ELSEVIER

Applications of random sampling to virtual screening of combinatorial libraries

Paul Beroza, Erin K. Bradley, John E. Eksterowicz,
Robert Feinstein, Jonathan Greene, Peter D. J. Grootenhuis,
Randal M. Henne, John Mount, William A. Shirley, Andrew Smellie,
Robert V. Stanton, and David C. Spellmeyer

DuPont Pharmaceuticals Research Laboratories, San Francisco, California, USA

We describe statistical techniques for effective evaluation of large virtual combinatorial libraries ($>10^{10}$ potential compounds). The methods described are used for computationally evaluating templates (prioritization of candidate libraries for synthesis and screening) and for the design of individual combinatorial libraries (e.g., for a given diversity site, reagents can be selected based on the estimated frequency with which they appear in products that pass a computational filter). These statistical methods are powerful because they provide a simple way to estimate the properties of the overall library without explicitly enumerating all of the possible products. In addition, they are fast and simple, and the amount of sampling required to achieve a desired precision is calculable. In this article, we discuss the computational methods that allow random product selection from a combinatorial library and the statistics involved in estimating errors from quantities obtained from such samples. We then describe three examples: (1) an estimate of average molecular weight for the several billion possible products in a four-component Ugi reaction, a quantity that can be calculated exactly for comparison; (2) the prioritization of four templates for combinatorial synthesis using a computational filter based on four-point pharmacophores; and (3) selection of reagents for the four-component Ugi reaction based on their frequency of occurrence in products that pass a pharmacophore filter. © 2000 by Elsevier Science Inc.

Keywords: combinatorial chemistry, combinatorial library design, virtual library screening, pharmacophore, computational filter

Corresponding author: David C. Spellmeyer, DuPont Pharmaceuticals Research Laboratories, 150 California Street, Suite 1100, San Francisco, CA 94111, USA. Tel.: 858-625-6701; fax: 415-732-7887.

E-mail address: david.c.spellmeyer@dupontpharma.com

INTRODUCTION

Combinatorial chemical synthesis offers the promise of large numbers, with a single synthetic strategy routinely providing access to millions of unique compounds. The potential availability of millions or even billions of compounds is particularly appealing in the area of drug discovery, where good lead candidates are rare,[1] and the chance of finding a lead candidate is generally thought to increase with the number of compounds synthesized and screened. However, large numbers can be a burden. Associated with each chemical synthesized in a combinatorial library is the overhead of purification, analysis, biological assay, and deconvolution of potential hits if mixtures of compounds are screened. The resources necessary can offset the benefits, and it quickly becomes desirable to limit and focus synthetic and screening efforts. To this end, computational models have helped direct drug discovery efforts using combinatorial synthesis.[2–5] Such models can be used to filter a large library, resulting in a smaller library that is enriched for a desired property. The resultant smaller library is taken forward to chemical synthesis, purification, and analysis, and finally assayed against one or more targets.

The computational methods applied to combinatorial library design are closely related to methods used in searching a virtual database of compounds. In screening applications, libraries of compounds, generated from internal synthetic efforts or purchased from external sources, are screened against newly identified targets. These collections are often large (10^5 to 10^6 compounds), and screening all of the compounds may not be practical. Computational methods can be employed at this stage to filter the compounds that are taken forward to screening. If active compounds are already known, the filter can take the form of a similarity search, i.e., identification of the compounds in the collection that are most chemically similar to the known actives.[6–8] A more general approach is to take forward a subset of the collection that is highly diverse with the hope that the chemical information in the subset is sufficiently close

Journal of Molecular Graphics and Modelling 18, 335–342, 2000
© 2000 by Elsevier Science Inc.
655 Avenue of the Americas, New York, NY 10010

1093-3263/00/$–see front matter
PIIS1093-3263(00)00059-0

to that of the entire collection so that no opportunities are missed.[3,9] Clustering methods,[10–13] in which chemically similar compounds are grouped, also can be useful, because corporate collections often consist of compounds from previous discovery efforts (against different targets) and, therefore, fall naturally into clusters. Representatives from each cluster can be taken forward, and, when hits are identified, the next screening round can be focused on the other members of the "active" clusters. The important point is that all these methods apply computational filters to the library. Each compound is represented on the computer and evaluated using algorithms designed to eliminate undesired molecules.

Computational filters, similar to those used for the searching of screening libraries, can be applied to combinatorial library design. Such libraries typically are derived from a scaffold, a set of chemical reactions, and selected lists of reagents. The resulting products can be represented *in silico* and selected for synthesis based on the criteria described earlier (i.e., similarity to known actives, chemically diverse from one another, etc.). However, combinatorial synthesis poses the following additional constraint. To reduce costs and to simplify the overall synthesis, the number of reagents at each diversity site should be kept to a minimum, which may conflict with the selections from the computational filter. Note that additional complications associated with combinatorial mixtures are not present in our designs, because our synthetic schemes result in purified, single compounds. This constraint creates an interesting optimization problem: How can we maximize the number of desired compounds (those that pass the computational filter), while keeping the number of reagents to a minimum? This problem has been addressed for chemistries with relatively small potential product spaces ($<10^6$ compounds) by methods that enumerate the complete set of products and then select a subset that optimizes the characteristics of the products while maintaining the constraints of combinatorial synthesis.[14,15] But what if the set of possible products is too large to allow for the complete computational examination of all possible products?

