12

ENUMERATION AND VISUALIZATION OF LARGE COMBINATORIAL CHEMICAL LIBRARIES

Sung-Sau So

Formerly of Hoffmann-La Roche Inc., Nutley, NJ, USA
Currently of Merck & Co., Kenilworth, NJ, USA

12.1 INTRODUCTION

The advent of combinatorial chemical library synthesis and, more recently, DNA-encoded chemical library synthesis offers drug discovery researchers access to billions of unique chemical entities that can be useful candidates for screening against therapeutic drug targets [1]. The ability to rapidly access new chemical space has stimulated considerable interest in the area of chemical diversity analysis [2–6] and also in the development of efficient methods to sample and compare large chemical libraries [7–11]. From the early days of combinatorial chemistry, chemists have long recognized that an astonishingly large number of library products can be synthesized from just several thousands of common blocking blocks. Yet, typical combinatorial libraries that are actually produced and screened for drug discovery programs nowadays seldom exceed 10^3 molecules. Synthesis, purification, analysis, and storage for a huge number of individual compound arrays are not economical or practical, and compound screening can also become very costly. In contrast, the use of mixture-based combinatorial libraries offers substantial saving in logistics, but this benefit is largely offset by the tremendous effort required for iterative deconvolution in hit identification and validation. So, the emphasis of computational library design in the past 20 years has primarily been the selection of optimal building blocks to produce chemical libraries at a modest scale [12–17]. In the case of exploratory or probe libraries that are not target dependent, the goal is often to make a

A Handbook for DNA-Encoded Chemistry: Theory and Applications for Exploring Chemical Space and Drug Discovery, First Edition. Edited by Robert A. Goodnow, Jr.
© 2014 John Wiley & Sons, Inc. Published 2014 by John Wiley & Sons, Inc.

minimal subset of diverse library compounds that best represent the entire library. For focused or biased libraries, one would optimize building block selection to produce compounds that exhibit the necessary pharmacophores to target specific protein families such as kinases, GPCRs, or proteases or to bias the synthesis of products with a desirable physicochemical property profile, such as compliance with the rule of five [18].

The recently developed DNA-encoded library technology is a disruptive technology that also changes the landscape of combinatorial library design. Not only is there now an effective way to produce and handle massive numbers of combinatorial products, but there are means to detect hits from such libraries with unparalleled sensitivity. A unique DNA tag attached to each library member serves as its identification bar code, from which a signal is subsequently amplified by PCR and analyzed by next-generation high-throughput sequencing technology. Computational library design of billions or even trillions of molecules is now of practical use and also a reality.

The purpose of this chapter is to provide a brief overview of methods and tools that are currently available to enumerate and analyze very large combinatorial chemical libraries of more than one billion members. Here, a combinatorial library produced from a four-component Ugi reaction serves to illustrate some practical considerations for enumeration and analysis [19]. In this reaction, isocyanides, aldehydes, amines, and carboxylic acids are mixed in one pot, and a very large array of combinatorial products is formed. The application of cheminformatics tools to select building blocks, enumerate library products, and also calculate several key molecular properties to profile the library is discussed. At the end of the chapter, several methods to compare chemical space for four compound libraries, either qualitatively through visualization or quantitatively, with molecular similarity calculations have been suggested.

12.2 ENUMERATION

In this section, the enumeration of a combinatorial chemical library for the four-component Ugi library is used as an illustrative example. A schematic representation of the reaction is shown in Figure 12.1.

12.2.1 Reagent Identification

The initial step in library enumeration is the selection of appropriate sets of reagents for the library. In this case, four classes of reagents—isocyanides, aldehydes, amines, and

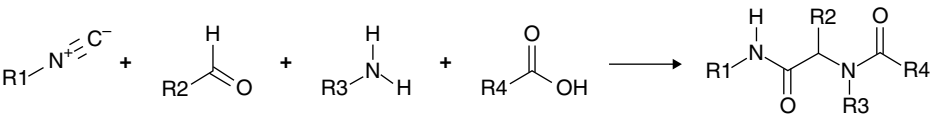

Figure 12.1. Four-component Ugi reaction. The four diversity sites (R1–R4) of the final products are introduced by isocyanides, aldehydes, amines, and carboxylic acids building blocks. The reaction introduces an uncontrolled stereocenter (the tetrahedral carbon attached to R2), and an epimeric mixture is produced.

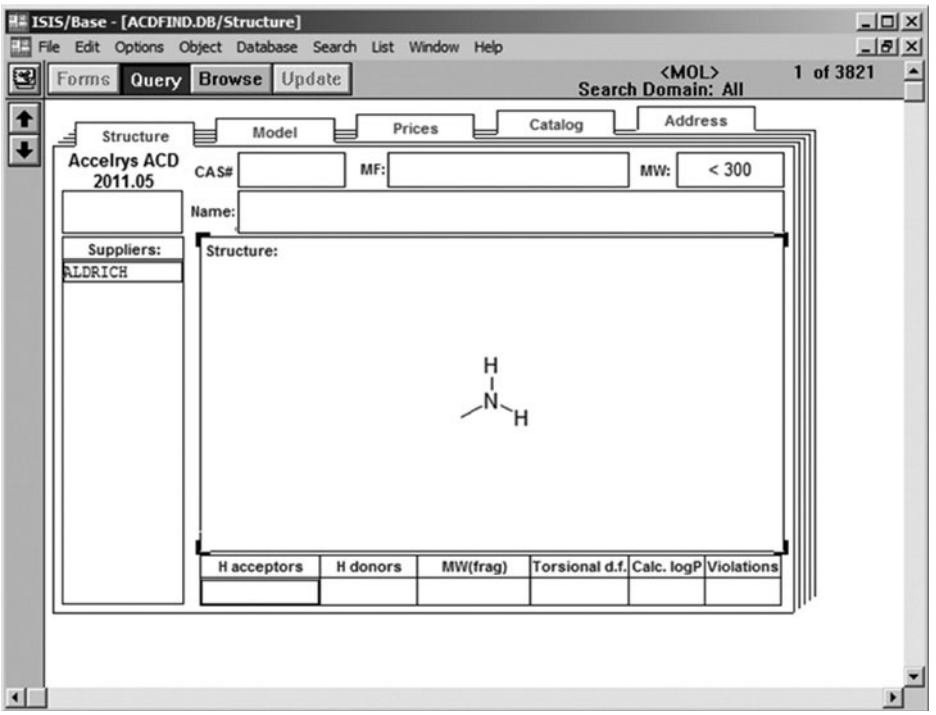

Figure 12.2. Example of a substructure query to search for low MW and commercially available primary amines from the ACD database through an ISIS database user interface.

carboxylic acids—are needed to build this library. One can start to assemble a short list of reagents from either an in-house repository or commercial suppliers. As shown in Figure 12.2, a substructure query was made in an ISIS/Base Available Chemicals Directory (ACD) [20] application to search for low molecular weight (MW < 300) primary amines that are commercially available from Aldrich. A major limitation of using a substructure-based query such as the one shown here is that the structure representation often lacks sufficient specificity. For example, primary carboxamides, which contain an amino group but cannot undergo the Ugi reaction, are also retrieved from this substructure query and become part of an initial selection of 3821 building blocks. It is therefore important that the list of building blocks is further refined so that only those that are chemically compatible with the reaction are selected prior to enumeration.

12.2.2 Reagent Filtering

The refining of the initial list of building blocks can be easily accomplished using tools from a cheminformatics platform such as Pipeline Pilot [21] or through scripting using Daylight [22] or OEChem toolkits [23]. In the current example, Pipeline Pilot was used for postprocessing the hits obtained from this primary amine ISIS query on the ACD [20].

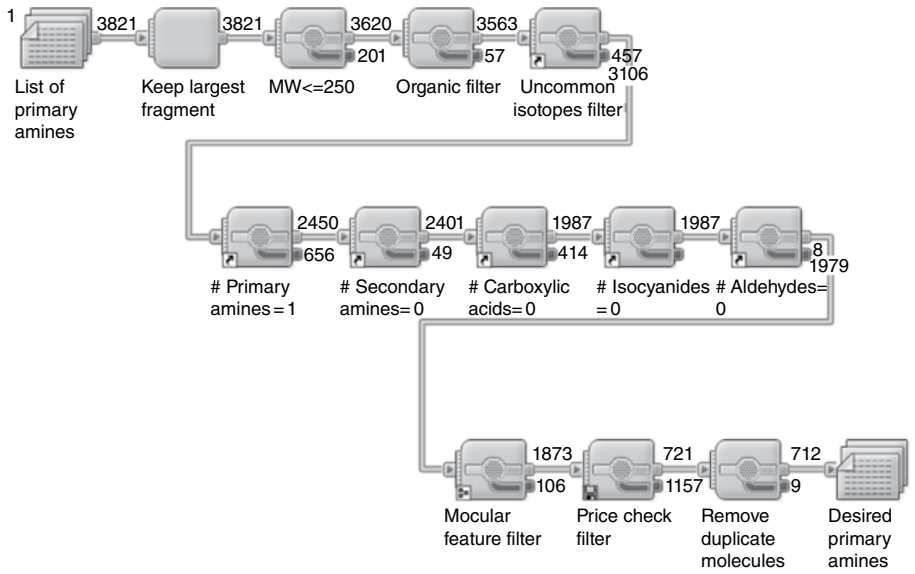

Figure 12.3. Pipeline Pilot workflow used to filter a list of primary amines obtained from the ACD.

In the workflow shown in Figure 12.3, the initial list of building blocks was processed through a chain of molecular filter and manipulator components to remove those that were incompatible or seemed less desirable. Here, the largest molecular fragment in each building block was kept (e.g., removal of any counterion). It was decided to apply an upper MW threshold, and only the building blocks with MW ≤ 250 were retained. An "organic" filter and an "uncommon isotope" filter eliminate organic molecules with unusual elements (i.e., molecules with H, C, N, O, P, S, F, Cl, Br, or I atoms are acceptable, but not those with B or Si), and also those without unusual isotopes were kept. During this initial stage of filtering, more than 700 building blocks were removed. The goal for the next stage of filtering is to remove reagents that contain incompatible or potentially problematic functional groups. This can be achieved through the use of a SMARTS expression, which is a powerful and versatile text-based cheminformatics language. The significant advantage of SMARTS over an ISIS-based query (such as the one shown in Figure 12.2) is that the substructure specification can be very precise. For example, a recursive SMARTS pattern such as

$$[NX3;H2;+0;!\$(NC=[O,S,N]);!\$(NS=O);!\$(N-N=C)]$$

identifies only primary amines, but does not identify secondary amines, primary carboxamides, or sulfonamides. This type of SMARTS-based molecular filter can be readily incorporated as a customized Pipeline Pilot component to detect the presence of specific chemical substructures. In this example, there are 2450 building blocks in the reagent list that contain exactly one primary amine in the molecule. The other 656 building

blocks, which either did not match the SMARTS pattern or matched the query more than once, were removed. Likewise, one can further remove other building blocks with chemical features that are considered problematic using a chain of SMARTS-based filter components. For example, one might prefer to exclude amino acids from this list of primary amine building blocks. The SMARTS patterns to identify aldehyde, carboxylic acid, primary amine, and isocyanide substructures are listed below:

Aldehyde	[CX3H1](=O)[#6]
Carboxylic acid	[CX3](=[OX1])[OX2H1,O-]
Primary amine	[NX3;H2;+0;!$(NC=[O,S,N]);!$(NS=O);!$(N-N=C)]
Isocyanide	[NX2;+1]#[CX1;-1]

It is beyond the scope of this chapter to discuss the SMARTS syntax in detail. Interested readers are referred to a manual and examples from Daylight for additional information [24].

