Journal of Computer-Aided Molecular Design, 9 (1995) 407–416
ESCOM

J-CAMD 313

Enhancing the diversity of a corporate database using chemical database clustering and analysis

Norah E. Shemetulskis, James B. Dunbar Jr., Bonnie W. Dunbar, David W. Moreland and Christine Humblet*

Parke-Davis Pharmaceutical Research Division, Warner Lambert Company, 2800 Plymouth Road, Ann Arbor, MI 48105, U.S.A.

Received 15 June 1995
Accepted 2 September 1995

Keywords: Molecular diversity; Structural databases; Jarvis–Patrick; Clustering; Octanol–water partition coefficient; Molar refractivity; Dipole moment

Summary

The contribution that the Chemical Abstracts structural database (CAST-3D) and the Maybridge database (MAY) would make to diversifying the structural information and property space spanned by our corporate database (CBI) is assessed. A subset of the CAST-3D database has been selected to augment the structural diversity of various electronic databases used in computer-assisted drug design projects. The analysis of the MAY database directly offers the potential to expand the CBI compound library, but also provides a source for structural diversity in a format suitable for computer-assisted database searching and molecular design. The analysis performed is twofold. First, a nonhierarchical clustering technique available in the Daylight clustering package is applied to evaluate the structural differences between databases. The comparison is then extended to analyze various structure-derived property spaces calculated from molecular descriptors such as the logarithm of the octanol–water partition coefficient (CLOGP), the molar refractivity (CMR) and the electronic dipole moment (CDM). The diversity contribution of each database to these property spaces is quantified in relation to our corporate database.

Introduction

The continuing development of high-throughput in vitro biological screening (i.e., high-volume screening) and computer-assisted drug design techniques has fostered an interest in new means of diversifying structural databases. Over the years, pharmaceutical companies have accumulated compound collections that evolved from lead optimization processes in various therapeutic areas. Such compound libraries are invaluable for drug discovery, as illustrated by the success that pharmaceutical companies have had in identifying new leads that originated from positive hits in high-volume screening assays. Because of their therapeutically oriented evolution, the libraries can span a limited space of structural and property diversity, thus narrowing the probability of identifying structural leads during high-volume screening.

Recent advances in computer technology now allow rapid processing of hundreds of thousands of electronic entries, enabling the manipulation and evaluation of structural properties offered in structural databases. A strategy has been developed to explore elements of molecular diversity within our corporate database, and to compare them to external databases considered for acquisition.

Two external databases, CAST-3D [1] and MAY [2], have been analyzed and compared to our corporate database (CBI). The Chemical Abstract Services (CAS) registry numbers provide a convenient index to retrieve literature references describing sources or synthetic schemes to obtain or prepare compounds that can supplement the CBI database. CAST-3D represents a selection of entries obtained from CAS. The MAY database is an electronic catalog, describing the chemical structures of commercially available samples [2]. MAY offers the direct option to select and acquire samples that can add diversity to existing collections.

A structural clustering technique has been applied to

*To whom correspondence should be addressed.

assess the chemical similarity presented by the CBI database in comparison to the CAST-3D or MAY databases. The application of similarity measures to cluster chemical structures is not new. The use of various hierarchical and nonhierarchical clustering techniques has been described [3]. Hierarchical techniques were not considered for the analyses described here, due to practical limitations imposed by the databases considered in this study [4]. The Jarvis–Patrick nearest-neighbor algorithm [5], a nonhierarchical technique, has been selected for its suitability to analyze chemical databases of the dimensions concerned in this particular study. Hundreds of thousands of compounds were clustered using this algorithm, which is available in the software package from Daylight Chemical Information Systems, Inc. [6]. The Jarvis–Patrick algorithm cannot be utilized indiscriminately. It has been applied to construct structurally representative subsets of chemical stores for use in high-volume screening [7]. In an analysis of molecular property data, however, the Jarvis–Patrick algorithm did not appear as a method of choice [8]. The strength of a given clustering methodology depends on individual properties and datasets to which it is applied and should be considered with caution [4].

The Jarvis–Patrick clustering technique groups similar compounds according to user-selected nearest-neighbor criteria. The method has been validated for its ability to group particular data sets in which cluster members present meaningful degrees of similarity [9]. Initially, clustering criteria and data-set size were investigated for consistent structural similarity assessments within various combinations of reduced subsets of the CBI and CAST-3D databases. The resulting stability of these subdatabase cross-comparisons was then assessed in analyses extended to the entire parent databases. Clustering results were found to generate consistent assessments in all cases. The analysis was then repeated to compare the structural similarity levels between CBI and MAY, using the same clustering nearest-neighbor criteria. The structural similarities resulting from these analyses were further quantified on the basis of the percentages of compounds that CAST-3D or MAY could contribute to expand the structural content of CBI.

Similarity measures derived from structural connectivity are just a few of the many parameters that can be computed to probe diversity. Physicochemical parameters derived from molecular structures represent alternate means, somewhat more tangible to chemists. Techniques available for calculating such molecular properties have been reported [10–12]. A computational approach to design diverse peptoid libraries recently described the use of a combination of structural and property parameters [13]. Given the size and nature of the databases considered in this study, database comparisons beyond structural similarity assessments were limited to the inclusion of physicochemical properties that can be rapidly computed. The octanol–water partition coefficients (CLOGP), molar refractivity (CMR) [10], and dipole moment (CDM) [11] data were calculated for individual compounds in each database as a measure of the lipophilicity, size, and electronic properties, respectively.

A graphic analysis of the space covered by each property was then designed and implemented for efficient scrutiny. A histogram algorithm was developed to provide yet another level of quantification, in order to assess the contributions that CAST-3D and MAY can add to the property space diversity of the CBI database.

Methodology

Analyses of structural similarity

Structural databases

The CAST-3D database represents a series of structural entries, compiled in the course of an initiative to convert two-dimensional structures to three-dimensional ones [1] using the CONCORD program [14]. From an initial set of over 600 000 compounds presenting primarily rigid structures, a database was retained after removing metal-containing compounds or pure hydrocarbon structures. The resulting CAST-3D database, containing 379 847 entries, thus represents a 3D electronic database that is better suited for drug discovery purposes. MAY is an electronic structural catalog that includes 41 912 chemical samples commercially available. CBI, a subset of our proprietary compound library, contains 117 459 structural entries.

