7). For each Φ the MCS list is screened for entries whose Tanimoto coefficient exceeds Φ. This allows performing this screening step in $O(N^2)$ instead of $O(N^3)$, where N corresponds to the number of MCS. The algorithm performs the following steps:

Step 0: Set Φ to 0.85, calculate fingerprints for all MCS in *m*.

Step 1: Select a MCS *i* from *m*.

Step 2: Select next MCS *j* from *m* whose Tanimoto coefficient exceeds Φ.

Step 3: Map *j* onto *i* using the substructure search algorithm (*vide supra*).[16] If *j* maps onto *i,* add *j* to the MCS cluster represented by *i* and continue with step 2.

Step 4: Calculate MCS between *i* and *j*.[16] The MCS forms the new representative MCS *k*. If *k* violates both rules (i) and (ii) *j* is put back to the pool. Otherwise all molecules represented by *j* are added to *i*, and *k* forms the new representative MCS *i* of the chemotype. If more MCS are present, continue with step 2 otherwise continue with step 1.

Step 5: If Φ has reached its original value, terminate. Otherwise lower Φ and continue with step 1. Throughout this publication Φ was lowered with a step-size of 0.025.

(4) Singleton Processing. From stage (3) of the algorithm, a list of MCS clusters of size one (*i.e.* singletons) will emerge. In this list, entries might still be present that can be fused to a chemotype. However they violate the above-mentioned parameters describing a chemotype because they are too small for fulfilling the rules. An example is the class with blue + in Figure 2. For such molecules, an extra definition of a chemotype was established in-house: (i) A chemotype contains Γ-2 rings, Ψ non-carbon heavy atoms, and Π_1-3 heavy atoms, (ii) the molecules of a chemotype have greater than Φ "similarity" to each other, and (iii) the MCS and the molecules show at least Ω percentage overlap between the heavy atoms. An example of the effect of these rules can be seen in Figure 4 for the HIV RT inhibitors 329623 and 329624. The calculated MCS represents a chemotype since it contains 2 rings, 4 non-carbon heavy atoms, and 20 heavy atoms and the MCS and both 329623 and 329624 have more than 80% of their atoms in common. The algorithm performs the following steps:

Step 0: Test rule *(i) above* for each singleton. Violating molecules are directly moved to stage (5).

Step 1: Select a singleton *i* from *m*.

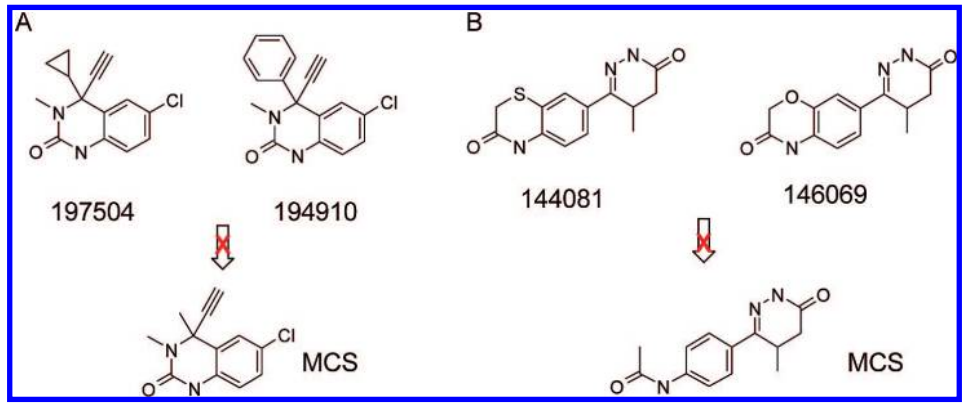

Figure 5. Two examples where a MCS does not represent a chemotype. A. 197504 and 194910 and the corresponding MCS. The MCS violates the ring criterion in stages 2/3 of the algorithm. Additionally 194910 and the MCS violate the percentage overlap parameter Ω in stage 4. B. 144081 and 146069 and the calculated MCS. The MCS violates the ring criterion in stages 2/3 of the algorithm.

Step 2: Select next singleton *j* from *m* whose Tanimoto coefficient to *i* exceeds Φ. Put *j* back to the pool if *j* is too large or too small with respect to *i* (Ω parameter violation).

Step 3: Map *j* onto *i* using the substructure search algorithm.[16] If *j* maps onto *i,* add *j* to the MCS cluster represented by *i* and continue with step 2.

Step 4: Calculate MCS between *i* and *j*.[16] The MCS forms the new representative *k*. If *k* violates either rule (i) or (iii), *j* is put back to the pool. Otherwise add *j* to the MCS cluster and continue with step 2. *k* forms the new representative MCS *i* of the chemotype. If no more singletons are present, continue with step 1.