This article describes how random statistical sampling of the possible products in a combinatorial library provides a simple way to estimate the properties of the overall library without the need to explicitly enumerate all of the library's possible products. This allows computational filters to assess combinatorial libraries that are vast (e.g., $>10^{10}$ possible products). Use of these statistical estimates allows us to evaluate and prioritize different libraries that are candidates for synthesis and screening. In addition, an extension of this technique facilitates selection of reagents for synthesis at a diversity site, as they can be ranked by the estimated frequency with which they appear in products that pass a computational filter (i.e., a "virtual screen"). The statistical methods presented have the advantage that they are fast and simple, and the amount of sampling required to achieve a desired precision is calculable.

We first discuss the software design that allows random product selection from a combinatorial library and the statistics involved in estimating errors from quantities obtained from such samples. We then describe three examples: (1) an estimate of average molecular weight for the several billion possible products in the four-component Ugi reaction, a quantity that can be calculated exactly for comparison; (2) the prioritization of four templates for combinatorial synthesis using a computational filter based on four-point pharmacophores; and (3) selection of monomers for the four-component Ugi reaction based on their frequency of occurrence in products that pass a pharmacophore filter.

METHODS

Virtual Combinatorial Libraries

Representation of molecules in computer programs is now commonplace: in two dimensions, molecules are represented as networks of elements and bonds; in three dimensions, coordinates of atomic nuclei are stored and the relative positions of atoms in a molecule are mostly determined by force-field approximations to the interatomic forces. In either case, it is impractical to construct and store all possible products for large combinatorial libraries in computer memory. However, it is possible to represent a combinatorial library as a set of reagents and reactions. A chemical reaction simulation program (the Cascader™) was developed internally to provide automation in the enumeration of combinatorial libraries. Briefly, the Cascader™ takes reactant molecules, reaction transformations, and synthesis schemes as input. The reactant molecules provide the building blocks for product enumeration. Reaction transformations provide details about what combinations of functional groups will react, along with the atom transformations for converting combinations of reactants into products. The synthesis schemes ("cascades") describe how reactions are chained together to simulate a multistep (or multicomponent, "one-pot") synthesis. The reaction-based product enumeration allows access to a very large population of products (millions to billions), which cannot be practically enumerated, and provides an implicit way to store them. Instead of storing the list of molecules that comprise the collection, we store a set of rules and constraints for generating such molecules from the much more easily stored reactants.

Random Sampling of Products

The combination of a cascade and sets of reactant molecules defines a population of virtual products, or a "virtual library." If each set of reactant molecules is thought of as a list, each combination of reactants (one per set) can be thought of as a coordinate in the virtual library, in which each dimension of the coordinate is an index into the corresponding reactant list. The Cascader can enumerate specific coordinates of a virtual library, each of which represents a particular combination of reactant molecules that result in a product structure. This provides a convenient mechanism for fully enumerating all products or sampling a subset of products from a specified virtual library. If only unique products are desired, a canonical representation of each product can be constructed and compared to the growing list of previously seen products.

These core library enumeration algorithms form the basis of several stand-alone tools that can produce arbitrary subsets of unique products of specified sizes. They also can be accessed via an extension module in a chemical scripting environment based on the Python programming language to provide a source of molecular products that can be sampled iteratively until some arbitrary termination condition is met. In this case, the termination condition of interest is a user-specified sampling accuracy.

Estimating Sampling Error

The reaction-based product enumeration provided by the Cascader allows us to store implicitly a large population of products. Although this population is too large to enumerate explicitly, any product can be readily constructed from a chosen combination of reagents and the rules for combining them. By storing our collection of molecules in this way, we can generate a uniform random sample (with or without replacement) and use the proportions measured in the sample to estimate the properties of the entire (implicit) collection.

Fortunately, the manner in which sample measurements approximate the total population effectively is well understood. We can, therefore, design our sample to guarantee that the measured proportion is within a given tolerance and probability. It has been shown previously (see, for example, Hoeffding[16]) that if we wish to measure a proportion to within an absolute error of ±x%, a sample of size k has a probability of giving an incorrect result of no more than[17]:

$$2e^{-kx^2/2{,}000}. \tag{1}$$

For example, a sample of size 1,000 is sufficient to guarantee a measurement that falls within a 10% absolute error of the true value with a probability $>$98.5%. It is important to note that 20% estimated to an absolute error of 10% is a number from 10% to 30%, not 18% to 22%. As long as the sample is drawn uniformly (and independently), this bound (Equation 1) is correct, independent of the size of the total population and independent of the unknown proportion.

Because Equation 1 is general, it tends to predict that a fairly large sample size is necessary. However, if we were to incorporate domain knowledge (such as known bounds on the true proportion and, to a lesser extent, the total population size), we can prove that smaller samples suffice. For this case, we can use the exact binomial formula instead of the Hoeffding estimate. When our sample is drawn with replacement, the exact odds of success become:

$$\sum_{i=\frac{(p-x)}{100}N}^{\frac{(p+x)}{100}N} \binom{N}{i} (p/100)^i (1-p/100)^{N-i} \tag{2}$$

where N is the size of the original population and p is the (unknown) true proportion (written as a percentage). As p will not be known, a good approximation can be obtained by replacing the unknown value of p with the worst-case value of 50% or with a user-supplied bound. Also, the exact value of N is not required, so any upper bound will do.