At this point, 1979 building blocks have been identified from the original list of 3821 containing exactly one primary amine (but no functional group from other deselected reagent classes) after the second stage of filtering. In the final stage of filtering, additional molecular feature filters were applied to remove other undesirable or reactive chemical groups (e.g., alkyl halides, sulfonyl halides), again using SMARTS patterns to flag such chemical substructures. So far, all filters being deployed were related to either chemical structures or molecular properties. It should be noted that other types of filters relating to logistical consideration (e.g., price of reagent) could also be used. In the current example, a customized "price check" filter was incorporated to positively select building blocks that satisfied a set of predetermined price limits (e.g., maximum of \$50/g or \$250/5 g). The more expensive reagents or those without pricing information were removed from further consideration. Finally, any duplicate building blocks from the list were dropped, resulting in a filtered list of 712 primary amines that are suitable for library enumeration. Shown in Figure 12.4 are some 40 examples from the list that have survived the three-stage filtering process. This procedure was repeated for the other three reagent classes by running substructure queries against the ACD and then filtering the substructure hits accordingly. In addition to the 712 primary amines, 687 carboxylic acids, 320 aldehydes, and 10 isocyanides were identified as the building blocks for this Ugi combinatorial chemistry library. In the subsequent sections of this chapter, the enumeration and property calculation of a 1.6 billion compound library derived from this reagent set are reported.

12.2.3 Building Block Selection

In some circumstances, there may be logistical considerations or budgetary concerns that prevent the sourcing and use of all available building blocks. From a practical point of view, smaller focused libraries produced from only a subset of building blocks could offer substantial cost saving in term of synthesis, storage, and screening against therapeutic targets. Indeed, depending on the purpose and the format of library synthesis,

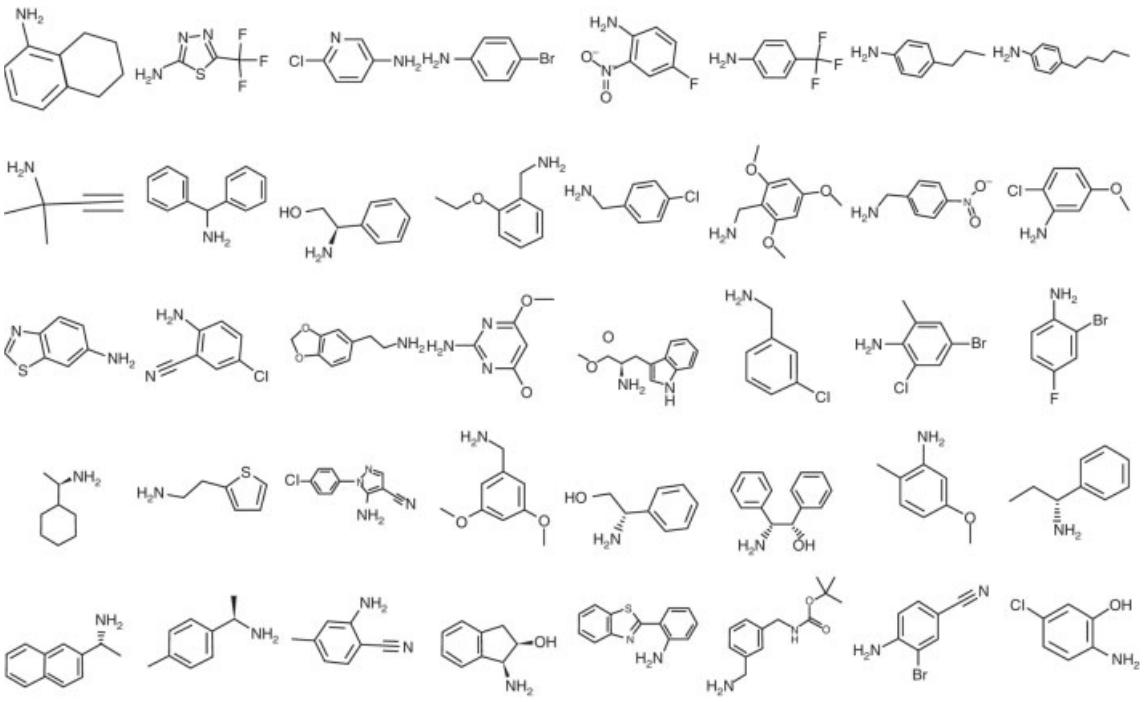

Figure 12.4. Examples of primary amine building blocks.

not all available building blocks should be included. For example, one may wish to eliminate building blocks that do not result in compounds containing a required pharmacophore feature (e.g., hydrogen donor/acceptor motif targeting protein kinases). If the goal is to produce general screening libraries that are rule-of-five compliant, then it would be beneficial to limit the number of highly lipophilic building blocks. Many powerful optimization algorithms originating from the field of computer science have also been adopted and deployed for optimal selection of building blocks. These included simulated annealing, Monte Carlo simulations, genetic algorithms, and Pareto sorting. The objective design functions vary according to the purpose of the focused library, with some examples including predicted activity from a QSAR model, molecular similarity against an active therapeutic ligand, or a docking score against a protein receptor.

One approach to select building blocks is the lockdown method proposed by Spellmeyer and coworkers [25]. The beauty of the lockdown method is that it is conceptually simple, easy to implement, and extremely efficient in execution. In this method, building blocks from different reagent classes are iteratively trimmed by strategically starting from the reagent class with the fewest candidates to ensure more efficient sampling. Through a series of random sampling of enumerated products, the subset of building blocks in the class that are associated with a minimum threshold of desirable compounds (e.g., matching a target molecular property profile or pharmacophore constraints) is retained. This subset of building blocks will be locked down as the preferred list in the class for the subsequent stages of sampling when the same procedure is applied to select another subset from other reagent classes. In the original publication, the authors demonstrated the use of this method to optimize for monomer selection of an Ugi library that maximized pharmacophore matching from a known thrombin inhibitor. Here, the use of this method to select a building block subset to optimally produce library products with no more than one Lipinski violations was demonstrated [18].

The result of the lockdown experiment on this current list of Ugi building blocks is summarized in Table 12.1. As discussed, the first lockdown stage was used to select building blocks from the reagent class with fewest candidates, which were the 10 isocyanides in this case. Random library products, 100,000 in number, were enumerated to sample how the 10 different diversity elements introduced by isocyanides could impact the overall physicochemical parameters of the final products. Only 19% of this initial sample of combinatorial library products passed with no more than one Lipinski violation. Four isocyanide building blocks produced at least 20%—the initial filter threshold—of their enumerated products, and they passed the desirability criteria. The four building blocks were subsequently locked down during the next stages of sampling. The procedure was reiterated to identify a subset of aldehydes (the class with second fewest number of building blocks) now using a reduced reagent set. Higher filter threshold values were set at the later stages so that the proportion of products passing the filtering criteria would steadily increase. After 10 lockdown stages, the number of building blocks for each reagent class was substantially reduced. The final selection of 290 amines, 201 acids, 41 aldehydes, and 4 isocyanides yielded a biased sublibrary containing approximately 10 million products, or about 0.6% of the size of the original library. This sublibrary was enumerated, and indeed, nearly all products (9,481,172/9,559,560 = 99.2%) had no more than one Lipinski violation. The lockdown

TABLE 12.1. Building blocks selection for the Ugi library using the lockdown approach

Stage	Class sampled	Amines	Acids	Aldehydes	Isocyanides	Possible products	Product sampled	Product passed	Filter threshold (%)
1	Isocyanides	712	687	320	10	1,565,260,800	100,000	18,848	20
2	Aldehydes	712	687	320	4	626,104,320	100,000	29,400	30
3	Acids	712	687	140	4	273,920,640	100,000	45,057	40
4	Amines	712	386	140	4	153,905,920	100,000	59,982	50
5	Aldehydes	468	386	140	4	101,162,880	100,000	73,778	60
6	Acids	468	386	127	4	91,769,184	100,000	75,636	70
7	Amines	468	234	127	4	55,632,096	100,000	84,125	80
8	Aldehydes	307	234	127	4	36,493,704	100,000	91,153	95
9	Acids	307	234	41	4	11,781,432	100,000	97,777	95
10	Amines	307	201	41	4	10,119,948	100,000	98,858	95
Final		290	201	41	4	9,559,560	9,559,560	9,481,172	

Successive stages of the lockdown process where the number of building blocks in each reagent class being sampled is shown. Product passed denotes the number of random enumerated products that have no more than one Lipinski violations. Initially, only 19% of the products from the full library would pass. One hundred thousand random samples were used at every stage to obtain statistics determining which building blocks (in the reagent class being sampled) have passed the minimum filter threshold set at that stage. Subsequently, this subset of building blocks in the reagent class would survive to the next lockdown stage. Note that the filter threshold criteria become increasingly more stringent at later stages to enforce more optimal subset selection. After 10 stages, a significantly reduced focused library was obtained, and more than 99% of its products have no more than one Lipinski violations.

procedure was computationally very efficient and the calculation required less than 10 min of CPU time for the current example.

As demonstrated, the lockdown procedure proves very useful for biased library design when an objective function is known. However, such knowledge is often not available, particularly when designing exploratory chemical libraries for hit identification purposes. To reduce the number of building blocks used in such a library, an alternative approach called molecular clustering can be applied to remove some building blocks that are already richly represented. In clustering, building blocks with similar chemical structures are grouped together to form a cluster. To exemplify each compound cluster, a representative structure (sometimes referred to as the cluster center) is picked from the group, while the remaining cluster members are discarded. A number of different clustering methods and fingerprinting types have been developed and reported, and interested readers should refer to several excellent reviews in this area [26–29].

To illustrate how clustering is applied to trim a list of building blocks, a Pharmacophore Graph (PG)-based method developed by DISCNGINE [30] was used to analyze and cluster the initial 712 primary amine data set. First, each amine was converted to an object called a PG using a decomposition algorithm that assigns a specific pharmacophoric feature to each of its constituent multiatom fragment units. Since multiple types of fragment units (which are structurally similar) can map to same pharmacophore feature, it follows that similar molecules can also be reduced to an identical PG. 167 unique PGs were found in this set of 712 amines, with the group of molecules that were converted to the same graph forming a compound cluster. One such cluster containing 11 amines is shown in Figure 12.5. The high pharmacophoric similarity shared by this set of compounds seems obvious here—all molecules have an amine functional group directly attached to a saturated atom of either an indane or tetrahydronaphthalene ring. In particular, the building block MFCD00003799 seems a good choice as a general representative for this cluster. This compound is a racemic mixture of two single enantiomer amines (MFCD00216669 and MFCD00216670) in the set, and it also shares close (sub)structural similarity with other group members. By selecting a single structure to represent each cluster, the size of the amine data set was significantly reduced from 712 to 167, leading to a fourfold reduction of the overall size of the combinatorial

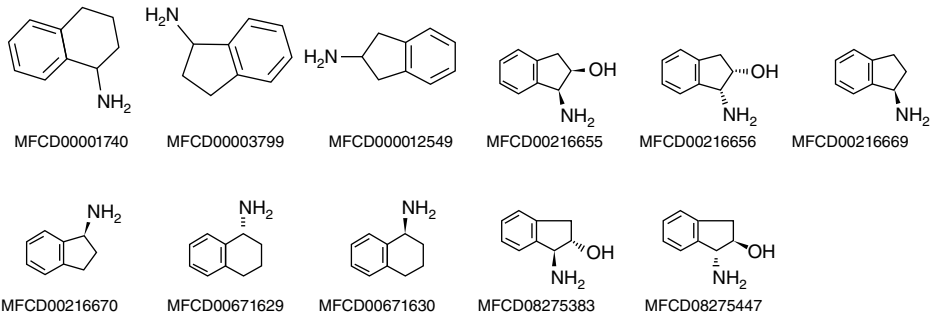

Figure 12.5. Structure of 11 amines grouped into the same cluster using a PG-based clustering method.

