A Cluster-Based Strategy for Assessing the Overlap between Large Chemical Libraries and Its Application to a Recent Acquisition†

Michael F. M. Engels,*,‡ Alan C. Gibbs,|| Edward P. Jaeger,§ Danny Verbinnen,‡ Victor S. Lobanov,§ and Dimitris K. Agrafiotis§

Johnson and Johnson Pharmaceutical Research and Development, Division of Janssen Pharmaceutica, Turnhoutsweg 30, 2340 Beerse, Belgium, Johnson and Johnson Pharmaceutical Research & Development, L.L.C., 665 Stockton Drive, Exton, Pennsylvania 19341, and Johnson and Johnson Pharmaceutical Research and Development, L.L.C., 8 Clarke Drive, Cranbury, New Jersey 08512

Received May 29, 2006

We report on the structural comparison of the corporate collections of Johnson & Johnson Pharmaceutical Research & Development (JNJPRD) and 3-Dimensional Pharmaceuticals (3DP), performed in the context of the recent acquisition of 3DP by JNJPRD. The main objective of the study was to assess the druglikeness of the 3DP library and the extent to which it enriched the chemical diversity of the JNJPRD corporate collection. The two databases, at the time of acquisition, collectively contained more than 1.1 million compounds with a clearly defined structural description. The analysis was based on a clustering approach and aimed at providing an intuitive quantitative estimate and visual representation of this enrichment. A novel hierarchical clustering algorithm called divisive k-means was employed in combination with Kelley's cluster-level selection method to partition the combined data set into clusters, and the diversity contribution of each library was evaluated as a function of the relative occupancy of these clusters. Typical 3DP chemotypes enriching the diversity of the JNJPRD collection were catalogued and visualized using a modified maximum common substructure algorithm. The joint collection of JNJPRD and 3DP compounds was also compared to other databases of known medicinally active or druglike compounds. The potential of the methodology for the analysis of very large chemical databases is discussed.

INTRODUCTION

There is a constant need in the pharmaceutical industry to explore novel chemistry.[1] New compounds are typically obtained either through internal chemical synthesis (traditional or high-throughput) or external acquisition. The latter has become particularly popular in recent years, as it offers the means to greatly expand the size and diversity of a compound collection in a relatively short period of time. Compound acquisitions often entail major capital investments, and many pharmaceutical companies have established safe-guarding mechanisms to maximize the utility of the acquired chemicals in relation to their own internal efforts.[2] Such mechanisms involve the extensive use of chemoinformatic techniques, including substructure and duplicate screening;[3−5] similarity, diversity, and quantitative structure−activity relationship analysis;[6−14] and lead- or druglike profiling.[15,16] However, many of these methods become computationally challenging, if not intractable, when the size of the collections exceeds a certain threshold.

In March of 2003, Johnson & Johnson Pharmaceutical Research & Development L. L. C. (JNJPRD)[17] acquired 3-Dimensional Pharmaceuticals, Inc. (3DP), a small biopharmaceutical company with a substantial library of small molecules.[18] The availability of this library and its relation to the existing JNJPRD collection posed several interesting questions. Both collections can be seen as the embodiment of diverging chemical discovery strategies. The 3DP library is predominantly the product of the company's chemogenomics strategy,[19] embodied in the DirectedDiversity platform.[20] This strategy involved the rapid generation of small molecule compounds that could be used both as tools to probe biological mechanisms and as leads for drug property optimization. The majority of these compounds were synthesized in an iterative fashion using combinatorial chemistry methods and were selected from a massive virtual database of synthetically accessible analogues. This virtual library was mined using multiobjective selection techniques to ensure that the selected compounds were not only optimized for target binding affinity but also possessed druglike characteristics that would allow them to be used directly as tool compounds in appropriate cellular or biological model systems. Additional compounds synthesized during the course of active drug discovery programs augmented this core library.

In contrast, the JNJPRD library represents the union of two major pharmaceutical company collections (those of the Janssen Research Foundation and those of the Robert Wood Johnson Pharmaceutical Research Institute), which were, in turn, agglomerations of historical collections from McNeil, Ortho, and other companies, each with its own unique

† Parts presented at the Third Joint Sheffield Conference on Chemoinformatics, April 21−23, 2004, Sheffield, United Kingdom.

* To whom all correspondence should be addressed. Current address: Grünenthal GmbH, Zieglerstrasse 6, 52078 Aachen, Germany. Tel.: +49 (0) 241-569-2570. Fax: +49 241 569 2949. E-mail: michael.engels@grunenthal.com.

‡ Johnson and Johnson, Beerse, Belgium.

§ Johnson and Johnson, Exton, Pennsylvania.

|| Johnson and Johnson, Cranbury, New Jersey.

10.1021/ci600219n CCC: $33.50
Published on Web 10/13/2006

perspective on the discovery of pharmacologically active agents. Therefore, the JNJPRD collection represents a rather diverse repository of compounds that have been produced for hit identification, lead exploration, and lead optimization but also for pharmacological tooling, chemical technology exploration, and other purposes.

The practical integration of these two chemical libraries was a complex task that involved several business-critical systems within our research organization. Besides the chemical registration and compound logistics systems, which were obviously affected first, the impact on secondary dependent systems such as external compound acquisition and high-throughput screening also had to be considered. To effectively assess that impact, questions such as "what is the number of duplicates?", "how diverse is the new collection?", "how well does it complement the existing collection?", and "to what extent does it explore druglike or biologically active space?" needed to be addressed. This information was used to support intelligent priority setting and resource planning for the laborious and time-consuming integration process, to deliver maximum incremental value to the organization (e.g., by determining the order in which these compounds should be registered and added to the global dispensary or by identifying the subset that should be included in the standard screening library for routine biological testing). To be effective, this type of analysis had to be executed in a timely manner and expressed using intuitive metrics.

In response to these requirements, we developed a procedure that could cope with the size and complexity of the data at hand and produce results that would be easily interpretable by a medicinal chemist. Our analysis was based on hierarchical clustering, a method with a long history in the analysis of chemical data sets.[21,22] Hierarchical clustering offers two significant advantages over alternative methods: (1) it provides a compact representation of chemical space by aggregating closely related molecules into clusters, and (2) it preserves the detailed relationships between these molecules in the form of a tree or dendrogram.[23] Despite these advantages, the application of hierarchical clustering methods to the analysis of large chemical data sets has not been possible because of their severe computational cost.[3,6]

In the present study, these limitations were mitigated through the use of a novel hierarchical clustering algorithm called divisive k-means, an algorithm originally introduced within the text-mining community[24] and later adopted within the chemoinformatics field.[25−27] Divisive k-means belongs to the class of divisive hierarchical clustering methods, which operate by progressively breaking the original data points into smaller and smaller clusters. The application of this algorithm in combination with Kelley's cluster-level selection method resulted in a simple description of the chemical space occupied by the JNJPRD and 3DP collections and served as a surrogate for subsequent diversity comparisons. Chemotypes representing the maximum common substructure present in the majority of compounds in each cluster were extracted and catalogued, providing a more intuitive interface to the medicinal chemists. Finally, the compatibility of the 3DP and JNJPRD databases to pharmacologically active and druglike space was investigated by mapping the compounds from the World Drug Index[28] onto the existing clusters, using similarity to the cluster centroids as a means for assigning membership. This report addresses only the first part of the problem. The second, namely, the prioritization of 3DP compounds for transfer into the JNJPRD compound inventory and the selection of compounds for inclusion into the corporate screening deck, will be described in a subsequent publication. These results are presented in the hope that future endeavors of this kind, both at JNJPRD and at other pharmaceutical companies, will benefit from the knowledge and experience gained during this exercise.