For samples without replacement (that is, samples with no repeated values), the bounds become slightly tighter (especially for small N); however, the explicit odds of success again are a simple series. We have implemented each of the methods described here and use them, as appropriate, to design our samples. For the more difficult problem of measuring very small proportions to a given relative error, see Mount.[18] These methods can be used to make informed choices between rare events (one of which might be much more desirable than others).

Computational Filters

In two of the examples of random sampling, we use computational filters constructed from pharmacophore-based 3D whole molecule descriptors. Pharmacophore descriptions of molecules and their application to virtual library searching and design have been described elsewhere,[19–21] and only the details pertaining to their use in evaluating random samples will be summarized here. The major component of our 3D descriptors is the "four-point pharmacophore," which consists of four chemical features and the six interfeature distances, and a chiral indicator. Standard feature types (i.e., hydrogen-bond acceptors and donors, hydrophobes, negative and positive charges, and aromatic groups) were identified on molecules by substructure query matches as described by others.[22,23] The potential number of pairwise feature distances is limited to a specific set of distance bins (e.g., interfeature distances between 3.5 and 5 Å would map to a single distance bin). We used 14 bins for the two- and three-point pharmacophore distances, spanning 1.6 to 13.2 Å, and eight bins for the four-point pharmacophore distances, spanning the same distance. Thus, the "pharmacophore space" (all possible combinations of two, three, and four features) is predetermined by interfeature distance bins and the specific set of features. Similar to Mason et al.,[20] we use a "molecular signature," a bit-string where the presence or absence of each of the two-, three-, or four-point pharmacophores is recorded. This resulted in a pharmacophore signature length of ~12 million bits.

Our computational filters consist of a specific subset of the bits in the pharmacophore signature (an ensemble) that is associated with a desired property in a chemical product (e.g., pharmacophores that are present in biologically active molecules). In the examples that follow, two different filters are used. The first consists of a randomly selected set of 100 pharmacophores from the ~12-million bit signature. The second consisted of an ensemble of 62 pharmacophores contained in the conformation of NAPAP bound to Thrombin (1ETS structure in the protein databank).[24]

To assess whether an individual molecule has the desired property (i.e., passes the computational filter), we generate the conformational model for the molecules using an in-house program CONAN (Conformational Analysis by intersection) described in greater detail elsewhere.[25,26] Then all two-, three-, and four-point pharmacophores that are present in the molecule's conformers are recorded as the molecule's pharmacophore signature. The molecule passes the computational filter if its signature contains a specified number of the pharmacophores in the ensemble.

It is important to point out that these computational filters were generated for purposes of demonstrating the sampling methods rather than constructing a predictive computational model for biological activity. Those interested in the construction of pharmacophore ensembles generated from activity data for a particular biological target should refer to the work of Bradley et al.[26]

Applications of Random Sampling

Physical property estimation The simplest application of random sampling is estimating the average physical properties of the products in a combinatorial library. Each compound chosen is constructed *in silico*, and its properties are calculated. As the compounds are sampled sequentially, running averages of properties are computed and convergence criteria are checked. When the estimated errors for the given sample size are acceptably small, sampling is terminated. Molecular weight, calculated octanol/water partition coefficient, and number of rotatable bonds are examples of physical properties that can be estimated and used to compare combinatorial libraries. These properties, although not well correlated with biological activity, are useful in determining which libraries follow observed properties in known drugs.[27,28]

Template evaluation A more challenging application for random sampling is to use the sampled compounds to estimate what fraction of a combinatorial library will pass a more complex computational filter (described earlier). After a sampled product is synthesized on the computer, its low-energy conformations determined, and its pharmacophore descriptor constructed, a score is assigned to the product based on the number of pharmacophores it has in common with those in the virtual filter. As each randomly sampled product in the virtual combinatorial library is sampled, a running average of scores is kept and an overall "pass rate" (i.e., fraction of products that pass some score threshold) is computed. Libraries then are compared based on their pass rates. The higher the pass rate, the more likely the library is able to provide products that pass the computational filter. We used a pharmacophore filter of 500 randomly selected pharmacophores. To pass the filter, a compound had to have $>$100 of these pharmacophores present in its signature.

Monomer selection Random sampling techniques can be extended to facilitate combinatorial library design for synthesis on a single template. Often combinatorial chemistries involve a single chemical scaffold, or template, on which pendant groups can be attached to "diversity sites" using various synthetic strategies. Each diversity site has a restricted set of chemicals that are appropriate, usually because they must have a reactive chemical moiety (e.g., if a template with an amine participates in an amide bond formation, the reagent must be an acid). Nevertheless, each reagent list for a diversity site can be quite large (hundreds or thousands of compounds), and it is desirable to limit consideration to those monomers that are more likely to be present in "successful" (i.e., model matching) products. This can be accomplished using a technique called "lockdown".

