Database Clustering with a Combination of Fingerprint and Maximum Common Substructure Methods

Martin Stahl* and Harald Mauser

F. Hoffmann-La Roche AG, Pharmaceutical Research, CH-4070 Basel, Switzerland

Received January 10, 2005

We present an efficient method to cluster large chemical databases in a stepwise manner. Databases are first clustered with an extended exclusion sphere algorithm based on Tanimoto coefficients calculated from Daylight fingerprints. Substructures are then extracted from clusters by iterative application of a maximum common substructure algorithm. Clusters with common substructures are merged through a second application of an exclusion sphere algorithm. In a separate step, singletons are compared to cluster substructures and added to a cluster if similarity is sufficiently high. The method identifies tight clusters with conserved substructures and generates singletons only if structures are truly distinct from all other library members. The method has successfully been applied to identify the most frequently occurring scaffolds in databases, for the selection of analogues of screening hits and in the prioritization of chemical libraries offered by commercial vendors.

INTRODUCTION

Automated clustering methods are useful tools to organize sets of chemical structures,[1] for example to visualize the content of databases or to select representative structures. In the context of biological high throughput screening, clustering may be applied at various stages.[2] Here we focus on the early hit validation process. A typical first step in hit validation is to check activities of compounds related to those identified as hits.[3−5] In this context, a preclustered screening library can yield important insights: Once active compounds are identified, they can be mapped onto the clustered library.[6,7] A comparison of active and related inactive structures provides first structure−activity relationships and helps to select compounds for follow-up screening. Our experience indicates that first structure−activity relationships are best understood if the compounds share a substructure of significant size. A clustering method suitable for this kind of application should therefore be substructure-conserving and ideally group together compounds with common scaffolds in one cluster. This also implies that the clustering method should generate singletons only when the respective compounds do not share a significantly large substructure with any other compound in the library. Of course, another requirement is that the method should be fast enough to be applicable to libraries of up to one million compounds on a regular basis, e.g. for every update of a screening library.

A large variety of chemical clustering methods has been proposed, and the interest in fast and robust methods has increased in recent years.[8−11] Typically, they are a combination of a similarity metric, often bit string or key-based,[12,13] and a clustering algorithm. Other established methods use catalogs of predefined substructures[14,15] or specific definitions of what constitutes a scaffold[16] to group structures into classes. Maximum common substructure algorithms[17] are increasingly used to analyze HTS results[18] but are too slow to be applicable to very large databases. Our quest for a method meeting the above specifications started with an analysis of clustering algorithms operating on Daylight fingerprints. These are attractive because of substructure-conserving properties and because of the low computational cost associated with their generation and comparison. While clusters generated by the original Jarvis−Patrick algorithm[19] and more refined variants thereof[12,20] were too heterogeneous for our purposes, we were intrigued by the tight clusters produced by the exclusion sphere (ES) algorithm.[21] However, ES clustering is prone to yield singletons that do have close neighbors in a cluster, and it often splits up sets of compounds sharing the same scaffold into several clusters. We realized that these limitations could be overcome by applying maximum common substructure (MCS) methods to the ES clustering results. The common scaffold typically shared by ES cluster members is extracted and used in a second clustering step to group closely related clusters. Special care is taken to merge singletons with clusters whenever they have common scaffolds. We describe the workflow of the combined fingerprint/MCS method and report on its application to a library of 56 000 drug-like structures.

METHODS

The workflow of our hybrid clustering method is outlined in Figure 1. An initial clustering step (1) by means of a modified ES algorithm based on Daylight fingerprint Tanimoto indices leads to a set of clusters and singletons. From the clusters, common substructures (scaffolds) are extracted (2). In two further steps, these scaffolds are compared with each other (3) and with the singletons from the initial step (4). In a final step (5), the neighbor lists obtained in steps 3 and 4 are combined to yield merged

* Corresponding author phone: +41 61 68 88421; e-mail: martin.stahl@roche.com. Corresponding author address: F. Hoffmann-La Roche AG, PRBD-CS 92/3.56B, CH-4070 Basel, Switzerland.