This last step of the algorithm scales also in $O(N^2)$ with N being the number of singletons. It should be noticed that in contrast to stage (3), Φ is not adjusted in order to start with the closest neighbors for the MCS calculation. The reason for this lies in the fact that no difference was observed in the resulting MCS clusters.

(5) Postclustering of the Chemotypes. The result of the MCS clustering process is a list of clusters (including singletons). Each cluster consists of a MCS and the molecules in the cluster. The molecules are sorted according to a modified version of the MaxMin algorithm:[28]

Step 1: Calculate the virtual centroid for the cluster. As descriptors, the fingerprint bits of the molecules are temporarily converted into double variables.

Step 2: Identify the closest neighbor to centroid. The calculation is performed using the numerical version of the Tanimoto coefficient in eq 1 (see stage (1) for the reason of selecting the numerical version of the Tanimoto coefficient). The selected entry forms the cluster representative and is moved at the top of the sorted result list.

Step 3: Select the next molecule whose maximum similarity coefficient (i.e. minimum Tanimoto coefficient) to the already selected molecules reaches a maximum. Here also eq 1 is applied. The molecule is moved to the next position in the sorted list.

Step 4: Repeat step 3 until all molecules have been moved to the sorted list.

During the evaluation of the algorithm, several observations were made:

A chemotype might violate the defined chemotype rules if it is smaller or contains fewer rings, etc. This will lead to an oversplitting of this particular chemotype into different MCS clusters. An example is shown in Figure 5A for the dihydroquinazolinones 197504 and 194910 and their calculated MCS. Both molecules cannot be assigned to one MCS cluster since the MCS violates the ring criterion in stages 2/3 of the algorithm. Additionally 194910 and the MCS violate the percentage overlap parameter Ω in stage 4.

In the database, molecules of a chemotype might be present which show variation at a specific position. This may also force the chemotype into different clusters. An example is shown in Figure 5B for the benzothiazinone 144081 and the benzooxazinone 146069. The calculated MCS violates the ring criterion in stages 2/3 of the algorithm.

When clustering large libraries, many chemotypes might emerge. This may require a guided graphical navigation in the data.

To allow such a guided navigation in the data, an additional hierarchical clustering of the representatives (*vide supra*) of the MCS clusters is performed. This has the additional advantage of bringing the above-mentioned oversplit chemotypes into the same regions of the hierarchical cluster dendrogram. An example is shown in Figure 2 for the class with blue +. For the clustering, the same modified version of the hierarchical *k*-means algorithm was used as in stage (1). No stop threshold was applied. Since for the clustering only the representatives are used, this clustering is much faster and less memory intense than the preclustering. For large libraries, it offers the additional possibility to use a more computation/memory-intensive hierarchical clustering algorithm like Ward's clustering.[9]

The final results are reported to an interactive graphical user interface (*vide infra*). The user can zoom in the dendrogram, enter clusters, and display the molecules in combination with the MCS. Additionally, biological results can be combined with the molecules. This facilitates the analysis of the SAR of each chemotype.

All clustering steps have been implemented in Java. In addition to the standard Java libraries, the Chemaxon libraries have been integrated into the program for performing the chemistry related calculations.[16] The application of the algorithm on the 118,691 molecules of the MDDR took 3.2 days on a Linux workstation with Dual Core AMD Opteron processor (2.4 megahertz). 957,900 MCS calculations were performed. This is about 8 *N* with *N* being the number of molecules.

Table 1. Number of Clusters and Singletons

	number of clusters	number of singletons
hierarchical *k*-means	28,285	20,583
MCS cluster creation	18,414	52,850
cluster fusion $\Phi = 0.85$	17,468	39,408
cluster fusion $\Phi = 0.825$	16,645	36,897
cluster fusion $\Phi = 0.8$	16,149	34,770
cluster fusion $\Phi = 0.775$	15,779	32,844
cluster fusion $\Phi = 0.75$	15,313	31,134
cluster fusion $\Phi = 0.725$	14,829	29,759
cluster fusion $\Phi = 0.7$	14,149	24,147
singleton processing	18,796	14,724

RESULTS AND DISCUSSION

The clustering algorithm described in this work has been developed for analyzing large data sets like HTS and virtual screening data sets. A filtered version of the MDDR database was used as an example. The MCS-based clustering was performed. At each stage the number of clusters and singletons was monitored (Table 1). It should be noted that for the initial clustering with the hierarchical *k*-means the terminal FP clusters (also referred to as leaves) are shown and that for the final clustering of the representatives the dissimilarity threshold was set so that only singletons were obtained (data not shown). The results are exemplified according to two classes of HIV reverse transcriptase inhibitors and some outliers of other activity classes presented in Figures 6 and 7. On the left side of both figures, the clusters from the original fingerprint-based clustering are shown. On the right side, the clusters are shown after application of the MCS-based clustering. In Figure 7 the hierarchy resulting from the clustering of the MCS representatives is schematically outlined.