In random sampling with "lockdown," products are generated until there are enough of them to gather statistics on the monomers at the different diversity sites. At that point, one diversity site is selected, and each possible monomer at that site is evaluated based on its prevalence in products that pass the virtual filter. Thus, each monomer has a "success rate" that is used to rank the monomer list. In this example, success was defined as passing a threshold of between 30 to 40 bits from the filter of 63 pharmacophore bits from a thrombin-inhibitor (described in the Methods).

Monomers that are seldom found in successful products are removed from the initial reagent list for that diversity site, leaving only those reagents whose products had a high success rate. This diversity site has then been "locked down," i.e., only a subset of the original reagent list remains. The process then is repeated for a second diversity site, only this time, the potential product space is reduced by the fraction of reagents that were purged from the first lock-down.

Each successive lockdown results in a smaller virtual product space that contains a higher fraction of successful products than were present before the lockdown took place. After the last lockdown, we are left with a sublibrary that has a high percentage of products that pass the virtual filter and that obey the constraints of matrix combinatorial synthesis. The lockdown can be carried out in a way that yields a library that, although still too large to take forward to combinatorial synthesis, can be fully enumerated on the computer. Once this is the case, other optimization methods can be applied to design an even smaller, synthetically practical, combinatorial library.[14,15,29,30]

RESULTS

Example 1: Estimating Physical Properties: Molecular Weight

The four-component Ugi reaction[31] (Figure 1) is a good example of a reaction that can be used to generate a very large combinatorial library. Computational analysis of the entire library, which in this example has $\sim 13 \times 10^9$ products, is impractical. However, its properties can be estimated from a random sampling of its products. The molecular weight is an example of such a property. Because the average molecular weight of the products is the sum of the average molecular weight of the reactants (minus the weight of the elements that are lost in the reaction), we have an exact solution for the average molecular weight of the $\sim 13 \times 10^9$ products. Figure 2 shows the estimate of this quantity based on random sampling of products. An average molecular weight calculated from only 600 products is within 0.5 amu of the actual value of 615. Moreover, the error in the estimate is easily calculated so that the number of samples necessary to obtain a given error is known.

Molecular weight was chosen because its value for the entire library was readily calculable. Other physical properties, such as the number of rotatable bonds or the calculated octanol/water partition coefficient, could be calculated using this method. The results from a computational assessment (e.g., a score indicating how well a product matches a computational

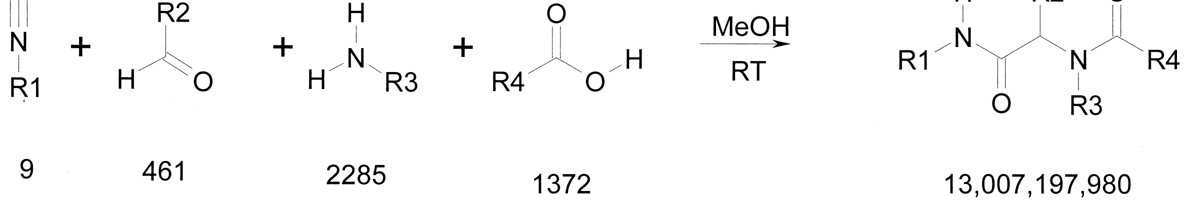

Figure 1. Four-component Ugi reaction. For the given number of isocyanides, aldehydes, amines, and carboxylic acids, there are (theoretically) more than 13 billion products.

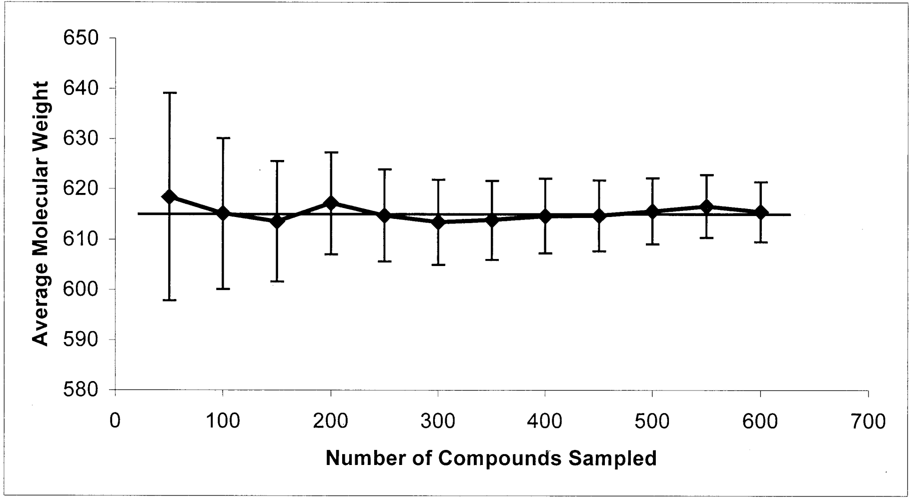

Figure 2. Estimate of the average molecular weight of Ugi reaction products as calculated by random sampling of the products. Shown are the estimates based on the number of samples with the associated estimated error (calculated from Equation 2). The solid horizontal line is the exact result.

model) of each sampled product can be the estimated quantity, as is the case in the following examples.