10.1021/ci050011h CCC: $30.25 © 2005 American Chemical Society
Published on Web 03/30/2005

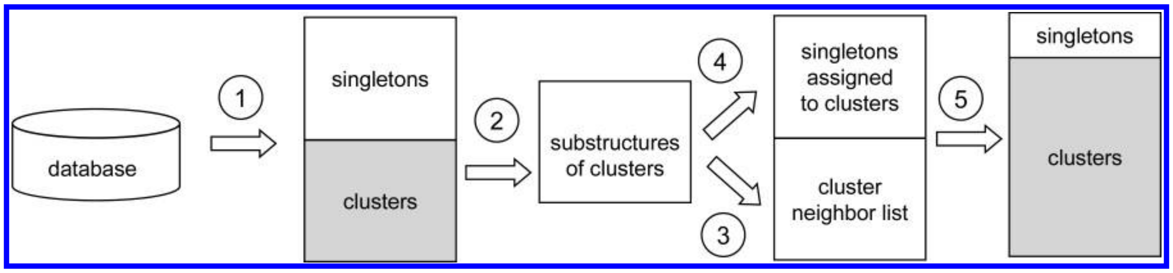

Figure 1. Workflow of the hybrid fingerprint/MCS clustering procedure. Step 1: Daylight/ES clustering, step 2: substructure extraction from clusters, step 3: generation of cluster neighbor list, step 4: comparison of singletons to cluster substructures, and step 5: combination of neighbor lists generated in step 3 and 4 in a second ES clustering step.

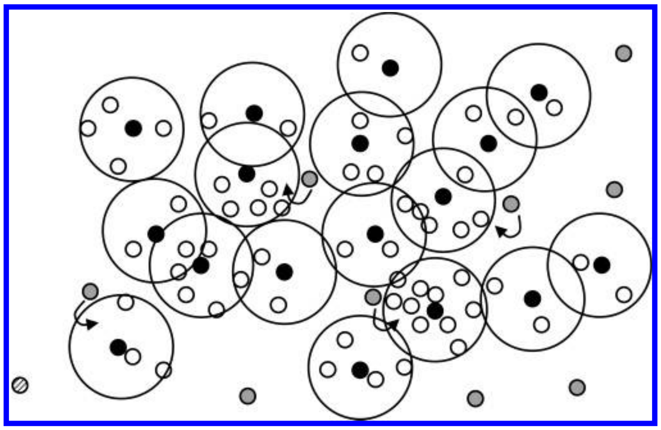

Figure 2. Schematic representation of exclusion sphere clustering in a two-dimensional descriptor space. Cluster centers (black circles) are surrounded by their members (open circles) within exclusion spheres (large circles). Singletons (gray) can be divided into true singletons and artificial ones, whose neighbors are removed during the clustering process by other clusters. Artificial clusters are marked by arrows pointing to the cluster they should be assigned to.

clusters and the remaining singletons. In the following, we discuss each of these individual steps in more detail.

1. Daylight ES Clustering. We implemented the ES algorithm as described by Butina,[21] using Daylight[22] fingerprints of 2048 bits fixed length and a conservative Tanimoto cutoff of 0.8.[23] For each library member, neighbors above a specified threshold Tanimoto value are collected. The library is then sorted according to the number of neighbors. The library member with most neighbors is defined as the first cluster center and its neighbors as cluster members. All cluster members are removed from the sorted library list. The algorithm continues defining clusters by moving down this list, generating clusters of decreasing size, and removing their members from the list. A representation of an ES-clustered database in a hypothetical 2-dimensional descriptor space is given in Figure 2, where small circles represent library members and large circles denote clusters. Cluster centers are colored black, and members are denoted by empty circles. We distinguish between two types of singletons: *True singletons* have no neighbors within the defined cutoff (hatched circles). *Artificial singletons* do have neighbors within the cutoff radius, but these have been previously assigned to other clusters by the algorithm (gray circles). It is an undesirable property of the ES algorithm to create such singletons. We have extended the algorithm by assigning such singletons to the cluster that contains its closest neighbor. This has the effect of creating a local "bulge" in the exclusion sphere defining the cluster, such that the cluster center no longer has a known minimum similarity to each of the cluster members. We find that this is no disadvantage, since these compounds typically are of the same chemical class as the other cluster members. According to our experience, between 10 and 30% of the singletons can be assigned to clusters by this procedure.