The clustering of the MDDR with the hierarchical *k*-means algorithm provided 28,285 clusters and 20,583 singletons. The initial MCS cluster creation step (stage (2)) led to a reduction of the number of clusters to 18,414. This directly translated into a pronounced increase in the number of singletons. Examples are shown in Figure 6 for compounds 205089 and 205092 in cluster 2 and compound 328023 in cluster 3 which are now part of the singleton list in the center of Figure 6. Another example for the origin of such singletons is the nonsingleton cluster in Figure 7. The main driving force of the separation was the size of the MCS and the predefined minimum number of three rings. This first clustering step was designed to remove the heterogeneity in the clusters. For the example in Figure 6, both 205089 and 205092 from cluster 2 are RT inhibitors, whereas the remaining three compounds (now forming cluster 5) are related to treatment of chronic obstructive pulmonary disease (COPD).[29] In cluster 3, compound 328023 is also involved in the treatment of COPD. The remaining molecules of this cluster are all RT inhibitors and are now grouped in cluster 6. In Figure 7 the MCS cluster creation led to the separation of a RT inhibitor (329622) from a molecule with antiangiogenic effect (337957). The examples demonstrate that this split of the compounds in the clusters provides smaller clusters. However, these clusters contain pure and at most isofunctional chemotypes.

The third stage of the MCS clustering algorithm is the fusion of MCS clusters and singletons. As shown in Table 1, this distance-dependent fusion provides a marked reduction of the number of clusters from 18,414 to 14,149. The singletons are also decreased from 52,850 to 24,147. The effect of this fusion step is best explained in Figure 6: After the MCS cluster creation, cluster 1 remained as it was and is now referred to as cluster 4. From clusters 2 and 3, compounds 205089, 205092, and 328023 were converted to singletons. The remaining entries formed the MCS clusters 5 and 6, respectively. The singleton (singleton 1) was also converted to a MCS singleton. The cluster/singleton fusion translated into clusters 7 and 8. Cluster 4 and cluster 6 were fused to form cluster 7. Additionally the three singletons 205089, 205092, and 197503 were added to this cluster. Cluster 8 resulted from the addition of the singleton 328023 to cluster 5. Another example is shown in Figure 7 for the FP-singletons 368905 and 368907 on the left side which now form a cluster on the right side. It should be noted that molecules like 368905 and 368907 are similar to each other and should not result as singletons out of a fingerprint based clustering algorithm. It demonstrates a shortcoming of the hierarchical *k*-means algorithm, which is due to its divisive nature: the iterative splitting of a data set into two partitions may translate into the separation of similar molecules.[14,15] Overall the results show that stage 3 of the algorithm markedly improves the homogeneity of the clustering.

The final singleton processing (stage (4), singleton processing in Table 1) shows a reduction in the number of singletons to 14,724 and an increase in the number of clusters to 18,796 (an average of 3 molecules per cluster). This is exemplified by molecules 329621, 329623, 329624, and 329622 in Figure 7 which now form a cluster. The results indicate that a high proportion of chemotypes is present in the data that escaped the original chemotype rules (3 rings, 2 non-carbon heavy atoms, and 13 heavy atoms). It thus makes this alternative grouping (based on the percentage of overlap of the molecules) a worthwhile and necessary step. It was investigated whether using only this setting would have been a better option for the complete MCS-based clustering. However the calculation time exceeded a week for less than 10,000 molecules, *i.e.* too long to be applicable to large data sets (data not shown). Moreover a pronounced oversplitting of the chemotypes occurred. An example would be clusters 7 in Figure 6 which would then form three different clusters (cluster 1: 203170, 205090, 205091, and 205092, cluster 2: 197503, 197501, 197502, 199613, and 197508 and cluster 3: 205089, 197504, and 197509) and two singletons (197510 and 296994).

In comparison to the fingerprint-based clustering, the MCS algorithm provides a superior clustering with reduced heterogeneity in the individual clusters in the overall clustering. Still, on the right side of Figure 7, two clusters and a singleton are present. The molecules in these clusters did not fuse into a single cluster even if they belong to one isofunctional chemotype. It indicates that the MCS clustering algorithm in combination with the predefined parameter settings has its limitations in its classification. To cope with such an oversplitting of a chemotype, the representatives of the MCS clusters are clustered using the fingerprint version of the hierarchical *k*-means algorithm. The assumption is that the representatives of such an oversplit chemotypes are colocated in the same subdendrogram. In addition it provides a guided graphical navigation through the chemi-