Table 1. Database Sizes

database	size
JNJPRD	729 829
3DP	403 528
WDI	68 780
USAN	2302

METHODS

Conceptual Framework. Our analysis aimed at obtaining three basic metrics of the relative content of the 3DP and JNJPRD collections: (1) the number of duplicate structures, (2) the overlap in chemical space, and (3) druglikeness.[29,30] Prior to the actual analysis, significant work had to be carried out in order to extract the appropriate compounds from the underlying chemical registration systems and put them in a common frame of reference. Data from the JNJPRD and 3DP collections were cleaned, normalized and canonicalized, and combined into a single database. Further details are provided in the Databases and Database Preparation section.

A central component of our approach was to partition the combined collection into a family of clusters and study the relative population and chemical composition of these clusters. The clustering was performed using Barnard Chemical Information's (BCI) implementation of the divisive k-means clustering algorithm,[31] and an optimal set of clusters was derived from the cluster hierarchy using Kelley's cluster-level selection method.[32−34] A population analysis of the optimal set of clusters led to the identification of overlapping regions between the two collections and to an assessment of the diversity contribution of each database to the combined cluster space. In addition, chemotypes novel to the JNJPRD collection were identified using an approximate maximum common substructure algorithm and were catalogued and depicted. The final part of the analysis aimed at assessing the distribution of the JNJPRD and 3DP collections with respect to the pharmacological and druglike chemical space represented by the World Drug Index and USAN libraries, respectively.

Databases and Database Preparation. Table 1 lists the sizes of the databases used in this study. The JNJPRD and 3DP corporate databases were obtained in SD file format from their respective chemical registration systems. Prior to comparison, the two collections were cleansed of partner data, for example, compounds that were not exclusively owned by either company. Although both chemical registration systems supported stereochemical annotation, there were slight differences in their conventions, and a direct one-to-one mapping could not be easily obtained. To minimize ambiguity, we decided to ignore stereochemical information and focus exclusively on the raw connection tables.

Also listed in Table 1 is the Word Drug Index (WDI), which at the time of the 3DP acquisition represented an

electronic catalog of about 76 000 drugs and pharmacologically active compounds, including all worldwide marketed drugs.[28] The compounds in this database have often been used to define "druglike" chemical space.[29,30] The database was obtained in SD file format and contained 68 470 unique entries, following the elimination of 7818 entries for which no structural information was available. To meet the normalization criteria of the JNJPRD and 3DP databases (see below), counterions were removed from the connection tables of the WDI entries. If the removal of counterions resulted in the generation of charged conjugated acids or bases, the corresponding groups were protonated or deprotonated, respectively, to obtain the neutral form.

A subset of 2302 entries was extracted from the normalized WDI database according to the procedure described by Lipinski et al.[35] This procedure, composed of both textual and structural filters, was used to identify compounds with preferred absorption behavior. Throughout this analysis, we refer to these 2302 compounds as the USAN database, following nomenclature used by Lipinski et al.[35]

To simplify data management and storage, each compound was converted into a SMILES string[36] and characterized by a 2048-bit Daylight fingerprint,[37] which was also used as input for the clustering calculations. Cleaning and refinement procedures were performed by scripts developed in-house based on DirectedDiversity[19,20] and Daylight toolkit functions.[38]

Duplicate Analysis. The number of unique structures in each database was calculated by (1) normalizing each molecule, (2) enumerating all tautomeric forms and deriving a canonical tautomeric SMILES for each compound,[39] and (3) counting the number of unique canonical tautomeric SMILES in each database. The number of unique canonical tautomeric SMILES resulting from this procedure corresponds to the number of different chemical entities in the respective database, and the difference between this number and the original number of input records represents the number of superfluous duplicates. Normalization included the adjustment of the protonation state (i.e., by protonating and deprotonating charged acidic and basic groups into their respective neutral forms) and the removal of smaller fragments (e.g., counterions). Cross-database duplication rates were obtained by computing the number of unique canonical tautomeric SMILES that were common between the two databases. Because stereochemical information was removed prior to this analysis, the reported duplication rate represents an upper bound of the true duplication rate between the two collections.

A duplicate analysis was performed using two different software packages: (1) Scitegic's Pipeline Pilot platform[40] and 3DP's DirectedDiversity suite.[20] As expected, the results were not identical, because the two systems employ different molecular perception and canonicalization algorithms. Compounds that were identified to be unique by one method and not unique by the other were set aside and visually inspected.

Divisive k-Means Clustering. The divisive k-means clustering method,[25,27,41] also known as bisecting k-means,[24] was used because of its favorable scaling characteristics. Divisive k-means is a hierarchical clustering method, which originated from the text-mining community for the processing of large sets of documents.[24] In its first described application,[24] the algorithm was projected to have a time complexity in the range of $O(N)$ to $O(N \log N)$, where N is the number of points, which makes it applicable to large data sets. By contrast, very efficient implementations of other hierarchical clustering methods such as Ward's algorithm[42] exhibit quadratic time complexity and can only be used with data sets of moderate size.[23] Because BCI's implementation of the divisive k-means clustering algorithm is not well-known in the chemical literature, we provide a brief outline of the method below.

The algorithm starts with a single cluster containing all of the points in the data set. This cluster is then split into two clusters, followed by further splitting. The cluster with the largest number of points is typically selected for splitting, although other criteria have also been suggested. In the original publication,[24] this step is repeated several times using different random seeds, and the split that generates the tightest subclusters is selected and added to the cluster hierarchy. These two steps, selecting an appropriate cluster for splitting and the splitting itself, are repeated until all clusters have become singletons or until some other termination criterion is met (see Kelly's algorithm below). In the BCI implementation,[31] the multiple seeding iterations are replaced by a single deterministic step, which uses two extreme points as seeds. These are the point most distant to the cluster centroid and the point most distant to that. The advantage of this approach is that it always results in the same cluster hierarchy.