Example 2: Template Evaluation

In this example, we rank order chemistries based on the ability of their products to pass a computational filter. Randomly chosen products from each chemistry can be evaluated and an estimate of an overall "pass rate" can be obtained. Combinatorial libraries with the best pass rates would receive priority for chemical synthesis or would be subject to more detailed computational analysis.

As an example of this, we chose four libraries whose products differ only by the chirality of the template that displays the monomers (Figure 3). Each library contained 1,280 compounds built around a chiral cyclopentane template. The products in each library were sampled and scored against a pharmacophore filter, as described in the Methods.

The score of a molecule is the fraction of the pharmacophores in the ensemble filter that are contained in the molecule's pharmacophore descriptor (a similar procedure is described in greater detail by others,[20,26] although in our case, each molecule is scored individually by its ability to present pharmacophores that are contained in the ensemble). If the score is greater than an established threshold, the molecule is a "hit." The fraction of sampled molecules in a library that passes the threshold is the estimate of the library's hit rate.

The results for the estimated hit rates for the four libraries are shown in Figure 4. As seen in Figure 4, the relative ranking of the libraries is quickly established. After fewer than 10% of the products of the libraries have been sampled, the libraries could be prioritized for further analysis or synthesis. Thus, very rapid library comparisons can be made from randomly sampled products of the virtual libraries.

Of course, four libraries that consist of only 1,280 compounds could be prioritized through explicit enumeration of all the compounds in the virtual libraries rather than by random sampling. A more practical example can be drawn from the template evaluation and prioritization in one of our therapeutic projects. In this project, a pharmacophore model was derived from known activity data. An ensemble of 50 pharmacophores was identified that was able to distinguish active from inactive molecules. These pharmacophores were used to score products from virtual libraries. Approximately 70 templates were proposed for synthesis, with each consisting of >250,000 possible chemical products. Each chemistry was evaluated from a random sample of ~5,000 of its products. A threshold of 60% was established for a compound to be model matching. With this threshold, seven of the chemistries contained >1% model-matching compounds and three contained >10% model-matching compounds. Thus, three chemistries were taken forward to synthesis based on this evaluation. This more practical example shows how chemistries can be prioritized from evaluation of a small fraction (in this case ~2%) of their possible products.

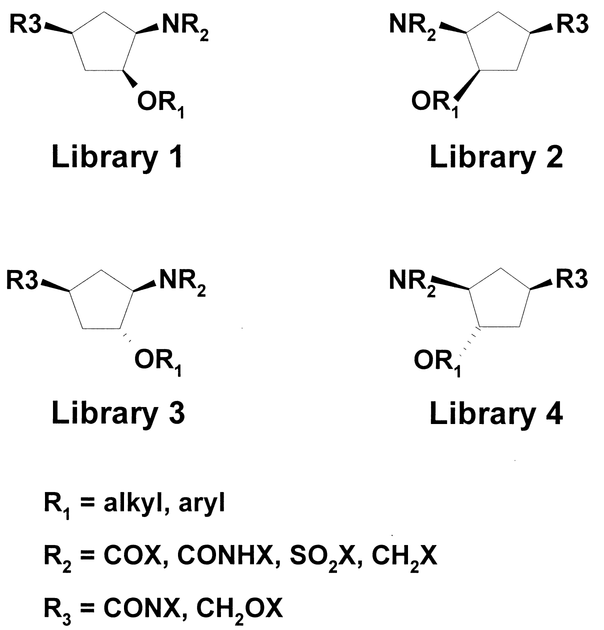

Figure 3. Four stereoisomeric templates and the types of monomers possible at each of three diversity sites. The diastereomer products were sampled randomly and scored to prioritize each template for chemical synthesis. The results of the evaluation are shown in Figure 4.

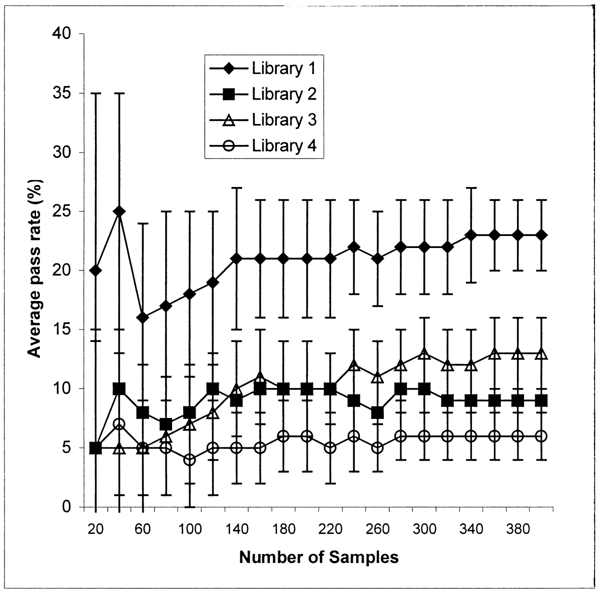

Figure 4. Cumulative pass rates for the four chemistries shown in Figure 3. Each randomly sampled product is a stereoisomer of those on the other three templates, so only the stereochemistry of the template differs at each sample. The first template clearly contains more products that pass the virtual filter and is, therefore, the leading candidate template for chemical synthesis.