2. Scaffold Extraction. In this step we exploit the tightness of the clusters produced by the Daylight/ES clustering method. For each cluster, the MCS common to all members is calculated. We omit those cluster members described as artificial singletons in step 1, because they may in a few cases unnecessarily reduce the size of the cluster scaffold. First, the MCS between the cluster center and an arbitrary member is determined. This substructure is stored as the cluster substructure and iteratively reduced in size as it is compared to the remaining cluster members that may not share the entire substructure. A minimum MCS size of 8 atoms is required. If the cluster substructure drops below 8 atoms, it is not updated. The rationale for choosing such a boundary is that the identification of the general cluster scaffold should not be impeded by the presence of a few structures whose scaffold is different from the one the majority of the members share. The generation of unrepresentative cluster scaffolds is avoided by requiring a minimum MCS size of 10 atoms in the subsequent step (step 3 below). We note that this substructure generation process is order-dependent, but by starting with the cluster center we minimize the danger of extracting a nonrepresentative scaffold. After the scaffold extraction process, each scaffold is checked for the presence of a phenyl ring (isolated or annulated ring system). The cluster IDs of all scaffolds containing a phenyl ring are stored in a separate list. This information is used in the next step to prune cluster neighbor lists.

3. Generation of Cluster Neighbor Lists. MCS algorithms are again used to compare the cluster scaffolds extracted in the previous steps. For each cluster, a neighbor list is constructed. For two scaffolds G and H to be within each other's neighbor list, a Raymond score of at least 0.6 must be achieved and the MCS must comprise at least 10 atoms. The Raymond score[24] R for two molecular structures G and H is defined as

$$R = \frac{(atoms(MCS) + bonds(MCS))^2}{(atoms(G) + bonds(G))(atoms(H) + bonds(H))}$$

The neighbor list generated in this manner is pruned by applying more stringent criteria for cluster neighborhood. Two clusters are kept on each other's neighbor list if $atoms(MCS) \geq 18$, or if $R > 0.85$ and $atoms(MCS) \geq 10$. If both G and H contain a phenyl ring as a substructure, the criteria are set to $R > 0.85$ and $atoms(MCS) \geq 14$ to avoid meaningless substructures (e.g. a phenyl ring with an amide substituent) to be regarded as a scaffold.

4. Assigning Singletons to Clusters. All singletons remaining in the Daylight ES-clustering step are compared to cluster scaffolds with the goal to find a cluster sharing a large common substructure with the singleton. For a singleton S to be assigned to a cluster, the scaffold SC must be contained within the structure of the singleton or vice versa, i.e., the ratio

$$\frac{atoms(MCS)}{\min\,(atoms(SC),atoms(S))}$$

must be equal to 1. The number of matching atoms between SC and S must be at least equal to $atoms(MCS) = 10$. If several alternative clusters fulfill these criteria for one singleton, the cluster with the largest matching substructure is chosen. The result of this procedure is a list of singleton IDs and matching cluster IDs.

5. Clustering the Clusters. All information gained in the previous steps is used as the basis for a procedure generating the final cluster IDs. Singletons from step 4 are assigned to their respective closest clusters. The cluster neighbor list from step 3 is then used as input for a second ES clustering carried out in the same manner as described above, i.e., with assignment of artificial singletons to clusters (note that the “singletons” referred to here are clusters generated in step 1). Clusters are then sorted according to the number of structures they contain.

MCS Calculations. All MCS routines employed in steps 2−4 are written in Python and based on Openeye’s OEChem library. This library offers a set of functions for pattern matching including the extraction of the maximum common substructures. We used the standard OEChem functions for clique detection and graph matching, which were found to be fast and robust. Within these functions, atom types can be differentiated by means of an integer value assigned to each atom (“colored graphs”). Our atom typing scheme is based on Tripos mol2 atom types. Carbon atoms are differentiated based on hybridization and aromaticity. Nitrogen atom types N.1, N.2, N.am, N.ar, and N.pl3 were treated as one atom type. N.3 and N.4 were also united in one group, i.e., we distinguish between potentially positively charged and “other” nitrogen atoms. Oxygen atoms are also divided into two classes based on whether they carry a hydrogen atom (donors) or not (acceptors). Acids are treated as acceptors. Sulfur atoms are either classified as oxidized (S.o, S.o2) or polarizable (S.2, S.3), and halogens are treated as one group. To ensure that ring membership could be treated as a conserved property even if a substructure contained an incomplete ring system, atoms in rings received separate atom types from acyclic atoms. Atom types were generated once for each input structure and stored in sd format along with the structures. The scripts *mcsextract.py*, to extract common substructures from a set of molecules, and *mcsim.py* for scoring pairs of structures take these sd files as input and do not change the atom typing at a later stage. In all calculations involving the determination of an MCS, bond types are not differentiated, and substructures with less than eight atoms are discarded. Within each MCS calculation, all remaining common substructures are ordered such that those with closed rings receive higher priority than others with incomplete rings.