The inner k-means loop involves an iterative refinement of the positions of the cluster centroids, starting from the original seeds. A cluster centroid represents the geometric mean of the points in that cluster. Throughout all iterations, each point is assigned to the nearest centroid and the positions of the two centroids are recomputed. This process is repeated until the centroids (and cluster memberships) no longer change. In the present study, we treat the 2048-dimensional binary fingerprints as real vectors and use the (squared) Euclidean distance to define the centroids. Thus, the centroids represent 2048-dimensional fractional bit sets, where each "bit" is a real value in the interval [0,1]. There are two major disadvantages to this approach, namely, the significant computational cost of distance calculations and the fact that the Euclidean distance is an unfamiliar metric for binary fingerprints (as opposed to, e.g., the Hamming or Tanimoto coefficients). An improved algorithm that addresses both of these limitations will be presented elsewhere.

The cluster analysis was performed on the combined JNJPRD and 3DP collections, totaling more than 1.1 million compounds with defined chemical connection tables. To our knowledge, this is the largest chemical data set that has been submitted to hierarchical clustering. Calculations were performed on an SGI Origin 300 using one 500 MHz MIPS processor. The divisive k-means calculation of the cluster hierarchy took 16 days of CPU time. In the meantime, further significant improvements have been made at our group and externally (G. Downs, private communication) in optimizing the original divisive k-means algorithm, reducing calculation times significantly.

Cluster-Level Selection Method. In hierarchical clustering, each level of the hierarchy defines a possible partitioning of the data set into a set of clusters. Given the complexity of the hierarchy and the size of data set, we applied Kelley's cluster-level selection method[32] to obtain an optimal set of

clusters. This method calculates, for all levels of the hierarchy, a penalty value that is a tradeoff between intracluster and intercluster variance.[32] The level of hierarchy that has the lowest penalty value is selected and used to partition the data set into the corresponding number of clusters.

Kelley's method has been validated for its ability to group chemical data sets into "natural" clusters and has been found to be superior to other methods, particularly when used in combination with high-dimensional binary representations such as Daylight fingerprints.[37] It is also computationally efficient, showing a time complexity of less than $O(N^2)$, which makes it applicable to large data sets.[33] In this work, we used the OPTCLUS implementation of the Kelley algorithm.[34]

Centroid Abstraction. The cluster centroids were used in the next phase of the analysis to summarize the contents of the clusters, as described by Downs and Barnard.[23] Three parameters were used to define the location and size of each cluster: (1) its centroid, (2) its extension, and (3) its radius.

Given a cluster comprised of N points in some vector space, $\{x_i, i = 1, ..., N\}$, the *centroid*, $\bar{x}$, is defined as their geometric mean:

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

The *extension* of a cluster is defined by the intracluster variance, v, which represents the average distance to the centroid $\bar{x}$:

$$v = \frac{\sum_{i=1}^{N} d(x_i,\bar{x})}{N}$$

where $d(x_i,\bar{x})$ is the distance between x_i and $\bar{x}$ [here, we used the squared Euclidean distance $d(x_i,\bar{x}) = (x_i - \bar{x})^2$].

Finally, the *radius* of the cluster, r, is defined as the maximum distance to the cluster centroid:

$$r = \max_{i-1}^{N}[d(x_i,\bar{x})]$$

These three parameters were used as a fast method to determine whether an external compound belonged to any of the clusters of the optimal set. An external compound was assigned to an existing cluster if the distance between the molecule and the cluster centroid was smaller than the radius of that cluster. If more than one cluster fulfilled that criterion, the compound was assigned to the cluster of the nearest centroid. If a molecule fell outside the radius of all existing clusters, it was considered diverse with respect to the existing collection and thought to occupy novel chemical space. Singletons, whose radius is by definition zero, were assigned an artificial radius equal to the average radius of all of the clusters in the optimal set.

Cluster Analysis. Several features were calculated for the clusters in the optimal set; these include (1) the size of the clusters and its distribution, (2) the intracluster variance and its distribution, (3) the cluster radius, and (4) the intracluster Tanimoto index. The intracluster Tanimoto index, T, is defined as the average Tanimoto distance between all pairs of compounds within a cluster:

$$T = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j<i}^{N-1} t(x_i,x_j)$$

where $t(x_i,x_j)$ is the Tanimoto coefficient between the ith and jth compounds in that cluster.

To assess the overlap between the 3DP and JNJPRD databases, we used a population analysis similar to the one introduced by Shemetulskis et al.[6] For each cluster, the ratio between 3DP and JNJPRD compounds belonging to that cluster was determined using the compound identification tags. Clusters were divided into three categories: (1) those populated exclusively by JNJPRD compounds, (2) those populated exclusively by 3DP compounds, and (3) those occupied by both JNJPRD and 3DP compounds. This categorization served as the basis for identifying complementary and overlapping regions of chemical space between the two databases.

Chemotype Extraction. A maximum common substructure (MCS) procedure was devised to offer insight into the chemical composition of these clusters and provide a list of novel chemotypes embodied in the 3DP collection. A chemotype is defined as a structural motif that is present in the majority of compounds within a cluster or group of related clusters. The MCS procedure was applied to only those clusters in the optimal cluster set which consisted exclusively of 3DP compounds. The calculations were carried out using the approximate MCS method in Pipeline Pilot.[43] This method is similar to the modal fingerprint approach by Shemetulskis et al.[44] and scales linearly with the number of compounds, in contrast to classical (exact) MCS approaches.[45,46] Clusters with less than four members were discarded from this analysis. The resulting maximum common substructure represents the largest structural fragment that is found in at least 80% of the members of a cluster. To obtain meaningful chemotypes of certain size, substructures consisting of less than eight atoms were discarded. This procedure resulted in a series of different but very closely related substructures of dubious value to medicinal chemists. Therefore, the obtained maximum common substructures were submitted to an additional refinement procedure, the removal of aliphatic carbon chains and substituents on aromatic rings (via the Daylight toolkit[38]). Substituents on heteroaromatic rings were retained. A canonical representation was generated for these refined MCSs, and duplicates were removed. The resulting canonical MCSs represent novel chemotypes in the 3DP collection that complement those already present in the JNJPRD library. To ensure the uniqueness of these chemotypes in the context of the JNJPRD collection, each chemotype was queried as a substructure against the JNJPRD corporate collection. If a chemotype was present in any molecule in the JNJPRD collection, it was removed from the set of novel chemotypes.

RESULTS AND DISCUSSION

Duplicate Analysis. Because the representation of stereochemistry and other chemical business rules differed between

Table 2. Structural Overlap between the JNJPRD and 3DP Collections

	number of entries	number of unique structures	
database	unprocessed input data	stereochemistry truncation	full normalization
JNJPRD	729 829	719 434	710 522
3DP	403 528	402 460	402 386
JNJPRD+3DP	1 113 357	1 099 205	1 086 934

the two chemical registration systems, we examined the effect of different normalization procedures on the individual databases as well as the combined collection. Table 2 summarizes the results of these calculations.

