ARTICLES

Data Mining a Small Molecule Drug Screening Representative Subset from NIH PubChem

Xiang-Qun Xie*[†,‡] and Jian-Zhong Chen[†]

Department of Pharmaceutical Sciences, School of Pharmacy, Pittsburgh Molecular Library Screening Center, Drug Discovery Institute, Pittsburgh, Pennsylvania 15260, and Departments of Computational Biology and Structure Biology, University of Pittsburgh, Pittsburgh, Pennsylvania 15260

Received June 5, 2007

PubChem is a scientific showcase of the NIH Roadmap Initiatives. It is a compound repository created to facilitate information exchange and data sharing among the NIH Roadmap-funded Molecular Library Screening Center Network (MLSCN) and the scientific community. However, PubChem has more than 10 million records of compound information. It will be challenging to conduct a drug screening of the whole database of millions of compounds. Thus, the purpose of the present study was to develop a data mining cheminformatics approach in order to construct a representative and structure-diverse sublibrary from the large PubChem database. In this study, a new chemical diverse representative subset, *rePubChem*, was selected by whole-molecule chemistry-space matrix calculation using the cell-based partition algorithm. The representative subset was generated and was then subjected to evaluations by compound property analyses based on 1D and 2D molecular descriptors. The new subset was also examined and assessed for self-similarity analysis based on 2D molecular fingerprints in comparing with the source compound library. The new subset has a much smaller library size (540K compounds) with minimum similarity and redundancy without loss of the structural diversity and basic molecular properties of its parent library (5.3 million compounds). The new representative subset library generated could be a valuable structure-diverse compound resource for in silico virtual screening and in vitro HTS drug screening. In addition, the established subset generation method of using the combined cell-based chemistry-space partition metrics with pairwised 2D fingerprint-based similarity search approaches will also be important to a broad scientific community interested in acquiring structurally diverse compounds for efficient drug screening, building representative virtual combinatorial chemistry libraries for syntheses, and data mining large compound databases like the PubChem library in general.

INTRODUCTION

As one of the major NIH Roadmap initiatives (http://nihroadmap.nih.gov), a public molecular information repository, PubChem database, was constructed to facilitate information exchange and sharing data deposits among the NIH funded ten Molecular Libraries Screening Centers Network (MLSCN).[1] The database system is maintained by the National Center for Biotechnological Information (NCBI) and can be accessed on the Internet at http://PubChem.ncbi.nlm.nih.gov. It consists of three dynamically growing primary databases: PubChem Compound (10 million entries of pure and characterized chemical compounds); PubChem Substance (15.8 million entries of mixtures, extracts, complexes, and uncharacterized substances); and PubChem Bioassay (bioactivity data from 337 HTS programs with several million records).

PubChem has now dynamically grown to 12 million compound records (vn. May 2007) containing mostly small molecules. The massive amounts of compound structures and molecular information stored in the PubChem database and bioactivity data deposited by all MLSCN centers have generated many research opportunities for the scientific community to exploit the available data in PubChem in order to optimize the structure−activity relationships and pharmacology, pharmacokinetics, metabolism, and toxicology profiles of target compounds both in vitro and in silico. However, it may be unrealistic for a typical academic laboratory to conduct direct high throughput screening (HTS) experiments to screen more than 10 million compounds for each biological target on a weekly or a monthly basis. Even for an NIH funded MLSCN center, it currently has a goal of conducting 20 assays per year screening about 100 000 compounds and depositing the screening data into the PubChem database. Of course, modern HTS technologies can screen millions of compounds more quickly and cheaply at costs much less than earlier HTS methods that were

* Corresponding author phone: 412-383-5276; e-mail: xix15@pitt.edu. Corresponding author address: Department of Pharmaceutical Sciences, School of Pharmacy, University of Pittsburgh, 10016 BST3, 3501 Fifth Avenue, Pittsburgh, PA 15260.

[†] Pittsburgh Molecular Library Screening Center, Drug Discovery Institute.

[‡] Departments of Computational Biology and Structure Biology, University of Pittsburgh.

10.1021/ci700193u CCC: $40.75
Published on Web 02/27/2008

estimated at a cost average of approximately $1 per well in pharmaceutical industries.[2] However, at screening rates of 100 000 compounds/week, it may still be a challenge for an academic institution to support on a weekly basis without the major funding support to cover the costs of HTS programs and resources as well as the assay development. It might be achievable for computer virtual high throughput screening (vHTS). It would still take tremendous computing resources and CPU time to perform 3D pharmacophore database searches for each defined pharmacophore query or 3D docking searches. Overcoming these limitations requires innovative approaches, either through development of fast- and low-cost HTS experiments or by employing a cheminformatics approach to taper down the large compound library without losing its information of original molecular properties as well as structural diversity. The latter is our current research to develop a generalized diverse subset selection strategy to filter out redundant or similar compounds and design optimal diverse subsets to enhance the structural diversity of the large databases such as the one found in PubChem. The basic idea underlying the diverse subset selection strategy is based on a central premise of medicinal chemistry that structurally similar molecules have similar physicochemical properties and possibly similar biological activities, which premise has been enunciated in many computational chemistry and similarity-based virtual screening drug design studies.[3−9]

There are several computational approaches for analyzing molecular similarity and diversity of compound libraries. These approaches have played a significant role in computer-aided drug design.[10] The available methods include the dissimilarity-based technique (e.g., OptiSim[11] and DivPick algorithms[12]); clustering techniques (e.g., JP clustering[13] and Ward clustering algorithms[14]); and partitioning cell-based techniques (e.g., BCUT metric algorithm[15]). Similarity and clustering techniques are principally based on the pairwise comparisons of molecules, whereas the cell-based partition methods do not depend on pairwise molecule and distance comparisons and are therefore in principle applicable to much larger data sets like the PubChem database. The cell-based partition methods rely on the characterization of chemistry space, which is defined as an *N*-dimensional "space" consisting of *N* different descriptors of molecular structure features and properties.[16] Therefore, the chemical diversity analysis is strongly dependent on how the molecular structure is measured and described or how a collection of compounds is represented digitally on a computer to best reflect their similarity or diversity, and the compounds are grouped accordingly.[10] There is a wide variety of different measures or descriptors of molecular structure, including topological descriptors, physical property descriptors, atom and bond counts, surface area descriptors, and charge descriptors.[17] The topology of molecular size, shape, and degree of branching could be described by a great quantity of information from the molecular connectivity parameters.[18,19] Different chemical reference spaces typically produce different compound distribution. No matter how chemistry space is defined, its generation for computational analysis strictly depends on the selection of descriptors of molecular structure and properties.