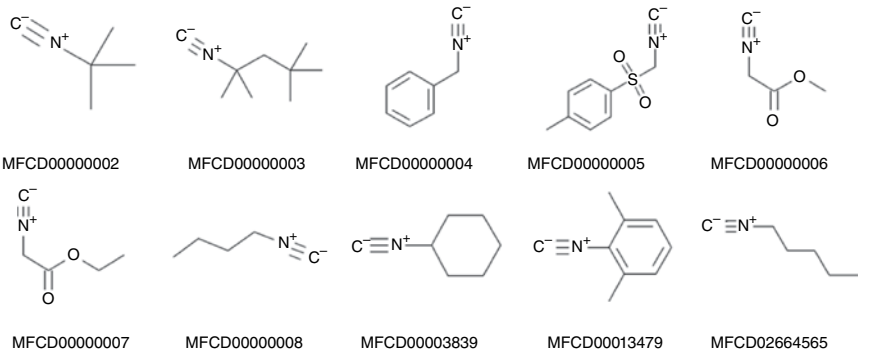

Figure 12.6. Structures of the 10 isocyanide building blocks used for this study. The PG-based algorithm groups MFCD00000008 and MFCD02664565 into the same cluster and perceives other eight building blocks as singletons. On the other hand, a human expert would likely expect an additional group containing the two isocyano-acetic acid esters (MFCD00000006 and MFCD00000007).

library. A similar factor of size reduction was also observed for the carboxylic acids (from 687 to 171) and also for the aldehydes (from 320 to 78) when this clustering method was applied. Interestingly, the majority of isocyanides (8 out of 10) in the set exist as singletons (i.e., they have a unique PG). Only two building blocks—1-isocyano-butane (MFCD00000008) and 1-isocyano-pentane (MFCD02664565)—were grouped into the same cluster (Fig. 12.6). This is an indication of either a greater level of structural diversity or paucity within this class of reagents. The isocyanide example also served as a useful demonstration that a fully automated clustering method does not always yield the type of result that is intuitive to a human expert. In addition to the 1-isocyano-butane/pentane pair, most medicinal chemists would also group isocyano-acetic acid methyl ester (MFCD00000006) and the corresponding ethyl ester (MFCD00000007) to the same cluster. However, the perception of the two molecules by computer is, in fact, different due to the algorithmic implementation for how PG are generated. The methyl ester is by itself a distinct pharmacophore unit, and any additional aliphatic group (i.e., the extra carbon in ethyl ester) attached is assigned to an additional aliphatic pharmacophore feature. This example is a reminder of the technical challenge in developing a tool that can produce a fully satisfactory clustering scheme.

In a similar manner to the use of the lockdown procedure, a molecular clustering method was also used for significantly trimming of the number of building blocks in a combinatorial library. The PG-based method used here led to a diversity-based selection of 167 amines, 171 acids, 78 aldehydes, and 9 isocyanides. An enumeration of this subset of building blocks yielded approximately 20 million products, which is approximately 1% of the original library. An interesting question arises: would this subset of compounds, which is derived from reduction of chemical space at reagent level, be able to approximate the great diversity of the overall product space? For example, can this sublibrary accurately reproduce the molecular property profile of the original library? In the next section, the enumeration of both the full 1.6 billion member library and this

sublibrary will be described to determine if there is a significant difference in the distributions of some key molecular properties.

12.2.4 Enumeration and Property Profiling

Enumeration and analysis of massive combinatorial libraries has been an active area of research in combinatorial library design. In spite of the rapid advancement in computational technologies, a library size of about one trillion (10^{12}) molecules remains a practical upper limit for explicit structure enumeration and property calculation today. In addition to the need for computing processing resources, storage requirements can quickly become an issue. A molecular data set containing 1 trillion molecules requires more than 20 terabytes of storage space even at a modest cost of 25 bytes per structure. To circumvent explicit enumeration, alternative approaches to evaluate the property profile of large combinatorial libraries have been developed by leveraging statistical sampling methods or even machine learning algorithms. Spellmeyer and coworkers published a binomial formula derived from statistical theory to estimate errors from quantities obtained from random sampling of a small number of library products [25]. Agrafiotis and Lobanov developed a novel approach called combinatorial neural networks, which is based upon the finding that most molecular descriptors that are commonly used in library design can be accurately predicted from the properties of their respective building blocks [31]. They were able to train a computational neural network to predict the coordinates of the enumerated products on a chemical space directly from properties of building blocks as inputs, circumventing expensive explicit structure enumeration steps.

In this section, the Ugi reaction has been used as an example for library enumeration and property calculation. A very simple method—random sampling on combinatorial products—was applied to estimate the statistical mean of several molecular properties. These properties include MW, calculated partition coefficient between octanol and water (*A*log*P*), and Polar Surface Area (PSA). It should be noted that the average value of a "decomposable" property such as MW is readily calculable without enumeration—it is simply the average MW of reactants subtracted by the MW of the elements lost in the reaction [25]. However, properties such as *A*log*P* or PSA are not easily derived from simple addition of fragment contributions from reagents alone. The value of the property often depends on molecular connectivity, which may be modified by formation of new bonds during a reaction.

With currently available computing power, it is feasible to enumerate and to compute molecular properties of a virtual combinatorial library with the size of 10^9. Several standard cheminformatics components available in the Pipeline Pilot program were used to enumerate the Ugi library products and calculate the properties. The structure enumeration was performed using an "Enumerate Combinatorial Reaction" component, and a combinatorial reaction scheme was encoded using Daylight SMIRKS reaction transform language [32]. For each molecule, a canonical SMILES string was obtained and then used to calculate four molecular properties: MW, *A*log*P*, PSA, and the number of Lipinski violations (NumLipinski). The data generated were stored as a text file in a compressed data format. In total, 1,565,260,800 library products were enumerated using

the full set of filtered reagents: 712 amines, 10 isocyanides, 320 aldehydes, and 687 acids. The calculation, including both the enumeration and property calculation, required approximately 8 days on a single core of an Intel Xeon CPU (E7340 @ 2.4 GHz); approximately 30 gigabytes of data were generated. This translated to a resource requirement of 7 min of CPU time and 19 megabytes of disk storage for every 10^6 molecules. It should also be pointed out that library enumeration is a class of computational problem that can be handled with an "embarrassingly parallel" workload, where very little effort is required to distribute the calculation to multiple computer processors to gain substantial speedup. In the present case, the calculation could be readily split to 10 parallel threads, with each thread handling the enumeration of 1 of the 10 isocyanides, and the calculation could be done in less than a day in real time.

From the enumeration and property calculation of the full 1.6 billion compound library, an exact solution was obtained for the mean value of each molecular property. For the entire library, the means are as follows: MW = 612.72, *A*log*P* = 6.119, and PSA = 118.24. With this information, it is possible to determine how the accuracy of mean estimation from random sampling is affected by parameters such as sampling size. This experiment was begun with a small sample size of 16 random enumerated products, which corresponds to 10^{-6}% of the library. The mean of MW, *A*log*P*, and PSA were estimated from these samples; this procedure was repeated 10 times to understand the variance within individual estimations. Sampling sizes belonging to 7 categories of sampling size, ranging from 10^{-6}% to 1% (i.e., 16–16 million samples), were investigated. The result of this sampling calculation is summarized in Table 12.2 and Figure 12.7a–c. The results from different percentage categories are grouped along the horizontal axis. The property mean computed from individual runs is plotted as black circles within each percentage group. The horizontal line in the middle of the "mean diamond" depicts the group mean, and the vertical apex provides the 95% confidence interval for the 10 runs in each group. The statistical mean for each molecular property, which was computed from the full 1.6 billion compounds library, is also shown on the plot as a dotted line. As expected, sampling accuracy steadily increases with higher sampling sizes along with lower variance among individual estimates. Consistent with the earlier reports [25, 33], this simulation suggests that statistical parameters such as mean value of a molecular property can be accurately determined using a small fraction of random enumeration products. For example, the result of sampling the 0.001% group (i.e., only 15,600 random samples) yielded mean estimates of MW = 612.86 ± 0.71, *A*log*P* = 6.119 ± 0.015, and PSA = 118.14 ± 0.30. These estimates are already accurate to three significant figures when compared to the exact solutions. Clearly, enumeration of all library products for this purpose would be unnecessary and wasteful in this case.

In addition to estimation of a single measure such as statistical means, it was also interesting to see whether the sampling of a small portion of the library would accurately reproduce property distributions such as a histogram profile. In the lower graph of Figure 12.8a, a histogram shows the percentage of compounds with various numbers of Lipinski violations in the full library (100% group, shown with black vertical bars) and compared with the corresponding percentages that were derived from just 15,600 random samples (0.001% group, shown with light grey bars). The two histogram profiles are virtually indistinguishable from each other. In the upper graph, the

histogram profile derived from a partial library enumeration (subset group, shown with dark grey bars) based on a subset of building blocks previously selected by PG-based clustering method is shown. Interestingly, the profile of this library subset is noticeably different from the full library, despite the fact that this sublibrary contains nearly 20

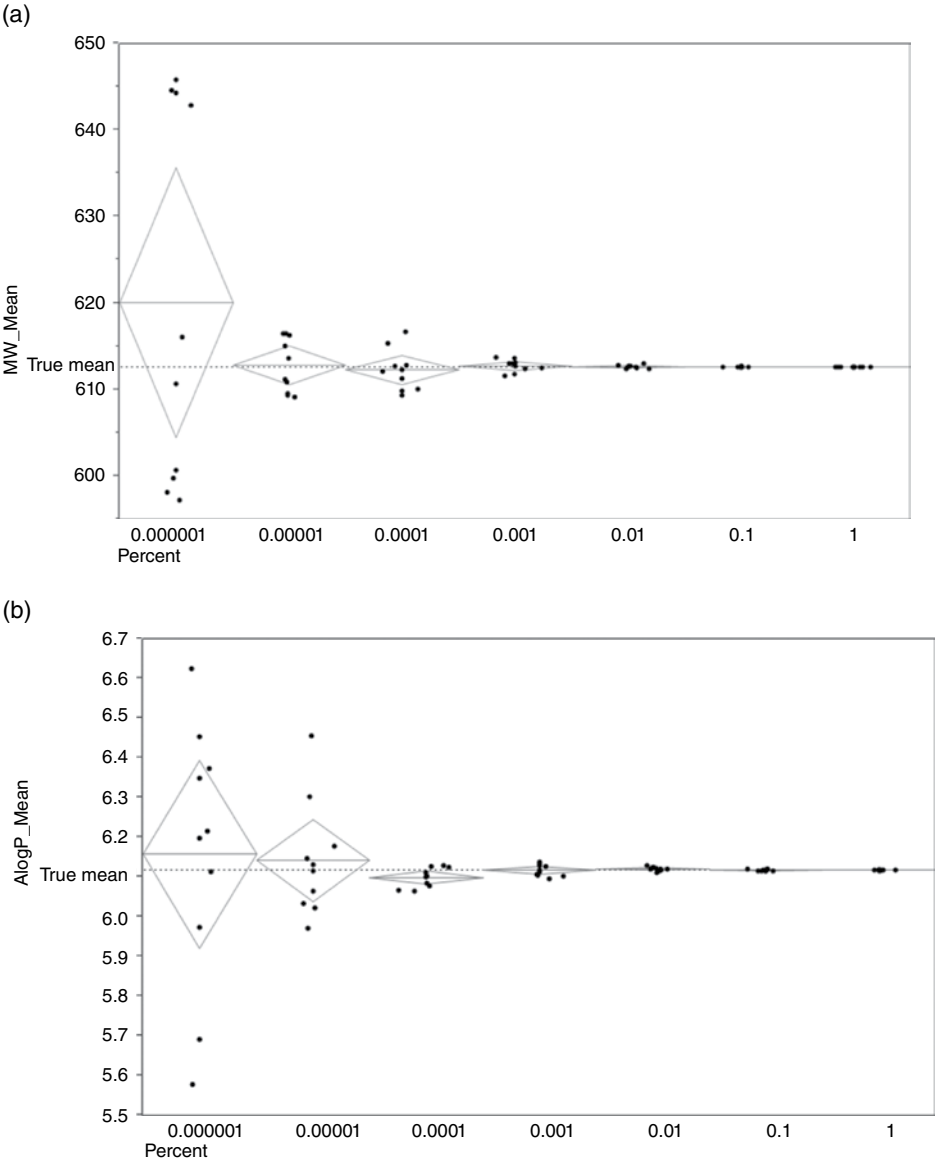

Figure 12.7. (a) Estimation of MW mean from random sampling of Ugi library products. (b) Estimation of *A*log*P* mean from random sampling of Ugi library products.