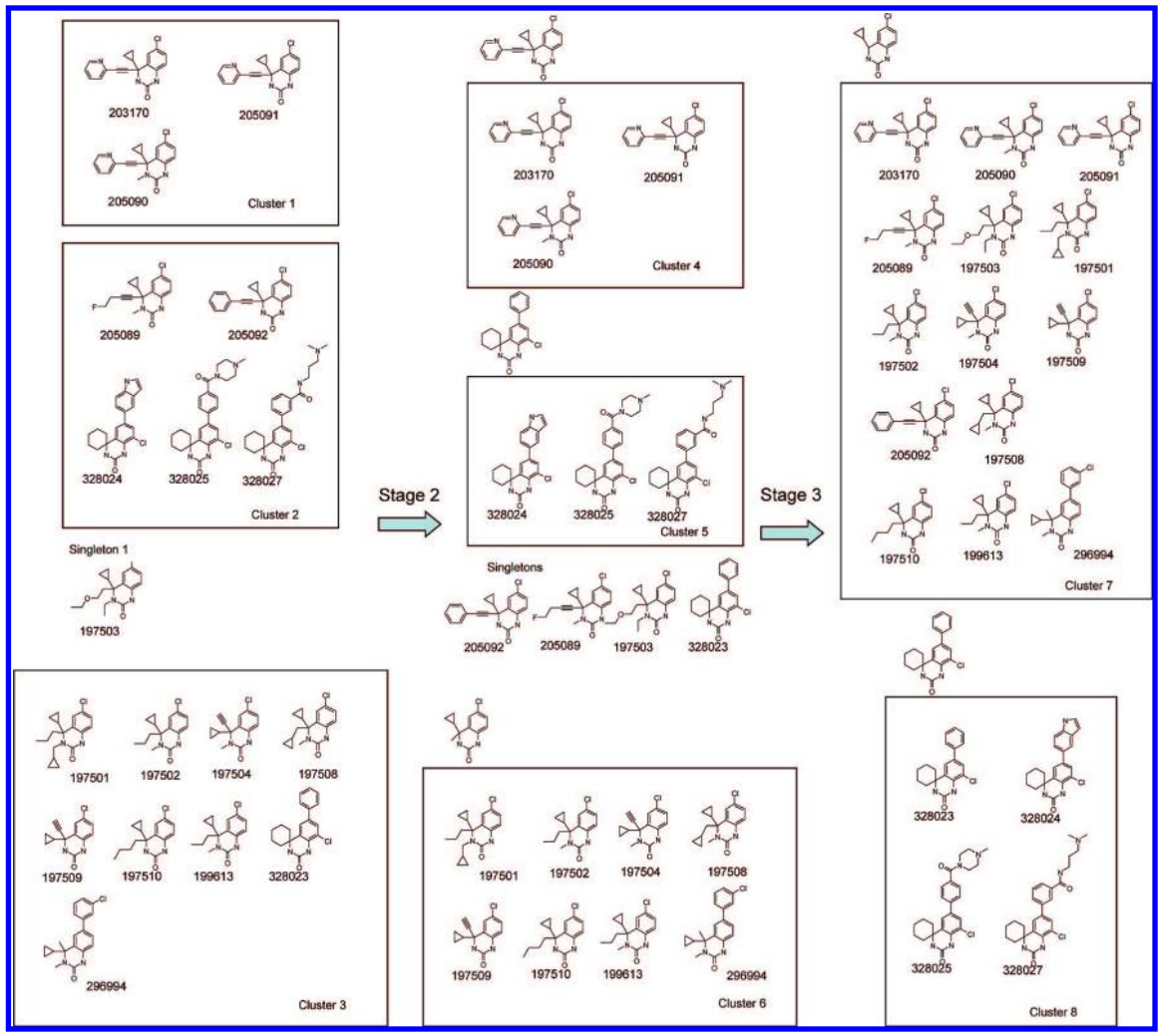

Figure 6. Application of stage (2) and stage (3) of the MCS clustering algorithm to a set of molecules with dihydroquinazolinone moiety from the MDDR data set. On the left side, the clusters from the original fingerprint-based hierarchical *k*-means clustering are shown (clusters 1−3, singleton 1). In the middle the clusters are presented after the MCS cluster creation (stage 2, clusters 4−6 and the singletons in the center). On the right side, the clusters are shown after application of stage 3 of the MCS-based clustering (clusters 7 and 8). All molecules of cluster 7 are RT inhibitors. The only exception is compound 296994 which has no defined target. All molecules in cluster 8 are implicated in the treatment of COPD.[27] The MCS are shown on top of clusters 4−8.

cal space, as shown by the phylogenetic-like representation in Figure 7.