Example 3: Monomer Selection Using Combinatorial "Lockdown"

After templates and chemistries have been identified as candidates for combinatorial synthesis, the task of selecting reagents remains. Depending on the template and the reaction, each diversity position may permit the use of hundreds or even thousands of possible reagents. Limits on resources for synthesis of combinatorial libraries require that the potential reagent lists be trimmed significantly. Random sampling of a virtual library could identify individual compounds that are model matching. Such "cherry-picked" molecules often are incompatible with matrix synthesis. Random sampling followed by "lockdown" as described earlier provides an efficient means to identify a subset of the possible products that are both model matching and consistent with constraints of matrix synthesis.

We will again use the four-component Ugi reaction (Figure 1) to illustrate library design using random sampling with lockdown. The Ugi reaction is a one-pot reaction and not subject to the constraints of efficient matrix synthesis. Nevertheless, if one seeks to minimize the number of reagents ordered while maintaining a high density of model-matching compounds, the requirements are identical to those of matrix synthesis: Given sets of reagents, which subsets of the lists will result in a high density of model-matching products?

In this example, we use the lockdown method to trim the number of products from $>10^{10}$ to $<10^{5}$. In the full virtual library, which has reagent list sizes of 9 isocyanides, 461 aldehydes, 2,285 amines, and 1,372 acids for the R1, R2, R3, and R4 positions, respectively, the fraction of model matching compounds is $<0.02\%$. After lockdown, the reagents list sizes have been trimmed to 9, 19, 20, and 19, and the final density of model matching compounds (i.e., those that contain 40 of the 50 preferred pharmacophores) is 100%.

Table 1 shows the number of reactants present in the virtual library at each stage of the lockdown. It is possible to limit the amount of sampling necessary by a judicious selection of the order of the reagents locked down. By selecting the shortest monomer list for the initial lockdown, the smallest number of products will have to be sampled to achieve the desired statistical accuracy (compared to the number of samples required if a larger list were chosen for the initial lockdown). In this way, when the statistically more challenging larger lists are examined, the virtual library size has already been greatly reduced, making the search easier. In the first step, approximately 400 aldehydes are filtered out. For each aldehyde, the sampled products that contain it are able to meet the requirements for success (matching 30 of the 50 pharmacophores in the model) more than 30% of the time. In the next stage of lockdown, only 80 aldehydes are allowed to participate in the random sampling of products. Based on ~40,000 random samples from this reduced product space, the acid list is reduced by a factor of 20 through application of the same filter with a more stringent cut-off of 36 of 50.

At each stage of the lockdown, the size of the library decreases and the quality of the products increases. This allows us to adjust the filters to make the requirements more stringent. The threshold for the fraction of products that pass the cut-off

Table 1. Stages of combinatorial lockdown

Lockdown stage	Monomer selected	R1	R2	R3	R4	Possible products	Products sampled	Filter pass rate[a]	Threshold for filter
1	R2	9	461	2,285	1,372	13,007,197,980	36,000	0.30 ± 0.20	30
2	R4	9	80	2,285	1,372	2,257,214,400	40,000	0.30 ± 0.16	36
3	R3	9	80	2,285	69	113,518,800	79,000	0.30 ± 0.08	38
4	R2	9	80	106	69	5,266,080	7,000	0.30 ± 0.13	40
5	R3	9	19	106	69	1,250,694	14,000	0.30 ± 0.06	40
6	R4	9	19	20	69	235,980	9,000	0.86 ± 0.05	40
Final	R1/R2/R3/R4	9	19	20	19	64,980	64,980	1.00	40

[a] Errors calculated from Equation 2.
See also Figure 1 for explanation of R1–R4.

is raised, as is the number of pharmacophores that must be hit by a successful product. The last stage of filtering creates an optimally dense set of model matching compounds based on an explicit enumeration of all the products in the virtual library.[30]

DISCUSSION

We have shown how random sampling is a very practical and useful tool in the computational evaluation of huge combinatorial libraries. It provides an efficient means to prioritize combinatorial chemistry strategies and can be used to select reagents for combinatorial synthesis on a single template.

The main advantage of library design using random sampling over other design methods derives from the reaction-based representation of proposed chemical synthetic strategies. Because of this, virtual products can be randomly chosen, constructed, and computationally evaluated without the need to fully enumerate all possible products. Thus, conclusions are based on samples that represent the full chemical product space available to the chemical synthesis at a small fraction of the computational cost that would be required to evaluate the entire library of products.

In our analysis of randomly sampled compounds from large virtual combinatorial libraries, we have concentrated on the estimated number of compounds that pass some computational filter and, as a result, did not concern ourselves with the estimated shape of the distribution. One could easily use the randomly sampled compounds to estimate other quantities of the distribution of the entire combinatorial library, such as higher moments or properties of the tail of the distribution using extreme value theory.[32]

The results from random sampling are approximations, but the strength of this method is that the errors associated with these estimates can themselves be estimated based on statistical theory. This permits calculation of the number of samples required to achieve a given accuracy in the results. Such estimates are critical when determining the necessary computational resources and time, and this ability is becoming increasingly important as computational methodologies are incorporated into mainstream combinatorial production pipelines.