The two libraries collectively contain more than 1.1 million entries. When stereochemistry was ignored, the number of unique structures dropped by approximately 14 000 to 1 099 205 structures. This reduction is largest in the JNJPRD database, which indicates a greater abundance of compounds with known, absolute or relative, stereochemistry. The number of unique entries dropped further when the compounds were normalized. Normalization involved the stripping of smaller fragments and the neutralization of acids and bases, the removal of stereochemical information, and the generation of canonical tautomers. Again, the number of unique compounds in the JNJPRD collection was reduced significantly by this procedure, dropping by approximately 19 000 entries from 719 434 to 710 522 entries. Next to the removal of stereochemistry, generating canonical tautomeric forms had the most drastic impact on the number of unique entries. The stripping of smaller fragments had a very small impact, because both registration systems use special codes to represent different salt forms and do not include counterions in the chemical structure.

The number of topologically equivalent duplicates shared between the two collections was estimated to be between 22 689 (402 460 + 719 434 − 1 099 205) and 25 974 (402 386 + 710 522 − 1 086 934). The number of superfluous duplicates in the 3DP library that overlap with compounds in the JNJPRD collection was slightly higher, estimated to be between 23 757 (22 689 + 403 528 − 402 460) and 27 116 (25 974 + 403 528 − 402 386). This corresponds to a duplication rate of 5.9−6.7% between the 3DP and JNJPRD databases, which is comparable to duplication rates of allegedly unrelated databases.[3] The origins of duplicates between the two databases were traced by source tags, via their respective chemical information systems. Not surprisingly, the majority of duplicates were acquired from external chemical suppliers, suggesting that the two companies followed similar compound acquisition strategies in the past.

The extent of structural novelty added to the JNJPRD collection from unique 3DP compounds is approximated at being between 379 771 (403 528 − 23 757) and 376 412 (403 528 − 27 116) entries. The following two points illustrate why these numbers are approximations. First, because of the removal of stereochemical information, the number of real duplicates will be smaller; as a consequence, the number of potential new entries will be higher. Second, diverging internal rules, implementations of chemoinformatic algorithms, and the robustness of the compound registration systems could have an impact on these estimates. Indeed, cross comparisons between three different chemoinformatics

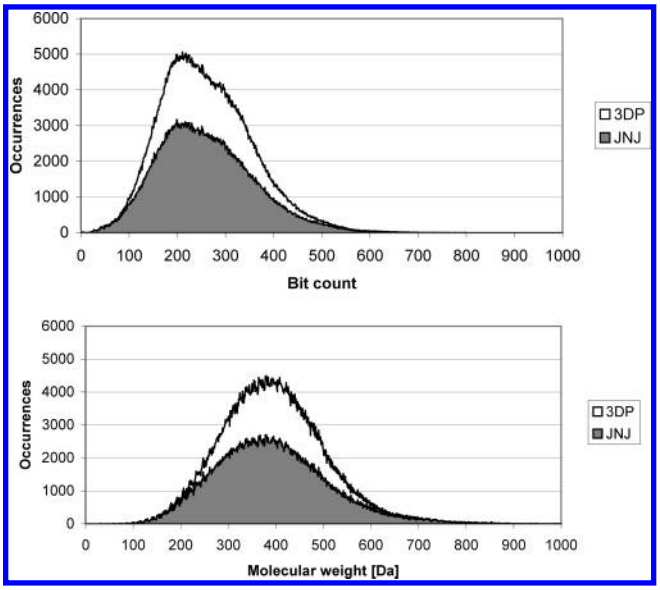

Figure 1. Distributions of fragment bit count (a) and molecular weight (b) in the 3DP and JNJPRD databases. The individual distributions of the 3DP and JNJPRD are shown in light and dark gray, respectively.

systems, Pipeline Pilot,[40] DirectedDiversity,[20] and the Daylight toolkit,[38] showed slight differences. While the Daylight toolkit and Pipeline Pilot calculated 402 460 unique elements for the stereochemically truncated 3DP data set, the DirectedDiversity toolkit computed 402 473 unique entries. Divergent aromatic perception algorithms are the primary cause of this discrepancy. For the type of approximate analysis presented here, this difference is negligible; nevertheless, this information was kept for final compound registration purposes.

Molecular Weight and Bit Count Profiles. As mentioned, the 3DP and JNJPRD collections can be seen as the embodiment of diverging drug discovery strategies. While 3DP's strategy employed combinatorial chemistry methods to a large extent, JNJPRD's strategy has been more traditional, though technologically very diverse. Of interest in this study was whether this difference was manifested, for example, in the utility and density of chemical fragments. Daylight fingerprints were used as computationally efficient surrogates of chemical fragments. These fingerprints characterize molecules by 2048-bit strings; each bit indicating the presence (1) or absence (0) of at least one linear substructure fragment.[37] Substructure fragments consist of one to eight atoms. Therefore, the bit count of a molecule is a summary of the density of these fragments within a structure. The higher the bit count is, the greater the number and diversity of structural motifs within a molecule. Because the bit count grows with the molecule size, the bit count is best presented in relation to the molecular weight of a molecule.

From the histograms in Figure 1a and b, it appears that the distributions of molecular weight and bit count are quite comparable in both databases. The average molecular weight within the JNJPRD collection is 392 Da, similar to the average molecular weight of the 3DP collection of 400 Da. Likewise, the average bit count for molecules in the JNJPRD collection is with 259 bits almost equivalent to the average bit count in the 3DP collection of 264 bits. However, while the increase in bit count does not lead to an increase in

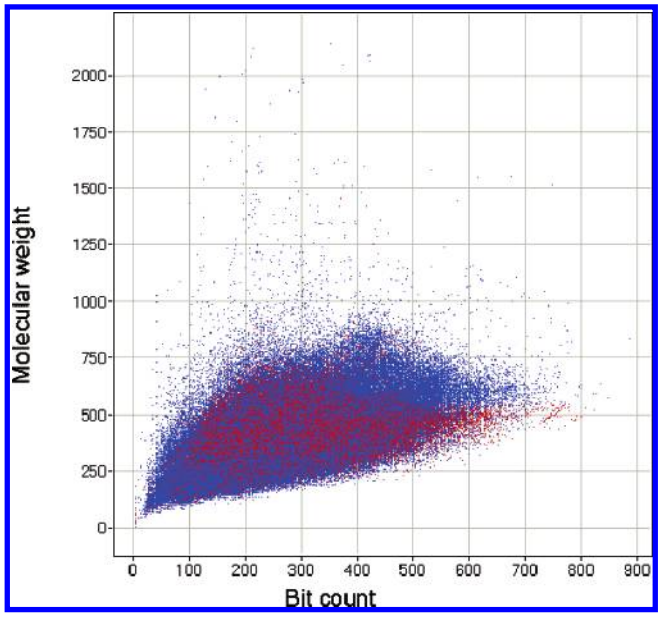

Figure 2. Distribution of molecular weight and fragment bit count for the 1.1 million compounds of the combined data set. JNJPRD and 3DP entries are indicated in blue and red, respectively.

molecular weight for 3DP compounds, the increase in bit count is accompanied by an increase in molecular weight for JNJPRD compounds (see Figure 2). This implies that the exploration of pharmacophore variety in the JNJPRD collection has been pursued by combining a limited number of small structural fragments, leading to an increase of the molecular weight. This trend, relatively speaking, is not observed in the 3DP collection. It remains the subject of further analyses whether this observation is linked to the diverging chemical discovery strategies of the two organizations or is due to the fingerprint representation, which artificially puts constraints on the size of fragments.