Compilation of Test Set. We combined all structures contained in the Medchem 2003 database[22] and in the 2004 release of the World Drug Index.[25] Duplicate structures were removed, as were compounds outside a molecular weight range of 150−700. In addition, a set of in-house SMARTS substructure exclusion filters were applied to remove inorganic salts, structures containing elements other than C, N, O, P, S, and halogens, reactive functional groups, and other groups generally not tolerated in drugs. The final test set, which we will refer to as the Wdimed library, consists of 56 415 structures. All format conversions from smiles strings to sdf format were carried out with the *csfc* program, which is part of the CACTVS toolkit.[26]

Run Times. The entire clustering procedure to the Wdimed library was run on Linux workstations and a cluster equipped with 2.6 GHz Intel Xeon chips and running the Redhat 7.3. Step 1 requires 45 min (single processor), step 2 requires 3 h (single processor), and step 3 requires 30 min (20 processors). Step 4 is the most time-consuming part and requires 2 h on 20 processors. The final clustering step is done interactively in a few seconds. We have also applied the full procedure to a 750 000 compound subset of our screening library. Timings are for 7 days for step 1 (single processor), 12 h for step 2 (single processor), 2.5 days for step 3 (40 processors), and 4 days for step 4 (40 processors). Overall, these are acceptable run times considering that these calculations need to be done only once per library update (approximately twice per year).

RESULTS

The mode of operation of the fingerprint/MCS clustering process is exemplified by the 275 benzodiazepine structures contained in the Wdimed library, which form its fourth largest cluster. A substructure search confirms that this cluster indeed comprises all 5-phenyl-benzodiazepines contained in this library. Representative structures are shown in Figures 3 and 4.

Standard Daylight/ES clustering of the Wdimed library leads to a total of 27 clusters and 23 singletons containing a benzodiazepine core structure. Of these singletons, 13 are artificial singletons in the sense of Figure 1 (see Methods section). They are included in the 27 clusters through our modified ES clustering routine. Compound **1** (Figure 3) is the center of one of the 27 clusters, compounds **2**−**9** are some of the cluster’s 18 primary members, and compounds **10**−**13** are artificial singletons assigned to this cluster. The primary cluster members—those lying within a similarity radius of 0.8 around the center—have in common that they contain aromatic fluorine substituents or CF_3 groups. The presence of these fluorine atoms sets a number of bond path bits in the Daylight fingerprint, making these compounds more similar to the center **1** than to benzodiazepines with different substituents. Benzodiazepines with e.g. nitro or alkoxy substituents form separate clusters. This clustering result is typical of the exclusion sphere algorithm. The separation into classes with different types of substituents is particularly obvious with small scaffolds, because with sparsely populated fingerprints the presence or absence of individual bits has a larger influence on the Tanimoto

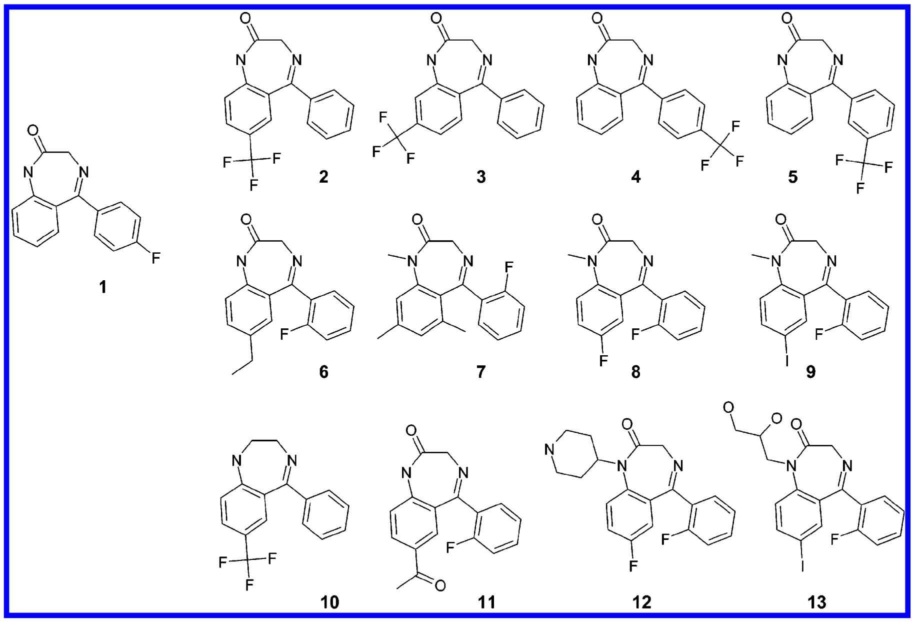

Figure 3. Members of One Cluster Generated by Daylight/ES Clustering. Structure **1** is the cluster center, structures **2**−**9** are primary cluster members within the exclusion sphere, and structures **10**−**13** are artificial singletons in the sense of Figure 1 that were assigned to the cluster.