A more novel application of random sampling is the combinatorial “lockdown” approach, which facilitates reagent selection for specific combinatorial reaction schemes that are based on a single chemical template or scaffold. By successively trimming away reagents that are seldom found in successful products, this technique identifies regions of product space that have a high density of desirable products and obey the constraints of matrix synthesis. In the Ugi reaction example, the number of products was reduced by five orders of magnitude within a handful of CPU days.

The lockdown method is flexible as well. The stringency of the filters applied to the randomly sampled compounds can be modulated at each stage of the lockdown as the quality of the surviving products in the virtual library improves. In addition, the process can be more iterative. Once all the monomer lists have been locked down, the constraints on any of the monomer positions can be relaxed, and the lockdown at that monomer site can be repeated to see if the chosen reagent lists change. That would provide a more robust, self-consistent lockdown procedure.

However, the lockdown method has some limitations. It is an approximation to a full evaluation of the entire library, and there is no guarantee that the final monomers chosen result in the best set of products. Moreover, even though the evaluation is based on fully constructed products, there is the possibility that the best monomers match the computational model (i.e., pass the virtual filter) by themselves. If this were the case, the result would be equivalent to independent computational evaluation of the reagents. However, the randomly sampled products that satisfy the computational model are known, and they can be examined to determine how they satisfy the model. For the pharmacophore models we used, we have found that whole products, rather than individual side chains, are necessary for success. In the Ugi lockdown example, partial products were constructed for each reagent list (by reacting the other diversity sites with a minimally small reagent). Only 2% of these compounds contained over 30 of the 62 pharmacophores in the ensemble filter, and none contained over 40. Thus, the enrichment shown in Table 1 could not have been obtained from analysis of the side chains alone.

The methods and applications that we have presented in this article illustrate the utility of random sampling in the evaluation and design of very large combinatorial libraries.* We have considered primarily computational filters based on three-dimensional descriptors, although other metrics for evaluating products, such as two- or three-dimensional diversity or even scores from docking calculations, could be used as well. However, three-dimensional descriptors require a complete conformational model of each compound analyzed and, as a result, are among the most computationally ambitious calculations in combinatorial library design. Thus, the statistical techniques outlined here allow the application of very complicated models to extremely large combinatorial libraries.

ACKNOWLEDGMENT

We thank Jennifer Miller for help in coordinating this research effort.

REFERENCES

1 Baxter, A.D. Synthesis utilizing insoluble polymers: New reactions and small molecules. *Curr. Opin. Chem. Biol.* 1997, **1,** 79–85

2 Spellmeyer, D.C., Blaney, J.M., and Martin, E.M. Computational approaches to chemical libraries. In: *Practical application of computer-aided design*, Charifson, P.S., Ed., Marcel-Dekker, New York 1997, pp. 165–194

3 Blaney, J.M., and Martin, E.J. Computational ap-

*After submission of this manuscript, Lobanov and Agrafiotis[33] published a similar application of random sampling for the computational evaluation of large combinatorial libraries. Like our study, theirs found that very large combinatorial libraries could be reliably evaluated based on computational assessment of randomly chosen products from the library. However, there are two key differences between the methods that are worthy of mention. They evaluated multiple random sublibraries and found that the averaged results agreed well with those calculated by exhaustive enumeration of the library. We have, through the use of statistical theory, calculated the smallest number of compounds that must be sampled to determine the desired accuracy in an estimated property of the library. Consequently, we are able to make our evaluation from a single randomly chosen sublibrary. They also selected monomer lists based on the frequency of their occurrence in successful products, which amounts to the first step in our lockdown procedure. The subsequent steps in our lockdown approach provide greater enrichment in the resulting sublibrary.

proaches for combinatorial library design and molecular diversity analysis. *Curr. Opin. Chem. Biol.* 1997, **1,** 54–59

4 Willett, P. *Similarity and clustering in chemical information systems*. Research Studies Press, Letchworth, England, 1987

5 Spellmeyer, D.C., and Grootenhuis, P.D.J. Recent developments in molecular diversity: Computational approaches to combinatorial chemistry. *Annu. Rep. Med. Chem.* 1999, **34,** 287–296

6 Martin, Y.C. 3D database searching in drug design. *J. Med. Chem.* 1992, **35,** 2145–2154

7 Downs, G.M., and Willett, P. Similarity searching in databases of chemical structures. In: *Reviews in computational chemistry, Volume 7*, Lipkowitz, K.B., and Boyd, D.B., Eds., VCH Publishers, New York, 1995, pp. 1–66

8 Cramer, R.D., Poss, M.A., Hermsmeier, M.A., Caulfield, T.J., Kowala, M.C., and Valentine, M.T. Prospective identification of biologically active structures by topomer shape similarity searching. *J. Med. Chem.* 1999, **42,** 3919–3933

9 Turner, D.B., Tyrrell, S.M., and Willett, P. Rapid quantification of molecular diversity for selective database acquisition. *J. Chem. Inf. Comput. Sci.* 1997, **37,** 18–22

10 Downs, G.M., and Willett, P. Clustering of chemical structure databases for compound selection. In: *Advanced computer-assisted techniques in drug discovery*, van de Waterbeemd, H., (ed.) VCH, Weinheim, 1994, pp. 111–130