To meet the data input requirements of the clustering software program, the database entries were converted into SMILES strings, which is a connectivity-based molecular representation [15]. The 379 847 compounds in CAST-3D [1], available in SYBYL multi-mol2 file format, were converted into SMILES strings using the program UNITY [11]. The MAY database is distributed by Daylight in a format readily compatible with the clustering programs. The MDL SDfile, which contains the connection tables for the compounds in the CBI database, was translated to yield 117 459 SMILES strings [11,16].

Cluster generation

The Daylight clustering package [6] was applied to generate the database clusters. As part of the clustering procedure, SMILES strings are encoded into molecular fingerprints [17], which is a binary representation of the connectivities. The molecular fingerprints result from the application of a hashing algorithm to substructural fragments that have themselves been generated from topological paths emanating from each atom in the molecule, through a maximum depth of seven bonds. Daylight molecular fingerprints differ from a structural key [16] representation of molecular structures in that the sub-

TABLE 1
CHARACTERISTICS OF THE SUBDATABASES GENERATED FROM CAST-3D AND CBI

Subdatabase	Clustering level	No. of compounds	No. of singletons	No. of centroids
CBI1	8/14	27 074	17 667	9 407
CBI2	9/15	29 545	19 650	9 895
CBI3	10/16	32 084	21 724	10 360
CAS1	8/14	101 772	68 277	33 495
CAS2	9/15	109 596	74 792	34 804
CAS3	10/16	117 084	81 410	35 674

structural fragments are derived directly from the structures and not from an a priori database of substructural definitions. The default criterion for fingerprint generation, which employs a maximum length of 2048 bits, was employed for all structural representations. Binary representations of molecular structures have their primary use in substructural and similarity searching of chemical databases, but their binary nature also offers an efficient means of generating similarity information on large data sets.

For all compounds in a given database, pairwise molecular similarities were determined using the Tanimoto coefficient (Tc), defined as [18]:

$$Tc = \frac{N(A\&B)}{N(A) + N(B) - N(A\&B)} \quad (1)$$

For two fingerprints A and B, Tc is a ratio of the number of bits in common between two fingerprints (N(A&B)), divided by the sum of the individual non-zero bits, N(A) and N(B), reduced by the number of bits in common between the two fingerprints. This measure ranges between zero and one, where one is the value obtained for identical compounds.

Once all pairwise comparisons have been made, a list of the 16 closest compounds for each structure in the database is generated. The calculation of this list, the nearest-neighbor list, scales as the square of the number of compounds, and is therefore the computationally limiting step in the analysis. For example, the nearest-neighbor list generation for the parent CAST-3D database required 64 CPU days on one processor of a Silicon Graphics 4D/480 workstation equipped with 64 Mb of memory [6]. The algorithm has recently been optimized to run on parallel processors, thus offering an option to increase the efficiency of nearest-neighbor list generation [19].

Given the computational requirements to analyze the CAST-3D and CBI databases, initial similarity analyses were performed on various database subsets, hereafter referred to as subdatabases. Three subdatabases each of CBI and CAST-3D were selected, using the Daylight clustering package [6] on the basis of the nearest-neighbor lists and applying the Jarvis–Patrick clustering algorithm. This algorithm has two criteria, which result in the grouping of two compounds. The first is that the two compounds being considered appear in each other's list of K nearest neighbors and the second is that they share $J < K$ nearest neighbors. In the Daylight application, the selection of J and K determines a clustering level called a need-versus-near criterion. The default options available in the Daylight program cluster a database at level 8/14 (i.e., 8-need-versus-14-nearest-neighbors). At this clustering level, two compounds form a cluster if they share eight out of the first 14 nearest neighbors and if each of the two belongs to the first 14 nearest neighbors. The clustering levels applied to select the subdatabases under study are listed in Table 1.

The structurally representative compound for a cluster is termed the cluster *centroid*. Compounds that do not cluster with any other compounds are termed *singletons*. For each level, a subdatabase comprises one compound from each cluster, either centroid or singleton, obtained from the clustering of the parent database. For example, CAS1 includes 33 495 cluster centroids added to the 68 277 singletons resulting from clustering the CAST-3D parent database at a Jarvis–Patrick level of 8/14. The

TABLE 2
CHARACTERISTICS OF MIXED SUBDATABASES FROM CAST-3D AND CBI

Mixed subdatabase	Clustering level	No. of compounds	% CAS compounds	
			Pure CAS[a]	>95% CAS[b]
CBI1/CAS1	8/14	128 846	80	86
CBI1/CAS2	8/14	136 670	80	86
CBI1/CAS2	9/15	136 670	77	82
CBI1/CAS3	8/14	144 163	79	85
CBI1/CAS3	10/16	144 163	85	90
CBI2/CAS1	8/14	131 317	82	87
CBI2/CAS1	9/15	131 317	79	85
CBI2/CAS2	9/15	139 141	82	87
CBI2/CAS3	9/15	146 634	82	87
CBI2/CAS3	10/16	146 634	80	84
CBI3/CAS1	8/14	133 856	79	84
CBI3/CAS1	10/16	133 856	84	89
CBI3/CAS2	9/15	141 680	84	89
CBI3/CAS2	10/16	141 680	82	86
CBI3/CAS3	10/16	149 173	84	88

[a] Percentage of CAS compounds in pure CAS clusters.
[b] Percentage of CAS compounds in clusters containing >95% CAS compounds.

sizes of all subdatabases generated from the CAST-3D and CBI parent databases are described in Table 1.

Following the clustering procedure just described, mixed databases were generated from pairwise combinations of each CBI and CAST-3D subdatabase. The number of compounds in each of the mixed subdatabases and the clustering levels applied to group the compounds into clusters from each of the mixed subdatabases are given in Table 2. The mixed subdatabase comprising CBI1 and CAS2, for example, was clustered at an 8/14 Jarvis–Patrick level, as had been used for CBI1, and a 9/15 level as in CAS2. The computational requirements for the nearest-neighbor list generation ranged from 10 CPU days for the CBI1/CAS1 combination to 14 CPU days for the CBI3/CAS3 pair.