The results of the postclustering are reported to a graphical user interface (GUI), allowing for results display, navigation through the dendrogram, and SAR analyses in the clusters (Figure 8A,B for dendrogram representation). Nodes in the dendrogram correspond to clusters. Clusters highlighted in red are enriched with RT inhibitors. The enrichment is defined by the enrichment factor (EF) scaled to the logarithm to base two of the dendrogram level k (eq 3). In this study, if the scaled EF (sEF) exceeded a value of 3, the cluster was colored red.

$$\mathrm{sEF} = \frac{\mathrm{EF}}{\log_2 k} \tag{3}$$

This scaling of the EF was introduced to obtain an enrichment factor that is independent of the dendrogram level (*i.e.* independent of the size of the data set in the cluster). It assumes that on each level the data are separated into equally sized proportions. Figure 8B shows a focused view on the RT inhibitor-containing subdendrograms indicated by the red oval in Figure 8A. In Figure 8C, the cluster indicated by the red oval in Figure 8B is further examined. The entries of the cluster are shown in combination with statistical information about the hits (RT inhibitors) and nonhits. In Figure 8D the molecules of the cluster are shown in combination with the MCS. This workflow exemplifies how to apply the navigation tool on a large data set and how to focus in a few steps on the chemotype of interest and draw preliminary conclusions about SAR.

The presented MCS clustering algorithm has been developed to provide a grouping of chemotypes in large data sets like HTS data. One might also think of applying the algorithm in a different context: the evaluation of descriptor-based similarity searching. Descriptor-based similarity searching comprises the calculation of a set of descriptors for the query and target molecules and the comparison between them. Optionally a descriptor selection step might be used as an intermediate step. When performing such a screening in a target database several questions arise. How effective is the screening in retrieving entries of the same chemotype? How effective is the screening in retrieving entries with alternative chemotype (chemotype hopping potential)? What

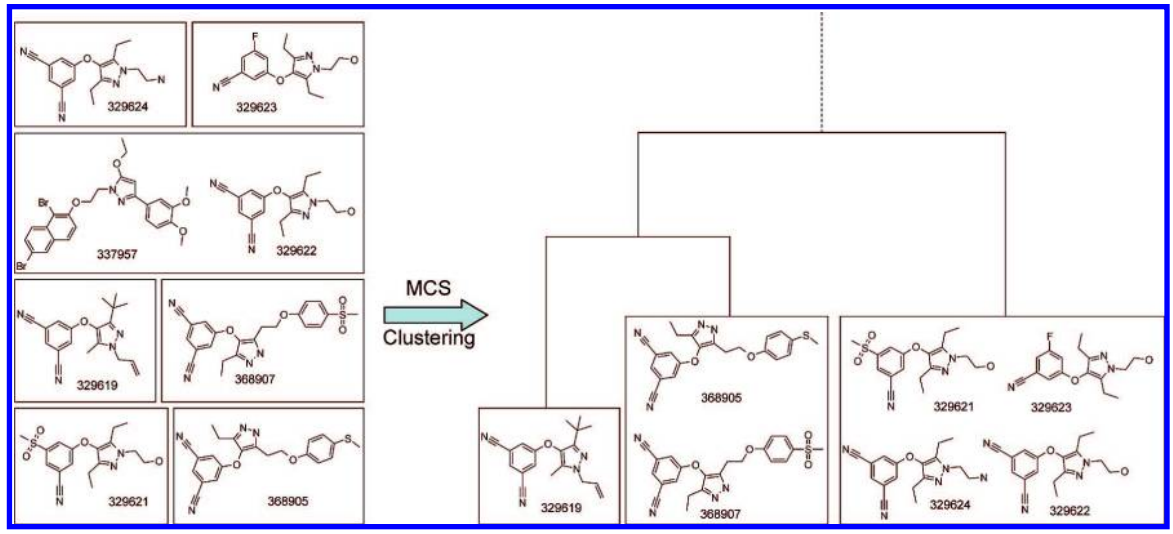

Figure 7. Example of the MCS clustering. On the left side, the clusters from the original fingerprint-based hierarchical *k*-means clustering are shown. On the right side, the clusters are shown after application of the MCS-based clustering. Additionally, the final hierarchical clustering of the representatives from the MCS clusters is schematically outlined. All molecules except compound 337957 have been described as RT inhibitors. 337957 is not shown on the right side since it belongs to a cluster that is far away in the final cluster dendrogram.