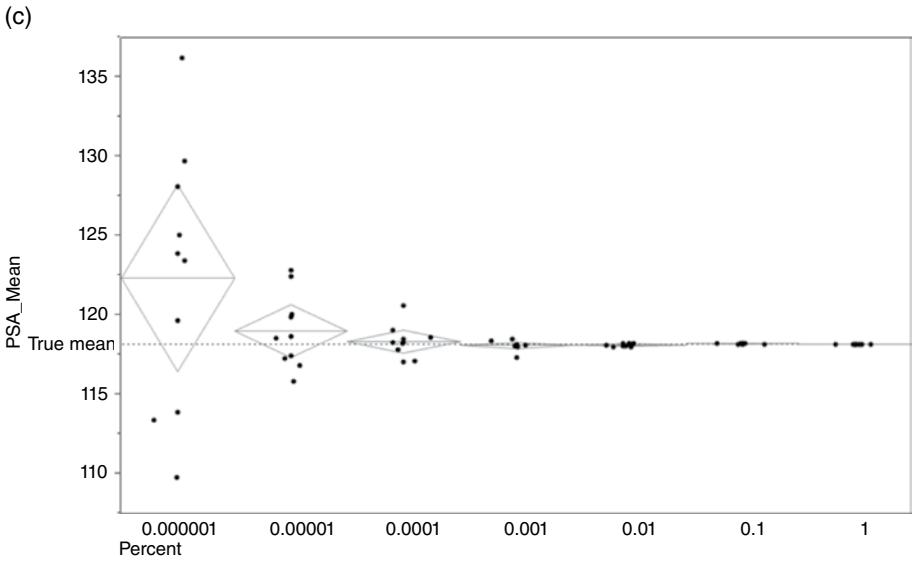

Figure 12.7. (*Continued*) (c) Estimation of PSA mean from random sampling of Ugi library products.

TABLE 12.2. Summary of the sampling runs using a small percentage (0.000001% to 1%) of enumerated library products to estimate various molecular properties

Run	MW	*A*log*P*	PSA
0.000001	620.14 ± 21.77	6.159 ± 0.330	122.38 ± 0.49
0.00001	612.91 ± 3.11	6.144 ± 0.144	119.04 ± 0.17
0.0001	612.37 ± 2.35	6.101 ± 0.024	118.41 ± 0.03
0.001	612.86 ± 0.71	6.119 ± 0.015	118.14 ± 0.02
0.01	612.75 ± 0.19	6.122 ± 0.005	118.19 ± 0.01
0.1	612.74 ± 0.05	6.119 ± 0.002	118.26 ± 0.00
1	612.72 ± 0.01	6.119 ± 0.001	118.24 ± 0.00
Full library	612.72	6.119	118.24
PG subset	639.01	5.927	132.89

The mean and standard deviation (from 10 sampling calculations) of MW, *A*log*P*, and PSA is reported. Full library indicates the exact solution obtained from full enumeration of the 1.6 billion compound library. The PG subset indicates the property average derived from a 20 million compound library subset enumerated using a diverse set of reagents obtained from the PG-based clustering method.

million samples, which are already more than 1% of the full library. A similar result is also found for the histograms for MW (Fig. 12.8b). The MW profile from 0.001% random sampling in product space (light grey bars) is very similar to the full library (black bars), whereas the library subset derived from reagent clustering (dark grey bars) has a distribution that is right shifted relative to the full library. This is not a surprising result. Reagent-based diversity library design is a useful strategy to simplify the logistics of library production and to minimize cost. However, there is no reason to

expect that a diverse set of reagents would necessarily lead to an optimally diverse set of products. There is already overwhelming evidence suggesting that a random selection based on products themselves can be substantially more diverse, perhaps by as much as 35–50% comparing with reagent-based selection [33–36]. In the current case, the estimated property means for MW, *A*log*P*, and PSA are 639.01, 5.927, and 132.89,

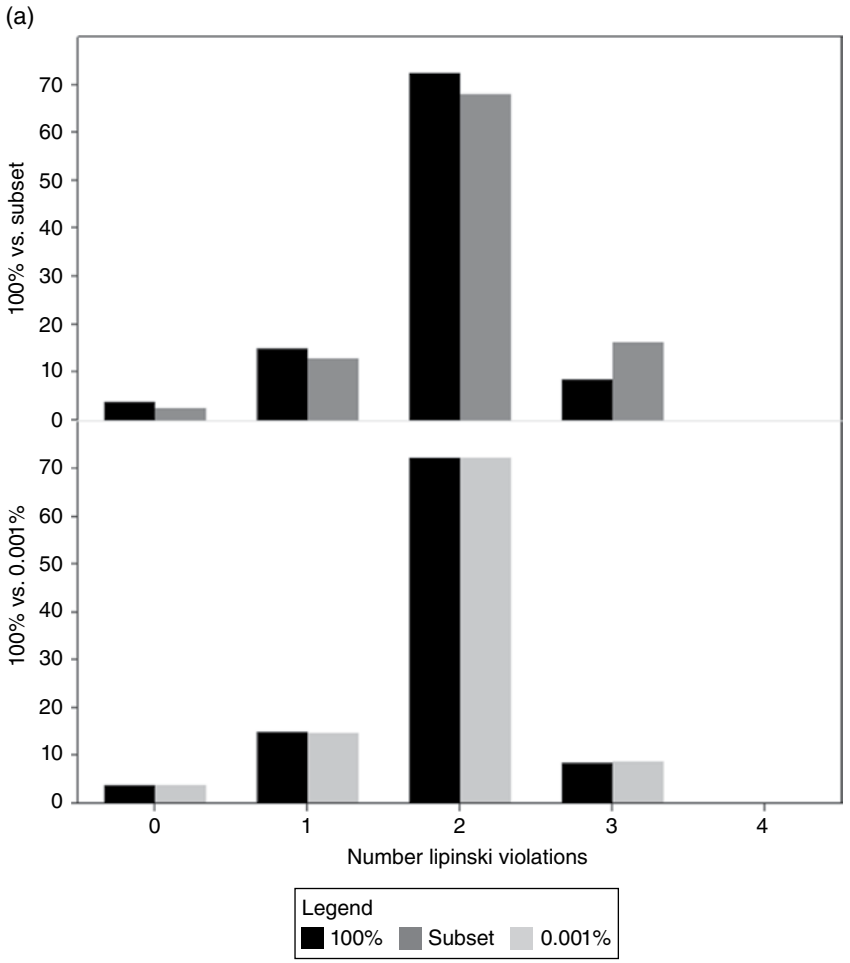

Figure 12.8. (a) Histogram showing distribution of Ugi library products according to the number of Lipinski violations. The lower plot shows a comparison of histograms derived from the full 1.6 billion compound data set (100%, black) versus 15,600 random samples (0.001%, light grey). The upper plot shows a comparison of histograms derived from the full data set (100%, black) versus a partial enumeration of the library (subset, dark grey) of 20 million compounds based on a subset of building blocks selected from a PG-based clustering method.

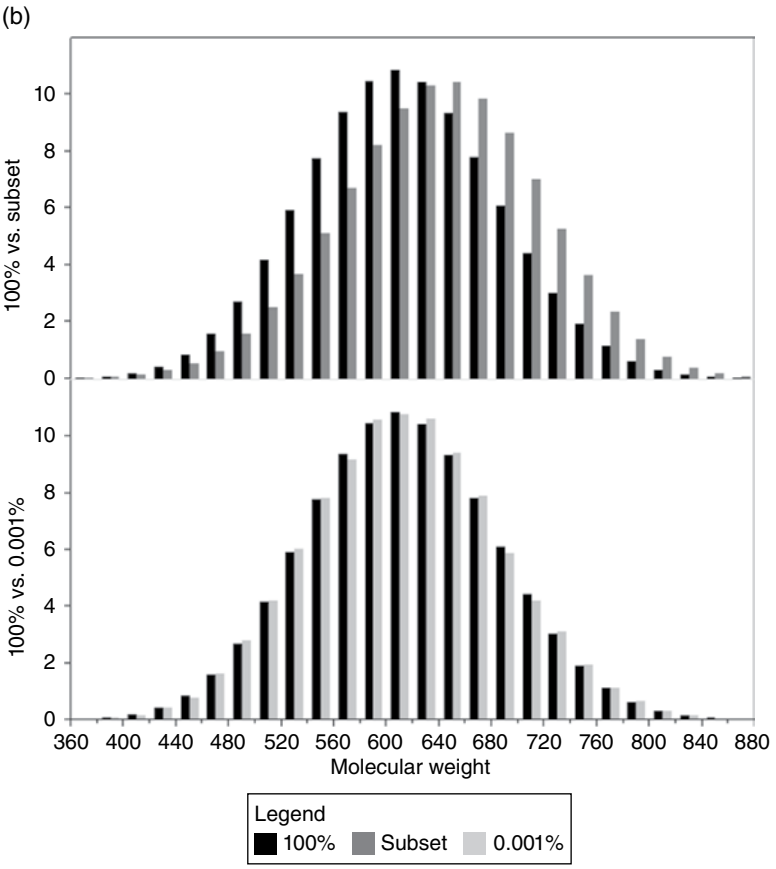

Figure 12.8. (*Continued*) (b) Histogram showing distribution of Ugi library products according to the MW. The lower plot shows a comparison of histograms derived from the full 1.6 billion compound data set (100%, black) versus 15,600 random samples (0.001%, light grey). The upper plot shows a comparison of histograms derived from the full data set (100%, black) versus a partial enumeration of the library (subset, dark grey) of 20 million compounds based on a subset of building blocks selected from a PG-based clustering method.

which deviate significantly from the grand means (612.72, 6.119, and 118.24) of the full library (Table 12.2).

12.3 CHEMICAL SPACE COMPARISON

The notion of chemical space is a concept in which chemists have a long-standing interest [37–42]. Different chemistry-space metrics have been developed to better characterize the broad range of chemical diversity in this extremely vast space [37, 40]. One central question is whether it is possible to compare the chemical space occupied by

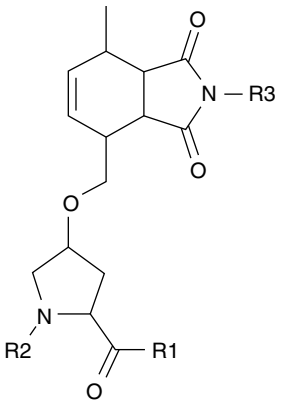

Figure 12.9. General scaffold for the DNA-encoded chemical library from Neri and coworkers. R1, R2, and R3 are the three diversity elements. A DNA tag is attached as part of R1 group.

different chemical libraries. To investigate this, three large compound libraries were used here and compared with the Ugi library enumerated in the previous section. The first library is the PubChem database [43], which is maintained by the National Center for Biotechnology Information (NCBI) and is a comprehensive source of chemical structures from the scientific literature. Currently, the database contains more than 30 million unique and chemically diverse compounds. The second library is the GDB-11 database, a (mostly) virtual library of small fragment structures generated by Reymond and coworkers [44]. This database contains about 26 million compounds that represent an exhaustive enumeration of small organic molecules containing five element types: C, N, O, F, and H, up to a total of 11 nonhydrogen atoms. A few simple chemical stability and synthetic feasibility rules were enforced during enumeration so that legitimate and stable organic molecules were produced. In essence, this data set represents the chemical space of small fragments, and similar to the PubChem database, it is very chemically diverse. The third library is a one million compound DNA-encoded chemical library published by Neri and coworkers [45]. In this third library, the key reaction step is a Diels–Alder cycloaddition between building blocks containing a hexadiene group and maleimides as dienophiles, leading to a 5,6-*cis*-fused bicyclic general scaffold with diversity elements shown in Figure 12.9. For the purpose of structure enumeration, the DNA tag that is attached to the R1 diversity element was replaced by a methyl group (i.e., *N*-methyl amide is the terminating group).