TABLE 3
RESULTS FOR CLUSTERING OF THE CAS AND MAY DATABASES WITH CBI

Database	Clustering level	% CAS and MAY compounds	
		Pure clusters[a]	>95% CAS or MAY[b]
CAS/CBI	8/14	78	80
CAS/CBI	9/15	80	82
CAS/CBI	10/16	82	84
MAY/CBI	8/14	53	56
MAY/CBI	9/15	56	59
MAY/CBI	10/16	59	62

[a] Percentage of CAS compounds in pure CAS clusters and percentage of MAY compounds in pure MAY clusters.
[b] Percentage of CAS compounds in clusters containing >95% CAS compounds; percentage of MAY compounds in clusters containing >95% MAY compounds.

Cluster analysis

To analyze the cluster populations within the mixed databases, a software program was developed to analyze clusters based on tag identifiers associated with each compound retrieved from the resulting clusters. Compound tags identify parent database membership and thus are used to trace and to count the number of CAST-3D compounds which reside in clusters containing, for example, only CAST-3D compounds or greater than a specified percentage of CAST-3D compounds. The percentages of CAST-3D compounds residing in pure CAST-3D clusters or in clusters that contain greater than 95% CAST-3D compounds for each of the mixed subdatabases are given in Table 2. The tagging procedure and percentage purity criteria were used to select 101 000 compounds from CAS1 to be licensed from the Chemical Abstracts Service. These compounds were chosen from CAS1, since its dimension was closest to the required size and it thus minimized compound elimination to less than one percent. The 'CAS subset' was selected from a structural clustering of CBI1/CAS1. In order to retain nearly the entire CAS1 database, compounds were chosen from clusters containing cluster purities of CAST-3D compounds from 100% down to 50%, until the required 101 000 compounds were attained. In total, 772 compounds were eliminated.

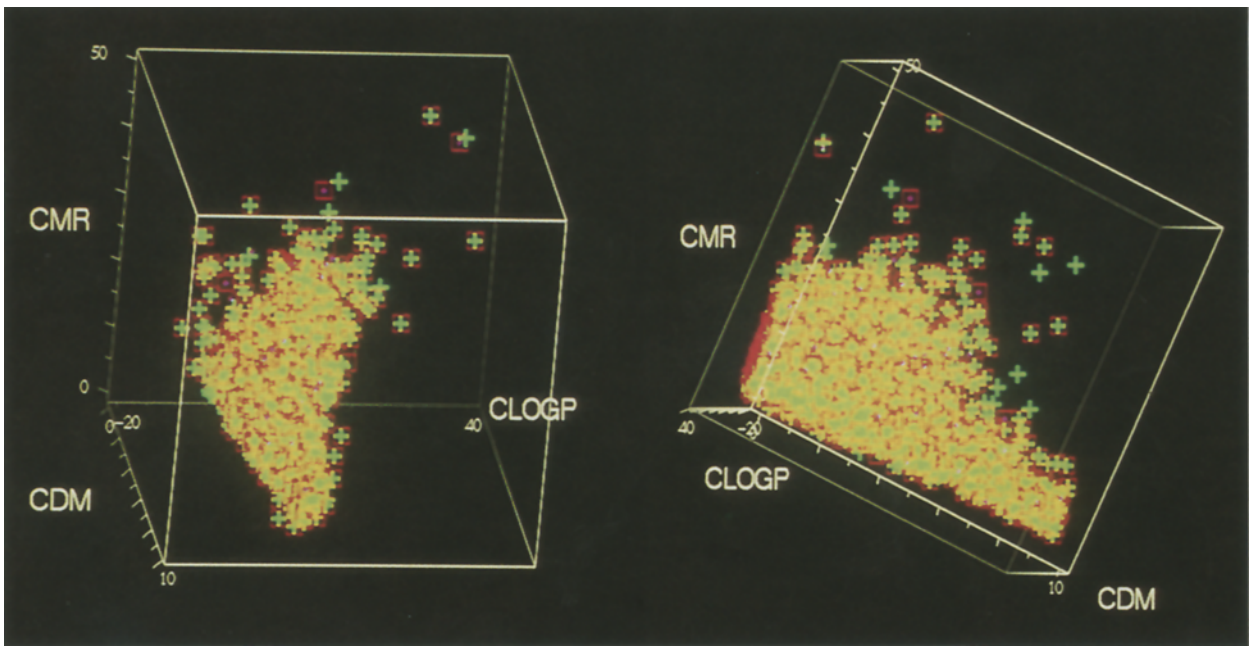

Fig. 1. A qualitative comparison of three subsets of the CBI database, i.e., CBI1 (red), CBI2 (magenta) and CBI3 (green), through a three-dimensional plot whose axes are defined by the calculated octanol–water partition coefficient (CLOGP), the calculated molar refractivity (CMR) and the calculated dipole moment (CDM). The molecular descriptor calculations resulted in some CLOGP values that could not be calculated. All compounds that did not have meaningful CLOGP, CMR or CDM values have been removed from these plots. The actual number of compounds for each subdatabase plotted here is: CBI1: 19 011; CBI2: 20 862; CBI3: 22 799.