In the present study, we chose to use the cell-based partition algorithm, i.e., the BCUT metrics (*B*urden *C*AS *U*niversity of *T*exas) method developed by Pearlman[15] on the basis of extensions of the molecular identification number originally reported by Burden.[20] This basically uses Burden descriptors (or eigenvalues of matrices) with a combination of the atomic number for each atom and a description of the nominal bond types for adjacent and nonadjacent atoms, to describe a compound simply by combining two or more measures of atom-based properties into a single matrix value. The BCUT matrices expand the number and types of atomic features that can be considered and also provide a greater variety of proximity measures and weighting schemes. Such a whole-molecule descriptor has shown significant utility in the measurement of molecular diversity and correspondent tasks.[17] In this manuscript, a structure-diverse subset was generated from the large PubChem database using the BCUT-based diversity analysis method. Combinations of BCUT parameters were first calculated and used to analyze and compare a wide variety of chemical structures in PubChem. The BCUT parameters and the defined BCUT matrices were then used for the diverse subset selection while maintaining a reasonable distribution of structures within the diversity space. Subsequently, the compound classification analyses were performed for both the parent and the representative databases by comparing their 1D and 2D molecular descriptors. The compound library comparison studies further confirmed that the molecules in the newly generated representative compound subset (540K) have the compound classifications and molecular property distributions similar to the parent PubChem database (5.3 millions compounds, May 2006 version). In addition, the pairwise similarity comparison studies were also conducted to evaluate the self-similarity and diversity properties of the representative subset and PubChem databases based on the 2D fingerprint and Tanimoto coefficient (Tc) calculation. By using BCUT matrices and cell-based partition chemistry-space calculation as the descriptions of the molecular structures, we have demonstrated that the calculated BCUT descriptors and matrices could provide unique information regarding molecular structure and diversity, and accordingly a representative and structurally diverse compound sublibrary was assembled.

METHODS

Cell-Based BCUT Metrics Calculation and Diverse Subset Selection. In order to generate the representative subset library, the cell-based partition algorithm was used to calculate the compound chemistry spaces of the parent library PubChem based on multiple molecular descriptors, including molecular connectivity, interatomic distances, and diverse molecular properties. A computation protocol was established using Tripos Diverse Solutions (DVS) program[21,22] to calculate atomic/molecular descriptors, to choose an optimum combination of these descriptors as a multidimensional chemistry space, and to provide the finest differentiation between compounds in order to build a representative compound or a diverse subset selection.[23] The protocol is given in the following: (i) clean up the source database by using the SELECTOR program[21] to filter out all metal-containing compounds and mixtures; (ii) perform the DVS method to calculate the matrices-composed chemistry space of the filtered PubChem compound database and identify the acceptable statistical significance and dimensionality in terms of which set of descriptors or combination

of BCUT values would form the chemistry space that best characterizes the chemical diversity of a given population of compounds;[15] and (iii) conduct chemical diversity and compound classification comparisons to analyze and evaluate both representative and parent libraries.

During the DVS calculation, molecular descriptors were first computed to express the molecular structure features and properties using the BCUT metrics approach reported by Pearlman.[20,15] In such a matrix calculation, a compound with a 'molecular ID number' was represented as the H-suppressed connection table of the molecule in terms of the highest and lowest eigenvalues of a matrix. In our established DVS protocol, multidimensional chemistry spaces were computed for four typical classes of matrices, which were defined by the diagonal elements representing the relevant atomic properties, including atomic Gasteiger−Huckel charges, polarizabilities, H-bond donor (HBD), and H-bond acceptor (HBA) descriptors. These four descriptors correspond to the electrostatic, dispersion, and H-bonding modes of bimolecular interactions. The matrix also encodes information about the 3D structure using various functions of interatomic distance and connectivity on the off-diagonal elements and applies atomic surface areas to weight the atomic properties on the diagonal and connectivity information on the off-diagonals with the defined scaling factors. Specifically, 0.1 times the nominal bond-type for two bonded atoms and 0.01 for two unbonded atoms were used as the off-diagonal molecular connectivity values. The scaling factors of 0.9, 1.5, and 2.5 were applied to weight the BCUT metrics of atomic charge, HBD/HBA, and polarity descriptor, which generated a total of 18 descriptor parameters under the highest and lowest eigenvalues.[21,22] These descriptors permit the generation of low-dimensional reference spaces consisting of typically no more than six orthogonal axes.[15] In our calculations, the low-dimensional matrices were identified to construct a small number of molecular descriptors (typically 3−6) and were defined by three factors: type of descriptors, number of descriptors, and range of values for each descriptor, in order to take advantage of the power and utility of cell-based diversity algorithms.

In order to evaluate the calculated BCUT matrices and identify the best set of molecular descriptors that provided the optimal distribution of the compounds, a powerful 'autochoose' algorithm developed by Pearlman[15] was used in our DVS studies. A systematic determination algorithm was then conducted to identify both the best dimensionality of the chemistry space and the best choice of matrices or matrix combination that would best represent the structural diversity of a given population of compounds. Such a combination requires that the axes of the chemistry space be mutually orthogonal.

Finally, the cell-based diverse subset was selected from the parent compound library on the basis of the matrices value and chemistry space created above.[23] Basically, a binning procedure is used to generate cells for compound partitioning from a multidimensional original descriptor space.[16] The multidimensional chemistry space is then defined into a small finite number of "bins" on each axis of that space. The bin definitions, in turn, define multidimensional "cells" which, altogether, cover the entire space. First, the subset selection was performed using a proportionally cell-based sampling process with a sample size corresponding to roughly 10% of the actual library. In accordance with the cell-based algorithm, each of the descriptor axes defined above was subdivided into M bins of equal size, yielding M^n hypercubic cells. Subsequently, the optimum values of M bins were defined for the selection of subset libraries using established sampling protocols with uniform or approximately equal sampling from each cell. The chemistry spaces for the entire library and the DVS-selected subset were graphically visualized and compared using Tripos Sybyl software and also plotted using the free software of Gnuplot (downloaded from http://www.gnuplot.info.).

Tanimoto Coefficient (Tc) Calculation. To further evaluate the internal diversity and self-similarity properties of the new subset generated as compared with the parent database, we have carried out pairwise Tc calculations based on 2D molecular fingerprints. Basically, a 2D fingerprint represents each molecule by a string of bits, 0's and 1's in a linear bitmap, and each bit indicates the absence or presence of the molecular features or particular fragments in the molecule.[14] The similarities between two compounds A and B were described by the Tanimoto coefficient $T(a,b)$[24] with the equation

$$T(a,b) = \frac{c}{a + b - c}$$

where a and b are the number of bits set in fingerprints of compounds A and B, and c is the number of bits set in common. Tc has values between 0 and 1. The dissimilarity is then defined as $1-T(a,b)$. A higher Tc value means two molecules have higher similarity. As far as the fingerprint descriptor is concerned, a Tc of 1 means that both structures have the same fingerprint, while a value of 0 indicates that there is no overlap between these two fingerprints. The computations were done by using the Tripos SELECTOR program. Molecular fingerprints were calculated using the standard Tripos UNITY 2D fingerprints, and the data processes were carried out using Sybyl Molecular Spreadsheet.[21]

Database Molecular Property. To further confirm the validity of representativeness of the sublibrary generated from PubChem, the Tripos SELECTOR program[21] was used to perform compound classification and molecular property distribution analyses based on the 1D or 2D molecular descriptors, including molecular weight (MW), ClogP, rotatable bond, hydrophobic center, H-bond donor, and H-bond acceptor. The definitions for each descriptor in the compound classification calculation are defined below.