12.3.1 Comparison of Exact Structures

One way to compare chemical space is to identify the overlap of exact chemical entities between two libraries. Interestingly, comparison of exact structures is a relatively easy and computationally efficient process even between two extremely large libraries. For example, one can simply create a relational database with two data tables, each containing a text field of canonical SMILES strings generated from compounds in standardized

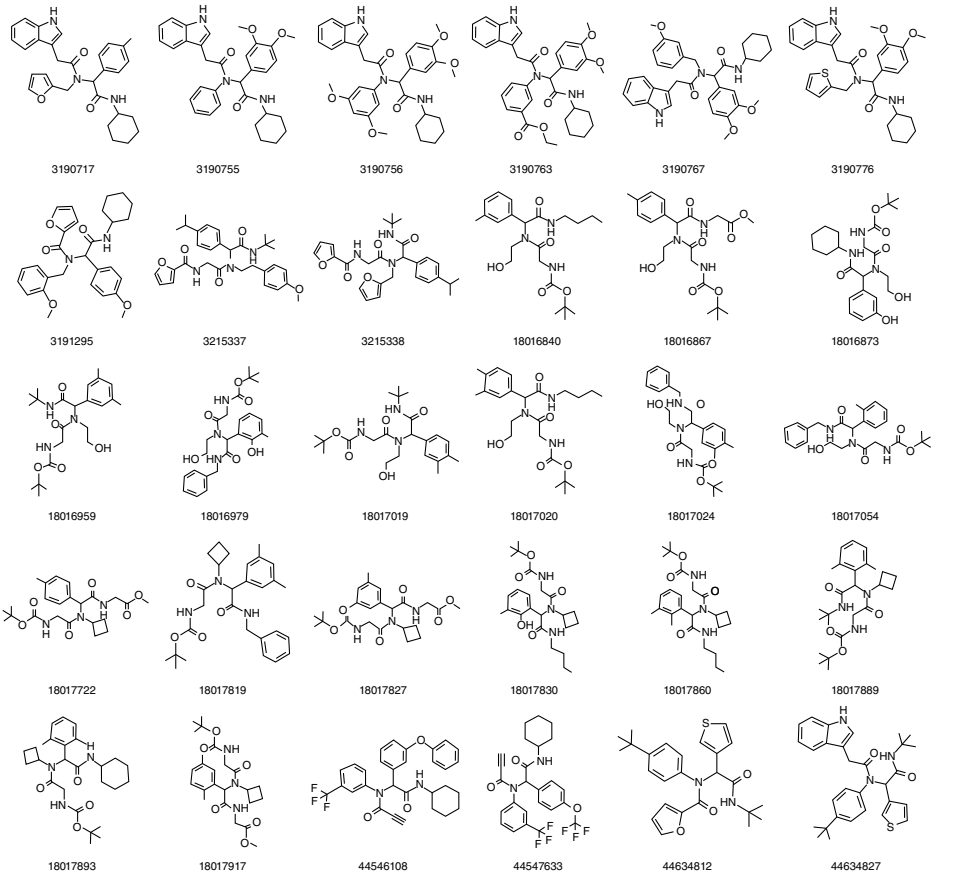

Figure 12.10. Examples of Ugi products from the current library that are found in the PubChem database. The label below each structure is the PubChem identifier.

tautomeric form from the library. Molecules that are common to both libraries can be quickly retrieved with an SQL join operation using the SMILES string as a key. Even for large data tables, this type of operation is very fast because the search can be indexed. The comparison of exact structures was done between the 1.6 billion compound Ugi library and the 30 million compound PubChem database. It is surprising that, given that the one-pot Ugi chemistry was published decades ago and the reaction has also been a benchmark for computational combinatorial library analysis, only 323 compounds from the current Ugi enumeration are known in PubChem. Selected examples of 30 such compounds are shown in Figure 12.10. It is intriguing to note, at least from this set of examples, the presence of some functional groups that are considered less common. For example, about one-quarter of compounds in this set contain the cyclobutane functionality (compared to the presence in less than 0.3% of compounds in PubChem overall). The prevalence of the *tert*-butyl functional group in this set of compounds is easy to comprehend. Many of them are synthetic intermediates derived from commercial Boc-protected amino acids.

The overlap of content between the GDB-11 and the PubChem databases was also analyzed. Compared to the Ugi library, the extent of exact structure matches for GDB-11 with PubChem is significantly higher. Approximately 100,000 GDB-11 molecules—but still only 0.4% of the library—have been synthesized and are known in the PubChem database. This fact gives evidence to the vastness and diversity of chemistry space. Even after more than a century of organic synthesis, only a small fraction of compounds belonging to this seemingly restricted chemical space of fragment-size molecules has been created and characterized. Finally, when comparing the structures from the Neri library and the PubChem database, it is not surprising to see that there was no structure match between the two as this library contains an uncommon general scaffold that is unlikely to be found in PubChem.

12.3.2 Chemical Space Heat Map

Clearly, chemical space is so huge that exact matching of chemical structures between even two very large libraries is rare. Different strategies to reduce molecular representation have been proposed, an interesting example of which is the 42-dimensional chemical space defined by Molecular Quantum Number (MQN) descriptors by Reymond and coworkers [40]. These are a set of integer-value descriptors encoding the counts of specific atomic, bond-type, or topological features (e.g., presence of three-membered rings) of a molecule. Molecules with identical MQN values would be regarded as "molecular isomers" in analogy to atoms with the same atomic number (isotopes). However, because of its high dimensionality, the MQN space itself is likely to be sparsely occupied, which makes comparison of chemical libraries challenging. In contrast to a chemical space encoded by intrinsic structural features, a small set of (ideally orthogonal) physicochemical properties (e.g., PSA) or pharmacophoric features (e.g., the number of aromatic hydrophobes) can be computed and used as coordinates of a low-dimensional property-based chemical space. Obviously, this type of chemical space is quite coarse and lacks any detailed description of the molecules. In addition, the space is likely to be highly degenerate as many different chemotypes can map to the same space by virtue of the similarities in molecular properties or pharmacophoric features that they share. To address such shortcomings, the BCUT metric, a set of descriptors that have been widely used to generate low-dimensional chemical space, was developed. The BCUT descriptors, which are related to the Burden index [46], were proposed and put into practical application by Pearlman and coworkers [37]. In essence, BCUT descriptors are eigenvalues of a modified adjacency matrix primarily encoding molecular connectivity information. Different types of BCUT descriptors result when atomic properties such as atomic charge, hydrogen bonding propensity, or polarizability are incorporated as the diagonal matrix elements. As a result, both chemical structural features and molecular property attributes of a molecule can be captured by just a few BCUT descriptors. One can then create an approximately orthogonal and low-dimensional chemical space by optimally choosing two or three descriptors that best distinguish the structural (or activity) differences between the compounds for the purpose of visualization and analysis. Alternatively, one can deploy dimensionality reduction methods to transform the full descriptor space to a lower dimension. Most

commonly, this is done using Principal Component Analysis (PCA), multidimensional scaling, nonlinear mapping, or Kohonen self-organizing maps [47]. This way, the areas of space that are occupied by compound sets with different characteristics (e.g., drug vs. nondrug, fragments vs. peptides) can be visualized and examined on a two-dimensional plot. An example of this type of approach to visualize and navigate chemical space is the Chemical Global Positioning System (ChemGPS) method proposed by Oprea and Gottfries [48, 49]. In their method, 423 small molecules are selected as spatial reference of a drug-like universe, and "chemographic" map coordinates are extracted from PCA of a standard fixed set of molecular descriptors. Because of the invariant nature of this reference system, this method is well suited for comparing multiple chemical libraries.

As an illustration, the BCUT metric was applied and followed by PCA to compare the chemical space occupied by the Ugi, Neri, PubChem, and GDB-11 data sets. For this analysis, the 30 million PubChem compounds were chosen as a reference chemical space due to their great chemical diversity. For each molecule, six BCUT descriptors, which were the highest and the lowest eigenvalues of the three adjacency matrices modified by atomic charge, hydrogen bond, and polarizability, were computed. A PCA was performed on this large data matrix (30 million by 6). The first (PC1) and second principal component (PC2) explained 27% and 20% of the variance of the BCUT metrics. The coordinates projected from the two principal components were used as a new low-dimensional chemical space for all data sets. For this calculation, the BCUT descriptors were obtained using a ChemAxon Calculator component in Pipeline Pilot. The PCA and the generation of heat maps were performed using standard components.

To further simplify visualization, a cell-based data aggregation approach was adopted. First, a two-dimensional grid was created with 500 equally divided bins along each principal component axis. For each principal component, a range of ±5 standard deviations from the mean was defined as relevant chemical space since nearly all compounds have descriptor values within this range. Each compound in the PubChem database was assigned to one of $500 \times 500 = 250{,}000$ individual cells in the grid based on the binning of PC1 and PC2 values. Finally, the total number of compounds that were mapped to each cell on the grid was counted. A key benefit of such a data aggregation procedure is the transformation of a large array (30 million by 6) of floating-point numbers to a more compact matrix containing only integer elements. More importantly, the size of the matrix depends only on a user-defined grid solution and is entirely independent of library size. Because of this attribute, the chemical space of any library can be summarized as a heat map in a consistent format and visually compared. In the heat map, the two principal components lie along the horizontal and vertical axes, and the population of each individual cell is depicted using color shading, where a darker color indicates a higher compound population in the cell. Figure 12.11a shows a heat map of the 30 million PubChem compounds in the BCUT-derived PC1/PC2 space. The highest compound density is found in the cells close to the center (i.e., within ±1 standard deviation from the mean) as evident by dark color shading in the middle of the plot. Figure 12.11b is a histogram showing the distribution of cells according to different population ranges (in log order of magnitude). Despite a fine grid resolution, fewer than one-third of the cells are empty, and nearly half of the cells contain at least 10 compounds. Only a small percentage (0.7%) of the cells have very high compound density with greater than 10^4 compounds.