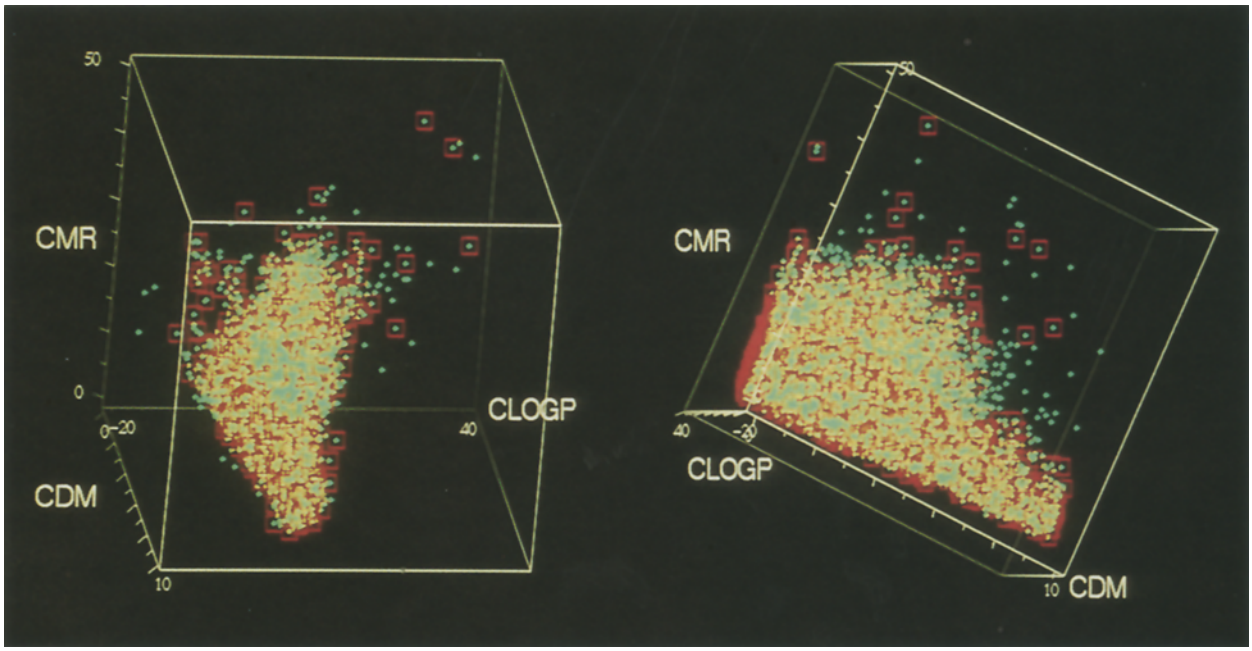

Fig. 2. A comparison of the CBI1 subdatabase (red) with the full CBI database (green) in the CLOGP, CMR and CDM property space. Two perspectives of a 3D plot are given, whose axes are defined as in Fig. 1 by CLOGP, CMR and CDM. We have eliminated compounds with missing fragments in their CLOGP and CMR or with CDM values greater than 10. The CBI database plotted here contains 86 908 compounds.

To determine the sensitivity of the structural assessment in the mixed subdatabase comparisons, the CAS subset was then clustered with the full CBI database. This computation required 12.8 days of CPU time on a Silicon Graphics ONYX workstation. A similar structural database analysis was also completed for the full Maybridge database by clustering it with 117 001 compounds from the CBI database. The results of the parent structural analysis for MAY and CAS are given in Table 3.

Analyses of the property space similarity

Molecular descriptor calculations

In the second phase of the analysis, database comparisons were carried out on the basis of three molecular descriptors. The physicochemical parameters CLOGP, CMR, and CDM were calculated for each compound in CBI1–CBI3, CAS1–CAS3, CBI, MAY, and CAS [10,11].

The MedChem software [10] computes CLOGP and CMR values from SMILES strings. Since this calculation is a fragment-based approach relying on substituent dictionaries, some compounds end up not having values for all of their substituents in the lookup tables. Consequently, CLOGP or CMR data cannot be determined and the software issues a missing fragment warning. The MedChem software also provides error codes reflecting on the accuracy of the CLOGP or CMR calculations. All compounds that had high MedChem error codes, reflecting uncertain results, were eliminated from the comparison.

The dipole moment calculation (CDM) was performed for all database entries, using the CONCORD-generated conformations and the Gasteiger atom charges calculated with the SYBYL software package [10,11]. The SYBYL multi-mol files for the CBI compounds were generated from a routine that translates an MDL SDfile into a SYBYL multi-mol output file. The CAST-3D compounds obtained from Chemical Abstracts [1] were already formatted in the appropriate SYBYL multi-mol input file. A SYBYL Programming Language (SPL) script was written to automate the CDM calculation process. Compounds which had a dipole moment greater than 10 D, a value beyond which the CDM becomes rather questionable, were eliminated from the comparisons.

The resulting databases included 43 877 CAS compounds, 86 908 CBI compounds and 23 176 MAY compounds. In all cases, the exclusion of compounds was primarily due to missing fragment data in the CLOGP calculation. Three-dimensional plots of the property distributions for each compound in the three CBI subdatabases, CBI1 (red), CBI2 (magenta), and CBI3 (green), are presented in Fig. 1. The axes of the plots are defined by CLOGP, CMR and CDM. Figure 2 presents the graphic comparison between the CBI1 (red) subdatabase and the full CBI (green) database.

Histogram technique

The previous graphic comparisons remain qualitative. A histogram approach was developed and coded to better quantify the contribution that an acquisition of compounds from CAS or MAY would make to diversify the CBI property space. The closest CBI compound to any given compound in an acquisition database is determined

from the histogram distance described as:

$$\text{distance} = \left[\left(\frac{\text{CLOGP}_{\text{ACQ}} - \text{CLOGP}_{\text{CBI}}}{\text{max CLOGP}} \right)^2 + \left(\frac{\text{CMR}_{\text{ACQ}} - \text{CMR}_{\text{CBI}}}{\text{max CMR}} \right)^2 + \left(\frac{\text{DM}_{\text{ACQ}} - \text{DM}_{\text{CBI}}}{\text{max DM}} \right)^2 \right]^{1/2} \quad (2)$$

$\text{CLOGP}_{\text{ACQ}}$, CMR_{ACQ}, and CDM_{ACQ} are the property values for the MAY or CAS acquisition compound. $\text{CLOGP}_{\text{CBI}}$, CMR_{CBI}, and CDM_{CBI} are the molecular descriptors for CBI compounds. MaxCLOGP, maxCMR, and maxCDM are the maximum descriptor values found when searching through all three databases MAY, CAS and CBI. To normalize the CAS/CBI and MAY/CBI comparisons, each component was scaled by the maximum value found from all databases in the comparison.

Once the histogram distance of an acquisition compound has been compared to all CBI ones, the minimum distance resulting from Eq. 2 is translated into a bin number. The nearest integer of the distance, multiplied by an arbitrary scaling factor of 5000, determines the bin number. Each compound under consideration contributes to the histogram as a member of the bin which represents the distance to the closest CBI compound. This quantification was performed for MAY and CAS, and the histograms are displayed in Fig. 6.