The rotatable bond is defined as any single bond except for a partial double bond such as an amide bond (C(═O)−N). A hydrogen bond is defined as the electrostatic attraction between a hydrogen bearing a partial positive charge (acceptor) and a high electronegative atom (usually O or N) bearing a partial negative charge (donor). The H-bond acceptor can either be an SP2 state O atom in the functional groups with high partial negative charge, such as COOH, COO, C═O, C═S, or R2N═O, R2S═O, and nitrone, or an SP3 state O atom in the alcohol, ethers, and aromatic oxygens (e.g., furan) or as the second oxygen atom in the functional groups of COOH, COO, and ester. The H-bond acceptors also include the lone pair position in nitriles, RN═C or HN═C, but exclude protonated amidines and guanidines, unpro-

tonated acyl, sulfonyl, cyano, and nitro amidines. Furthermore, a nitro group and sulfur or phosphorus with an oxygen atom can also be considered as an H-bond acceptor. The H-bond donor normally is a proton atom attached to a highly electronegative atom. Therefore, the H-bond donor includes amide or amidine NH, pyrrole type NH other than tetrazole, or imine NH, any hydroxyl group excluding COOH or SOOH, and the basic imine nitrogen. The N atom in an aromatic system (such as protonated pyridine N except for pyridone tautomers, 1,2- or 1,3-tautomerizable NH diazoles other than tetrazoles, basic N-substituted imidazole) and the O atom in an aromatic system (such as the hydroxyl azole tautomers of lactams, tautomerizable hydroxyl azoles, 2- or 4-pyridone tautomerizable from hydroxyl pyridine) are also recognized as the H-donor sites.

ClogP values were computed in this study in order to estimate the molecular partition coefficient, i.e., the ratio of equilibrium concentration of solute (chemical) in nonpolar solvent (n-octanol) and polar solvent (water). The logarithm of the partition coefficient, logP, is a very important physical chemical property parameter, is used to measure the hydrophilic ("water loving") or hydrophobic ("water hating") properties of a chemical substance, and is useful in estimating the distribution of drugs within the body. ClogP is a fragment-based method of calculating logP.[25] The detail calculation algorithm is described in the Sybyl manual and literature.[26] ClogP calculation methods have been widely used for predicting drug logP values in virtual screening studies.[27] The hydrophobe or the hydrophobic center was defined as hydrophobic rings (e.g., 4-, 5-, or 6-member alicyclic or aromatic rings), and aliphatic chains (more than three hydrophobic atoms except for electronegative atoms). The calculated molecular properties and comparisons between the new subset and the parent libraries are described in the Results and Discussion section.

RESULTS AND DISCUSSION

We have applied a cell-based partition method to generate a representative subset from the PubChem compound database by the BCUT metrics approach. A diversity analysis of the new subset was performed on the basis of chemistry-space matrix computation. The new subset was then evaluated in comparison with its parent database in terms of their chemical space and compound classification analyses. To prepare the source library, 5.3 million compounds in PubChem (the May 2006 version when this work was performed) were downloaded from the NIBC Web site in SDF (standard data format) file format and were converted to 2D SLN (Sybyl Line Notations) format (or Unity hitlist format). Subsequently, a filtering process with the default parameters was carried out by using the Tripos SELECTOR program to identify the metal-containing compounds and mixtures in the PubChem compound library. After filtering out the 60 688 metal-containing compounds and 82 291 mixture compounds, 5 172 573 compounds were obtained. In addition, the CONCORD program[21] was used to automatically generate 3D coordinate files. Both 2D SLN files and 3D molecular files were used in later diversity analysis and molecular property distribution computation.

Representative Library Generation and Diversity Analyses. The first step of the diverse-subset-selection computational study was to identify the optimum percentage of the compound-filled cells in terms of the different percentage outliers, the defined bins per axis (dividing the cubic chemistry space), and the maximum chemistry-space dimensions (including Gasteiger−Huckel atomic charge, H-bond acceptor, H-bond donor, and polarizability). This was accomplished using the Diverse Solutions (DVS) program (Tripos implementation), as illustrated in Figures 1 and 2. By definition, outliers are the compound structures having extreme values of computed properties and may represent unusual molecules of little interest as drug candidates. Theoretically, it is recommended that outliers up to 10% are allowed during chemistry-space generation.[23] However, the percentage of outliers chosen is related to the number of bins divided along each axis and the maximum number of molecular descriptors (i.e., the maximum chemistry-space dimensionality) as described in the Experimental Section, which will affect the optimum percentage (12−16%) of the cells filled in the matrix for DVS compound selections. In DVS calculation, the concept of "outlier" was defined as the compounds which may have metric values outside the limits of the chemistry space (or the acceptable ranges for each axis represented by a typical molecular descriptor). To characterize and control the number of cells and the cell occupancy as well as outlier percentage, we conducted computational studies to search for and adjust the maximum number of molecular descriptors and the bins of each axis in a diversity analysis calculation.

In general, although 18 chemistry-space descriptors were calculated for this study, the cell-based algorithms could only be applied in low-dimensional chemistry spaces as restricted by the availability of suitably low correlated metrics and the limitations in the BCUT approach.[15,16] Therefore, it is important to identify the maximum number of molecular descriptors and the optimal number of bins per axis and the percentage of outliers in terms of the resulting percentage of cells filled by compounds. In our calculations, smaller subsets of BCUTs (4−6) were automatically selected by a χ-squared-based "autochoose" algorithm[15] to best represent the structural diversity of a given compound collection. This defined low dimension chemistry space was then partitioned into cells for compound selection and database comparison.[28] An illustration is given in Figure 1A,B to reveal how to determine and validate variations in descriptor ranges and percent cell populations based on the given structural set. The results demonstrated compound chemistry spaces generated under the different maximum number of molecular descriptor ranges and the equal-size bins of 4−10 per descriptor, along with the generated outliers of 0%, 2.5%, 5%, 7.5%, and 10%. Figure 1 also shows that cell occupancy would be increased by changing the percentage of outliers from 0% to 10%. The results also show that poor cell occupancy (or the compound cell population is over 40% of percent cells filled) was obtained with different combinations of axis/bins and outliers when the maximum chemistry-space dimension was set to four (Figure 1A). However, when the maximum chemistry-space dimensionality was set to five (Figure 1B), the recommended ideal cell occupancy range of 12−16% was achieved under the selected 10 bins of each axis in the case of 0%−10% outliers. With the identified optimum chemistry-space parameters above, further diversity analysis calculation studies were conducted as follows.