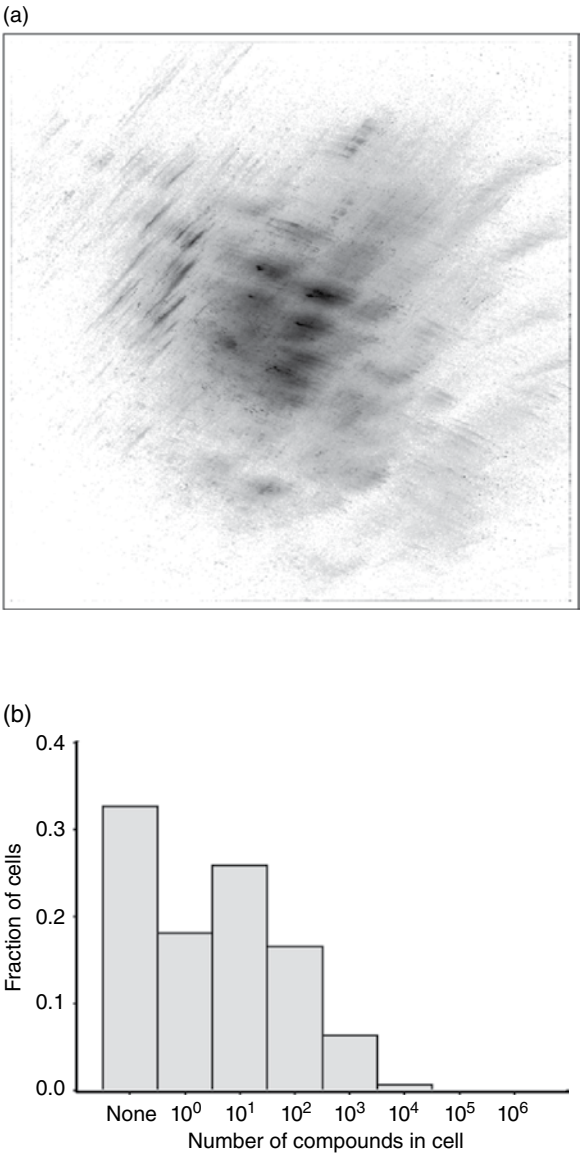

Figure 12.11. (a) Heat map showing a representation of the chemical space of the PubChem library. (b) Distribution of cells according to different population ranges for the PubChem database.

Overall, the excellent coverage of compounds across the cells along the two principal axes makes this a good reference chemical space for the purpose of library comparison.

Based on the principal component parameters obtained from the PubChem data set, another heat map (that has the same horizontal and vertical scales as Fig. 12.11a) was

generated using a random selection of 30 million Ugi products (Fig. 12.12a). Not surprisingly, the chemical space of the Ugi library compounds is considerably more well defined when compared to the structurally diverse PubChem database. This also reflects a markedly different population distribution (shown in Fig. 12.12b). More than 80% of

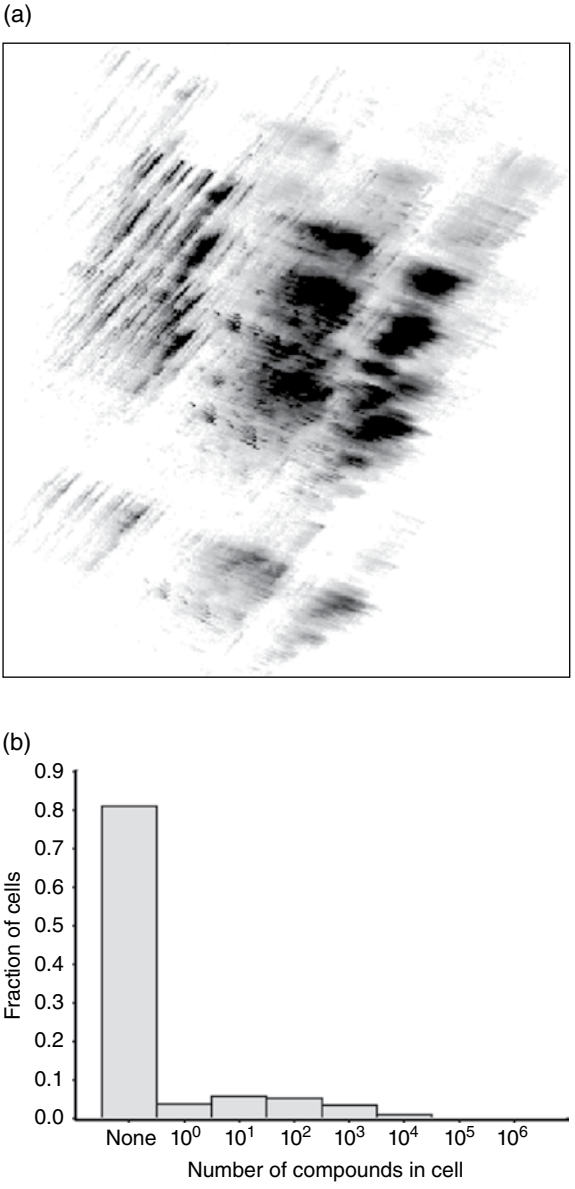

Figure 12.12. (a) Heat map showing a representation of the chemical space of the Ugi library. (b) Distribution of cells according to different population ranges for the Ugi library.

the cells are unoccupied, approximately 5% of the cells contain compounds in the 10^3 range, and a few cells contain more than 10^5 compounds.

Projection of the 23 million GDB-11 compounds onto the PubChem reference chemical space shows that these small molecular fragments occupy primarily the upper right quadrant (Fig. 12.13a), an area distinctly different from the location of the Ugi library. In fact, a significant portion of compounds from GDB-11 occupy a region of space that is sparsely represented by the PubChem compounds. Figure 12.13b shows a different heat map where this "virtual" chemical space (i.e., upper middle part of the plot) unique to the GDB-11 compounds is characterized. About 60% of the grid cells are

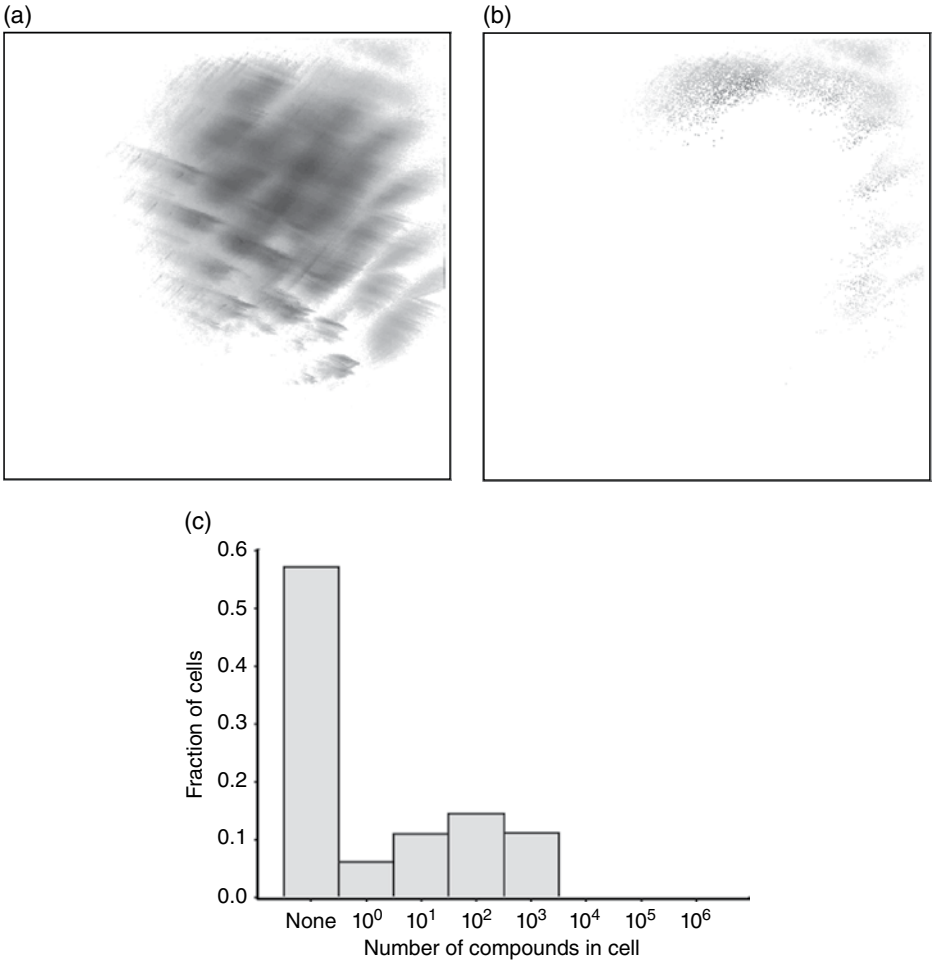

Figure 12.13. (a) Heat map showing a representation of the chemical space of the GDB-11 database. (b) Heat map highlighting the areas of chemical space that are represented by molecules in the GDB-11 database but not by any compound in the PubChem database. (c) Distribution of cells according to different population ranges for the PubChem database.

empty, but unlike either PubChem or Ugi libraries, the GDB-11 population is distributed quite evenly across the rest of the chemical space, and none of the cells have population density greater than 10^3 molecules (Fig. 12.13c).

Figure 12.14a shows the heat map that corresponds to the chemical space of the Neri library. In general, the location of chemical space occupied by the Neri library seems quite similar to the Ugi library (Fig. 12.14a). But due to very high structural similarity among its members, the associated chemical space seems quite confined. This

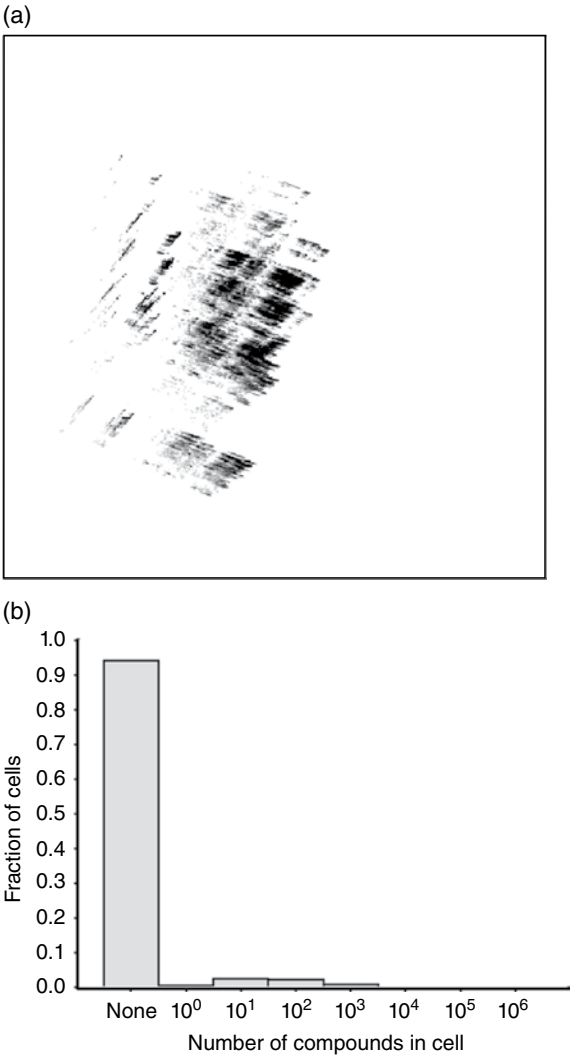

Figure 12.14. (a) Heat map showing a representation of the chemical space of the Neri DNA-encoded chemical library. (b) Distribution of cells according to different population ranges for the Neri DNA-encoded chemical library.

finding is also consistent with the fact that only 6% of the grids are occupied, and also a significant proportion of its population resides inside just a handful of cells.

12.3.3 Library Similarity Calculation

The heat maps shown in the previous section represent a useful qualitative tool that provides comparative visualization of how different compound libraries occupy distinct regions of chemical space. A method that yields a quantitative measure of library overlap using molecular similarity concept can provide complementary information. The Tanimoto coefficient is arguably the most commonly used formula (Eq. 12.1) for molecular similarity calculation, and these coefficients are readily computed based on the comparison of molecular fingerprints between a pair of molecules:

$$\text{Tanimoto}\ (\text{A,B}) = \frac{N_{\text{AB}}}{N_{\text{A}} + N_{\text{B}} - N_{\text{AB}}} \tag{12.1}$$

where

N_{A} and N_{B} are the number of bits in the fingerprint set in molecule A and molecule B
N_{AB} is the number of bits set common to both molecules

Tanimoto coefficient ranges between 0 and 1, where a higher value indicates a molecular pair that is structurally more similar. Here, the use of this concept is extended, and a mean Tanimoto coefficient value is derived from similarity calculations of all molecule pairs between two compound libraries. This new parameter is referred to as "library similarity," and the following simple worked example is used to illustrate the principle (Fig. 12.15).