Results and Discussion

Structural similarity analysis

The clustering analyses first applied to the various CBI and CAST-3D subdatabases, CBI1-3 and CAS1-3, were derived at varying Jarvis–Patrick clustering levels. As can be seen in Table 1, the resulting numbers of compounds selected represent similar fractions of the parent database populations. The CBI subdatabases all contain approximately 25% of the CBI database, while the CAST-3D subdatabases comprise nearly 30% of the CAST-3D parent database. The analyses thus appear to yield consistent fractionations in the parent database groupings. Subdatabases represent the different resolution produced from the application of various Jarvis–Patrick clustering thresholds to a parent database. As indicated in Table 2, the subsequent cross-comparisons between the combined subdatabases were performed at each Jarvis–Patrick clustering level that had been previously applied to select individual subdatabases. The structural similarity between the CBI and CAS subdatabases is best assessed from the counts of cluster members that originated from the CAST-3D parent database. Column four in Table 2 lists the percentage of CAST-3D compounds that reside in clusters containing only CAST-3D compounds. This calculation includes CAST-3D singletons as well as pure CAST-3D clusters. In the CBI2/CAS2 comparison, for example, 82% of the CAS2 compounds appear structurally different from the CBI2 compounds. The percentages of CAST-3D membership listed in column five of Table 2 were calculated from the clusters that contained more than 95% CAST-3D compounds. As expected with a 5% purity tolerance, the dissimilarity between CBI2/CAS2 is reflected by an increase in the value given in column four to 87% in column five. Comparisons of all mixed subdatabases lead to consistent clustering overall, as indicated by the average percentages of 81% for pure CAS membership and 86% for >95% CAS compounds.

Using the tagging procedure and the percentage purity criteria outlined in the methodology section, the 'CAS subset', which contains 101 000 compounds from CAS1, was combined with the full CBI database and a cluster analysis was performed to assess the accuracy of the subdatabase approach. As seen in Table 3, the clustering results obtained at Jarvis–Patrick clustering levels of 8/14, 9/15, or 10/16 for percentage purity criteria of 100% and 95% again predict an average of 80% structural differences between the 'CAS subset' and the CBI database. The 'CAS subset' database thus significantly adds to the electronic structural diversity of our corporate database.

Similar clustering analyses applied to the MAY database led to the results presented in Table 3. On average, it appears that an acquisition of 56% of the MAY database would add to the structural diversity of the CBI compound library.

Property similarity analysis

Dissimilar structural content provides but one aspect of the available information an acquisition database can contribute to diversifying a corporate database. Chemical structures found to be structurally similar to corporate database compounds may contribute to other desirable diversity components. In the present studies, the database comparisons were expanded to explore the distributions of three molecular descriptors, CMR, CLOGP, and CDM. These descriptors were chosen to measure overall size, hydrophobicity, and electronic properties of each molecule. Initially, the descriptor distributions were compared to see to what extent a subdatabase, selected on the basis of structural criteria, is representative of its parent database. As illustrated in Fig. 1, the three CBI subdatabases derived from the three distinct Jarvis–Patrick clustering levels, CBI1 (red), CBI2 (magenta), and CBI3 (green), lead qualitatively to similar distributions in the property spaces considered. In Fig. 2, CBI1 (red) is compared to its parent database CBI (green) to visualize the three property space distributions. Qualitatively, the subdatabase appears to encapsulate a reasonable amount of the property space covered by its parent database.

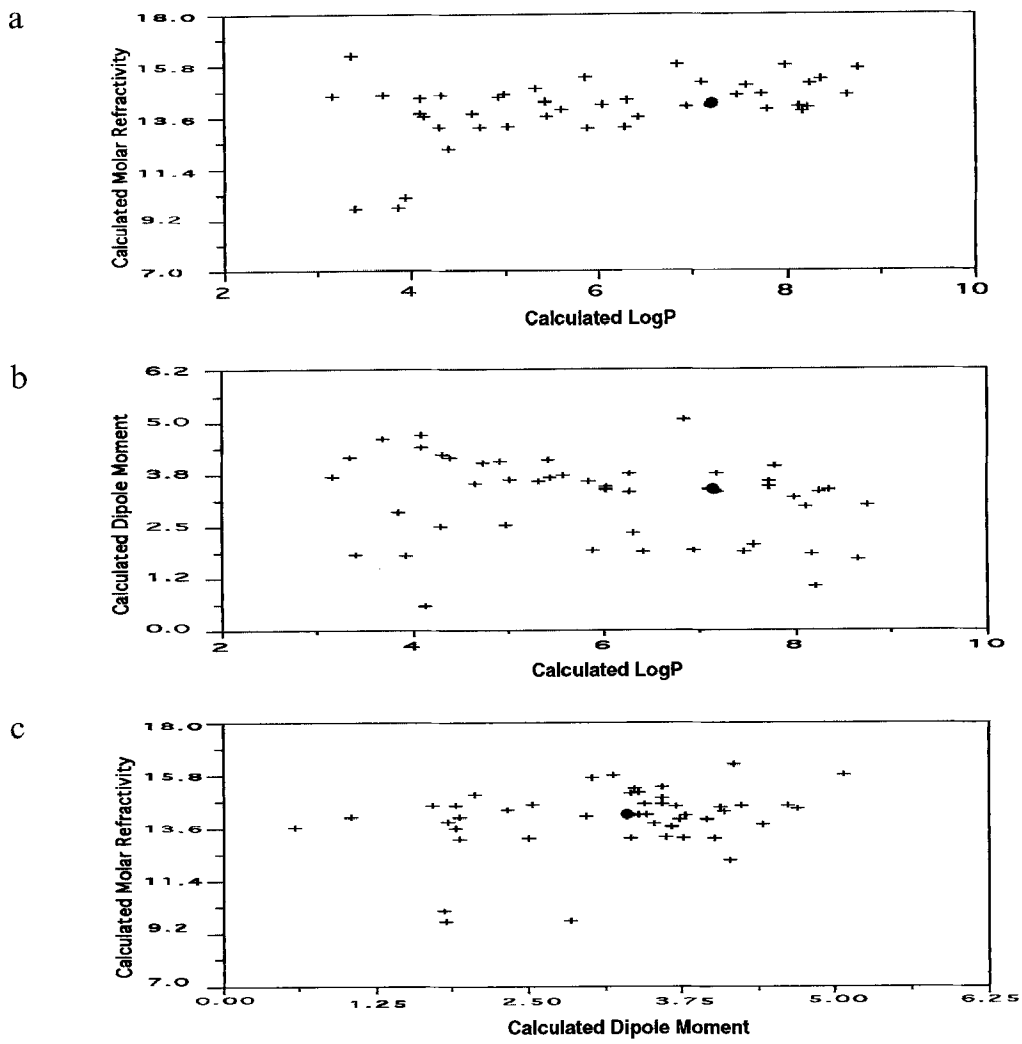

Fig. 3. Projections of a 3D plot of members from one cluster of CBI containing 271 compounds (black cross), resulting from a clustering of CBI at level 8/14. (a) CMR/CLOGP comparison with CDM = 0; (b) the CDM/CLOGP cross section with CMR = 0; and (c) the CMR/CDM projection in the CLOGP = 0 plane. The cluster centroid, which was the one member of this cluster selected for the subdatabase CBI1, is also plotted as a black circle. The cluster centroid's molecular descriptor values are: CLOGP = 7.2, CMR = 14.1 and CDM = 3.3 D. The average molecular descriptor values for this cluster are: CLOGP = 5.9, CMR = 13.9 and CDM = 3.2 D. This plot reveals that for this property space the centroid of a structural cluster does not necessarily represent the property space spanned by the cluster members.