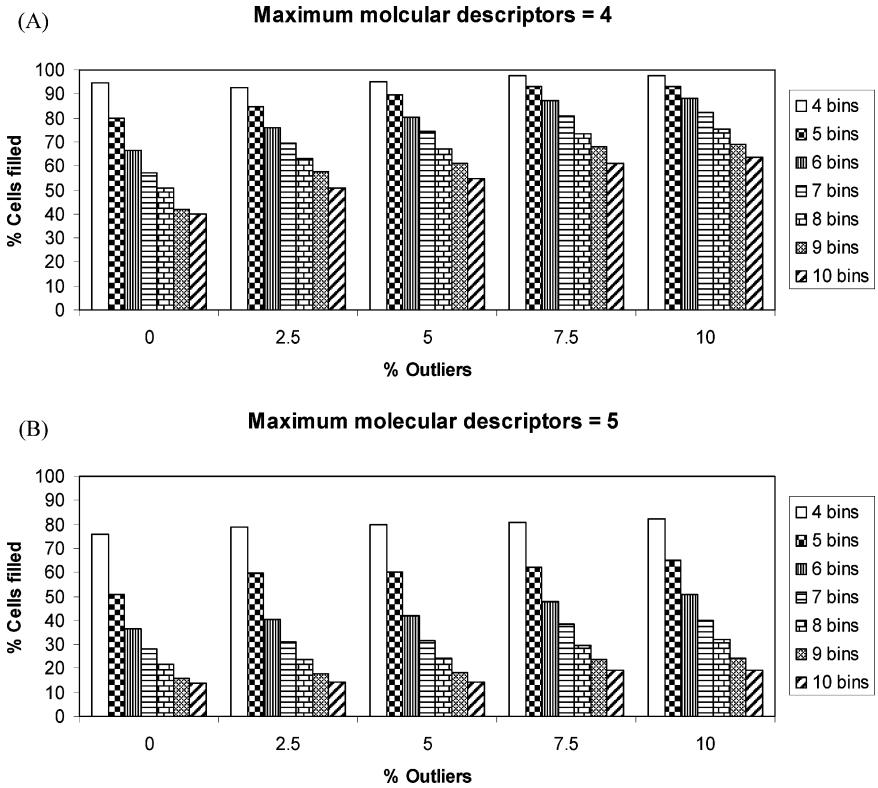

Figure 1. Illustration of the compound chemistry-space cell population (% cells filled) vs percent outliers for various bins per axis under different maximum numbers of the molecular descriptors or the maximum chemistry-space dimensionality.

Figure 2 gives an example of the chemistry-space BCUT matrix analysis results for the parent and new subset libraries under the defined maximum chemistry space dimensionality of 5, the 10 bins per axis and 0% outliers. The 3D BCUT matrix plot in Figure 2 (It was difficult to visualize the point-by-point comparison by plotting 5 million matrix dots. Thus only 900K compound data sets are displayed as an example for a plotting illustration.) exhibits the overlapping matrix dots of the compounds from the source library (red dots: 900K compound data sets) and the representative subset generated from the DVS study (blue dots: 90K). An expansion plot of the zoomed-in cubic region is displayed as an insert in Figure 2 (top). As shown, the blue chemistry-space matrix dots (compounds in the representative subset) are smoothly spread out and evenly distributed on top of the red matrix dots (compounds in the parent library) in the 3D plots. In this study, each of these dots, representing a chemistry-space matrix, was calculated for a given compound in the database by using DVS tools based on the generated BCUT molecular descriptors, including atomic charge, H-bond donor, H-bond acceptor, and polarizability.

Subsequently, the selection of the cell-based diverse subset was performed on the parent compound database. We set the desired size of the subset to be a maximum of 10% of the original source compound database, a conservative "empirical" percentage in order to avoid missing too many dissimilar compounds or retaining too many similar compounds in the new subset. For example, in our trial calculation, the subset of 89K compounds was generated from an original 900K compound database to compose the representative library. The DVS program tool chose the compounds closest to the center of cells and ensured that the compounds from the neighboring cells are as distant as possible. Such a procedure is referred to as an "unbiased" subset selection. The uniform sampling protocol,[21] which selects the same number of compounds from each cell throughout chemistry space, was set to select a maximum of one compound from each cell throughout chemistry space.

In our diversity calculation, we selected 10 bins on each molecular descriptor axis to divide the metric chemistry space into smaller cell units with a maximum chemistry-space dimension of 5 defined above, which gave the 10 bins per axis of a 12−16% cell occupancy (Figures 1 and 2). As discussed above, the number of cells and the cell occupancy is another important factor in the diverse analysis using the DVS program. It is important to point out that the larger cells (or less number of cells) give the illusion of "tighter clustering" or fewer unit clusters required to contain the structure diverse compounds, whereas each unit cluster will also contain more similar compounds in this case. On the other hand, smaller cells (or more number of cells) give the illusion of "poor clustering", or more unit clusters required to contain diverse compounds, but each unit cluster will contain fewer similar compounds.[21] Empirical calculations have demonstrated that the most meaningful results are obtained when the number of bins/axis results in 12−16% compound cell occupancy,[23] i.e., 12−16% of all cells contain one or more compounds. Such cell occupancy would keep the number of empty cells at a reasonable level and limit the number of compounds per occupied cell, which gives a more diverse compound distribution for compound selection.

Compound Database Self-Similarity Analyses and Comparisons. Furthermore, a pairwise similarity comparison was also conducted to evaluate the self-similarity and diversity properties of the representative subset and PubChem databases based on the 2D fingerprint and Tanimoto coefficient (Tc) calculation. It should be pointed out that pairwise distance-based diversity algorithms consider only the dis-