In this example, there are three "libraries" of compounds, and the goal is to obtain a quantitative measure to determine whether Library 2 or Library 3 is structurally more

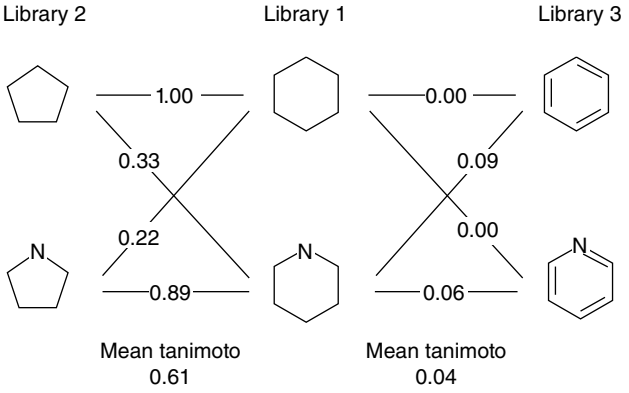

Figure 12.15. A simple example showing the calculation of mean Tanimoto coefficient (or library similarity) between library sets. The result indicates that Library 1 is structurally more similar to Library 2 according to library similarity calculation.

similar to Library 1. Library 1 contains cyclohexane and piperidine, Library 2 contains cyclopentane and pyrrolidine, and Library 3 contains benzene and pyridine. To compare Library 1 and Library 2, the Tanimoto coefficients of all compound combinations between the two library sets are computed based on a user-selected molecular fingerprint (FCFP_4 fingerprints in this example). This resulted in a 2×2 similarity matrix. The four individual Tanimoto coefficients in the matrix were averaged to give a library similarity value of 0.61 that indicates a relatively high degree of similarity between these two libraries. It seems reasonable as both set of compounds are cyclic aliphatic molecules. Repeating the procedure to compare Library 1 and Library 3 yielded a much lower library similarity (0.04), a reflection of the obvious difference in aromaticity between the two sets of compounds. This example illustrates that this type of calculation is quite easy to set up but can quickly become prohibitively expensive when comparing large numbers of input compounds because calculation of pairwise similarity is an $O(N^2)$ process. For example, it is estimated that the computation of a full similarity matrix between the 1.6 billion compound Ugi library and the 30 million compound PubChem database would require 30,000 years on a Xeon CPU (E7340 @ 2.4 GHz). So, a strategy to obtain an exact result of the full similarity matrix by using a brute-force approach (e.g., massive parallelization with a large array of computer processors) seems intractable and, most likely, wasteful. Instead, one would attempt to estimate the library similarity value between two compound libraries using small, multiple sets of random samples taken from each library. To test the effectiveness of this sampling approach, a subset of 50,000 compounds from the Ugi library was randomly selected and along with a subset of 50,000 compounds from the PubChem database. It was possible to determine an exact solution (0.1822) for the library similarity between the two library subsets by performing 2.5 billion similarity calculations, which took 13 CPU hours. Next, a series of sampling calculations was run on the subsets by varying both (i) the sample size (N_S) used for the similarity calculations and (ii) the number of repeated sampling runs (N_R) so that different sets of random samples can be drawn from the two 50,000 compound subsets. Table 12.3 shows the estimates of library similarity for a range of N_S and N_R parameters. First, it is remarkable that nearly all of the runs yielded very good estimates with the exception of those involving very few samples (i.e., N_S or $N_R \leq 5$). The N_S or N_R combinations shown in grey cells in the table produced estimates that were accurate to three significant figures to the exact result (0.1822). It is also interesting to compare the following pairs of result that are derived from same number of individual similarity calculations. For example, a 10×10 matrix repeated 100 times (library similarity estimate = 0.1832) and a 100×100 matrix done just once (0.1849) both took 10,000 similarity calculations, a 10×10 matrix repeated 1,000 times (0.1824) and a 100×100 matrix repeated 10 times (0.1830) both took 100,000 calculations, and a 50×50 matrix repeated 1,000 times (0.1824) and a 500×500 matrix repeated 10 times (0.1832) both took 2,500,000 calculations. In all cases, the observed trend strongly suggests that, given the same number of similarity calculations, a more rapid convergence of library similarity can be achieved by a higher number of repeated runs at the expense of smaller sample size drawn from the libraries.

To investigate the convergence behavior, the sampling calculation for the 10×10 matrix set was continued in order to draw additional random samples until the simulation reached 100,000 iterations. The value for the library similarity estimate obtained at each iteration step was tracked plotted against the number of iterations taken (Fig. 12.16).

TABLE 12.3. Variation of sampling accuracy according to sampling size of the similarity matrix ($N_S \times N_S$) and also the number of repeated sampling runs (N_R)

	Number of sampling runs (N_R)								
Matrix size ($N_S \times N_S$)	1000	500	250	100	50	25	10	5	1
1000 × 1000	0.1822	0.1822	0.1823	0.1824	0.1823	0.1822	0.1824	0.1832	0.1830
500 × 500	0.1822	0.1823	0.1823	0.1823	0.1823	0.1824	0.1832	0.1830	0.1850
250 × 250	0.1823	0.1823	0.1823	0.1822	0.1824	0.1828	0.1830	0.1840	0.1850
100 × 100	0.1825	0.1824	0.1823	0.1825	0.1832	0.1830	0.1830	0.1850	0.1849
50 × 50	0.1824	0.1823	0.1825	0.1834	0.1832	0.1842	0.1852	0.1849	0.1820
25 × 25	0.1823	0.1824	0.1829	0.1832	0.1842	0.1840	0.1851	0.1854	0.1656
10 × 10	0.1824	0.1833	0.1832	0.1832	0.1847	0.1841	0.1839	0.1810	0.1795
5 × 5	0.1835	0.1835	0.1843	0.1854	0.1844	0.1855	0.1826	0.1658	0.1858
1 × 1	0.1838	0.1872	0.1865	0.1871	0.1807	0.1622	0.1787	0.1679	0.1343

The exact solution of library similarity between the two library subsets is 0.1822.

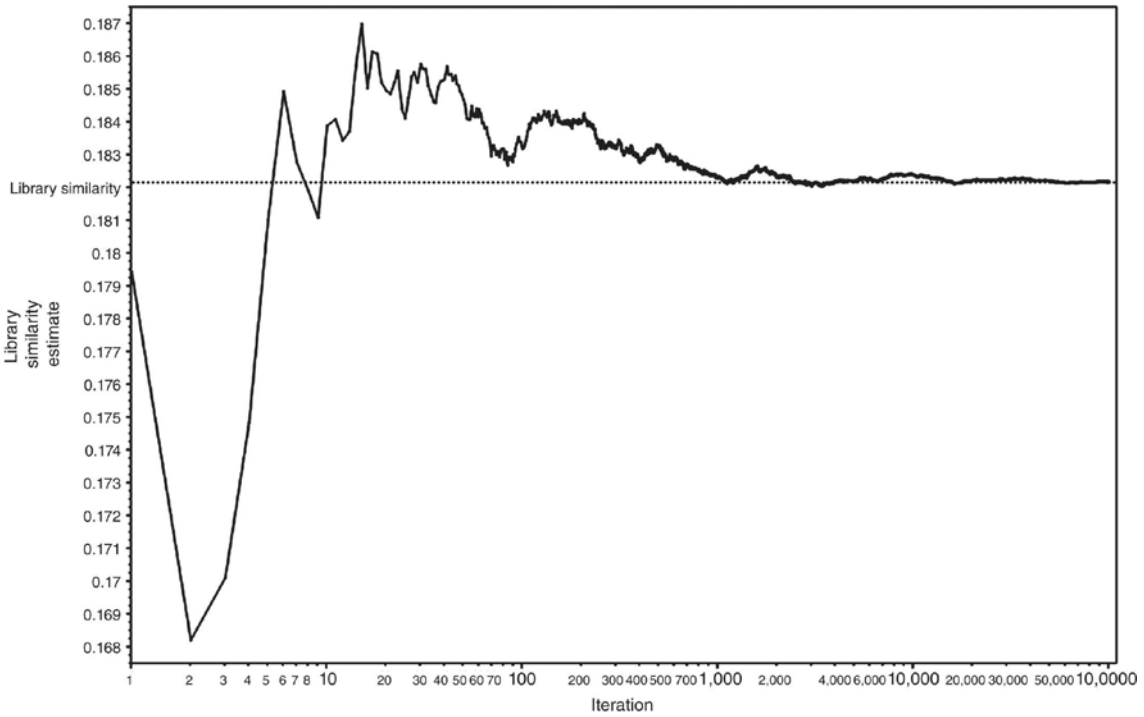

Figure 12.16. Estimate of library similarity as a function of iteration steps between the Ugi and the PubChem library subsets using a 10 × 10 sampling matrix. The exact solution for library similarity (0.1822) is represented by the horizontal dotted line. The horizontal axis of the figure is shown in log scale to highlight convergence behavior at the early stages of the calculation. A sampling accuracy to two, three, and four significant figures was reached after 100, 1,000, and 50,000 iterations, respectively.

The dotted line shows the exact solution of library similarity calculated from the full matrix. As expected, the estimate for library similarity fluctuated quite significantly at the early phase due to insufficient sampling of data. It was very encouraging that after only 100 iterations, the sampling calculation has already produced an estimate that is already accurate to two significant figures. At the 1000th iteration, sampling accuracy to three significant figures—a level of accuracy that is probably not needed for most practical purposes—was reached. In an actual application, one can define their desired level of sampling accuracy and terminate the simulation once this condition of interest is met. For this current case, the total number of similarity calculations that had been performed at this stage is $10 \times 10 \times 1{,}000 = 100{,}000$, a very small number (about 0.0004%) comparing to the 2.5 billion individual calculations needed for the full matrix calculation.

After this initial validation on smaller subsets, the same sampling protocol was used to estimate library similarity values between the Ugi, Neri, PubChem, and GDB-11 libraries. For the full runs, 100 random samples were taken from each library at a time, and the similarity matrix calculation was repeated 50,000 times (i.e., $100 \times 100 \times 50{,}000 = 500$ million similarity pairs). Based on this test run, this should provide sufficient sampling to generate reasonable library similarity estimates, and further sampling would yield only marginal improvement. Each library comparison calculation took approximately 4 CPU hours to complete.

The library similarity values from these pairwise library comparisons are shown in Table 12.4. The low values obtained from the comparisons between the various libraries against GDB-11 confirm earlier qualitative assessment from the PCA-based heat maps. The GDB-11 seems to be the most unique of the four libraries as these molecular fragments were occupying a distinct region of chemical space. Not surprisingly, the Neri library has the highest intralibrary similarity (0.457) of the four libraries. This is due to a high level of structure homogeneity originating from a unique scaffold that is present in every compound. The Neri library seems less structurally similar (library similarity = 0.170) to PubChem in general than the Ugi library to PubChem (library similarity = 0.182). This difference in library similarity value is significant from a computational library design point of view. For example, if the purpose of library building is to produce chemical moieties that are generally more novel than compounds that currently exist, one may want to prioritize the production of the Neri library over the Ugi library since most of its library members are considered more unique and chemically differentiated than those found in PubChem.

TABLE 12.4. The library similarity values from the comparison of four large chemical libraries (Ugi, Neri, PubChem, and GDB-11) used in this study

	Ugi	Neri	PubChem	GDB-11
Ugi	0.289 ± 0.006	0.216 ± 0.003	0.182 ± 0.004	0.099 ± 0.002
Neri		0.457 ± 0.007	0.170 ± 0.004	0.108 ± 0.003
PubChem			0.176 ± 0.006	0.099 ± 0.003
GDB-11				0.147 ± 0.005

The diagonal values are the intralibrary similarity values that reflect compound heterogeneity within a given library. The off-diagonal elements are the interlibrary similarity values between two different compound libraries.