11 Schemetulskis, N.E., Dunbar, J.B., Dunbar, B.W., Morel, D.W., and Humblet, C. Enhancing the diversity of a corporate database using chemical database clustering and analysis. *J. Comput. Aid. Mol. Design* 1995, **9,** 407–416

12 Poetter, T., and Matter, H. Random or rational design? Evaluation of diverse compound subsets from chemical structure databases. *J. Med. Chem.* 1998, **41,** 478–488

13 Stanton, D.T., Morris, T.W., Roychoudhury, S., and Parker, C.N. Application of nearest-neighbor and cluster analyses in pharmaceutical lead discovery. *J. Chem. Info. Comput. Sci.* 1999, **39,** 21–27

14 Gillet, V.J., Willet, P., and Bradshaw, J. The effectiveness of reactant pools for generating structurally-diverse combinatorial libraries. *J. Chem. Info. Comput. Sci.* 1997, **37,** 731–740

15 Leach, A.R., Bradshaw, J., Green, D.V.S., and Hann, M.M. Implementation of a system for reagent selection and library enumeration, profiling, and design. *J. Chem. Info. Comput. Sci.* 1999, **39,** 1161–1172

16 Hoeffding, W. Probability inequalities for sums of bounded random variables. *Am. Stat. Assoc. J.* 1963, **58,** 13–33

17 Habib, M., McDiarmid, C., Ramirez-Alfonsin, J., and Reed, B. *Probabilistic methods for algorithmic discrete mathematics*. Springer, New York

18 Mount, J.A. Estimating the range of a function in an online setting. *Info. Proc. Lett.* 1999, **72,** 31–35

19 Van Drie, J.H., and Nugent, R.A. Addressing the challenges of combinatorial chemistry: 3D databases, pharmacophore recognition and beyond. *SAR QSAR Env. Res.* 1998, **9,** 1–21

20 Mason, J.S., Morize, I., Menard, P.R., Cheney, D.L., Hulme, C., and Labaudiniere, R.F. New 4-point pharmacophore method for molecular similarity and diversity applications: Overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *J. Med. Chem.* 1999, **42,** 3251–3264

21 McGregor, M.J., and Muskal, S.M. Pharmacophore fingerprinting. 1. Application to QSAR and focused library design. *J. Chem. Inf. Comput. Sci.* 1999, **39,** 569–574

22 Bush, B., and Sheridan, R. PATTY: A programmable atom typer and language for automatic classification of atoms in molecular databases. *J. Chem. Inf. Comput. Sci.* 1993, **33,** 756–762

23 Greene, J., Kahn, S., Savoj, H., Sprague, P., and Teig, S. Chemical functional queries for 3D database search. *J. Chem. Inf. Comput. Sci.* 1994, **34,** 1297–1308

24 Brandstetter, H., Turk, D., Hoeffken, H.W., Grosse, D., Stuerzebecher, J., Martin, P.D., Edwards, B.F., and Bode, W. Refined 2.3 A X-ray crystal structure of bovine thrombin complexes formed with the benzamidine and arginine-based thrombin inhibitors NAPAP, 4-TAPAP and MQPA. A starting point for improving antithrombotics. *J. Mol. Biol.* 1992, **226,** 1085–1099

25 Teig, S.L., and Smellie, A.S. Method and apparatus for conformationally analyzing molecular fragments. US patent no. W09859306, 1998

26 Bradley, E.K., Beroza, P., Penzotti, J.E., Grootenhuis, P.D.J., Spellmeyer, D.C., and Miller, J.L. A rapid computational method for lead evolution: Description and application to the alpha 1-adrenergic antagonists. *J. Med. Chem.* 2000, **43,** 2770–2774

27 Lipinski, C.A., Lombardo, F., Dominy, B.W., and 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

28 Ajay, A., Walters, W.P., and Murcko, M.A. Can we learn to distinguish between "drug-like" and "nondrug-like" molecules? *J. Med. Chem.* 1998, **41,** 3314–3324

29 Good, A.C., and Lewis, R.A. New methodology for profiling combinatorial libraries and screening sets: Cleaning up the design process with HARPick. *J. Med. Chem.* 1997, **40,** 3926–3936

30 Stanton, R.V., Mount, J., and Miller, J.L. Combinatorial library design: Maximizing model-fitting compounds within matrix synthesis constraints. *J. Chem. Inf. Comput. Sci.* 2000, **40,** 701–705

31 Ugi, I., Lohberger, S., and Karl, R. The Passerini and Ugi Reactions. In: *Comprehensive organic synthesis: Selectivity for synthetic efficiency, Volume 2,* Trost, B.M., and Heathcock, C.H., Eds., Pergamon Press, Oxford, 1991, pp. 1083–1109

32 Young, S.S., Sheffield, C.F., and Farmen, M. Optimum utilization of a compound collection or chemical library for drug discovery. *J. Chem. Info. Comput. Sci.* 1997, **37,** 892–899

33 Lobanov, V.S., and Agrafiotis, D.K. Stochastic similarity selections from large combinatorial libraries. *J. Chem. Inf. Comput. Sci.* 2000, **40,** 460–470