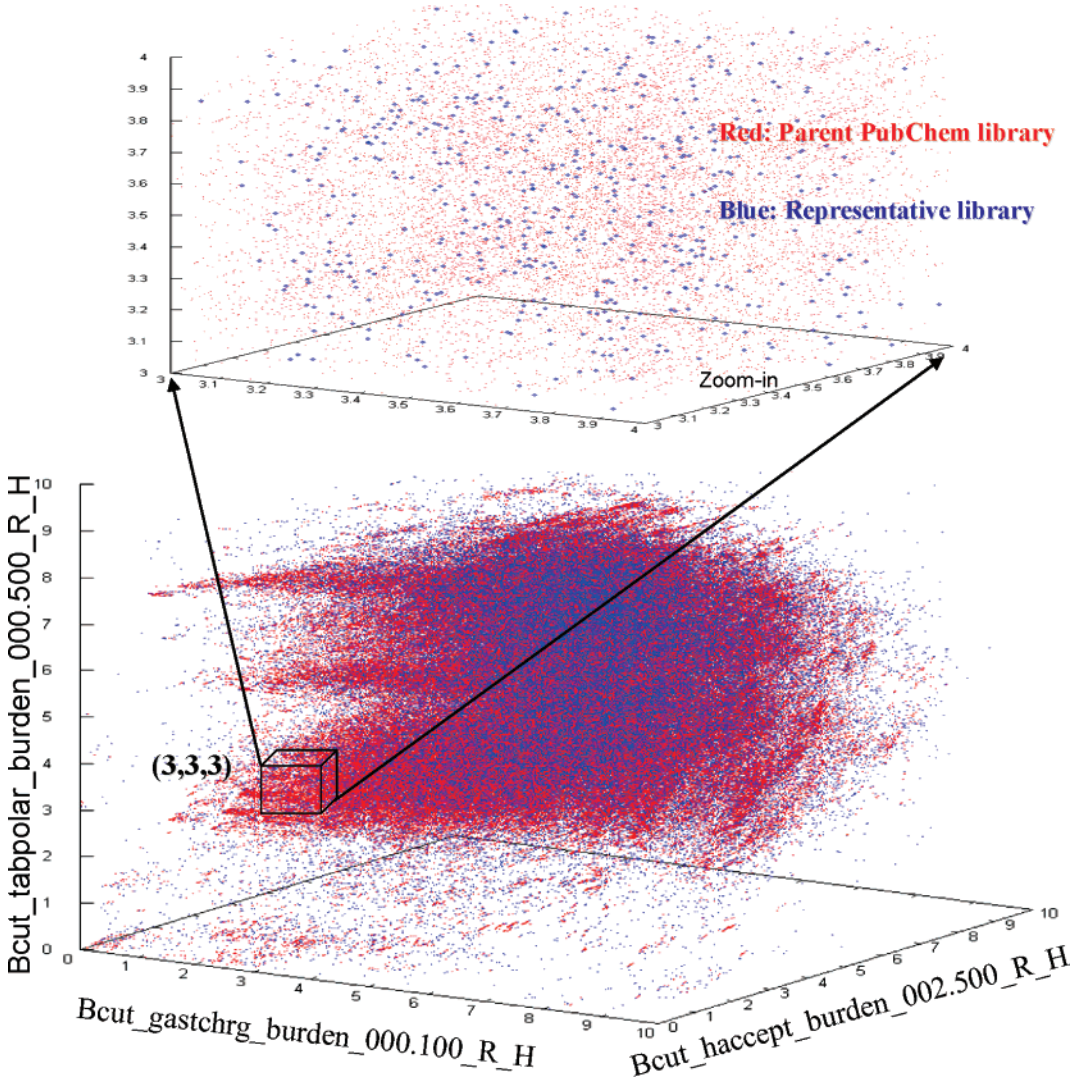

Figure 2. Graphic representation of a representative sublibrary created from the parent library by using cell-based diversity analysis based BCUT metrics calculation. The red dot represents the chemistry-space matrix of a compound in the PubChem database, and the blue dot is the matrix dot of a compound in the representative subset generated from its parent PubChem database by using DVS.

tances between compounds not the absolute positions of compounds in chemistry space. Thus, the algorithm is inherently limited and is ill-suited for many other diversity-related tasks such as locating diversity voids in chemistry space. In contrast, the cell-based diversity algorithm partitions chemistry space into a lattice of multidimensional hyper cubes and, therefore, considers not only the intercompound distance but also the absolute position of compounds in chemistry space.[15] On the other hand, although the Tanimoto coefficient has been widely applied in similarity studies or virtual screening, the Tc values are often not evenly distributed in the diversity analysis, and there are some statistical limitations in the use of the Tc metrics based on the reports.[29] Furthermore, regarding the varied specifics of the fingerprints, a Tc value of 1.0 means two similar molecules but does not necessarily mean that the two molecules are identical. Therefore, the pair-wised Tc similarity calculation was only used in this study as a secondary evaluation tool in order to confirm the self-similarity and diversity properties for the target databases. This was done by computing mean Tanimoto coefficient (mean Tc or mTc) values based on the 2D UNITY fingerprints of all molecules in a compound database in order to confirm that the DVS-generated diverse subset samples a wide range of diverse and nonredundant compounds from the parent database.

Figure 3 shows the internal-similarity calculation plots for the percentage of compounds (or percentage of compound population within the same Tc range) versus the Tanimoto coefficient value for the PubChem database (A) and its DVS-created representative subset library (B). Basically, each compound in the database was compared to the rest of the database, and the Tc values were recorded as a function of the percentage of compounds (Figure 3). Such a self-comparison of each compound in the database generates a histogram of the similarity index distribution and the associated mean Tc value, which is used to capture the scope of chemical diversity of the given database, i.e., the smaller the mean Tc, the greater the range of database structure-diversity. Our data show that the new subset constructed by the cell-based partition method above has a mean Tc value of 0.76 (Figure 3B), which, as expected, is lower than that of the source PubChem database (0.91) (Figure 3A). The data are also reflected in the changes of percentage of compounds distribution in the similarity index histograms as shown in Figure 3A,B. Our Tc value calculation showed a good Gaussian distribution of compound percentage for

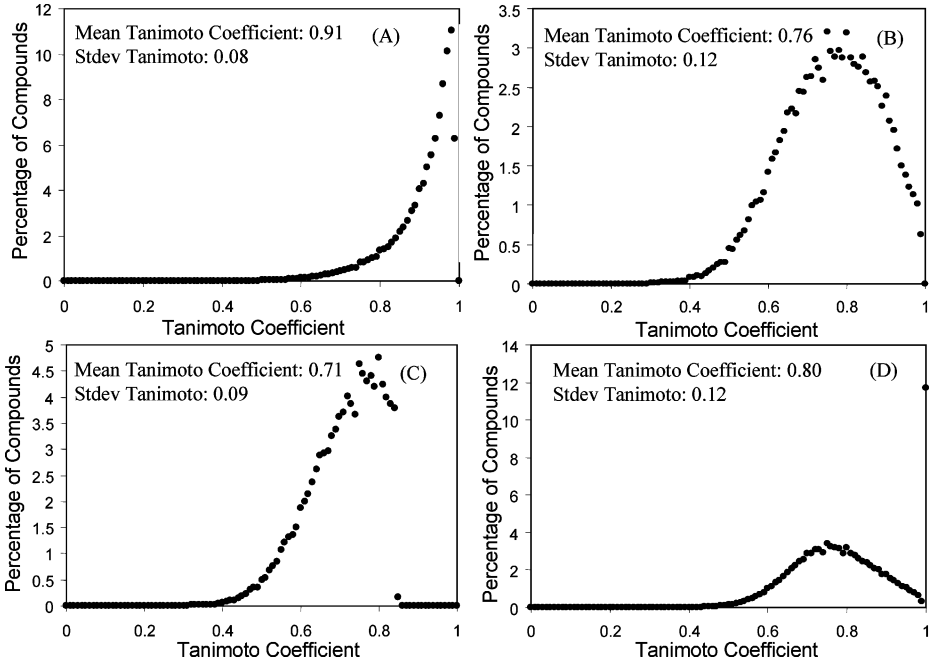

Figure 3. Database internal-similarity plots of the calculated Tanimoto coefficients (Tc) versus the correspondent percentage compound population for the PubChem database (A) and the DVS-selected representative subset library (B), the Tc < 0.85 selected representative subset (C), and the database comparison between the DVS-selected and its parent PubChem libraries (D). The compound Tanimoto coefficient values were calculated with a pairwise 2D fingerprints metrics method. High mean Tanimoto coefficients indicate a high degree of similarity of the compounds in the data set.

the representative subset with a mean value of 0.76 and a standard deviation of 0.12) (Figure 3B), whereas a skewed compound percentage distribution was observed for the source PubChem database with a mean value of 0.91 and a standard deviation of 0.08 (Figure 3A). Our data also revealed that there are still a few percent of compounds with Tc > 0.85 (Figure 3B), a reported statistic value being that if a compound has a Tc similarity ≥ 0.85 to an active compound, then the compound has an 30% chance of itself being active in the same assay.[30] This low fraction occurrence may be attributed to the different structure representations (a property-derived matrix versus a fingerprint) and the different modes of processing (cell-partitioning versus similarity calculation) used in these two algorithms as discussed above. We tried and used the secondary compound-selection method to remove these compounds with Tc > 0.85 by using 2D fingerprint-based Tc similarity computation, and the results are given in Figure 3C. However, the potential risk is that the Tc dissimilarity diversity selection method may have resulted in the loss of some important "similar' compounds that have certain key hydrophobic or halogenated functional groups as discussed below.