Each subdatabase was generated from the collection of centroids and singletons resulting from a structural clustering of its parent database. It is thus of interest to see whether the CLOGP, CMR, and CDM properties of a structural cluster centroid are representative of the property space spanned by the cluster. The molecular property plots shown in Fig. 3 address this question. A cluster centroid (solid black circle) and its cluster members (crosses) were selected at random from the CBI database clustered at a level of 8/14. While the centroid belongs to the CBI1 subdatabase, the other cluster members do not. Figure 3 reveals that the centroid has CLOGP = 7.2, CMR = 14.1 and CDM = 3.3 D, while the cluster members present property values ranging from 3.2 to 8.8 for CLOGP, 9.7 to 16.2 for CMR and 0.59 to 5.1 D for CDM. The average CLOGP, CMR, and CDM molecular descriptor values for the 47 members of this cluster are 6.0, 14.0 and 3.2 D, respectively. An examination of several clusters found the centroid's molecular descriptor values close to the cluster average values, but not reflecting the range of the property values spanned by the cluster members. From the clusters examined, it is prudent to conclude that the cluster centroid, although structurally representative of the cluster, is not necessarily a good indicator of the property range spanned by the other cluster members. Given the feasibility of property space comparisons of parent databases, the analyses and comparisons of property distributions were thus carried out for the parent databases rather than using averaged molecular descriptors or cluster centroid molecular descriptor values derived from the subdatabases.

The property comparisons between the CBI (green)

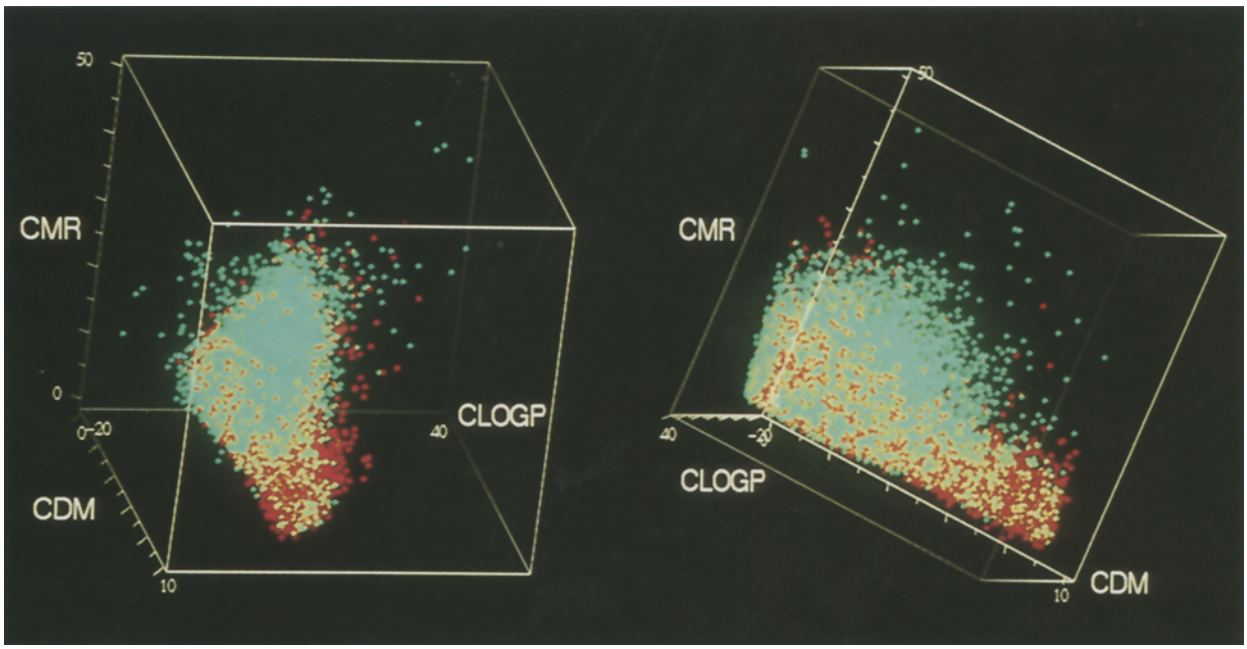

Fig. 4. A comparison of the CBI database (green) and the CAS database (red) in CLOGP, CMR and CDM space. Compounds which gave high error codes for the CLOGP and CMR calculations have been eliminated from the comparison, as well as those with a CDM greater than 10. The number of compounds in the CAS database resulting after these criteria are applied is 43 877.

and CAS (red) databases are described in Fig. 4. Overall, a high degree of overlap is seen for each property space between these two datasets. CAS compounds appear to populate the dipole moment range toward higher values. The comparison of the CBI (green) and MAY (magenta) databases is shown in Fig. 5. In this case, the MAY property distributions seem to be mostly contained within smaller ranges than those covered by CBI.