Overlap Analysis. One key attribute of Kelley's cluster-level selection method is that it provides a partitioning which optimally balances intracluster and intercluster variances for a given clustering method and structure representation.[32,33,47] In that sense, we can use the suggested number of clusters as an approximate measure of a compound collection's structural diversity, or better, the extent of local exploration of different structural motifs. The Kelley plot (Figure 3) illustrates that the optimal partitioning of the combined data set is between 89 570 and 89 630 clusters. The minimum penalty value is observed at 89 615 clusters, and that was the number used for our subsequent analyses. We refer to this set of clusters as the optimal set of clusters.

The Venn diagram in Figure 4 illustrates the distribution of the two collections within the optimal set of clusters. The JNJPRD collection provides a relatively greater contribution to the cluster-based chemical space than the 3DP collection. JNJPRD compounds are observed in 73 768 clusters, which correspond to 82% of all clusters, whereas 3DP compounds are distributed over 44 562 clusters, which correspond to 50% of all clusters. The number of clusters occupied exclusively by JNJPRD and 3DP compounds is 45 043 and 15 837, respectively (see Table 3).

A total of 32% of the clusters are mixed clusters, containing both JNJPRD and 3DP compounds. Approximately 50% of the 3DP compounds have been assigned to

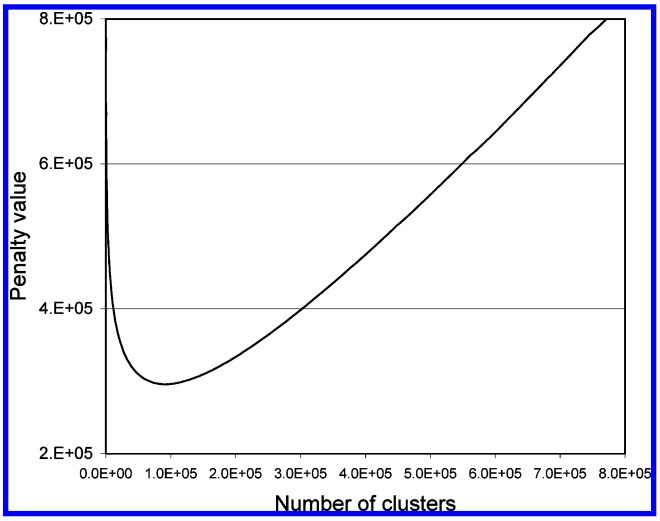

Figure 3. Kelley plot describing the course of Kelley's penalty value as a function of the number of clusters.

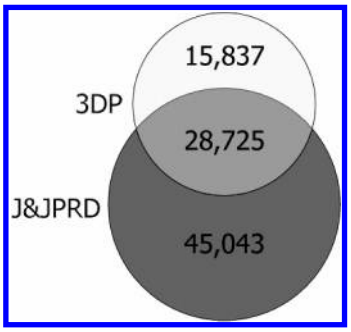

Figure 4. Venn diagram illustrating the number of clusters occupied exclusively by JNJPRD (dark gray) and 3DP (light gray) compounds and those occupied by both (medium gray).

Table 3. Extent of Overlap and Unique Contributions of the JNJPRD and 3DP Collections[a]

database segment	number of clusters	number of singletons
3DP clusters	15 837	1118
JNJPRD clusters	45 043	4628
mixed clusters	28 725	NA
sum	89 615	5746

[a] 3DP clusters refers to clusters consisting of 3DP compounds only; JNJPRD clusters refers to clusters constituted exclusively by JNJPRD compounds; mixed clusters refers to clusters constituted by JNJPRD and 3DP compounds; NA = not available.

this overlapping region. This is significantly greater than one would expect on the basis of the duplicate analysis, which showed that approximately 23 000 compounds have identical structures. This divergence between structural similarity and structural redundancy can be explained by a high degree of internal similarity within the 3DP database.

Around 6% of the clusters in the optimal cluster set are singletons. Singletons are clusters that contain a single member and represent outliers in a given partitioning. The percentage of outliers within the JNJPRD database is significantly greater than in the 3DP database, again, demonstrating the higher structural compactness of the 3DP collection.

Mixed clusters are the largest and exhibit the highest degree of local structural diversity. More than 430 000 compounds are grouped into these clusters. Each of these clusters contains on average 15.1 members, and their average

Table 4. Average Properties of Clusters

category	average cluster size[a,b]	average intracluster Tanimoto index[b,c]	average intracluster variance[d]
3DP clusters	12.6 ± 7.2	0.87 ± 0.07	19.9 ± 12.4
JNJPRD clusters	11.3 ± 7.6	0.76 ± 0.15	38.9 ± 31.6
mixed clusters	15.1 ± 6.6	0.73 ± 0.15	41.1 ± 32.9

[a] Cluster size corresponds to the number of members within a cluster. [b] Singletons excluded. [c] Intracluster Tanimoto index corresponds to the average of pairwise Tanimoto distances within a cluster. [d] Cluster variance is defined as the sum of squared Euclidean distances of all cluster members to the centroid of the cluster.

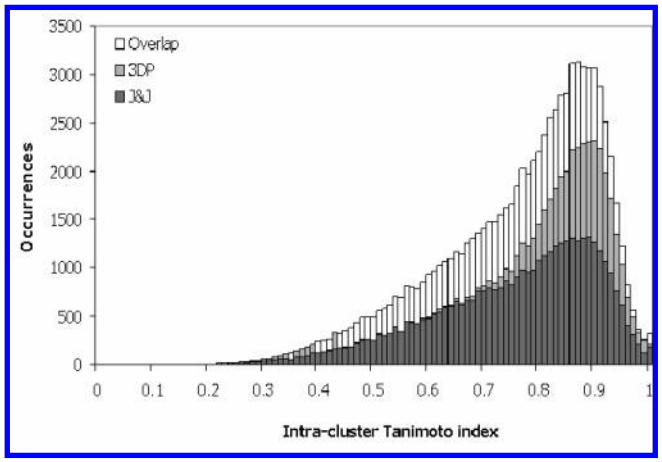

Figure 5. Compactness of clusters expressed as the intracluster Tanimoto index. The histogram shows clusters located in the overlapping region (white) and clusters occupied exclusively by JNJPRD (light gray) and 3DP (dark gray) compounds.

intracluster Tanimoto index is 0.73 (see Table 4 and Figure 5). In contrast, clusters populated exclusively by 3DP compounds are very dense. Their average intracluster Tanimoto index is 0.87, which is significantly higher than that of mixed clusters and clusters populated exclusively by JNJPRD compounds. In addition, as Figure 5 shows, the range of intracluster Tanimoto indexes explored in a typical 3DP cluster is much narrower.