Furthermore, we compared the DVS-selected sublibrary with the parent database using the pairwise Tc comparison algorithm described above in order to examine the representativeness of the DVS-generated subset with respect to the parent database. The results of the two database diversity comparison are given in Figure 3D. The data show no compounds in the two databases with a pairwise Tc lower than 0.36 and approximately 11% of the compounds with a Tc value of 1.0, which is to be expected as the representative compound subset was selected from the parent database. A well-defined Gaussian distribution pattern (a mean value of 0.80 and standard deviation of 0.12) (Figure 3D) suggests that the compound diversity of the subset is evenly distributed in the parent database with a mean Tc of 0.80.

In order to further verify the diversity and representativeness of the new subset with respect to the source PubChem database, we randomly selected nine compounds from the representative library under the Tc ranges of 0.95−0.99, 0.85−0.86, and 0.75−0.76, respectively. These nine compounds were used as the target compounds to examine the searched hitlists from the representative and PubChem databases. As illustrated in Figure 4 for a search against the compound CID743200 (PubChem compound ID: CID), the similarity search gave three hits (one is itself) from the representative library and 29 similar hits (including itself) from PubChem databases using the Tripos SELECTOR similarity function with Tc > 0.85. The remaining target compounds also gave similar results in hitlist comparisons of similarity searches between the representative and PubChem databases. Again, the 2D fingerprints-based pairwise Tc calculation further confirmed that the cell-based DVS selection produced a smaller and less redundant compound subset library with higher structure diversity. It should be pointed out that the 2D fingerprint-based Tc similarity search considers the compounds CID564003 and CID743213 "similar" to CID743200 based on Tc > 0.85 (Figure 4), whereas the cell-based partition BCUT matrix calculation treated them as "dissimilar" to the target compound (or in different cells) based on the defined molecular descriptors (Br charge or Ph hydrophobic properties). If 2D fingerprint similarity is used to generate a subset, these compounds will be eliminated because of their high similarity Tc values. However, the bromine compound (CID564003) and amphipathic compound (CID743213) may have important biological values for drug discovery. Phenyl and bromine analogs are structurally diverse compounds and may in some biological tests behave similarly

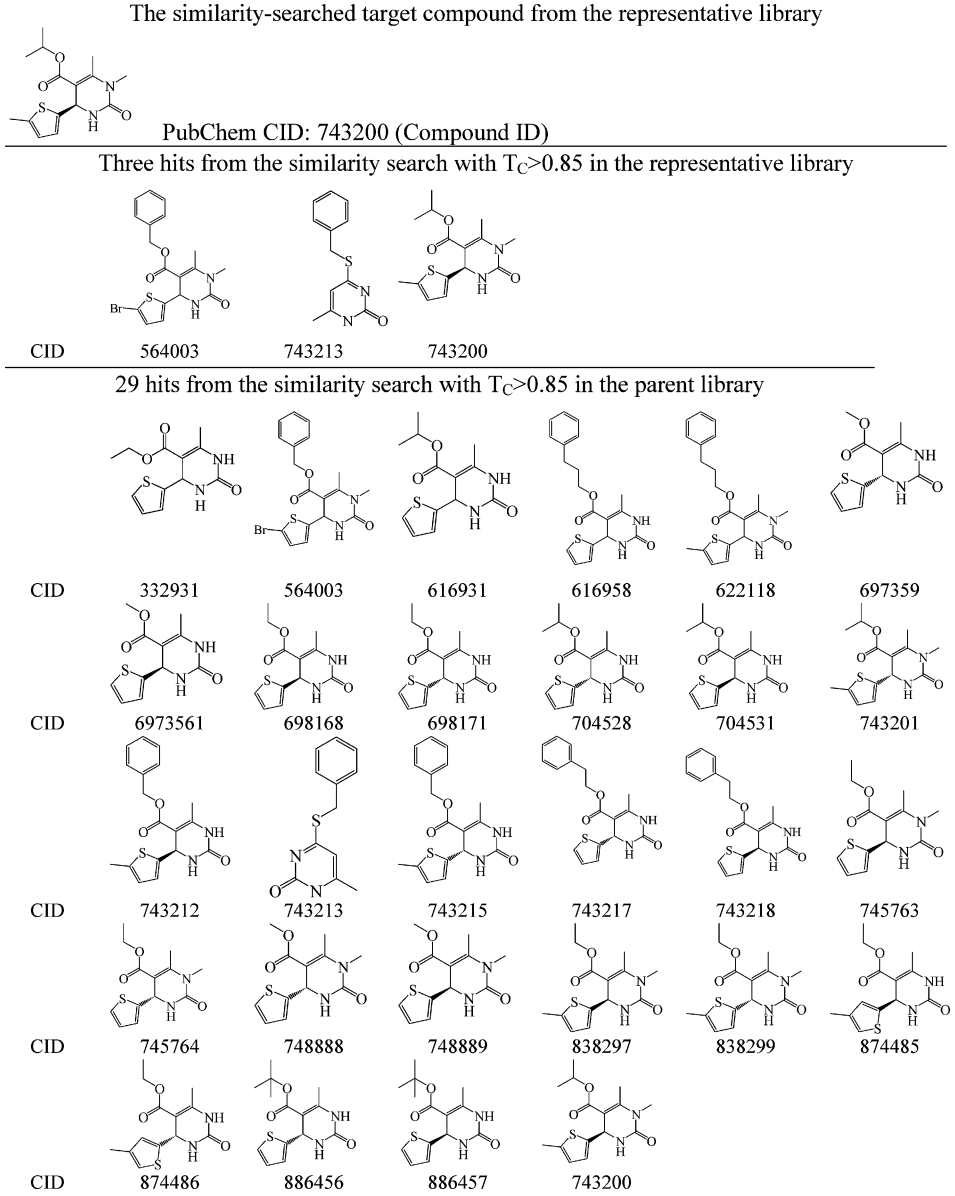

Figure 4. The compound (PubChem CID743200) similarity search comparison of the hitlists from the representative and parent libraries by using the Tripos SELECTOR program with Tanimoto coefficients larger than 0.85.

and in others differently. Again, the data also show that the cell-based partition chemistry-space approach uses different algorithms and different descriptors versus pairwise Tc similarity calculation and gives unique results in diverse compound selection using the present computational programs.

Compound Classification and Molecular Property Analysis Comparison. To further validate the representative sublibrary generated from the PubChem database, we have also carried out systematic comparison studies of both the parent database (5.3 million compounds) and the DVS-generated subset (540K) in terms of the defined molecular property descriptors.

Figure 5A−F summarizes the results of the compound classification and the molecular property distribution analyses of the PubChem and the representative databases and shows similar peak distribution patterns for the two databases in terms of molecular weight (MW), rotatable bond, ClogP, H-bond donor and acceptor, and hydrophobe. As shown in Figure 5A, the MW distribution analyses of both compound libraries indicated that approximately 87% of molecules in both databases have an MW less than 500. Figure 5B shows a similar peak distribution for the molecules with different numbers of rotatable bonds in both databases, indicating that 90% of the compounds in both databases have 1−9 rotatable bonds. Figure 5C summarizes the ClogP distribution plot comparison for PubChem and representative databases, showing that 73% of the molecules in both libraries have ClogP values less than 5. Figure 5D gives a comparison plot of H-bond donor distribution analyses. A very similar distribution pattern was observed between the two databases, showing 98% of the compounds in the databases having less than 5 H-bond donors. Figure 5E presents the distribution analysis for H-bond acceptors of the compounds in the two databases, showing 84% of the compounds with 1−7 H-bond acceptors. Only 6% of the compounds in the PubChem library have no H-bond acceptor or over seven H-bond acceptors.