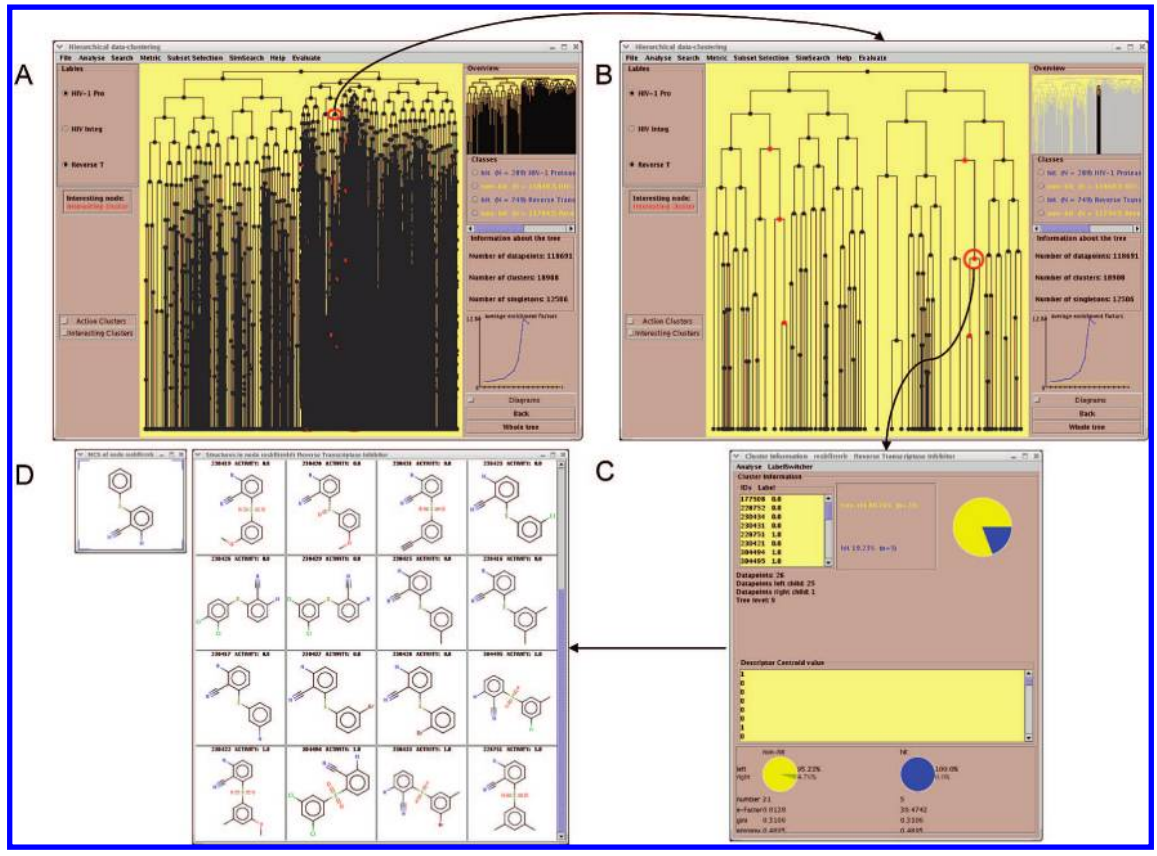

Figure 8. Example of navigation in the graphical user interface resulting from the MCS clustering algorithm.

data preparation steps should be used? What metric should be used? How far should we look? Answering these questions in a systematic manner is not a trivial task and depends on the target under investigation. Here an approach is outlined to answer some of these questions with the combined FP based and MCS based clustering algorithm. Using this algorithm, the MDDR database was clustered into 18,796 clusters and 14,724 singletons (*vide supra*). The activity description of the MDDR was added to the entries. Only activities were considered with defined target and defined mode of activation (*i.e.* dopamine D_3 receptor antagonist, HIV RT inhibitor, etc.). If more than one activity description has been assigned to a molecule, all descriptions were considered. From each cluster and each target class, a representative was randomly selected. As a prerequisite at least 4 additional entries of the same activity class had to be present in the same MCS cluster (same chemotype) and at least 5 additional entries of the same activity class had to be present within other MCS clusters (alternative chemotype). In total 2387 different MDDR molecules were selected. For

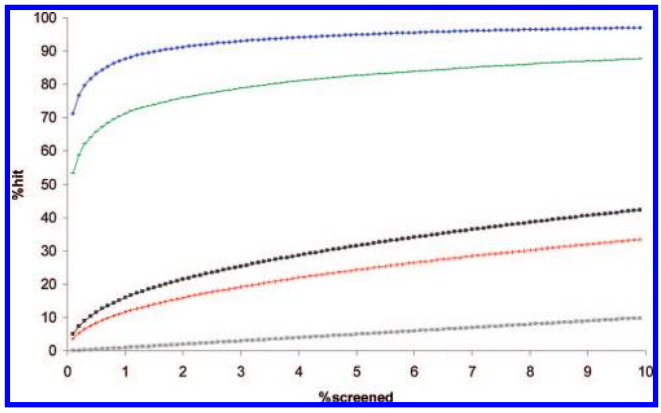

Figure 9. Average enrichment plot (first 10%) obtained for 2387 queries. The descriptor-based similarity searching was performed using MDDR_MOE2D_NORM (red and green curve with plus and minus markers, respectively) and MDDR_MOE2D_SCALED (black and blue curve with square and diamond markers, respectively). The gray straight line with star markers represents a random selection. Screening for the same chemotype corresponds to the blue and green curve. Screening for alternative chemotypes corresponds to the black and red curve.

each of the molecules a descriptor-based similarity search was performed in the MDDR.