As with the duplicate analysis, the origin of the 3DP compounds in the mixed and 3DP-exclusive clusters was traced. Again, the majority of compounds in the overlapping region originated from external compound acquisition campaigns at 3DP. Compounds originating from in-house combinatorial chemistry programs are mainly found in the 3DP-exclusive clusters.

Cataloging Novel 3DP Chemotypes. One of the key goals of this study was to determine the extent to which the 3DP library contributes to the diversity of the JNJPRD corporate library in terms of novel chemotypes. In the previous section, we identified clusters that complement the JNJPRD collection. Nevertheless, this representation remained quite intangible to our medicinal chemists. Therefore, a catalog of chemotypes was derived from the cluster analysis, which also included their chemical depiction.

In total, around 8000 different chemotypes were obtained after cleaning, filtering, and final validation against the JNJPRD library. This is significantly less than the number of clusters occupied exclusively by the 3DP collection (15 567). However, many of these clusters consisted of less than four compounds or resulted in chemically equivalent chemotypes after application of the MCS and subsequent refinement procedures. More than 98% of these chemotypes could be represented as valid SMILES strings. This is a surprisingly high percentage, given the fact that most of these chemotypes were intended to encode substructures rather than exact structures. Structural variation within those clusters involved primarily acyclic aliphatic functionalities; ring scaffolds remained largely invariant.

A database of the chemotype SMILES along with information associated with their pharmacological activities was created and was indexed for substructure and similarity searching. The combination of structure and pharmacological activity was intended to provide JNJPRD medicinal chemists with a practical tool for assessing the content of the 3DP library.

Table 5. Statistics on Molecular Weight and Bit Counts

database	molecular weight (Da)	bit count
JNJPRD	392 ± 122	259 ± 99
3DP	400 ± 89	264 ± 87
WDI	451 ± 342	234 ± 127
USAN	292 ±120	193 ± 117

Table 6. Overlap of the WDI and USAN Libraries with the Joint Collection of 3DP and JNJPRD Databases

	WDI		USAN	
category	compounds	clusters	compounds	clusters
JNJPRD	29 141	4916	1167	641
3DP	93	57	7	6
overlap	27 040	4981	1056	531
none	12 196		72	

Overlap with Biologically Active Chemistry Space. To determine whether either corporate library contained pharmacologically active entities or entities with preferred intestinal absorption potentials, the combined collection was compared to two reference databases of bioactive substances: (1) the WDI and (2) the USAN library. The WDI represents a database of pharmacologically active compounds, which are either on the market or in development. The USAN library is a 2300-compound subset of the WDI, representing not only biologically active molecules but also molecules with desirable absorption behavior.

Table 5 summarizes the distribution of molecular weight and bit count of the four databases. The WDI and the USAN libraries strongly deviate in both properties from the 3DP and JNJPRD collections. The average molecular weight of WDI is at least 50 Da higher; however, the bit count is 25 bits lower. On the other hand, the average molecular weight of molecules in the USAN library is 292 Da, approximately 100 Da lower than either the JNJPRD or 3DP libraries. This observation can only be rationalized by the fact that the WDI consists of a very heterogeneous set of pharmacologically active compounds covering different substance classes, from relatively small molecules such as drugs, peptides, steroids, and metal-containing agents to large biopolymers such as proteins and DNA. The relatively low average bit count of the WDI compounds may be explained by the fact that these biopolymers are constructed from a limited set of building blocks, which do not contribute much to chemical fragment diversity encoded by the local bit strings. Compounds in the USAN subset of the WDI library are significantly smaller

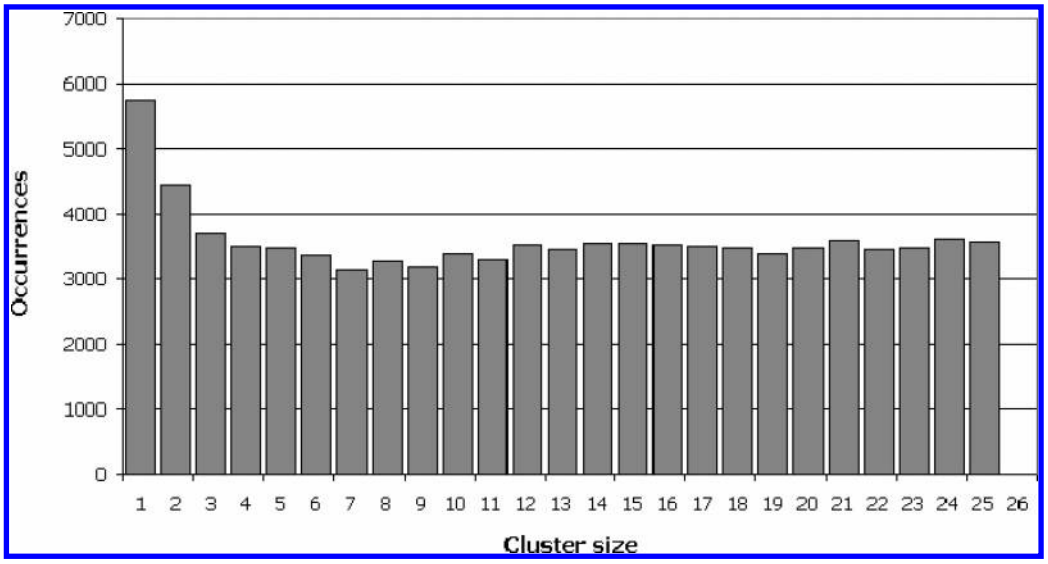

Figure 6. Distribution of cluster sizes for the preferred set of 89 615 clusters.

and have lower bit counts. To quantitatively assess the structural overlap between the 3DP/JNJPRD and WDI/USAN libraries, we used the centroid method described in the Methods section. The overlap was estimated by counting the number of WDI or USAN compounds that were assigned to the clusters occupied by the JNJPRD and 3DP collections, and the number of clusters that had WDI/USAN compounds assigned to them.

Overall, there is a high degree of overlap between the cluster space spanned by the 3DP and JNJPRD databases and that of the WDI and USAN libraries (Table 6). More than 80% of the 68 470 WDI compounds were assigned to a cluster occupied by the in-house collections. There is a clear preference for WDI compounds to overlap with clusters occupied exclusively by JNJPRD compounds. A total of 93 WDI compounds were assigned to clusters occupied exclusively by 3DP compounds. Compounds in these clusters are primarily products of combinatorial synthesis. The number of clusters occupied exclusively by the WDI database is relatively small; only 12 196 WDI compounds could not be assigned to any of the pre-existing clusters and are therefore diverse with respect to the in-house collections. Similar results are obtained for the USAN library. More than 95% of the USAN compounds are represented by similar analogues in the JNJPRD and 3DP collections, and only 72 could not be assigned to any pre-existing JNJPRD/3DP clusters.