In addition, we have conducted a distribution analysis of hydrophobic features, or hydrophobe for both libraries using

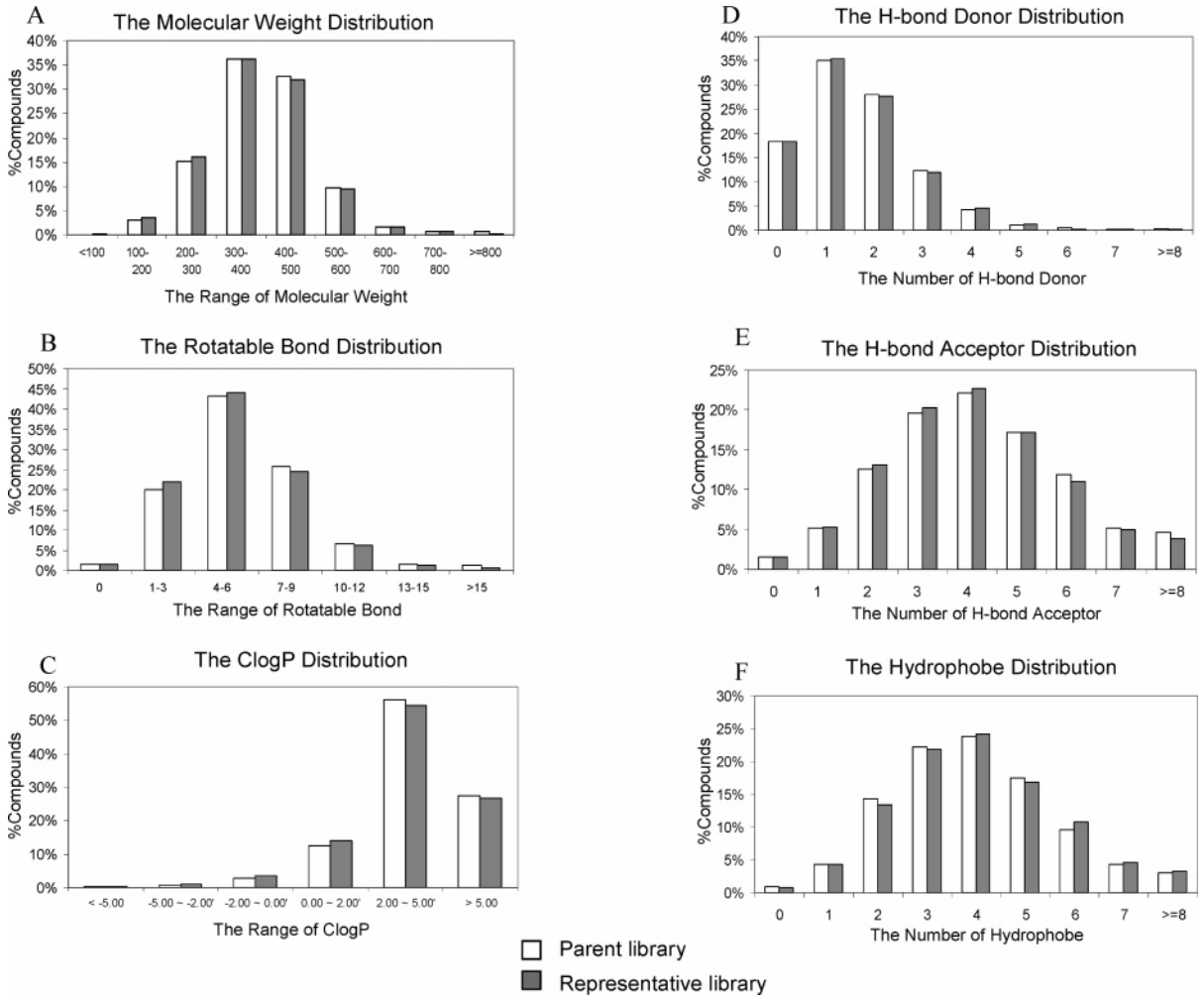

Figure 5. Plots of the compound classification and molecular property distribution comparisons between the parent PubChem database (unfilled bars) and the DVS-generated new representative subset database (filled bars): (A) the molecular weight (MW) distribution; (B) the rotatable bond distribution; (C) the ClogP distribution; (D) the H-bond donor distribution; (E) the H-bond acceptor distribution; and (F) the hydrophobe distribution comparisons.

the Tripos SELECTOR program,[21] as shown in Figure 5F. Although Lipinski's rule of five did not include directly critiques for the drug hydrophobic feature (only via clogP), the hydrophobic property is another important factor that is often considered in pharmacophore-based virtual screening. Hydrophobic features are calculated on the basis of the hydrophobic groups defined in the Method section. As shown in Figure 5F, both databases have a similar peak distribution of hydrophobes, showing approximately 96% of the compounds with 1−7 hydrophobes and 87% of the compounds with 2−6. The representative subset has hydrophobe peak distribution patterns similar to the parent library.

As all of the data above have shown, the distribution analyses using 1D (MW, ClogP) and 2D descriptors (H-bond donor, H-bond acceptor, rotatable bond, and hydrophobe) confirm that similar peak distribution patterns can be observed between the PubChem and representative databases. The results also demonstrated that most compounds in the PubChem library meet Lipinski's rule of five:[31] an orally active drug should generally have the molecular properties of a molecular weight under 500, a partition coefficient (logP) under 5, hydrogen bond donors less than 5, and hydrogen bond acceptors less than 10. However, the partition coefficient descriptor (or ClogP) was a little off from the ranges in the Lipinski's rule of five, showing about 76% of compounds in the databases having a ClogP value less than 5 for both databases. As described in the Method section, the LogP value is often used for measuring hydrophobicity and is related to absorption, bioavailability, ligand−receptor interactions, metabolism, and toxicity. Apparently, the ClogP < 5 cutoff rule was not used to prefilter out some of the compounds from the parent PubChem library because the wider range of ClogP or expanded Lipinski's rule may offer more choices of structurally diverse compounds for drug screening. Actually, in virtual screening and in silico drug design, larger ClogP values (e.g., up to 6) are often used in lead hit selection since the lead optimization could be carried out to improve the logP value and other ADME (Absorption, Distribution, Metabolism, and Excretion) values. Nevertheless, the H-bond donor and acceptor descriptors are in the range of Lipinski's rule, indicating that most of the compounds in PubChem may have already gone through a prefiltering process to remove the abnormal compounds before they were deposited into the repository.