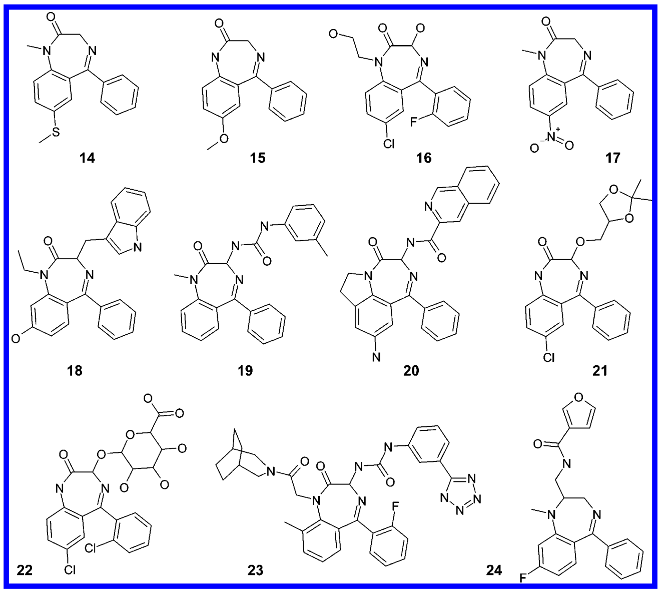

Figure 4. Members of the Benzodiazepine Cluster in the Wdimed Library. Structures **14**−**21** are centers of clusters formed in the Daylight/ES clustering step, and structures **22**−**24** are singletons added to this cluster in step 4 of the process (see Figure 2).

coefficient.[27] The artificial singletons added to the cluster in Figure 3 can be regarded as "nearest neighbor neighbors" of the cluster center. They possess the substitution pattern of a cluster member but carry additional substituents or lack a substituent present in the cluster member. For example, structure **10** is identical to **2** but lacks the carbonyl group. Structures, **11**−**13** carry additional substituents but are otherwise identical to **6**, **8**, and **9**, respectively.

Structures **14**−**21** in Figure 4 are centers of some of the clusters combined to one large cluster by the MCS/ES clustering procedure operating on the extracted cluster scaffolds. These structures share the benzodiazepine scaffold but carry different sets of substituents. Structure **22** is the center of a cluster added to the large cluster as an "artificial singleton", which also contains the benzodiazepine core structure. Finally, structures **23** and **24** are true singletons from the Daylight/ES clustering step which are added to the large benzodiazepine structure through the comparison of the cluster scaffolds with the cluster centers. Structures added to a cluster in this way carry particularly large or unique substituents which make them appear as singletons in the Daylight/ES clustering step.

The composition of the other major clusters of the Wdimed library can be described in a similar way as the benzodiazepine cluster. Most large clusters are formed through the assembly of several larger and smaller Daylight/ES clusters but share the same substructures. The largest cluster consists of 385 1H-quinoline-4-ones, a class of gyrase-inhibiting antibiotics with ciprofloxacin as a prominent representative. The second largest cluster contains 365 steroids with partially unsaturated A-rings, and cluster 3 contains adenosine derivatives (348). Further major classes of structures among the top 15 clusters are chromen-4-one derivatives (236), 4-phenyl-dihydropyridine dicarboxylic acid derivatives (225), uracil derivatives (225), N-phenyl-benzamides (194), diaminopyrimidines (145), pteridine derivatives (133), and several further steroid clusters. The list continues with clusters of quickly decreasing size; the top 10 clusters contain 4.6% of the Wdimed library, the top 100 clusters about 16%, and the top 1000 clusters 47% of the library.