Divisive k-Means Clustering. In this study, we present a framework for comparing large chemical libraries. The approach involves partitioning the compounds into separate families of chemotypes and using the number of families as a measure of diversity. An external library is compared to a reference collection (in this case, the 3DP and JNJPRD libraries, respectively) by clustering their union and computing the number of clusters introduced by the external library. An advantage of clustering methods is that they are intuitive to experts and nonexperts alike. They allow one to easily express an increase in diversity in terms of the number of new "chemotypes" (clusters) introduced by the external library, which is a far more appealing alternative than, for example, a reduction of, for example, 0.01 in the mean nearest neighbor distance in some abstract high-dimensional descriptor space. This advantage, however, comes at a significant computational cost. Clustering methods, and in particular hierarchical ones, are notoriously slow and do not scale well with the size of the data set.

We chose to employ the divisive k-means clustering approach because of its attractive time complexity, which was reported by Steinbach to be between $O(N)$ and $O(N \log N)$. However, we cannot confirm these optimistic projections. Internal benchmarking studies of the divisive k-means implementation by BCI using increasing subsets of the NCI database[48] showed a time complexity close to $O(M \log N)$ with $M \sim N$. Although we are not familiar with the internals of the BCI implementation, we believe that this is due to the use of reciprocal nearest neighbors.

We note that the quality of the clusters generated by divisive k-means differs significantly from those obtained by Ward's algorithm. This is based on a visual inspection of several clusterings obtained in a structure−activity relationship analysis of high-throughput screening results, which are typically processed by Ward's hierarchical algorithm.[42] Divisive k-means tends to generate clusters of relatively uniform size (see Figure 6), while regular k-means or Ward's clustering tends to produce clusters of widely different sizes.[48] Smaller clusters are often of higher quality, but this does not contribute much to the overall quality measure because each cluster's contribution is weighed by it size. Larger clusters, on the other hand, tend to be of lower quality and make large negative contributions to cluster quality.

CONCLUSIONS

The present analysis evolved from an initial interest to develop robust strategies for characterizing the overlap between two compound collections. These strategies were developed not only to simplify the integration of large external collections into our corporate file but also to support the evaluation of smaller libraries before they are purchased from external suppliers. In response to that need, several strategies have been developed, some of which have been presented in this study. Central to our approach is the use of a clustering method that can work on a large scale.[22] Unlike

the most commonly used Ward's,[42] k-means,[48] and Jarvis−Patrick[49] clustering methods, the bisecting k-means clustering method implemented in the BCI clustering package provided the means to perform these studies in a tolerable time frame.

A population analysis of these clusters provided a simple means for assessing structural overlap and complementarities. On the basis of this analysis, we identified approximately 27 000 duplicates and 200 000 close analogues in the 3DP collection, most of which could be traced back to external chemical suppliers. Clear structural novelty with respect to the JNJPRD collection was identified in a specific combinatorial chemistry subset of the 3DP collection. On the basis of the cluster description, it was estimated that this subset will extend JNJPRD's corporate chemical space by ∼20%. More importantly, this subset is rich in novel chemotypes, which makes it particularly attractive in the highly congested intellectual property environment in which modern pharmaceutical companies must operate. The methods employed herein can be easily extended to support the design of diverse, focused, or tailored libraries for high-throughput screening and the selection of compounds for filling in the depleted portions of a corporate collection.

Finally, we note that there is an increasing demand in analyzing and exploring large chemical data sets. The availability of ZINC,[50] a free database of commercially available compounds for virtual screening, stresses the importance of tools that can cope with the size and complexity of nowadays data sets. The BCI implementation of the bisecting k-means clustering demonstrates its advanced level of applicability and robustness. Nevertheless, this study also showed that there is room for improvement, in particular with regard to speed performance. Efforts for improving the speed performance of the bisecting k-means algorithm are already underway and will be presented shortly.

ACKNOWLEDGMENT

We thank Geoff Downs and John Barnard from Digital Chemistry, Ltd., formerly Barnard Chemical Information, Ltd., for early access to the divisive k-means program and Tom Tabruyn and Yves Wetzels from Ordina for scripting and database management. In addition, we would like to acknowledge discussions with several JNJPRD and 3DP colleagues during this project, including Frederik Deroose, Frank Brown, Peter Roevens, Eddy Freyne, Trevor Howe, Mark Player, Carl Illig, Nalin Subasinghe, Jim Rinker, Roger Bone, and F. Raymond Salemme.