12.4 SUMMARY

A major lesson learned from early days of combinatorial library synthesis and screening is that without proper library design, many chemical libraries either failed to improve hit rate against therapeutic targets or the poor drug-like characteristics of resulting hits had discouraged further medicinal chemistry follow-up [12, 50, 51]. Clearly, the ability to make millions or even billions of compounds is not sufficient to ensure success; a good library design is still a critical factor. In this chapter, a practical overview for reagent selection, enumeration, property profiling, and library comparison of some very large combinatorial libraries was shown. Cheminformatics workflow packages offer many useful tools for both the selection and filtering of building blocks. SMARTS-based queries are simple to implement and can quickly identify very specific substructures that are reaction compatible. Enumeration of combinatorial library with 10^9 members can now be done on a routine basis, although libraries with greater than 10^{12} members still pose technical challenges. Also demonstrated was that the mean and statistical distribution of several key properties commonly used in molecular design can be accurately estimated using a small fraction of random library products. These simulations suggested that random sampling on product space is deemed much more effective than on partially enumerated library products based on reagent-based sampling.

Heat maps derived from the first two principal components of six BCUT metrics were used to compare the chemical space of different libraries, using the PubChem data set as a spatial reference. The molecular similarity concept can be extended to provide a quantitative parameter, library similarity, to estimate overlap between two chemical libraries. This author suggests that library similarity with other libraries is a factor to be considered in the prioritization and production of large combinatorial chemical libraries.

ACKNOWLEDGMENTS

The author gratefully acknowledges the careful review of the manuscript and helpful comments by Drs. Robert Goodnow and Paul Gillespie.

REFERENCES

1. Kennedy, J. P., Williams, L., Bridges, T. M., Daniels, R. N., Weaver, D., Lindsley, C. W. (2008). Application of combinatorial chemistry science on modern drug discovery. *J. Comb. Chem.*, 10, 345–354.
2. Martin, E. J., Blaney, J. M., Siani, M. A., Spellmeyer, D. C., Wong, A. K., Moos, W. H. (1995). Measuring diversity: experimental design of combinatorial libraries for drug discovery. *J. Med. Chem.*, 38, 1431–1436.
3. Holliday, J. D., Ranade, S. S., Willett, P. (1995). A fast algorithm for selecting sets of dissimilar molecules from large chemical databases. *Quant. Struct.–Activity Relationships*, 14, 501–506.
4. Bender, A., Glen, R. C. (2004). Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.*, 2, 3204–3218.

5. Maldonado, A. G., Doucet, J. P., Petitjean, M., Fan, B.-T. (2006). Molecular similarity and diversity in chemoinformatics: from theory to applications. *Mol. Diversity*, 10, 39–79.
6. Gillet, V. J. (2011). Diversity selection algorithms. *WIREs Comput. Mol. Sci.*, 1, 580–589.
7. Ashton, M. J., Jaye, M. C., Mason, J. S. (1996). New perspectives in lead generation. II Evaluating molecular diversity. *Drug Discov. Today*, 1, 71–78.
8. Lewis, R. A., Mason, J. S., McLay, I. M. (1997). Similarity measures for rational set selection and analysis of combinatorial libraries: the diverse property-derived (DPD) approach. *J. Chem. Inf. Comput. Sci.*, 37, 599–614.
9. Flower, D. R. (1998). DISSIM: a program for the analysis of chemical diversity. *J. Mol. Graph. Model.*, 16, 239–253.
10. Lobanov, V. S., Agfrafiotis, D. K. (2002). Scalable methods for the construction and analysis of virtual combinatorial libraries. *Comb. Chem. High Throughput Screen*, 5, 167–178.
11. Gillet, V. J. (2008). New directions in library design and analysis. *Curr. Opin. Chem. Biol.*, 12, 372–378.
12. Gillet, V. J., Khatib, W., Willett, P., Fleming, P. J., Green, D. V. (2002). Combinatorial library design using a multiobjective genetic algorithm. *J. Chem. Inf. Comput. Sci.*, 42, 375–385.
13. Jamois, E. A., Lin, C. T., Waldman, M. (2003). Design of focused and restrained subsets from extremely large virtual libraries. *J. Mol. Graph. Model.*, 22, 141–149.
14. Schneider, G., Schueller, A. (2009). Adaptive combinatorial design of focused compound libraries. *Methods Mol. Biol.*, 572, 135–147.
15. Yu, N., Bakken, G. A. (2009). Efficient exploration of large combinatorial chemistry spaces by monomer-based similarity searching. *J. Chem. Inf. Model.*, 49, 745–755.
16. Sciabola, S., Stanton, R. V., Johnson, T. L., Xi, H. (2011). Application of Free-Wilson selectivity analysis for combinatorial library design. *Methods Mol. Biol.*, 685, 91–109.
17. Truchon, J.-F. (2011). GLARE: a tool for product-oriented design of combinatorial libraries. *Methods Mol. Biol.*, 685, 337–346.
18. Lipinski, C. A., Lombardo, F., Dominy, B. W., Feeney, P. J. (1997). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.*, 23, 3–25.
19. Ugi, I., Lohberger, S., Karl, R. The Passerini and Ugi reactions. In: *Comprehensive organic synthesis: selectivity for synthetic efficiency*, Vol. 2, Trost, B., Fleming, I., Eds. Pergamon Press, Oxford, 1991, pp. 1083–1109.
20. Accelrys Available Chemicals Directory (ACD). Accelrys, Inc. 10188 Telesis Court, Suite 100, San Diego, CA 92121, USA.
21. Pipeline Pilot, version 8.5. Accelrys, Inc. 10188 Telesis Court, Suite 100, San Diego, CA 92121, USA.
22. Daylight toolkit. Daylight Chemical Information Systems, Inc., PO Box 7737, Laguna Niguel, CA 92677, USA.
23. OEChem TK. OpenEye Scientific Software, 9 Bisbee Court, Suite D, Santa Fe, NM 87508, USA.
24. A tutorial and examples of SMARTS can be found following this URL: http://www.daylight.com/dayhtml_tutorials/languages/smarts/index.html. Accessed on November 28, 2013.
25. Beroza, P., Bradley, E. K., Eksterowicz, J. E., Feinstein, R., Greene, J., Grootenhuis, P. D. J., Henne, R. M., Mount, J., Shirley, W. A., Smellie, A., Stanton, R. V., Spellmeyer, D. C. (2000). Applications of random sampling to virtual screening of combinatorial libraries. *J. Mol. Graph. Model.*, 18, 335–342.

26. Hert, J., Willett, P., Wilton, D. J., Acklin, P., Azzaoui, K., Jacoby, E., Shuffenhauer, A. (2004). Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.*, 44, 1177–1185.
27. Schuffenhauer, A., Brown, N., Ertl, P., Jenkins, J. L., Selzer, P., Hamon, J. (2007). Clustering and rule-based classifications of chemical structures evaluated in the biological activity space. *J. Chem. Inf. Model.*, 47, 325–336.
28. Steffen, A., Kogej, T., Tyrchan, C., Engkvist, O. (2009). Comparison of molecular fingerprint methods on the basis of biological profile data. *J. Chem. Inf. Model.*, 49, 338–347.
29. Willett, P. (2011). Similarity searching using 2D structural fingerprints. *Methods Mol. Biol.*, 672, 133–158.
30. DiscNgine. DISCNGINE S.A.S., Parc Biocitech, 102 route de Noisy, 93230, Romainville, France.
31. Agrafiotis, D. K., Lobanov, V. S. (2001). Multidimensional scaling of combinatorial libraries without explicit enumeration. *J. Comp. Chem.*, 22, 1712–1722.
32. A tutorial and examples of SMIRKS can be found following this URL: http://www.daylight.com/dayhtml_tutorials/languages/smirks/index.html. Accessed on November 28, 2013.
33. Lobanov, V. S., Agfrafiotis, D. K. (2000). Stochastic similarity selections from large combinatorial libraries. *J. Chem. Inf. Comput. Sci.*, 40, 460–470.
34. Gillet, V. J., Willet, P., Bradshaw, J. (1997). The effectiveness of reactant pools for generating structurally-diverse combinatorial libraries. *J. Chem. Inf. Comput. Sci.*, 37, 731–740.
35. Jamois, E. A., Hassan, M., Waldman, M. (2000). Evaluation of reagent-based and product-based strategies in the design of combinatorial library subsets. *J. Chem. Inf. Comput. Sci.*, 40, 63–70.
36. Gillet, V. J. (2002). Reactant- and product-based approaches to the design of combinatorial libraries. *J. Comput.-Aided Mol. Des.*, 16, 371–380.
37. Pearlman, R. S., Smith, K. M. (1998). Novel software tools for chemical diversity. *Perspect. Drug Discov.*, 9–11, 339–353.
38. Fink, T., Bruggesser, H., Reymond, J.-L. (2005). Virtual exploration of the small-molecule chemical universe below 160 daltons. *Angew. Chem. Int. Ed.*, 44, 1504–1508.
39. Medina-Franco, J. L., Martinez-Mayorga, K., Giulianotti, M. A., Houghten, R. A., Pinilla, C. (2008). Visualization of the chemical space in drug discovery. *Curr. Comput. Aided Drug Des.*, 4, 322–333.
40. Nguyen, K. T., Blum L. C., van Deursen, R., Reymond, J.-L. (2009). Classification of organic molecules by molecular quantum numbers. *ChemMedChem*, 4, 1803–1805.
41. Reymond, J.-L., Ruddigkeit, L., Blum, L., van Deursen, R. (2012). The enumeration of chemical space. *WIREs Comput. Mol. Sci.*, 2, 717–733.
42. Lopez-Vallejo, F., Giulianotti, M. A., Houghten, R. A., Medina-Franco, J. L. (2012). Expanding the medicinally relevant chemical space with compound libraries. *Drug Discov. Today*, 17, 718–726.
43. Bolton, E. E., Wang, Y., Thiessen, P. A., Bryant, S. H. (2008) PubChem: integrated platform of small molecules and biological activities. In: Annual reports in computational chemistry, Vol. 4, pp. 217–241. American Chemical Society, Washington, DC.
44. Fink, T., Reymond, J.-L. (2007). Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physico-chemical properties, compound classes and drug discovery. *J. Chem. Inf. Model.*, 47, 342–353.

45. Buller, F., Steiner, M., Frey, K., Mircsof, D., Scheuermann, J., Kalisch, M., Buehlmann, P., Supuran, C. T., Neri, D. (2011). Selection of carbonic anhydrase IX inhibitors from one million DNA-encoded compounds. *ACS Chem. Biol.*, 6, 336–344.

46. Burden, F. R. (1997). A chemically intuitive molecular index based on the eigenvalues of a modified adjacency matrix. *Quant. Struct.–Activity Relationships*, 16, 309–314.

47. Reutlinger, M., Schneider, G. (2012). Nonlinear dimensionality reduction and mapping of compound libraries for drug discovery. *J. Mol. Graph. Model.*, 34, 108–117.

48. Oprea, T. I., Gottfries, J. (2001). Chemography: the art of navigating in chemical space. *J. Comb. Chem.*, 3, 157–166.

49. Oprea, T. I. (2002). Chemical space navigation in lead discovery. *Curr. Opin. Chem. Biol.*, 6, 384–389.

50. Martin, E. J., Crichlow, R. E. (1999). Beyond mere diversity: tailoring combinatorial libraries for drug discovery. *J. Comb. Chem.*, 1, 32–45.

51. Valler, M. J., Green, D. (2000). Diversity screening versus focussed screening in drug discovery. *Drug Discov. Today*, 5, 286–293.