It is worthwhile to study the effect of the two main clustering steps on the number and size of clusters. The Daylight/ES clustering step yields 8423 clusters. 3568 of these consist of 2 structures only, the remainder are larger clusters. The original ES clustering algorithm would generate 14 486 singletons (close to 26% of the library). Our modified procedure to assign "artificial singletons" to clusters reduces this number to 12 209 singletons (21% of the library). The effect of the MCS/ES clustering step is to reduce the number of small clusters (with 11 members and less) and to increase the number of larger clusters. At the end of the clustering process, 6773 clusters remain, of which 2367 are 2-member clusters. Thus, the MCS/ES clustering step relocates ap-

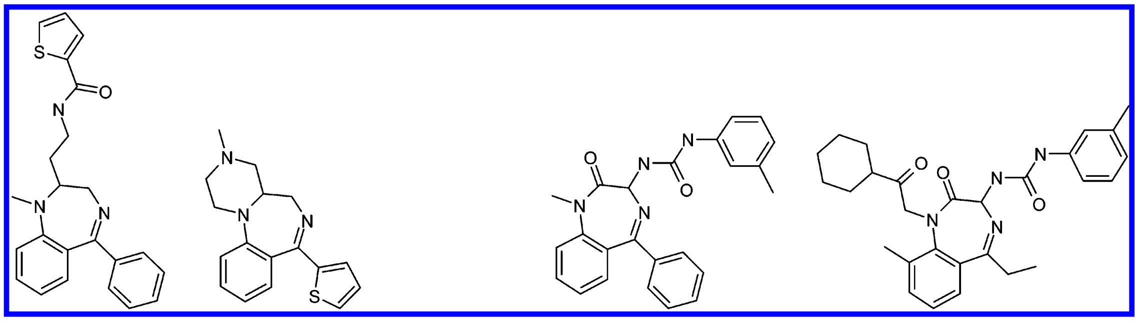

Figure 5. If the first ES clustering step is replaced by JP clustering, substructures extracted from each cluster are smaller and merged clusters become larger and more heterogeneous. Two pairs of compounds are shown. Only the left hand structure of each pair is a phenylbenzodiazepine, which is a member of the large, homogeneous benzodiazepine cluster discussed in the Results section. Both members of each pair belong to the same JP cluster, effectively reducing the size of the common substructure by a phenyl ring.

proximately 1500 clusters, corresponding to about 17% of the structures in the Wdimed library. The number of singletons is reduced to 8488 through the MCS-based assignment of singletons to clusters. That is, the overall number of singletons has been reduced from 26% of the library to about 15%.

DISCUSSION

When applied to screening libraries or other large sets of drug-like structures, the combination of Daylight/ES clustering with a second clustering step typically yields a small number of large clusters and a large number of small clusters. To some extent this may be a consequence of the exclusion sphere algorithm, which creates clusters in densely populated areas first. It is also a manifestation of the history of medicinal chemistry research yielding a number of heavily explored compound classes and many significantly less well represented structural classes.

The method significantly reduces the number of singletons relative to the original ES clustering algorithm. Our experiences with clustering compound collections offered by many external vendors and several screening libraries indicate that the number of singletons produced with the fingerprint/MCS method is typically below 10%. This is an important aspect in particular with respect to the application to HTS hit lists. Single hits are much more difficult to follow up than hit clusters. We do not regard it as a disadvantage that in some cases, singletons assigned to clusters in one of the three steps may deviate structurally from the majority of the cluster members. It is easier to discard compounds that are not similar enough to be of interest than to identify suitable compounds for follow-up screening. We note that our method can be used as a two-level clustering scheme with the original Daylight/ES cluster IDs as a more fine-grained compound classification below the merged cluster IDs that may be useful to explore the immediate environment of a hit. A further reduction of singletons may be possible if singletons were not only compared to cluster scaffolds but also to each other. We have refrained from implementing this option, because such calculations would be very time-consuming and would in practice only lead to small changes in the overall clustering.

The sequential application of fingerprint and MCS-based ES clustering delivers results that are close to a manual substructure-based clustering process performed by a medicinal chemist. Key structural classes contained in a database are quickly discovered. No preconceived notion of the terms "structural class" or "scaffold" is used here. Rather, we use an operative definition of the term: Two structures have the same scaffold if they share a large substructure. The definition of "large" in this context is given by the empirically determined clustering parameters given in the Methods section. It is an advantage of the method that cluster IDs may be quickly regenerated with slightly different parameter settings once all neighbor lists (up to step 4 in the Methods section) are generated. Clearly, the substructures common to a cluster may not be equally meaningful to a medicinal chemist's eye. While the benzodiazepine cluster described above contains many compounds synthesized through the same chemical reactions and with similar biological activity, there are other clusters whose substructures, however large, may seem trivial or uninteresting to a chemist's eye. One relatively large Wdimed cluster, for example, has a glycosylated phenol ring as a common cluster substructure. Clustering results can become meaningless in particular when scaffolds contain relatively large but frequently occurring substructures. We had noted that a simple phenyl ring is such a substructure (approximately 70% of the Wdimed clusters contain a phenyl ring). To avoid that a simple benzyl ether or a related recurring structural element would be the common denominator of a cluster, we introduced more stringent clustering criteria for pairs of clusters containing phenyl rings. This kind of empirical correction works well in our hands, and it may be interesting to explore the usefulness of a list of substructural elements that should not be used to merge clusters. Another way of avoiding "uninteresting" clusters might be the use of a fuzzy clustering method that allows clusters to be members of several structurally related superclusters.