The results above indicate again that the distributions of important molecular property descriptors of the subset are congruent with the ones of its parent PubChem compound database in terms of molecular weight, rotatable bond, ClogP, H-bond donor and acceptor, and hydrophobe. Each of these key molecular descriptors has very similar Gaussian distribu-

tion patterns. Therefore, the sublibrary selection procedure established in this study does not change molecular properties of the compounds in the database, whereas the representative data set has a higher structure diversity showing a mean Tc < 0.76 and covering the same chemistry space of the parent database. Such results further demonstrate that the DVS-generated compound sublibrary has similar molecular properties but a smaller compound subset with better chemical diversity in comparison to its large parent PubChem.

CONCLUSION

We have presented a strategy to establish the cell-based chemistry-space partitioning algorithm in selecting a diverse compound subset from the large PubChem database. Our studies were based on the known similarity principle that structurally similar molecules may exhibit similar physicochemical and biological properties. It has been rationalized that the use of very similar molecules for drug screening does not enhance the probability of finding new bioactive leads that have novel chemical scaffolds, but rather using structurally dissimilar molecules should increase the probability of finding interesting leads. Our results have demonstrated that the established representative compound selection approach and diversity analysis method were able to select a smaller subset (540 K, ~10%, mTc =0.76) that best represents the full range of chemical diversity or entire molecular property space, including MW, ClogP, H-bond acceptor/donor, and the hydrophobic feature, presented in the parent PubChem database (5.3 million, mTc=0.91) with minimal similarities or redundancies. The new representative PubChem database generated, or *rePubChem*, will be available online on our laboratory Web site and could be used as an initial drug screening library using either in silico or in vitro HTS methods to avoid the high cost in time and money to screen similar compounds in a large database. Once the novel leads are identified in the representative database, one can go back to the source PubChem library to find all of their analogs or derivatives for further in silico or in vitro HTS experiments in order to build the important SAR (structure−activity relationships) or identify the more potent compounds. It should be pointed out that there is no absolute correlation of biological activities to 2D structural similarity.[30,32] It is advantageous to have a representative subset by which the screening can be carried out rapidly and less time wasted on unsuitable hits. However, no diverse subset (or no universal method) can identify all the hits that could be found from screening the whole database of compounds. In addition, dependent upon the type of similarity metric calculations used, very similar compounds can sometimes have quite different overall profiles of biological activities and properties. Currently, further studies of the representative compound library are being carried out on the newly available 12 million compounds PubChem database using the protocol developed. In silico virtual screening studies are also being carried out to examine the representative database.

ACKNOWLEDGMENT

This project is supported by grants from NIH R01 DA015417 and U54 MH074411. We thank Dr. John Lazo for reading this manuscript and professional discussion. We also acknowledge Drs. Bill Curtiss, Tom Jones, and Lei Wang at Tripos Company for technical support.

REFERENCES AND NOTES

(1) Lazo, J. S. Roadmap or roadkill: a pharmacologist's analysis of the NIH Molecular Libraries Initiative. *Mol. Interv.* **2006**, *6*, 240−243.
(2) Armstrong, J. W. A review of high-throughput screening approaches for drug discovery. *Am. Biotechnol. Lab.* **1999**, *17*, 26, 28.
(3) Lajiness, M. S.; Johnson, M. A.; Maggiora, G. M. Implementing drug screening programs by using molecular similarity methods. *Prog. Clin. Biol. Res.* **1989**, *291*, 173−176.
(4) Lewis, R. A.; Pickett, S. D.; Clark, D. E. Computer-Aided Molecular Diversity Analysis and Combinatorial Library Design. *Rev. Comput. Chem.* **2000**, *16*, 1−51.
(5) Pozzan, A. Molecular descriptors and methods for ligand based virtual high throughput screening in drug discovery. *Curr. Pharm. Des.* **2006**, *12*, 2099−2110.
(6) Johnson, M. A.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; John Wiley: New York, 1990.
(7) Johnson, M. A.; Maggiora, G. M.; Lajiness, M. S.; Moon, J. B.; Petke, J. D.; Rohrer, D. C. Molecular similarity analysis: applications in drug discovery. *Methods Principles Med. Chem.* **1995**, *3*, 89−110.
(8) Alvesalo, J. K. O.; Siiskonen, A.; etc. Similarity Based Virtual Screening: A Tool for Targeted Library Design. *J. Med. Chem.* **2006**, *49*, 2353−2356.
(9) Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006**, *11*, 1046−1053.
(10) Kitchen, D. B.; Stahura, F. L.; Bjorath, J. Computational techniques for diversity analysis and compound classification. *Mini. Rev. Med. Chem.* **2004**, *4*, 1029−1039.
(11) Clark, R. D. OptiSim: An Extended Dissimilarity Selection Method for Finding Diverse Representative Subsets. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1181−1188.
(12) Nilakantan, R.; Bauman, N.; Haraki, K. S. Database diversity assessment: New ideas, concepts, and tools. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 447−452.
(13) Jarvis, R. A.; Patrick, E. A. Clustering using a similarity measure based on shared near neighbors. *IEEE Trans. Comput.* **1973**, *22*, 1025−1034.
(14) 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.
(15) Pearlman, R. S.; Smith, K. M. Novel software tools for chemical diversity. *Perspect. Drug Discovery Des.* **1998**, *9*, 339-353.
(16) Stahura, F. L.; Bajorath, J. Partitioning methods for the identification of active molecules. *Curr. Med. Chem.* **2003**, *10*, 707−715.
(17) Stanton, D. T. Evaluation and Use of BCUT Descriptors in QSAR and QSPR Studies. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 11−20.
(18) Kier, L. B.; Hall, L. H. *Molecular connectivity in structure-activity analysis*; John Wiley & Sons: New York, 1986.
(19) Randic, M. On characterization of molecular branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609−6615.
(20) Burden, F. R. Molecular identification number for substructure searches. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 225−227.
(21) *Sybyl*, *version 7.3.3*; TRIPOS, Inc.: St. Louis, MO 63144, 2007.
(22) Pearlman, R. S.; Smith, K. M. Software for chemical diversity in the context of accelerated drug discovery. *Drugs Fut.* **1998**, *23*, 885−895.
(23) 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.
(24) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(25) Leo, A.; Jow, P. Y.; Silipo, C. Calculation of hydrophobic constant (log P) from pi and f constants. *J. Med. Chem.* **1975**, *18*, 865−868.
(26) Tripos-cLogP/CMR. *Tripos Sybyl 7.3.3*; TRIPOS, Inc.: St. Louis, MO 63144, 2007. http://www.tripos.com/data/SYBYL/ClogP_CMR_072505.pdf (accessed May 2007).
(27) Eros, D.; Kovesdi, I.; Orfi, L.; Takacs-Novak, K.; Acsady, G.; Keri, G. Reliability of logP predictions based on calculated molecular descriptors: a critical review. *Curr. Med. Chem.* **2002**, *9*, 1819−1829.
(28) Pirard, B.; Pickett, S. D. Classification of Kinase Inhibitors Using BCUT Descriptors. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1431−1440.
(29) Flower, D. R. On the properties of bit string-based measures of chemical similarity. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 379−386.
(30) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity? *J. Med. Chem.* **2002**, *45*, 4350−4358.

(31) 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.

(32) Stanton, R. V.; Cao, Q. Biological fingerprints. *Comp. Med. Chem. II* **2006**, *4*, 807−818.

CI700193U