Screening was performed using the two data sets MDDR_MOE2D_NORM (all descriptors normalized to the range of 0 and 1) and MDDR_MOE2D_SCALED (all descriptors centered to the mean and scaled to unit variance). As a metric, the cosine coefficient was used,[25] since it gave the best results compared to other metrics and coefficient used in this study (data not shown). For each screen, the entries of both data sets were sorted in decreasing similarity coefficient to the query. Enrichment plots were created, whereas hits were either defined as entries from the same activity class and the same cluster or as entries from the same activity class and an alternative cluster (chemotype). Finally averages were calculated over all queries. These average results are shown in Figure 9 for the first 10% of the screened data set. The blue curve and the black curves were obtained with MDDR_MOE2D_SCALED. The green curve and the red curves were obtained with MDDR_MOE2D_NORM. The gray line represents a random selection. Screening for the same chemotype corresponds to the blue and green curve, and screening for alternative chemotypes corresponds to the black and red curve. Three conclusions can be drawn from Figure 9: (i) screening for entries with the same chemotype is more effective than screening for molecules with alternative chemotype, (ii) the extended MOE descriptors have chemotype hopping potential since the enrichment curve for both data sets and the alternative chemotype is above the random curve, and (iii) mean centering of the descriptors and scaling to unit variance is the better choice in this context compared to the normalization of the descriptors since both average enrichment curves are significantly higher (the significance test was performed using the comparison circle algorithm in Spotfire DecisionSite 8.2.1).[30] The same trends were observed when plotting receiver operator characteristic (ROC) curves (data not shown). It has to be emphasized that these results represent an average screen and that large standard deviations were observed (data not shown). For a particular screen/target a different parameter setup might be better suited. Moreover such a different setup might provide an alternative view on the data. Still, given the scenario of a prospective screen and a toolbox with 5 different descriptor sets, 10 different metrics, several descriptor selection routines, and scaling routines requires one to have a guide on what setting can be used as a starting point.

CONCLUSION AND OUTLOOK

A new clustering algorithm has been developed. The main stages include the following: (1) preclustering a molecule data set with the hierarchical *k*-means algorithm, (2) creation of MCS-based clusters from the preclustering, (3) fusion of the MCS-based clusters and singletons, (4) processing of the remaining singletons, and (5) hierarchical postclustering of the representatives of the identified chemotypes with the hierarchical *k*-means algorithm. Results are reported to an interactive GUI. The clustering algorithm was exemplified with two classes of HIV RT inhibitors in the MDDR. It was demonstrated that the initial MCS cluster creation process is capable of reducing the heterogeneity in the clusters obtained by FP clustering. The MCS cluster/singleton fusion translated into a reduction of the number of clusters and singletons. It indicates that the homogeneity in the clustering results is enhanced. The final cluster creation based on the proportion overlap between the molecules was able to reduce the number of singletons. This step provided again an improvement of the homogeneity of the clustering. The final clustering of the MCS cluster representatives was applied to group oversplit chemotypes into the same subdendrogram. Although superior parameter settings might exist and an oversplitting of a chemotype may still be observed, these data show that all calculation steps are indeed necessary and that the algorithm is able to provide a clustering of high quality. In combination with the interactive GUI, the algorithm enables clustering of large data sets on a routine basis and to immediately draw conclusions about SAR. It also allows posing new questions to the data set as exemplified in the virtual screen validation example.

Additional applications of the algorithm might exist for the clustering of complete HTS data sets, for example in viewing the hits in the context of the nonhits and providing a more complete picture of the data. Moreover it might enable the selection and/or prioritization of representatives from each chemotype, a typical scenario in postprocessing of large virtual screening hit sets. Finally, by mapping the MCS obtained for a particular library on the compounds of a second library, an analysis may be possible as to the extent of chemotype overlap between the libraries. To sum up, I propose that this MCS clustering algorithm will add a new and useful dimension to the chemoinformaticians toolbox.

Abbreviations: COPD: chronic obstructive pulmonary disease, EF: enrichment factor, FP: fingerprint, GUI: graphical user interface, HIV: human immunodeficiency virus, HTS: high-throughput screening, ID: identifier, MCS: maximum common substructure, RT: reverse transcriptase, SAR: structure activity relationship, TPSA: topological polar surface area, UFS: unsupervised forward selection, VS: virtual screening.