ES clustering has the advantage to yield tight clusters with a clearly defined centroid. On the other hand, it typically yields many separate clusters of compounds with the same scaffold. Whether or not these are merged to larger clusters by our procedure depends on the threshold values applied in steps 2−5. To assess the effect of the primary clustering method on the scaffold extraction process, we replaced the Daylight/ES clustering step by a Daylight/Jarvis−Patrick (JP) clustering step. We used identical Daylight fingerprint settings in both cases, a neighbor list length of 14 members; 8 common cluster members were required for cluster merging. As expected, we obtained fewer and larger clusters than with ES clustering (5209 clusters and 7444 singletons). However, the common substructures that can be extracted from these clusters are in many cases much smaller than

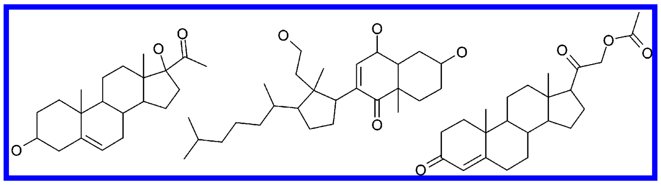

Figure 6. Members of a Daylight/ES Cluster Formed by Large Hydrocarbons with Few Functional Groups. No informative cluster scaffold can be extracted from these three structures.

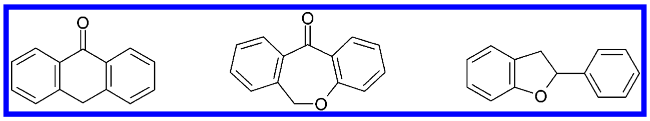

Figure 7. The central structure is related to each of the others through a different common substructure. Heterogeneous clusters can be formed in this way.

with ES clustering. An example is given in Figure 5: Several JP clusters contain mixtures of benzodiazepin(on)es and phenylbenzodiazepin(on)es, such that a pure phenylbenzodiazepine cluster cannot be generated any more. As a result, the JP clusters could either not be merged to larger clusters unless one would admit the formation of a much larger and very heterogeneous cluster. We conclude, therefore, that Daylight/ES clustering is a better basis than Daylight/JP for the subsequent clustering steps.

Finally, it should be mentioned that the clustering procedure we present here has two potential pitfalls. First, there are cases where heterogeneous sets of structures form one cluster in the Daylight/ES clustering step. This happens with compounds bearing the same functional groups attached to otherwise featureless but topologically different hydrocarbon skeletons. Daylight fingerprints are not suitable to distinguish between such structures, because they only note the presence or absence of bond paths, not their frequency. Heterogeneous clusters will not yield large and representative cluster scaffolds. In clustering the Wdimed library, this happens in particular with clusters containing steroids. Three sample structures from one Wdimed cluster are depicted in Figure 6. Fortunately, molecules of interest in medicinal chemistry typically consist of a mosaic of carbon and heteroatoms distributed over the entire structure, in which case different topological arrangements of the same functional groups yield distinct fingerprints. The second potential pitfall is the nontransitivity of maximum common substructure calculations. If molecules A and B are related through a large common substructure, then molecules B and C may be related through a different large common substructure. It follows that A and C may not be related at all. An example is given in Figure 7, where the central structure is related to each of others through a large structure incorporating both phenyl rings but different linker fragments between them. In practice, this problem plays a role only with particularly large compounds that can "mediate" similarity between two smaller ones and with polycyclic compounds.

CONCLUSIONS

Since the seminal work of Willett on chemical clustering methods,[28] much faster computer hardware has become available, allowing the application of more elaborate clustering methods to larger data sets. However, a full analysis of a large screening library based on MCS methods is still out of reach. The stepwise combination of fingerprint and MCS clustering thus represents a compromise between speed and accuracy. It is a chemically intuitive approach that will be useful whenever compounds with related substructures—not necessarily with related physicochemical properties—should be grouped together. The method is able to identify key substructure classes in chemical libraries, to select compounds related to hits for follow-up screening, and to understand relationships between chemical structures and biological activities obtained from HTS.