The stacked histogram procedure offers a more quantitative alternative to assess the differences in property space coverage. The results presented in Fig. 6 illustrate the CAS compound membership for each bin in the histogram in red and the MAY compound contributions in green. The bin numbers of the histogram indicate the nearest distance to a CBI compound in property space. The ordinate indicates how many CAS or MAY com-

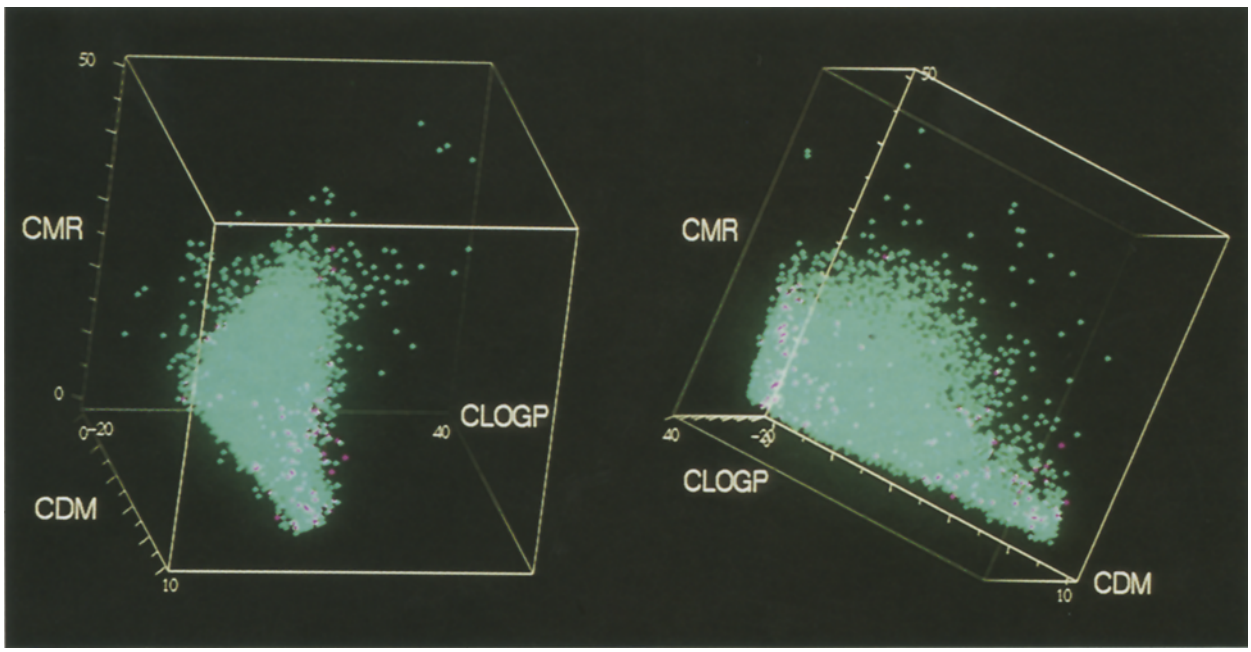

Fig. 5. The same comparison that was made in Fig. 4 is shown for CBI (green) and the MAY (magenta) database. The MAY database had 23 176 compounds with reasonable values for CMR, CLOGP and CDM.

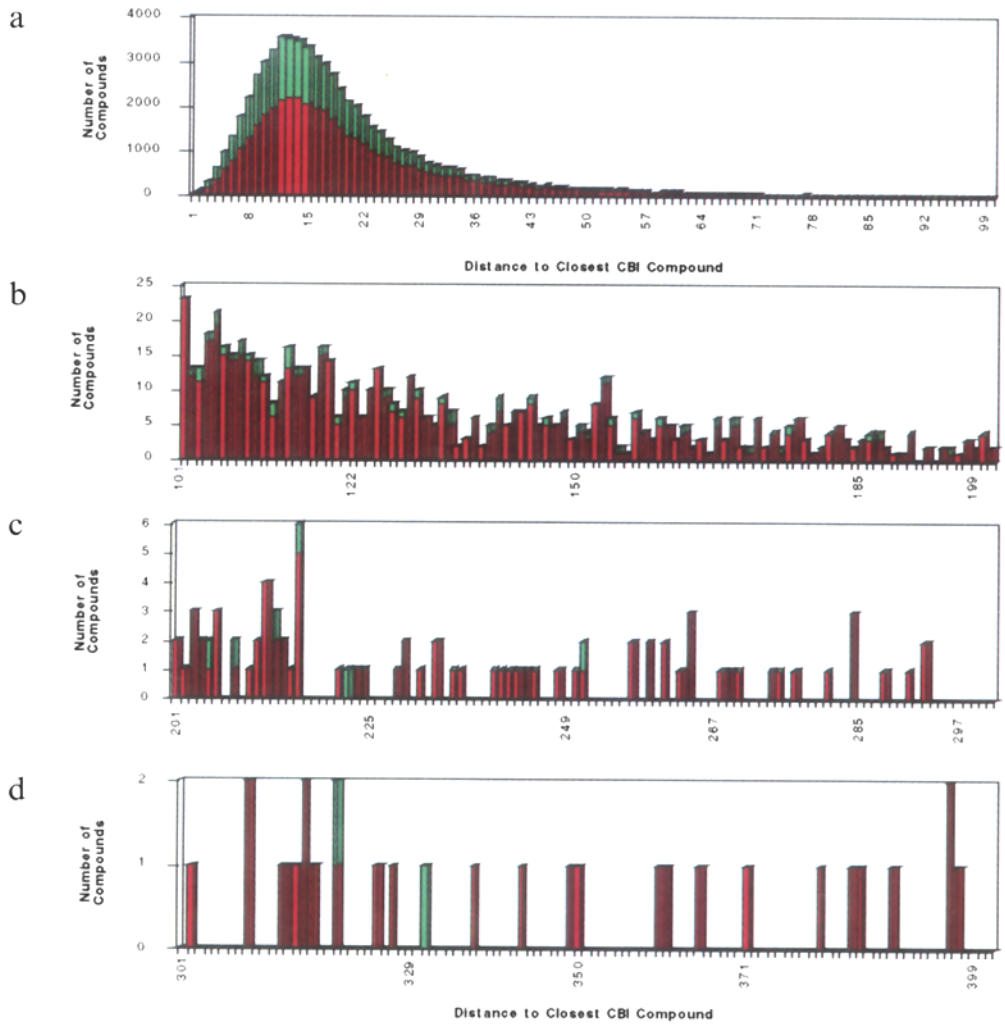

Fig. 6. A quantitative measure of the differences between the MAY database (green) and the CAS database (red), displayed as stacked histograms. Each compound in each comparison database is placed into a bin that represents the distance to the closest CBI compound to it. The measurement is done in Euclidean space, with each component scaled by the maximum value for all three databases. The compounds in the largest bin numbers are those that would add to the property space diversity of CBI. Compounds in small bin numbers would supplement the current property space spanned by CBI compounds.

pounds share a common distance to a CBI compound. The stacked histogram technique was chosen to determine whether the property space diversity of a corporate database can be enhanced through the selection of compounds that are different in this property space from those in the CBI database. Compounds which are in high bin numbers are thus good candidates to add to the property space diversity, since they are less similar to any compounds in the CBI database. Figures 6c and d show that the CAS database would provide a greater contribution to the property space diversity than would the MAY database.