ACKNOWLEDGMENT

I would like to thank Britta Sasse, Anne-Marie Faucher, and Pierre Bonneau for their useful comments on the manuscript and many helpful discussions.

REFERENCES AND NOTES

(1) Duda, R. O.; Hart, P. E.; Stork, D. G. *Unsupervised Learning and Clustering. In Pattern Classification*, 2nd ed.; Duda, R. O., Hart, P. E., Stork, D. G., Eds.; John Wiley & Sons, Inc.: New York, NY, 2001; pp 517−599.

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

(3) 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.

(4) Engels, M. F. M.; Gibbs, A. C.; Jaeger, E. P.; Verbinnen, D.; Lobanov, V. S.; Agrafiotis, D. K. A Cluster-Based Strategy for Assessing the Overlap between Large Chemical Libraries and Its Application to a Recent Acquisition. *J. Chem. Inf. Model.* **2006**, *46*, 2651–2660.

(5) Selzer, P.; Ertl, P. Applications of Self-Organizing Neural Networks in Virtual Screening and Diversity Selection. *J. Chem. Inf. Model.* **2006**, *46*, 2319–2323.

(6) Clark, D. E.; Higgs, C.; Wren, S. P.; Dyke, H. J.; Wong, M.; Norman, D.; Lockey, P. M.; Roach, A. G. A Virtual Screening Approach to Finding Novel and Potent Antagonists at the Melanin-Concentrating Hormone 1 Receptor. *J. Med. Chem.* **2004**, *47*, 3962–3971.

(7) Kellenberger, E.; Springael, J.-Y.; Parmentier, M.; Hachet-Haas, M.; Galzi, J.-L.; Rognan, D. Identification of Nonpeptide CCR5 Receptor Agonists by Structure-based Virtual Screening. *J. Med. Chem.* **2007**, *50*, 1294–1303.

(8) Jain, A. k.; Murty, M. N.; Flynn, P. J. Data Clustering: A Review. *ACM Comput. Surv.* **1999**, *31*, 265–323.

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

(10) 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.

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

(12) Kohonen, T. Self-Organized Formation of Topologically Correct Feature Maps. *Biol. Cybern.* **1982**, *43*, 59–69.

(13) *MDL Drug Data Report, Version December 2006*; Symyx Technologies Inc.: Santa Clara, CA, 2006.

(14) Böcker, A.; Derksen, S.; Schmidt, E.; Teckentrup, A.; Schneider, G. A Hierarchical Clustering Approach for Large Compound Libraries. *J. Chem. Inf. Model.* **2005**, *45*, 807–815.

(15) Böcker, A.; Schneider, G. Teckentrup. NIPALSTREE A new Hierarchical Clustering Approach for Large Compound Libraries and Its Application to Virtual Screening. *J. Chem. Inf. Model.* **2006**, *46*, 2220–2229.

(16) *JChem, version 3.2.8*; ChemAxon Ltd.: Budapest, Hungary, 2006.

(17) 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.

(18) 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.

(19) 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.

(20) Stahl, M.; Mauser, H. Database Clustering with a Combination of Fingerprint and Maximum Common Substructure Methods. *J. Chem. Inf. Comput. Sci.* **2005**, *45*, 542–548.

(21) Gardiner, E. J.; Gillet, V. J.; Willett, P.; Crosgrove, D. Representing Clusters Using a Maximum Common Edge Substructure Algorithm Applied to Reduced Graphs and Molecular Graphs. *J. Chem. Inf. Model.* **2007**, *47*, 354–366.

(22) *Pipeline Pilot, version 6*; Accelrys, Inc.: San Diego, CA, 2006.

(23) *Molecular Operating Environment (MOE), version 2006.08*; Chemical Computing Group Inc.: Montreal, Canada, 2006.

(24) 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.

(25) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 986–996.

(26) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do Structurally Similar Molecules have Similar Biological Activity. *J. Med. Chem.* **2002**, *45*, 4350–4358.

(27) Daylight Chemical Information Systems, Inc. http://www.daylight.com/dayhtml/doc/theory/theory.finger.html (accessed July 14, 2008).

(28) Schmuker, M.; Givehchi, A.; Schneider, G. Impact of Different Software Implementations on the Performance of the Maxmin Method for Diverse Subset Collection. *Mol. Divers.* **2004**, *8*, 421–425.

(29) Barns, P. J. New Treatments of COPD. *Nat. Rev. Drug Discovery* **2002**, *1*, 437–446.

(30) *Spotfire DecisionSite, version 8.2*; TIBCO Software Inc.: Palo Alto, CA, 2006.

CI8000887