ACKNOWLEDGMENT

We thank Stefanie Bendels, Paul Gerber, Wolfgang Guba, Manfred Kansy, Connie Oshiro, and Olivier Roche for fruitful discussions on clustering methods and Daniel Parish for pioneering work on exclusion sphere algorithms at Roche.

REFERENCES AND NOTES

(1) Downs, G. M.; Barnard, J. M. Clustering methods and their uses in computational chemistry. *Rev. Comput. Chem.* **2004**, *18*, 1−40.

(2) Böcker, A.; Schneider, G.; Teckentrup, A. Status of HTS data mining approaches. *QSAR Comb. Sci.* **2004**, *23*, 207−213.

(3) Stanton, D. T.; Morris, T. W.; Roychoudhury, S.; Parker, C. N. Application of nearest-neighbor and cluster analyses in pharmaceutical lead discovery. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 21−27.

(4) Bajorath, J. Integration of virtual and high-throughput screening. *Nature Rev. Drug Discovery* **2002**, *1*, 882−894.

(5) Shanmugasundaram, V.; Maggiora, G. M.; Lajiness, M. S. Hit-directed nearest-neighbor searching. *J. Med. Chem.* **2005**, 240−248.

(6) Engels, M. F. M.; Thielemans, T.; Verbinnen, D.; Tollenaere, J. P.; Verbeeck, R. CerBeruS: A system supporting the sequential screening process. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 241−245.

(7) Engels, M. F. M.; Venkatarangan, P. Smart screening: Approaches to efficient HTS. *Curr. Opin. Drug Discuss. Dev.* **2001**, *4*, 275−283.

(8) Reynolds, C. H.; Druker, R.; Pfahler, L. B. Lead discovery using stochastic cluster analysis (SCA): A new method for clustering structurally similar compounds. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 305−312.

(9) Nicolaou, C. A.; Tamura, S. Y.; Kelley, B. P.; Bassett, 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.

(10) Holliday, J. D.; Rodgers, S. L.; Willett, P.; Chen, M.-Y.; Mahfouf, M. et al. Clustering files of chemical structures using the fuzzy k-means clustering method. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 894−902.

(11) Ott, T.; Kern, A.; Schuffenhauer, A.; Popov, M.; Acklin, P. et al. Sequential superparamagnetic clustering for unbiased classification of high-dimensional chemical data. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1358−1364.

(12) 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.

(13) McGregor, M. J.; Pallai, P. V. Clustering of large databases of compounds: Using the MDL "keys" as structural descriptors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443−448.

(14) Roberts, G.; Myatt, G. J.; Johnson, W. P.; Cross, K. P.; Blower, P. E. Leadscope: Software for exploring large sets of screening data. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1302−1314.

(15) Cross, K. P.; Myatt, G.; Yang, C.; Fligner, M. A.; Verducci, J. S. et al. Finding discriminating structural features by reassembling common building blocks. *J. Med. Chem.* **2003**, *46*, 4770−4775.

(16) Xu, J. A new approach to finding natural chemical structure classes. *J. Med. Chem.* **2002**, *45*, 5311−5320.

(17) Raymond, J. W.; Willett, P. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 521−533.

(18) Shen, J. HAD: An automated database tool for analyzing screening hits in drug discovery. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1668−1672.

(19) Jarvis, R. A.; Patrick, E. A. Clustering using a similarity measure based on shared nearest neighbours. *IEEE Trans. Comput.* **1973**, *C-22*, 1025−1034.

(20) Barnard, J. M.; Downs, G. M. Chemical fragment generation and clustering software. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 141−142.

(21) Butina, D. Unsupervised data base clustering based on Daylight's fingerprints and Tanimoto similarity: A fast and automated way to cluster small and large data sets. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 747−750.

(22) http://www. daylight.com.

(23) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have simialr biological activity? *J. Med. Chem.* **2002**, *45*, 4350−4358.

(24) Raymond, J. W.; Gardiner, E. J.; Willett, P. Heuristics for similarity searching of chemical graphs using a maximum common edge subgraph algorithm. *J. Chem. Inf. Comput. Sci*. **2002**, *42*, 305−316.

(25) http://thomsonderwent.com/products/lr/wdi/.

(26) http://www.xemistry.com.

(27) Flower, D. R. On the properties of bit string measures of chemical similarity. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 379−386.

(28) Willett, P.; Winterman, V.; Bawden, B. 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.

CI050011H