As determined from the cluster analyses (see Table 2), the degree of structural similarity between CAS and CBI was lower than in the case of the MAY database. This trend is also seen in the property space comparisons. The CAS database comprises rigid chemical structures selected from the CAST-3D database. The MAY database is primarily a database of chemical intermediates. It is therefore not surprising to see that the CAS database can potentially add more diversity to the structural and property space of the CBI database.

Conclusions

The present analysis evolved from an initial interest in structural database similarity assessments. Strategies for acquisition of two external databases, CAS and MAY, in reference to our proprietary subset CBI were then explored. The initial mandate to choose and extract approximately 100 000 CAS structures from the six times larger CAST-3D database started our exploration into methods that would filter the CAST-3D and MAY databases in the context of both structural and property information.

The Daylight software package offered multiple computational tools for our structural assessments. Computa-

tional advantages that binary fingerprints offer as an efficient means to manipulate large databases were determined. The nonhierarchical Jarvis–Patrick clustering technique provided a straightforward path to cluster generation and subsequent identification of *cluster centroids* and *singletons*. Tracking identification tags and cluster purity criteria were shown to be useful mechanisms for similarity assessment.

Curious to explore other factors that contribute to molecular diversity, we considered physicochemical parameters that describe hydrophobicity, size and polarity of compounds. Graphic plots of the descriptor spaces provided useful representations to visualize general trends observed within or across databases. The histogram algorithm was developed to better quantify the similarities or differences observed in the molecular descriptor data.

Determining which structures in a chemical acquisition database will increase the chances for novel drug discovery is a difficult task. We elected to focus on structural and property assessments as an initial step to approach this problem, while simultaneously gaining a feel for the diversity represented in our corporate database. The methodology implemented to measure both structural and property aspects could be easily expanded.

The selection of compounds based on structural similarity was directed particularly at clusters that contain high percentages of acquisition compounds. Additional criteria, such as cluster size or cluster composition, could readily be taken into account to augment the selection process. For example, a cluster of size two, containing one CBI compound and one CAST-3D compound, contains only 50% CAST-3D compounds. If the selection threshold is greater than 50%, this CAST-3D compound, although an important structural supplement to the in-house database, would not be retained. Consideration of cluster size and parent database content would thus appear critical in such an instance.

With the large number and efficiency of high-volume screens now available, and also the sometimes limited amounts of samples available, particular compounds can be rapidly depleted. This can affect the representation of key structural members in a corporate library. The present methodology can be readily applied to address such depletion concerns. Thus, the acquisition database could be clustered with a subset of the corporate database, containing compounds which are nearly depleted. Here, the selection criterion would need to be adjusted to select compounds that cluster with corporate compounds. For this purpose, lower percentages of cluster purity would allow for similarity-based supplementation of a compound library.

Hundreds of connectivity-based molecular descriptors can be calculated for databases of the sizes discussed here. Such an explosion of data raises the need to isolate key descriptors that are relevant to the drug discovery process. These descriptors could be used to incorporate compounds that were eliminated from the present property space comparisons, and could provide insight into diversity contributions. Isolation of key descriptors and further structural validity assessments of clustering criteria are the focus of current work.

In conclusion, the methodology presented here provides an initial approach to the diversity assessment and enhancement of an acquisition database. The structural and property-based approach applied to assess the MAY and CAST-3D databases has provided us with a straightforward initial assessment of their diversity contribution to our corporate database. The methods employed are expandable to incorporate the particular need for compound selection in the context of corporate compound library depletion. They also find a logical extension in series design projects, as they relate to molecular design or substituent selection for automated compound libraries.

References

1 CAST-3D database, Chemical Abstracts Services, Columbus, OH.
2 MayBridge 94 Database, Daylight Chemical Information Systems, Irvine, CA, 1994.
3 Barnard, J.M. and Downs, G.M., J. Chem. Inf. Comput. Sci., 32 (1992) 644.
4 Brown, R.D., Bures, M.G. and Martin, Y.C., manuscript in preparation.
5 Jarvis, R.A. and Patrick, E.A., IEEE Trans. Comput., C-22 (1973) 1025.
6 Weininger, D., clustering package, v. 4.3, Daylight Chemical Information Systems Inc., Irvine, CA, 1993.
7 Willett, P., Winterman, V. and Bawden, D., J. Chem. Inf. Comput. Sci., 26 (1986) 109.
8 Downs, G.M., Willett, P. and Fisanick, W., J. Chem. Inf. Comput. Sci., 34 (1994) 1094.
9 Dubes, R. and Jain, A.K., Pattern Recogn., 11 (1979) 235.
10 Leo, A., CLOGP and CMR, v. 4.54E, BioByte Corporation, Claremont, CA, 1994.
11 Tripos Associates Inc., St. Louis, MO, 1994.
12 Hall, L.H., Molconn-X, v. 2, Hall Associates Consulting, Quincy, MA, 1993.
13 Martin, E.J., Blaney, J.M., Siani, M.A., Spellmeyer, D.C., Wong, A.K. and Moos, W.H., J. Med. Chem., 38 (1995) 1431.
14 Pearlman, R.S., Balducci, R., Rusinko, A., Skell, J.M. and Smith, K.M., CONCORD, Tripos Associates, St. Louis, MO, 1994.
15 Weininger, D., J. Chem. Inf. Comput. Sci., 28 (1988) 31.
16 Molecular Design Inc., San Leandro, CA.
17 James, C.A. and Weininger, D., Daylight Theory Manual, Daylight Chemical Information Systems Inc., Irvine, CA, 1995, pp. 43–50.
18 Willett, P., Winterman, V. and Bawden, D., J. Chem. Inf. Comput. Sci., 26 (1986) 36.
19 Weininger, D. and Delany, J., clustering package, v. 4.4, Daylight Chemical Information Systems Inc., Irvine, CA, 1995.