REFERENCES AND NOTES

(1) Editorial. *Nat. Rev. Drug Discovery* **2004**, *3*, 375.
(2) Dunbar, D. B. Compound Acquisition Strategies. *Pac. Symp. Biocomput.* **2000**, 555−565.
(3) Voigt, J. H.; Bienfait, B.; Wang, S.; Nicklaus, M. C. Comparison of the NCI Open Database with Seven Large Chemical Structural Database. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 702−712.
(4) Baurin, N.; Baker, R.; Richardson, C.; Chen, I.; Foloppe, N.; Potter, A.; Jordan, A.; Roughley, S.; Parratt, M.; Greaney, P.; Morley, D.; Hubbard, R. E. Drug-like Annotation and Duplicate Analysis of a 23-Supplier Chemical Database Totalling 2.7 Million Compounds. *J. Chem. Inf. Comput. Sci.* **2003**, *44*, 643−651.
(5) Merlot, C.; Domine, D.; Cleva, C.; Church, D. J. Chemical Substructures in Drug Discovery. *Drug Discovery Today* **2003**, *8*, 594−602.
(6) Shemetulskis, N. E.; Dunbar, J. B.; Dunbar, B. W.; Moreland, D. W.; Humblet, C. Enhancing the Diversity of a Corporate Database Using Chemical Database Clustering and Analysis. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 407−416.
(7) Agrafiotis, D. K.; Rassokhin, D. N.; Lobanov, V. S. Multidimensional Scaling and Visualization of Large Molecular Similarity Tables. *J. Comput. Chem.* **2001**, *5*, 488−500.
(8) Eichler, U.; Ertl, P.; Gobbi, A.; Poppinger, D. Addressing the Problem of Molecular Diversity. *Drugs Future* **1999**, *24*, 177−190.
(9) Menard, P. R.; Mason, J. S.; Morize, I.; Bauerschmidt, S. Chemistry Space Metrics in Diversity Analysis, Library Design, and Compound Selection. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1204−1213.
(10) Rhodes, N.; Willett, P. Bit-String Methods for Selective Compound Acquisition. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 210−214.
(11) Higgs, R. E.; Bermis, K. G.; Watson, I. A.; Wikel, J. H. Experimental Designs for Selecting Molecules from Large Chemical Databases. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 861−870.
(12) Cummins, D. J.; Andrews, C. W.; Bentley, J. A.; Cory, M. Molecular Diversity in Chemical Databases: Comparison of Medicinal Chemistry Knowledge Bases and Databases of Commercially Available Compounds. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 750−763.
(13) Nilakantan, R.; Bauman, N.; Haraki, K. S. Database Diversity Assessment: New Ideas, Concepts, and Tools. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 447−452.
(14) Turner, D. B.; Tyrrel, S. M.; Willett, P. Rapid Quantification of Molecular Diversity for Selective Database Acquisition. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 18−22.
(15) Muegge, I. Selection Criteria for Drug-like Compounds. *Med. Res. Rev.* **2000**, *23*, 302−321.
(16) Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D. Is There a Difference between Leads and Drugs? A Historical Perspective. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1308−1315.
(17) Johnson & Johnson Completes Acquisition of 3-Dimensional Pharmaceuticals, Inc. http://www.jnj.com/news/jnj_news/20030328_125532.htm (accessed Dec 14, 2004).
(18) Johnson & Johnson Pharmaceutical Research & Development. http://www.jnjpharmarnd.com/ (accessed Dec 14, 2004).
(19) Agrafiotis, D. K.; Lobanov, V. S.; Salemme, F. R. Combinatorial Informatics in the Post-Genomics Era. *Nat. Rev. Drug Discovery* **2002**, *1*, 337−346.
(20) Agrafiotis, D. K.; Bone, R. F.; Salemme, F. R.; Soll, R. M. System and Method for Automatically Generating Chemical Compounds with Desired Properties. U.S. Patent 5,574,564, 1996. Agrafiotis, D. K.; Bone, R. F.; Salemme, F. R.; Soll, R. M. System and Method for Automatically Generating Chemical Compounds with Desired Properties. U.S. Patent 5,574,656, 1996.
(21) 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.
(22) Engels, M. F. M.; Venkatarangan, P. Smart Screening: Approaches to Efficient HTS. *Curr. Opin. Drug Discovery Dev.* **2001**, *4*, 275−283.
(23) Downs, G. M.; Barnard, J. M. Clustering Methods and Their Uses in Computational Chemistry. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; John Wiley & Sons: Hoboken, New Jersey, 2002; Vol. 18, pp 1−40.
(24) Steinbach, M.; Karypis, G.; Kumar, V. A. Comparison of Document Clustering Techniques. Technical Report #00-034; University of Minnesota, Computer Science & Engineering: Twin Cities, MN, 2000. http://www.cs.umn.edu/tech_reports/ (accessed Mar, 2003).
(25) *Divisive K-Means Cluster Manual*, version 1.0; Barnard Chemical Information Ltd.: Sheffield, U.K.
(26) Engels, M. F. M. Overlap Analysis of Compound Collections − Strategies From a Recent Acquisition. Presented at the Third Sheffield Conference on Chemoinformatics, Sheffield, U. K., April 2004. http://cisrg.shef.ac.uk/shef2004/talks/MEngels.pdf (accessed Jan 5, 2005).
(27) 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.
(28) Thomson Scientific, 3501 Market Street, Philadelphia, PA 19104, U. S. A. http://www.derwent.com/products/lr/wdi/ (accessed Dec, 2004).
(29) Ajay; Walters, W. P.; Murcko, M. A. Can We Learn to Distinguish between "Drug-like" and "Nondrug-like" Molecules? *J. Med. Chem.* **1998**, *41*, 3314−3324.
(30) Sadowski, J.; Kubinyi, H. A Scoring Scheme for Discriminating between Drugs and Nondrugs. *J. Med. Chem.* **1998**, *41*, 3325−3329.
(31) *BCI Clustering Package: PROG_DIVKM*; Barnard Chemical Information: Stannington, Sheffield, U. K.; info@bci.gb.com.
(32) Kelley, L. A.; Gardner, S. P.; Sutcliffe, M. J. An Automated Approach for Clustering An Ensemble of NMR-Derived Protein Structures Into Conformationally-Related Subfamilies. *Protein Eng.* **1996**, *9*, 1063−1065.
(33) Wild, D. J.; Blankley, C. J. Comparison of 2D Fingerprint Types and Hierarchy Level Selection Methods for Structural Grouping Using Ward's Clustering. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 155−162.

(34) *OPTCLUS Cluster Manual*; Barnard Chemical Information: Stannington, Sheffield, U. K.; info@bci.gb.com.
(35) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.
(36) Weininger, D. Smiles I, Introduction and Encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(37) *Daylight Theory Manual: Fingerprint (v4.73)*; Daylight Chemical Information Systems, Inc.: Mission Viejo, CA; info@daylight.com.
(38) Daylight Toolkit Programs, Daylight Chemical Information Systems, Inc., 27401 Los Altos, Suite 360, Mission Viejo, CA 92691, info@daylight.com.
(39) Sayle, R.; Delany; J. Canonicalization and Enumeration of Tautomers; presentation at the EuroMUG meeting 1999. http://www.daylight.com/meetings/emug99/Delany/taut_html/index.htm (accessed Dec 14, 2004).
(40) *Pipeline Pilot*, v.4.0; SciTegic Inc.: San Diego, CA. www.scitegic.com (accessed Dec, 2004).
(41) Duda, R. O.; Hart, R. E. *Pattern Classification and Scene Analysis*; John Wiley & Sons: New York, 1973.
(42) Ward, J. H. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.* **1963**, *58*, 236−244.
(43) *Maximum Common Substructure Analysis*; Documentation by D. Rogers, SciTegic Inc., 9665 Chesapeake Drive, Suite 401, San Diego, CA 92123-1365. http://www.scitegic.com/ (accessed Dec, 2004).
(44) Shemetulskis, N. E.; Weininger, D.; Blankley, C. J.; Yang, J. J.; Humblet, C. Stigmata: An Algorithm To Determine Structural Commonalities in Diverse Datasets. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 862−871.
(45) Levi. G. A Note on the Derivation of Maximal Common Subgraphs of Two Directed or Undirected Graphs. *Calcolo.* **1972**, *9*, 341−352.
(46) Cone, M. M.; Venkataraghavan, R.; McLafferty, F. W. Molecular Structure Computer Program for the Identification of Maximal Common Substructures. *J. Am. Chem. Soc.* **1977**, *99*, 7668−7671.
(47) 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.
(48) Barnard, J. Better Clusters Faster. Presentation held at the Third Joint Sheffield Conference on Chemoinformatics, 2004. http://cisrg.shef.ac.uk/shef2004/abstracts.htm (accessed Jan, 2005).
(49) Jarvis, R. A.; Patrick, E. A. Clustering Using a Similarity Based on Shared Near Neighbors. *IEEE Trans. Comput.* **1973**, *C-22*, 1025.
(50) Irwin, J. J.; Shoichet, B. K. ZINC − A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177−182.

CI600219N